

L'ADT¹ au service des humanités numériques :

*Cas des représentations des combattants britanniques et américains dans la propagande de la Première Guerre mondiale à l'aide de l'application Dataviv'*²

Oula Bayarassou¹, Noémie Haché-Chiasson², Jean Moscarola³,

¹ Scbs | South Champagne Business School – oula.bayarassou@gmail.com

² Université de Moncton – enh3291@umoncton.ca

³ Université Savoie Mont Blanc – jmoscarola@gmail.com

Abstract

The objective of this paper is to illustrate the possibilities presented by the Dataviv' software in the exploration, analysis and visualization of text corpora by an example of Textual Data Analysis (ADT), corpus composed of articles and books dealing with American and British fighters during the First World War. The functionalities presented by Dataviv' software and its ergonomics are very useful particularly in research and teaching in digital humanity fields. For this we have chosen in our paper to combine theory and application to show the interest of the software for each analysis whether it is lexical, semantic or thematic.

Keywords: digital humanities, ATD, historical corpus, teaching and research, textual analysis, data visualization, *Dataviv'*

Résumé

L'objectif de ce papier est d'illustrer les possibilités que permet le logiciel Dataviv' dans l'exploration, l'analyse et la visualisation de corpus textuels par un exemple d'Analyse de Données Textuelles (ADT) provenant d'articles et d'ouvrages traitant des combattants américains et britanniques durant la Première Guerre mondiale. Les fonctionnalités que présente le logiciel Dataviv' et son ergonomie sont des outils qui peuvent être très utiles notamment en recherche et en enseignement dans les humanités numériques. Pour cela, nous avons choisi dans notre article de combiner théorie et application pour montrer l'intérêt du logiciel pour chaque analyse qu'elle soit lexicale, sémantique ou thématique.

Mots clés : humanités numériques, ADT, corpus historique, enseignement et recherche, analyse textuelle, data visualisation, *Dataviv'*

1. Introduction

La recherche en sciences humaines et sociales (SHS) a basculé dans l'univers numérique : le numérique comme outil de recherche; comme moyen de communication et comme objet de recherche. Cette influence des modèles computationnels a donné naissance à une nouvelle discipline appelée aujourd'hui « les humanités numériques » (HN), relevant de l'automatisation de l'analyse de l'expression humaine (musique, peinture, texte, etc.). Ce croisement entre « l'humain » et « le numérique » prend la forme d'un oxymore associant deux mondes opposés : la tradition des arts et des lettres et l'innovation technologique. Toutefois, il serait réducteur de nier le rôle premier de la forte mobilisation des moyens de

¹ L'Analyse de Données Textuelles (ADT)

² <https://www.lesphinx-developpement.fr/logiciels/data-visualisation-reporting-dataviv/>

l'informatique dans l'avancement de l'état des connaissances scientifiques dans le domaine SHS et dans l'établissement des liens entre les diverses spécialités. Des disciplines comme la sociologie (dérivant des SHS) et la géographie (positionnée à la frontière des SHS) interagissent avec le mouvement des humanités numériques à travers, par exemple, la représentation des données, *Data visualisation* (Dacos et Mounier, 2015)

Un bref historique des humanités numériques nous permettra de mettre en exergue différents aspects que recouvre ce terme : la mise à disposition numérique de grands corpus, les méthodes d'analyses statistiques développées par le courant de l'ADT, et des services web ou outils logiciels disponibles pour l'enseignement et la recherche en littérature ou en histoire. Parallèlement à cela, le web met à disposition de multiples sources de corpus disponibles pour des exercices d'initiation à la recherche plus traditionnels effectués dans le cadre des masters. Les jeunes générations d'étudiants, plus familières au web et aux outils numériques, se trouvent ainsi face à des méthodes d'une recherche construite sans outils numériques, qui demandent plus d'exigences quant à la qualité de l'indexation rigoureuse des sources telles que l'habituelle numérotation des lignes. Cette pratique se développe, en particulier, par l'exercice de la lecture et de la citation. À de rares exceptions, cette herméneutique exclue tout recours à l'approche statistique de l'ADT et à fortiori des récents développement de l'IA³.

Ainsi, dans cet article on se propose de montrer l'utilité du logiciel *Dataviv'* dans l'exploration, l'analyse et la visualisation de corpus textuels en ligne/numérisés. A partir d'une application à une étude en histoire, nous présentons les fonctionnalités du logiciel. En s'appuyant sur l'analyse d'articles de journaux et de livres traitant des combattants britanniques et américains dans la propagande de la Première Guerre mondiale on utilisera le logiciel *Dataviv'*² pour le traitement et la visualisation des données textuelles. Combinant théorie et application, on montrera l'intérêt du logiciel pour chaque analyse qu'elle soit lexicale, sémantique ou thématique. Cet exemple permettra à d'autres chercheurs en sciences sociales ainsi qu'aux enseignants dans différentes disciplines en sciences humaines de se familiariser avec ce logiciel et d'entrevoir la possibilité de la mettre en pratique dans l'analyse de leurs données textuelles.

2. Construction du sens grâce à l'ADT avec *Dataviv'*: Étude de l'image des alliés de la France dans la propagande de la Première Guerre mondiale (1914-1918) via *Dataviv'*

2.1. Contexte de l'étude

L'historiographie de la Grande Guerre, surtout en France, s'est orientée depuis les trente dernières années vers une approche culturelle en mettant au cœur de l'histoire de la guerre les soldats et les civils qui ont participé à l'effort de guerre (Becker, 2005). La « culture de guerre » est considérée comme l'élément unificateur de cette nouvelle histoire de la guerre par les chercheurs associés à l'Historial de Péronne. Ces derniers considèrent que les populations ont été immergées dans un conflit total où les représentations étaient profondément intégrées (Becker, 2005).

La représentation de l'ennemi est devenue un thème central de cette nouvelle historiographie (Guillot, 2012), toutefois, l'image des alliés de la France semble négligée. Dans ce cadre, il nous a paru intéressant d'interroger les représentations des combattants britanniques et américains dans la propagande de la Première Guerre mondiale (Haché-Chiasson, 2020). Ces alliés ont été choisis en raison de leur importance dans l'issue victorieuse du conflit. Nous

³ L'intelligence artificielle (IA)

avons ainsi rassemblé un corpus contenant tous les articles qui traitent de ces combattants dans deux magazines illustrés français auxquels nous avons ajouté les ouvrages de deux collections à vocation patriotique. Il s'agit des magazines *Lectures pour tous*, le *Supplément illustré du Petit Journal* et les collections *Patrie* et les *Livres roses* pour la jeunesse.

Notre étude considère la propagande comme objet de parole. Ainsi, parallèlement à la méthode historique, nous avons décidé d'analyser ce discours descriptif et narratif axé sur les qualifications des combattants alliés à l'aide des méthodes d'analyse de données textuelles assistée par ordinateur via l'application *Dataviv*'.

Les résultats obtenus sont significatifs : par exemple, la lecture des mots qui composent le dictionnaire de notre corpus nous a permis de repérer et d'identifier les différents concepts clés des stéréotypes représentant les alliés britanniques et américains. Les résultats doivent être analysés finement et interprétés à l'aide du contexte et des connaissances historiques, mais l'utilisation des deux approches s'avère féconde. D'ailleurs, la version *Dataviv*' en ligne du logiciel *Sphinx* offre la possibilité aux lecteurs de la thèse à partir d'un lien hypertexte inséré ci-après, de consulter l'ensemble des analyses de notre corpus. Le lecteur est invité à vivre l'expérience de découvrir l'aspect révélateur des résultats.

<https://aspdev.ergole.fr/reporting/report/7e26b54f-1e42-4277-6409-08d7522e211e>

2.2. Exploration lexicale et sémantique : découvrir le corpus à partir des entrées lexicales, sémantiques ou thématiques

2.2.1. Analyse lexicale

L'Analyse de Données Textuelles (ADT) fait appel à des procédés qui reposent sur une méthodologie de regroupement, et l'exploitation de différents textes et productions linguistiques dans un corpus (Poudat et Landragin, 2017) qui nécessite, tout d'abord, la transformation du matériel textuel en données : tables de mots, de phrases, d'expressions. Une telle organisation facilite la lecture hyper textuelle (comme on peut prendre connaissance d'un livre à partir de son index) et à la compilation statistique (comme on se fait une idée des sujets principaux du livre par le nombre de pages auxquelles les éléments de l'index renvoient) (Moscarola, 2018).

Explorer un corpus relève premièrement de l'approximation lexicale (Gavard-Perret et Moscarola, 1998), c'est à dire une démarche qui renvoie à l'utilisation de l'ADT pour prendre connaissance de ce corpus en ne considérant que les résultats d'une analyse statistique des mots qui le constituent : *l'analyse lexicale*. Deux idées simples se trouvent à la base de cette approche : d'abord, le sens d'un texte qui est réductible aux mots qui le composent pris isolément ou en association avec d'autres mots, pour ensuite s'intéresser à la fréquence qui fait sens. Plus la fréquence d'un mot est importante, pris isolément ou en association avec d'autres, plus la signification qu'il porte est marquée (Moscarola, 2018). Techniquement, comment cela se concrétise à l'aide des logiciels d'ADT comme *Sphinx iQ²* et *Dataviv*' ?

Les mots qui constituent le corpus sont automatiquement réduits à leur forme canonique appelée *lemme*. Cette opération consiste à mettre les verbes à l'infinitif et les substantifs et les adjectifs au singulier. Le processus de lemmatisation consiste, ensuite, à regrouper les déclinaisons d'un *lemme*. Concrètement, ce processus associe, par exemple, « guerre » (singulier) et « guerres » (pluriel) au sein d'une même unité : guerre (singulier), ce qui permet d'analyser ces deux mots comme une seule unité et non comme deux unités distinctes (Moscarola, 2018) et de réduire ainsi la variété⁴ lexicale du contenu textuel. Le nombre de mots différents est bien inférieur à la taille du corpus. Les opérations de lemmatisation, de

⁴ Nombre de mots différents d'un corpus

suppression des mots outils, de regroupement des synonymes réduisent encore cette variété, ce qui met en évidence des rapprochements sémantiques (Siounandan et al., 2013). En plus de la lemmatisation, *il* est possible de dénombrer automatiquement les différentes formes du lexique en les classant du plus fréquent au moins fréquent. *In fine*, le résultat de cette analyse est un substitut lexical que l'on peut condenser en quelques pages contenant des listes de mots (*lexiques*), tableaux et représentations visuelles (*nuages de mots-clés*). Il est produit de manière automatique, objective et reproductible et réduit à l'exposé de signifiants (mots simples ou composés, expressions, univers lexicaux). C'est un révélateur de structures linguistiques et statistiques.

Avec l'arrivée de *Dataviv'*, logiciel d'analyse et de visualisation de données en ligne de *Sphinx*, l'analyse textuelle est devenue plus simple et plus automatisée. Son originalité réside dans sa manière intuitive de construire des *reporting* interactifs et des infographies dynamiques. Il est organisé selon l'objectif de l'analyse : l'exploration et la contextualisation du corpus (analyse lexicale et sémantique), l'identification de l'orientation des verbatims (analyse des sentiments et des opinions) et la construction de grilles thématiques (analyse de contenu thématique).



Figure 1 : Analyse textuelle et sémantique - *Dataviv'*

L'affichage des données textuelles sur *Dataviv'* trouve sa source dans le découpage du corpus. Il pourrait se faire sous différentes formes : *verbatim*, *mots*, *regroupement de réponses*, *orientations*, *concepts*, *classification thématique* etc. Dans la figure 2, nous affichons nos données sous forme de verbatim en précisant l'année de la publication de chaque texte. En effet, il est possible d'ajouter une ou plusieurs variable(s) dites de signature⁵ (année, source, etc.). Ce type d'affichage sert de conducteur à la lecture « flottante » de l'ensemble des textes, étape essentielle de préanalyse.

| | N° | 1. Texte | 3. Annees |
|---|----|--|-----------|
| 🔍 | 1 | La victoire de la mame | 1914 |
| 🔍 | 2 | Elle brillera d'un éclat magnifique parmi les victoires françaises, cette bataille de la Marne qui, changeant soudain la face des chose... | 1914 |
| 🔍 | 3 | Dans l'émouvant ré qu'on va lire on suivra les péripéties de cette formidable lutte de sept jours | 1914 |
| 🔍 | 4 | On aura l'impression de ce qu'est une bataille moderne soit, sur un front démesurément étendu, la bataille se fractionne en une séri... | 1914 |
| 🔍 | 5 | Par-dessus tout, on éprouvera une admiration sans limites pour la bravoure et l'endurance nos troupes et pour l'esprit de décision et... | 1914 |

Figure 2 : Table de données (verbatim)

Lorsque l'on choisit d'afficher nos données sous forme de mots, *Dataviv'* nous propose un tableau de mots (*lemmes*) avec par défaut l'effectif (la fréquence ou l'occurrence de chaque lemme). En activant le mode « graphique », un nuage de mots-clés s'ajoutera à la page

⁵ Elle identifie individuellement les éléments du texte.

d'analyses en cours. On peut remarquer que cet affichage renvoie, en effet, à l'approximation lexicale puisqu'il réduit le texte aux mots qui le composent et s'intéresse à leur énumération.

Nous nous sommes intéressés, ici, à ce type d'affichage afin de réaliser une analyse lexicale des représentations de Tommy et Sammy⁶ dans les médias français entre 1914 et 1918. Les résultats ont montré que les textes sont de plus en plus abondants en termes de nombre de mots. Celui-ci varie⁷ entre 9953 mots en 1914 et 34816 mots en 1918.

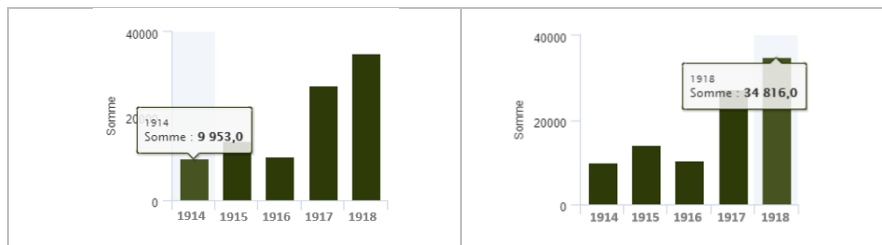


Figure 3 : Variation du nombre de mots en fonction de l'année (amplitude temporelle)

Ces mots s'articulent autour du qualificatif « allemand » qui, en se référant au verbatim⁸ auxquels il est associé, renvoie à l'ennemi commun (forces, armée, etc.).

| | |
|--|---|
| <p>Nuage de mots-clés à partir de l'ensemble du corpus</p> | |
| <p>Consultation des verbatim par mot clé (allemand)</p> | <p>lien de visualisation : https://aspdex.ergole.fr/reporting/report/7°26b54f-1°42-4277-6409-08d7522e211e/1981</p> <div style="border: 1px solid gray; padding: 5px;"> <p>“ Parce que les forces allemandes cherchaient et parvenaient constamment, depuis la frontière belge, à déborder notre aile gauche, la seule manœuvre qui pût enrayer les progrès de ce débordement avant qu’il ne développât, c’était d’aller chercher l’appui de l’armée et des forts de Paris ”</p> <p>“ Efforts hasardeux, résistance compromise, semblait-il, le grand état-major allemand n’en doutait point : sur le rapport de ses espions, il escomptait notre découragement ; il croyait pouvoir en toute sécurité tenter l’attaque brusquée, la « grande trouée » contre le secteur de l’Est ”</p> <p>“ On apprenait que l’extrême droite allemande — l’aile von Kluck qui, déjà, frôlait Paris — venait d’accomplir brusquement une conversion à droite et de s’étendre vers Meaux et Coulommiers ; on apprenait en même temps l’arrivée autour de Paris de troupes fraîches, celles notamment d’une division du Maroc ”</p> <p>“ Dût-il supposer, en trouvant brusquement en face de lui vingt mille hommes de troupes fraîches, que le général Gallieni — ce fut un trait de génie dépeché dans des taxi-autos au-devant des lignes allemandes, qu’ils étaient l’avant-garde d’une armée inattendue et puissante ”</p> <p>“ Ce qu’il y a de certain c’est que le 5, au lever du jour, le régiment allemand qui avait poussé jusqu’à Chantilly déguerpissait hâtivement du château où il s’était installé avant de poursuivre plus avant, et se rejetait, après s’être fait ouvrir la grille du parc situés derrière les étangs, sur la route de Senlis, tandis qu’à l’Est les autres contingents remontaient dans la direction de Meaux ; et de Lucy-sur-Ouisq ou ils se retranchaient ”</p> </div> |

Figure 4 : Nuage de mots-clés (global) et consultation des verbatim

De plus, parmi les avantages de l'interface de *Dataviv*, on compte la possibilité de visualiser les nuages de mots sur une amplitude de variabilité temporelle (exemple : entre 1914 et 1918).

⁶ Le soldat américain était nommé Sammy, de (Uncle) Sam, sur le modèle de Tommy (abréviation de Thomas), le soldat britannique.

⁷ Via le lien de visualisation sur *Dataviv*, il est possible de visionner la variation du nombre des mots en fonction de l'année, il suffit de passer le curseur de la souris sur l'une des barres du graphique.

⁸ Via le lien de visualisation sur *Dataviv*, il est possible de consulter les verbatim en rapport avec le mot « allemand » en cliquant dessus.

Nous remarquons ici que les mots-clés⁹ diffèrent d'une année à l'autre, ce qui nous invite à faire un retour au texte pour interpréter cette variation selon le contexte.

L'analyse lexicale via *Dataviv* ne se limite pas à la visualisation des termes les plus fréquents à partir de l'ensemble du corpus. En effet, il est possible d'appliquer un filtre d'affichage à un nuage de mots selon la catégorie grammaticale (verbe, adjectif, nom propre) ou les expressions-clés¹⁰. Bien que ces substituts lexicaux (mots, expressions) soient isolés de leur contexte, cet affinement de l'analyse lexicale contribue au commentaire général du corpus. Par exemple, à partir d'un nuage de mots illustrant les verbes les plus cités, nous pouvons distinguer les verbes d'action¹¹ et les verbes d'état¹² ce qui permet de justifier la nature du texte (descriptif, narratif ou expressif).

La figure ci-après expose les noms propres et les adjectifs clés qui composent notre corpus. Sur le premier nuage de mots, par exemple, nous remarquons que la France se positionne au centre de la représentation graphique. Cité 277 fois, ce terme fait référence à la patrie. Voici quelques exemples de verbatim qui apparaissent sur le lien de visualisation des résultats en cliquant sur ce nom propre : « *Je ne connais que la consigne : mort aux alboches et vive la France* », « *Et, grâce à l'héroïsme de ses enfants, la France vit se lever pour son sol envahi l'aube de la délivrance* ».

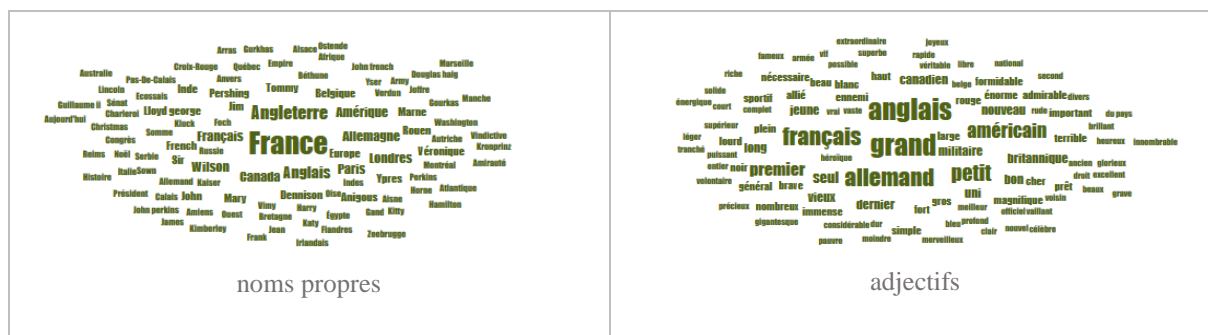


Figure 5 : Nuages de mots-clés des noms propres et des adjectifs

Partant du même principe, en choisissant d'étudier la tendance d'une série de données dans un intervalle de temps, il est possible de visualiser les adjectifs des textes publiés en 1916 en les comparant à ceux des textes publiés en 1918.

Grâce au *thésaurus*¹³ général Larousse du français intégré dans le logiciel, un nuage¹⁴ de concepts-clés peut être affiché d'une manière automatique. Ces concepts font références à des champs lexicaux de même nature sémantique. Dans notre cas, « armée », « armement », « combat » se caractérisent par une forte fréquence, reflétant un contexte de guerre. Mais, est ce que ce premier essai d'analyse sémantique automatisée est suffisant pour donner du sens à la data ?

⁹ Nous désignons par mot clé un mot ou groupe de mots qui caractérisent le contenu d'un corpus.

¹⁰ Analyse morphosyntaxique

¹¹ Expriment une action faite ou subie par le sujet

¹² Décrivent un état ou une manière d'être du sujet

¹³ Répertoire structuré de termes (mots-clés) pour l'analyse de contenu et le classement de documents.

¹⁴ Nous pouvons visualiser les nuages de concepts-clés par année.



Figure 6 : Nuages de concepts-clés

2.2.2. Catégorisation selon les structures du corpus

Réduite à l'examen des mots ou expressions clés, l'approximation lexicale souffre d'isoler les mots de leur contexte. En revanche, les méthodes d'analyses factorielles font la synthèse entre l'extrême rigidité de la recherche des segments répétés et la focalisation excessive des cartes cognitives centrées sur un seul mot (Moscarola, 2018). Elles permettent d'étudier systématiquement les associations lexicales en examinant toutes les relations de proximité entre les mots situés à l'intérieur des unités de signification (phrase, proposition) et de dégager ainsi des affinités lexicales entre termes qui se retrouvent fréquemment associés. L'analyse de ces associations peut être effectuée grâce à une analyse factorielle des correspondances (Benzécri, 1973). Les axes factoriels sont construits par les mots qui les définissent statistiquement. Ils forment des univers lexicaux, reflet de leur propension à se retrouver associés dans une même unité de signification. Ces ensembles peuvent ainsi se lire comme le résumé des unités de significations dont ils proviennent. Pour leur donner sens, l'analyste s'engage dans le travail du triangle sémiotique en mettant en relation signifiants (les mots), référents (les univers lexicaux), signifiés (les idées et connaissances du chercheur). De manière duale à la mise en évidence des univers lexicaux, définis selon les axes factoriels, par les mots situés en colonnes dans la table de données, on peut créer une partition des lignes de cette table pour regrouper les unités de signification (réponses, phrases ou séquence de mots, etc.) qui se ressemblent du point de vue des mots qui les composent. Reinert (2007) à la suite de Benzécri (1973) a proposé une méthode pour créer ce type de partition en opérant une classification hiérarchique descendante consécutive à une séquence d'analyses factorielles utilisée pour progressivement définir des classes homogènes. Les classes ainsi obtenues peuvent être caractérisées par les mots qui s'y trouvent surreprésentés. On obtient, ainsi, une autre représentation des univers lexicaux. Il appartient à l'analyste de nommer les thèmes auxquels ces catégories font référence. Il s'appuie pour cela sur l'examen des verbatim de chaque classe et sur sa connaissance du domaine.

En adoptant cette méthode (classification hiérarchique descendante), nous avons choisi de présenter deux typologies à 5 classes et à 12 classes regroupant les phrases en catégories similaires. Pour que l'analyse soit pertinente, il est souhaitable d'opter pour un nombre de classes élevé. Nous présenterons dans la figure ci-dessous les deux typologies en question sous forme de « *Treemap* ». Ce graphique représente hiérarchiquement des données en effectifs ou en pourcentages (exemple : la classe « allemand homme tranchée » comporte 1256 phrases, soit 15% du corpus). Il se lit de gauche à droite et de haut en bas. Dans un souci de donner sens à l'analyse, cette représentation graphique permet de faire un retour au texte de manière interactive en cliquant sur l'une des classes la page qui s'affiche présente les mots clés et verbatim correspondant à l'élément sélectionné.

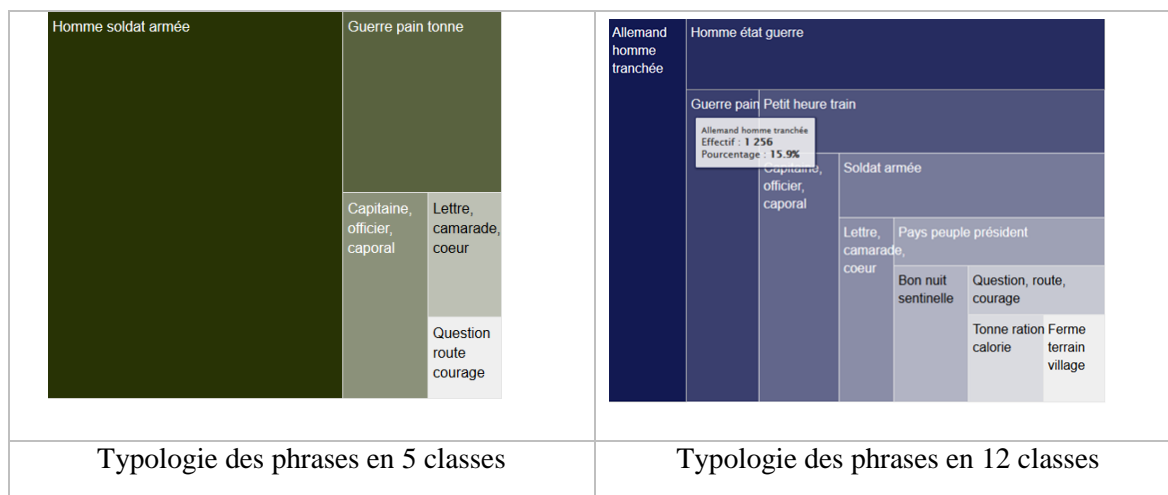


Figure 7 : Les thèmes révélés par le regroupement des phrases en classes de similitude

Comme le montrent les figures 7 et 8, les résultats de la classification hiérarchique descendante peuvent être présentés sous d'autres formes : un tableau récapitulatif des mots clés et concepts associés à chaque thème ou une carte factorielle.

Si l'on observe les deux tableaux présentés ci-dessous, nous remarquons que des classes comme « Homme soldat armée » et « Guerre pain tonne » (selon la typologie en 5 classes) se révèlent importantes. Elles sont respectivement composées de 64,9% et 16,4% de l'ensemble des mots et concepts clés qui forment le corpus.

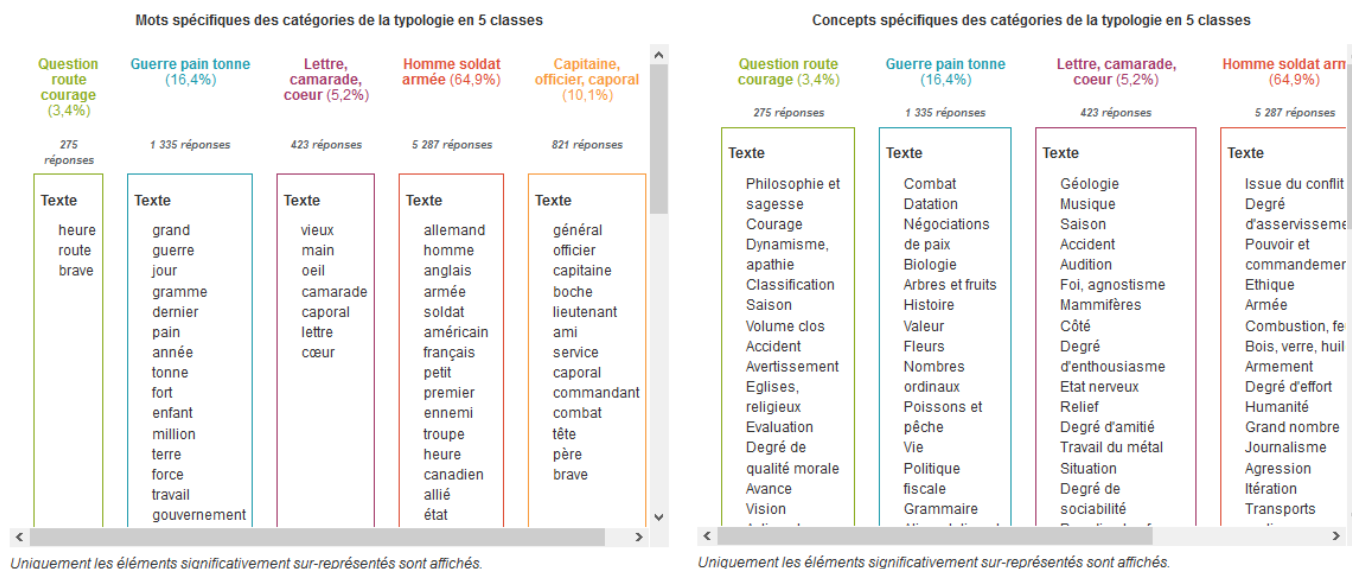


Figure 8 : Les mots clés et concepts associés à chaque classe de la typologie en 5 classes

Sur cette carte, les classes sont représentées par des *ronds bleus* et les concepts-clés par des *ronds verts* dont la taille indique l'importance du substitut lexical (en termes de fréquence). Nous présentons ici une carte factorielle des concepts-clés composant les 12 classes identifiées ci-haut. Remarquons que la taille des classes « allemand homme tranchée » (à droite de la carte), « homme état guerre », « guerre pain » est importante par comparaison aux autres classes. Plus la distance qui sépare une classe et un concept-clé est étroite plus la surreprésentation est importante. Des classes comme « homme état guerre » et « soldat armée » sont caractérisées par des concepts comme « humanité », « guerre », « état », « domicile ».

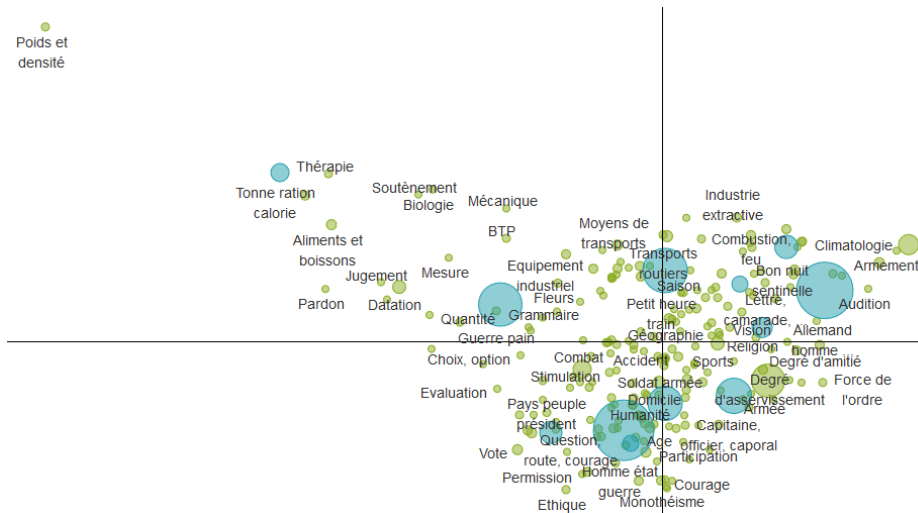


Figure 9 : Mapping des concepts spécifiques de la typologie en 12 classes

Pour apporter des éclaircissements à la lecture de cette carte factorielle, l'analyste peut faire un retour à l'analyse lexicale réalisée en amont en cliquant sur élément de la carte. Une page s'affichera dans laquelle figurent les mots clés (sous forme d'un graphique : nuage de mots-clés ou tableau en mentionnant l'occurrence de chaque terme) et les verbatim correspondants. Ci-après le lien de visualisation permettant la manipulation et le commentaire la carte factorielle issue de la classification hiérarchique descendante des concepts-clés :

<https://aspdev.ergole.fr/reporting/report/7e26b54f-1e42-4277-6409-08d7522e211e/1982>

2.2.3. Recherche de contenu : « Tommy et Sammy et les thèmes associés »

Plaçant le texte au centre de l'analyse, nous avons cherché à identifier les thèmes qui se rapportent aux représentations des « Tommy et Sammy » grâce à la création d'un dictionnaire thématique à partir de la liste des mots composant le corpus.

Les thèmes identifiés (voir figure 7) se réfèrent aux alliés de la première guerre mondiale (les acteurs).

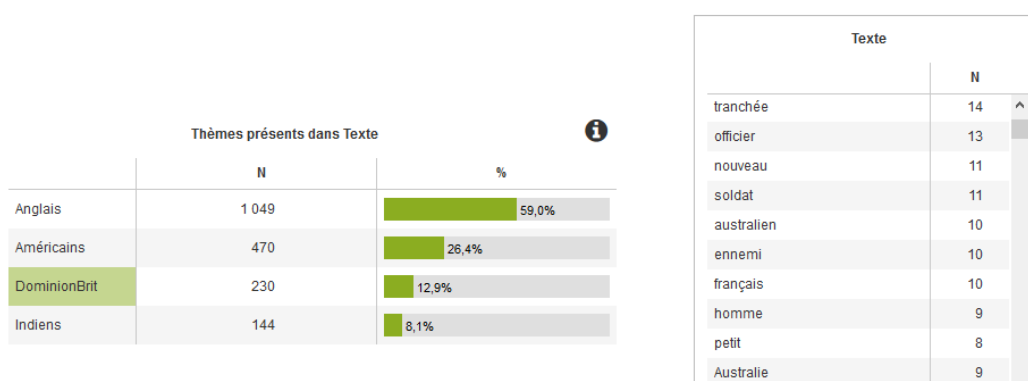


Figure 10 : Les thèmes présents dans le corpus

Le poids des thèmes correspond au nombre de fragments de texte contenant au moins un mot du thème considéré. Nous retrouvons en première position les anglais (1049 mots) suivis par les américains (470 mots).

L'évocation de ces thèmes est variable selon l'année. Sur le diagramme ci-dessous, nous comparons cette variation entre les anglais et les américains (les deux premiers thèmes). Nous remarquons que l'évolution de l'évocation des Sammy prend la forme de courbe

exponentielle contrairement à celle des Tommy. Celle-ci était presque stable avec une légère diminution entre 1916 et 1917 alors que l'évocation des américains durant cette année était en évolution positive et continue, profil qui s'explique par l'entrée officielle en guerre des États-Unis, évènement phare dans l'histoire de cette guerre.

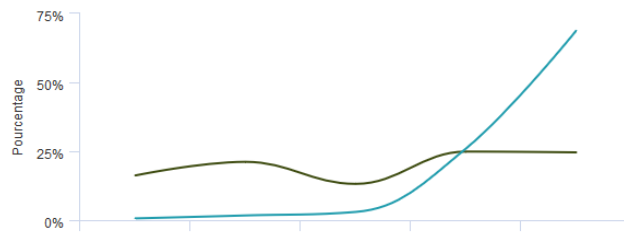


Figure 11 : Evolution de l'évocation des américains et des anglais

Ces résultats (Figure 10, thèmes des alliés) ont ensuite fait l'objet d'une Analyse Factorielle des Correspondances qui a consisté à les croiser simultanément avec ceux de la typologie thématique émergente (Figure 7, typologie en 5 classes) et les concepts du thésaurus intégré dans Sphinx (Figure 6). Cette AFC que nous présentons dans la carte ci-dessous, a permis de caractériser chaque allié d'après les concepts-clés et les classes identifiées.

Pour interpréter cette carte, il est important de noter que « *la proximité entre deux points (corrélation) n'est pas établie par la distance qui les sépare, mais bien par la mesure de l'angle entre leur vecteur respectif, à partir de la jonction centrale* » (Daigneault et Pétry, 2017, p.113). A titre d'indication, les ronds en rose correspondent aux thèmes qui reflètent les acteurs ciblés par la thématique (les alliés), les ronds verts sont les classes identifiées à partir d'une typologie et les ronds bleus renvoient aux concepts-clés. Les anglais, par exemple, sont caractérisés par des classes comme « capitaine officier caporal », « allemand homme tranchée » et des concepts-clés comme « armée », « armement ».

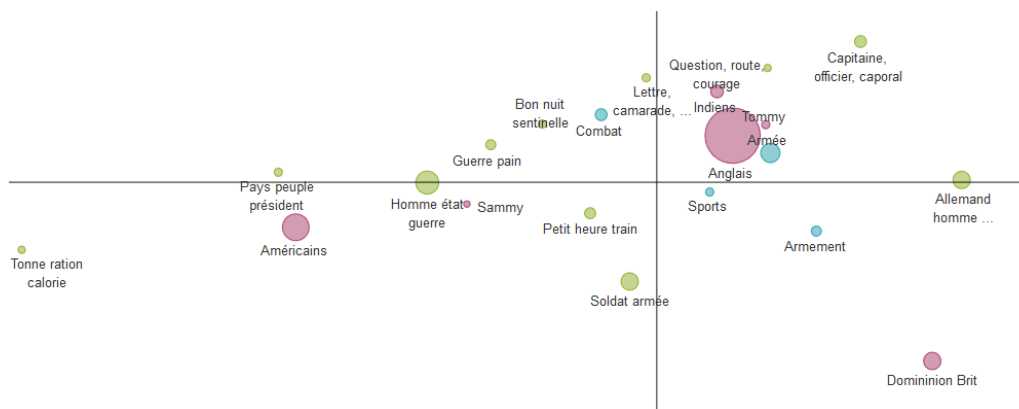


Figure 12 : Correspondance entre les

acteurs ciblés par l'analyse thématique, les catégories de la typologie et les concepts-clés

3. Conclusion

La rigueur et l'efficacité des méthodes de l'ADT via des logiciels d'analyse et de visualisation de données textuelles comme Datatviv', permettent de répondre au mieux aux attentes des humanités numériques. Aujourd'hui avec le développement des technologies de data visualisation, l'analyste ainsi que son lecteur sont tentés par une expérience de communication originale. En effet, le recours à l'ADT en ajoutant à la tradition des humanités, la puissance du numérique enrichit la lecture. « L'infographie met en scène les données et raconte l'histoire proposée par l'auteur. L'interactivité en prise directe sur les données libère la curiosité et l'esprit critique du lecteur. » (Moscarola, 2018, p. 240).

Au travers de cet article, nous avons présenté un exemple d'analyse de corpus traitant des représentations des combattants britanniques et américains dans la propagande de la Première Guerre mondiale en utilisant le logiciel *Dataviv'* de Sphinx. L'objectif était de montrer l'utilité de l'ADT automatisée dans l'exploration des textes historiques suivie de différents types d'analyse : lexicale et sémantique, typologie, recherche de contenu, etc. Ce travail pourra servir d'exemple dans l'enseignement et la recherche en matière d'ADT en histoire et en d'autres disciplines liées aux SHS.

On notera que l'approximation lexicale est une sorte d'automatisme exploratoire sur lequel le chercheur n'exerce pas de prise, se laissant simplement guider par les nuages de mots clés, les lexiques et les expressions proposées par le logiciel comme substituts à la lecture du corpus. En revanche, les typologies permettent de simplifier des données et présenter des descriptions plus synthétiques. Elles établissent et structurent les univers lexicaux caractéristiques du corpus (Moscarola, 2018). Grâce à des outils d'analyse et de data visualisation comme *Dataviv'*, l'ADT va au-delà de la description, elle permet de dessiner et caractériser des liens de significativités entre les différents résultats obtenus des analyses lexicale, sémantique et thématique. Ainsi, en offrant une autre vision du corpus, indépendante des connaissances, ou des attentes du lecteur, l'ADT peut renouveler les méthodes d'enseignement et de recherche. Le chercheur y gagne un point de vue différent pour fonder ses interprétations et ses commentaires. Mais surtout, grâce au web, il a la possibilité de partager cette expérience avec ses lecteurs. Reste l'obstacle des habitudes et des préventions qui freine la pleine adoption de ces nouveaux outils, souvent considérés comme réducteurs et illégitimes.

Bibliographie

- Benzécri J. P. & Coll. (1973). *L'analyse des Données*. Dunod.
- Becker J.-J. (2005). *Histoire culturelle de la Grande Guerre* (dir.), Paris, Armand Colin.
- Dacos M. et Mounier P. (2015). *Humanités numériques : état des lieux et positionnement de la recherche française dans le contexte international*, OpenEdition
- Daigneault P M. et Pétry F. (2017). *L'analyse textuelle des idées, du discours et des pratiques politiques*. Presses de l'Université Laval.
- Gavard-Perret M L et Moscarola J. (1998). Enoncé ou énonciation ? Deux objets différents del'analyse lexicale en marketing. *Recherche et Applications en Marketing* 13 (2) : 31-47.
- Guillot H (2012). L'image officielle du soldat allemand pendant la Grande Guerre, *Revue Historique des armées*, (269) pp.1-11, [en ligne] consulté le 18 janvier 2018. URL : <http://rha.revues.org/7575>
- Haché-Chiasson N. (2020). *Tommy et Sammy : figures des alliés britanniques et américains dans les médias français (1914-1918)*, maîtrise en histoire, Université de Moncton
- Moscarola J. (2018). *Faire parler les données*. Editions EMS - Management et société.
- Moscarola J. et Moscarola B. (2018). ADT et visualisation, pour une nouvelle lecture des corpus : Les débats de 2ème tour des Présidentielles (1974-2017) in *proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, pp. 563-570.
- Poudat C. et Landragin F. (2017). *Explorer un corpus textuel. Méthodes – pratiques – outils*. Louvain-la-Neuve, Belgique : De Boeck Supérieur, 240 p. ISBN : 978-2-8073-0563-2. Née, É. (dir.) (2017). *Méthodes et outils informatiques pour l'analyse des discours*. Rennes : PUR, 250 p. ISBN : 978-2-7535-5499-3.
- Reinert M. (2007). Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage & Société*, (121-122), pp. 189-202.
- Siounandan, N., Hébel, P., & Colin, J. (2013). Va-t-on vers une frugalité choisie ? Cahier de recherche. CREDOC.