Les humanités numériques : Contribution de l'ADT[1] à la recherche et à l'enseignement en histoire avec *Dataviv'*[2]

Cas des représentations des combattants britanniques et américains dans la propagande de la Première Guerre mondiale

Oula Bayarassou[1], Joceline Chabot[2], Noémie Haché-Chiasson[3], Jean Moscarola[4], Sylvia Kasparian[5]

[1] Université Grenoble-Alpes – oula.bayarassou@gmail.com

[2] Université de Moncton – joceline.chabot@umoncton.ca

[3] Université de Moncton – enh3291@umoncton.ca

[4] Université Savoie Mont Blanc – jmoscarola@gmail.com

[5] Université de Moncton – sylvia.kasparian@umoncton.ca

Abstract

The objective of this paper is to illustrate an example of Textual Data Analysis (ADT) from articles and books dealing with American and British fighters during the First World War. The exploration of the corpus formed from these sources was carried out using Dataviv analysis and data visualization software in order to popularize the use of these tools in research and teaching in history. We have chosen to combine theory and application on software to show the interest of 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 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. L'exploration du corpus formé à partir de ces sources a été menée à l'aide du logiciel d'analyse et de *data* visualisation Dataviv' dans le but de vulgariser l'utilisation de ces outils dans la recherche et l'enseignement en histoire. Nous avons choisi de combiner la théorie et l'application sur logiciel pour montrer l'intérêt de 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; le numérique comme moyen de communication et le numérique comme objet de recherche. Cette influence des modèles computationnels a donné naissance à une nouvelle discipline que l'on appelle les humanités numériques (HN), qui relèvent 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.

[1] L'Analyse de Données Textuelles (ADT)

[2] https://www.lesphinx-developpement.fr/logiciels/data-visualisation-reporting-dataviv/

Toutefois, il serait absurde de nier le rôle premier de la forte mobilisation des moyens de 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'IA[3].

Ainsi, ce papier se propose de montrer l'utilité de l'ADT dans l'interprétation des données textuelles issues de sources littéraires et historiques en ligne et/ou numérisées. Il s'appuie sur l'analyse d'articles et de livres traitant des combattants britanniques et américains dans la propagande de la Première Guerre mondiale en utilisant le logiciel *Dataviv'²* pour le traitement et la visualisation des data. Il ouvrira la voie aux chercheurs, enseignants et étudiants en histoire et en littérature pour se familiariser avec l'ADT et la mettre en pratique dans leurs différents travaux et projets.

2. L'ADT et l'analyse du corpus en histoire

2.1.Intérêt de l'ADT pour l'étude des corpus historiques

Nous nous sommes intéressés à l'ADT tout d'abord pour la facilité du traitement que permettent les outils dans l'analyse de gros volumes de textes et la capacité d'en dégager du sens, des sens impossibles à percevoir à l'œil nu, étant donné leur volume. Ainsi, la rapidité d'exécution et le gain de temps inestimable que proposent ces outils statistiques et informatisés pour la lecture et le traitement des données textuelles sont non négligeables.

Ces outils permettent l'exploration, de surcroît, de tout genre de volumes de textes sans idée préconçue, pour en dégager des pistes de recherche. Ils ont donc un pouvoir suggestif, mettant à jour des particularités, des anomalies, offrant des pistes de réflexion ou d'analyse auxquelles on n'aurait pas pensé à priori.

L'approche multidimensionnelle de l'analyse des données textuelles grâce à l'apport de l'Analyse Factorielle de correspondance développée par Benzécri (1973) et implantée dans la plupart des logiciels[4] est un apport très important de l'ADT à l'analyse de textes. Cette analyse descriptive des données textuelles permet de traiter à la fois plusieurs variables en offrant des vues synthétiques éclairant ainsi des relations invisibles entre les modalités de ces variables (variables qui peuvent être textuelles ou associées aux textes).

[3] L'intelligence artificielle (IA)
[4] Alceste, Iramuteq, Sphinx

Enfin les logiciels développés permettent de traiter de différents niveaux de l'organisation de textes ou de discours, à la fois les niveaux phonétique-orthographique, morphologique et lexical, sémantique et pragmatique et permettent de rendre saillant l'organisation thématique des textes. L'évolution chronologique d'une unité textuelle (mot, lemme, thème, segment, etc…) par exemple, devient accessible sur de longues périodes de temps (100-200 ans ou plus), analyse impensable encore il y a quelques décennies.

2.2. Apport de l'ADT à l'analyse des corpus historiques

Traditionnellement, la méthode historique est basée sur une lecture critique des sources. Les chercheurs s'approprient les sources, les analysent et, au terme d'un processus souvent assez long, proposent une interprétation basée sur leur connaissance intime du corpus. L'analyse de données textuelles (ADT) nous permet d'éclairer les sources d'une autre manière et sous un autre angle. D'abord, elle oblige l'historien et l'historienne à une réflexion préalable sur la constitution même du corpus en tant qu'objet de recherche. De plus, elle constitue un outil précieux lorsqu'il s'agit de traiter de gros volumes de textes, on peut penser ici à des corpus de presse, des textes littéraires, etc. Les résultats statistiques obtenus grâce à l'ADT offrent des pistes de réflexion qui viennent compléter, nuancer, voire bonifier la lecture traditionnelle des sources. Par exemple, cela peut contribuer à affiner la problématique et à ouvrir de nouvelles perspectives de recherches. Bien sûr, les chercheurs ne doivent pas être naïfs en imaginant que soumises à l'ADT les sources parleront d'elles-mêmes. C'est pourquoi ils doivent se livrer à une lecture éclairée des résultats statistiques sur lesquels s'appuiera une interprétation plausible et pertinente qui n'en sera que plus féconde.

2.3. Contributions du logiciel de visualisation et de reporting, « Dataviv' », à l'explicitation des résultats

Un premier contact avec *DataViv'* nous donne d'abord la possibilité de consulter les corpus *in vivo* et d'accéder aux analyses qui se font en direct. Accessibilité du corpus qui permet de vérifier les analyses, de les partager, donc de les reproduire : fonctions utiles à la communauté scientifique et en situation de salle de classe où les données peuvent être partagées. L'utilité de cet outil au niveau de l'enseignement est indéniable.

L'aspect visuel et l'ergonomie de l'outil sont à souligner. *DataViv'* est visuellement très agréable et stimulant. Les résultats sont présentés par de beaux graphiques épurés et faciles à lire par des étudiants en sciences humaines et sociales.

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

3.1. 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. 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.

La représentation de l'ennemi est devenue un thème central de cette nouvelle historiographie. 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 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. À l'occasion de la participation de Jean Moscarola au séminaire de la professeure de linguistique Sylvia Kasparian, nous nous sommes familiarisés avec le logiciel *Sphinx*. Dans ce cadre, Jean Moscarola, un des développeurs historiques de *Sphinx*, a accepté de nous aider à traiter la masse importante des données de notre corpus avec l'analyse 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

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

3.2.1. 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 est propice à 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'* ?

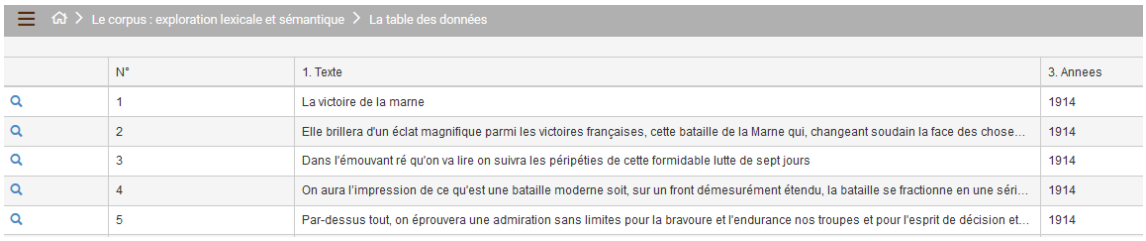
Les mots qui constituent le corpus sont automatiquement réduits à leur forme canonique appelée *lemme*. Cette opération consiste à mettre les verbes à l'infinitif et les substantifs et les adjectifs au singulier. Le processus de lemmatisation consiste, ensuite, à regrouper les déclinaisons d'un *lemme*. Concrètement, ce processus associe, par exemple, « guerre » (singulier) et « guerre**s** »

(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 variété[5] lexicale du contenu textuel. Le nombre de mots différents est bien inférieur à la taille du corpus. Les opérations de lemmatisation, de 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 son origine 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 signature[6] (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 marne	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)

[5] Nombre de mots différents d'un corpus
[6] Elle identifie individuellement les éléments du texte.

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 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 Sammy[7] 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 varie[8] entre 9953 mots en 1914 et 34816 mots en 1918.

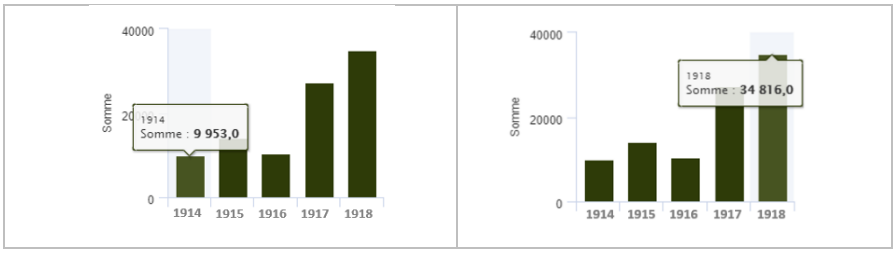


Figure 3 : Variation du nombre de mots en fonction de l'année (amplitude temporelle)

Ces mots s'articulent autour du qualificatif « allemand » qui, en se référant au verbatim[9] auxquels il est associé, via le lien de visualisation, renvoie à l'ennemi commun (forces, armée, etc.).

Nuage de mots-clés à partir de l'ensemble du corpus	
Consultation des verbatim par mot clé (allemand)	 lien de visualisation : https ://aspdev.ergole.fr/reporting/report/7e26b54f-1e42-4277-6409-08d7522e211e/1981

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

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

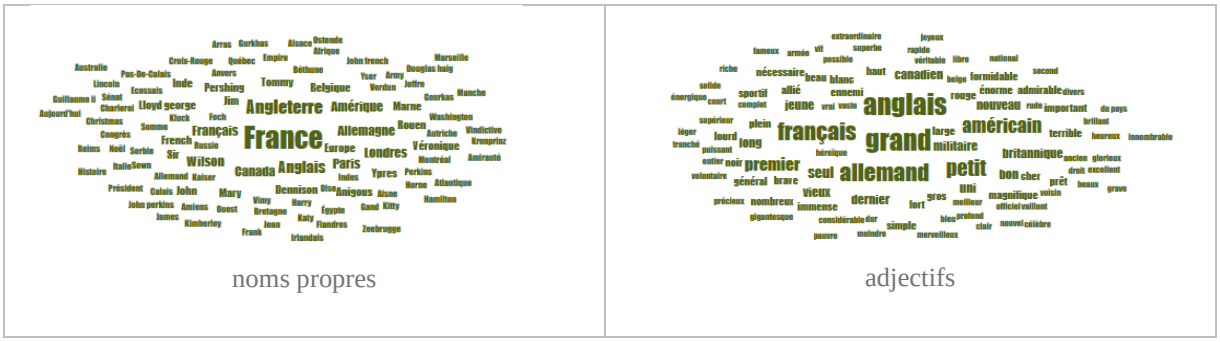
[8] Via le lien de visualisation sur Dataviv', il est possible de visionner la variation du nombre des mots en fonction de l'année, il suffit de passer le curseur de la souris sur l'une des barres du graphique.

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

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). Nous remarquons ici que les mots-clés[10] diffèrent d'une année à l'autre. Ceci 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-clés[11]. 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'action[12] et les verbes d'état[13] 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* ».



noms propres | adjectifs

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*[14] général Larousse du français intégré dans le logiciel, un nuage[15] 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* ?

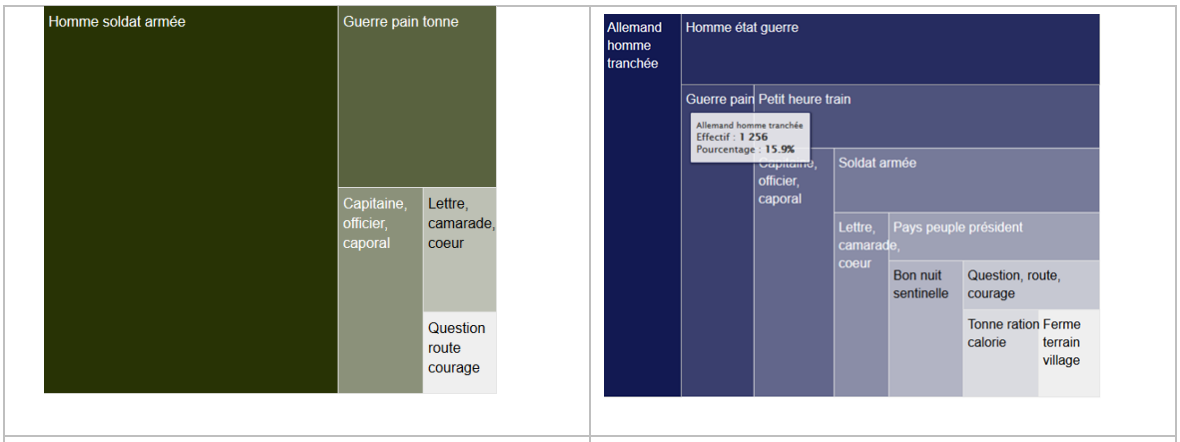
[10] Nous désignons par mot clé un mot ou groupe de mots qui caractérisent le contenu d'un corpus.
[11] Analyse morphosyntaxique
[12] Expriment une action faite ou subie par le sujet
[13] Décrivent un état ou une manière d'être du sujet
[14] Répertoire structuré de termes (mots-clés) pour l'analyse de contenu et le classement de documents.
[15] Nous pouvons visualiser les nuages de concepts-clés par année.

Figure 6 : Nuages de concepts-clés

3.2.2. 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 entres 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 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 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.



Typologie des phrases en 5 classes | Typologie des phrases en 12 classes

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.

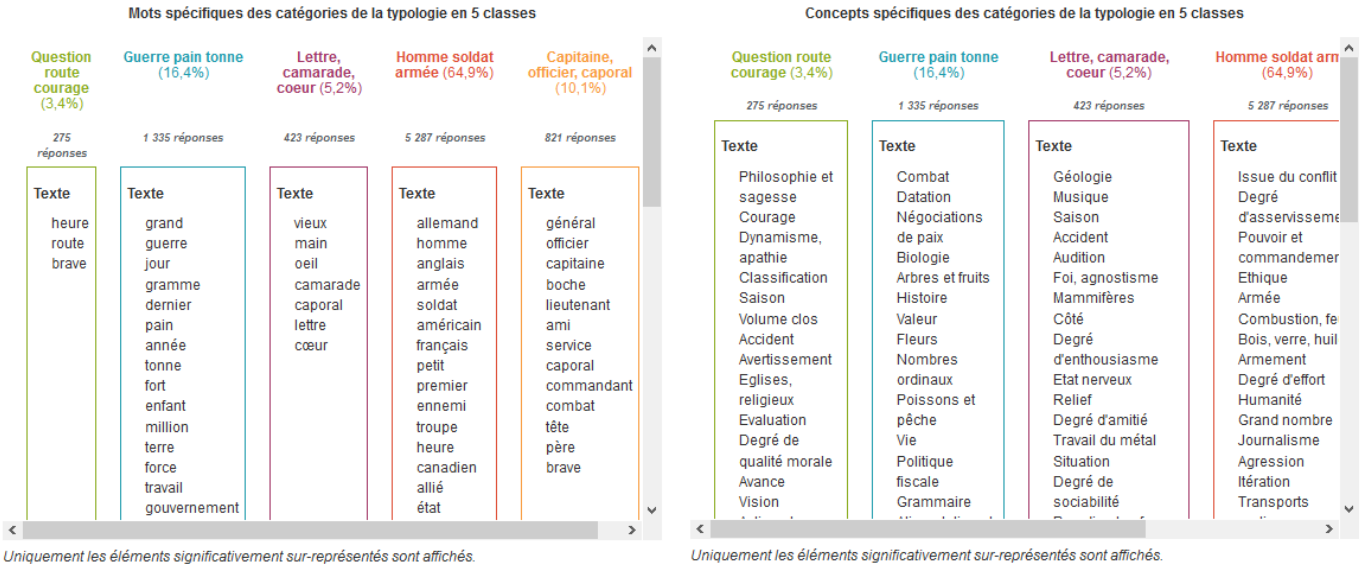


Figure 8 : Les mots clés et concepts associés à chaque classe de la typologie en 5 classes

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 ».

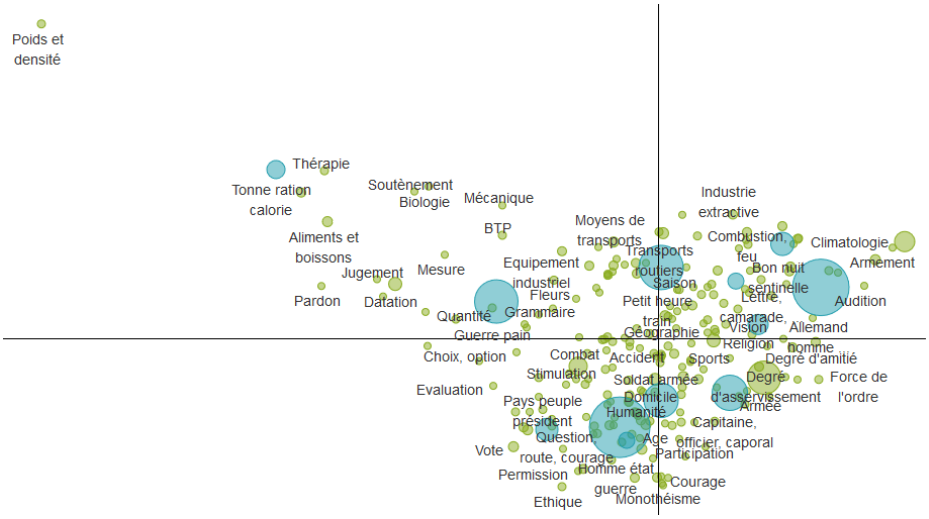


Figure 9 : *Mapping* des concepts spécifiques de la typologie en 12 classes

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

3.2.3. 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).

Thèmes présents dans Texte

	N	%
Anglais	1 049	59,0%
Américains	470	26,4%
DominionBrit	230	12,9%
Indiens	144	8,1%

Texte

	N
tranchée	14
officier	13
nouveau	11
soldat	11
australien	10
ennemi	10
français	10
homme	9
petit	8
Australie	9

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.

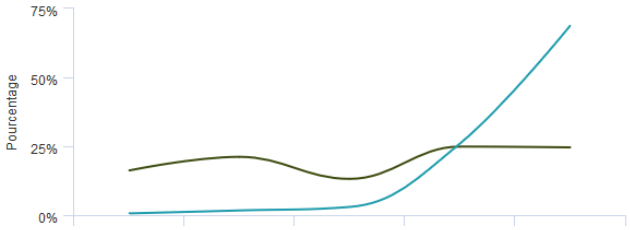


Figure 11 : Evolution de l'évocation des américains et des anglais

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 ».

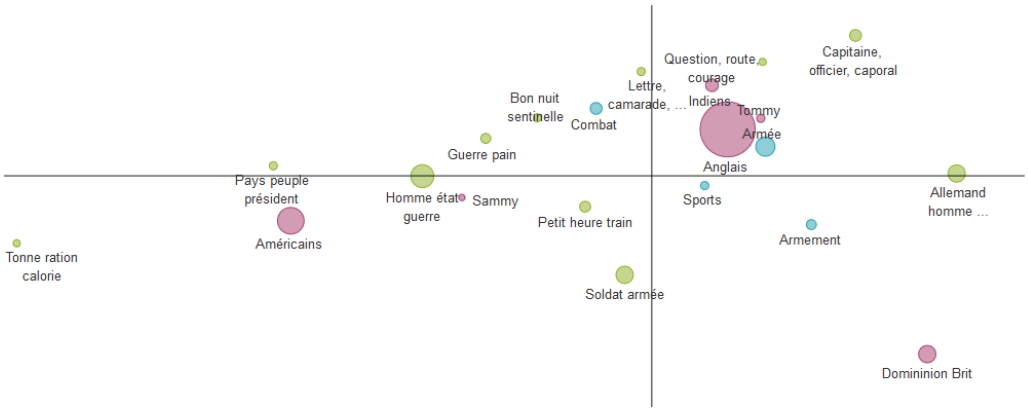


Figure 12 : Correspondance entre les acteurs ciblés par l'analyse thématique, les catégories de la typologie et les concepts-clés

4. 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 ce papier, 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.

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.

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.