Analyse de contenu avec Tropes

L'analyse de contenu est une méthode d'étude scientifique dans les Sciences Humaines et Sociales. C'est un champ de pratique large avec des méthodes, des méthodologies et des techniques scientifiques diverses. Elle est utilisé en recherche mais plus particulièrement dans les domaines des Sciences de l'Information et de la Communication (SIC) et des Sciences de l'Information et du Document (SID). L'analyse de contenu « *se définit comme une technique permettant l'examen méthodique, systématique, objectif et, à l'occasion, quantitatif du contenu de certains textes en vue d'en classer et d'en interpréter les éléments constitutifs, qui ne sont pas totalement accessibles à la lecture naïve* » (Robert, Bouillaguet, 2007). Le terme de « texte » peut désigner aussi bien des productions orales ou écrites, textuelles ou sonores. Dans notre cas, nous nous focaliserons sur les textes d'articles de presse écrite.

Dans l'analyste de corpus de presse, la posture de l'analyste doit être différente de la posture du lecteur de journal. Pour le projet ANR MémoMines, l'analyste doit étudier les numéros du journal papier La Voix du Nord téléchargé sous format HTML, donc sous format numérique. Ce qu'il faut comprendre dans l'analyse de corpus de presse est que les textes qui ont été donnés à étudier pour l'analyste ne sont uniquement les textes que le journal a donné à lire. De plus « *ce que lit l'analyste n'est pas ce qu'à lu le lecteur, ce que lit l'analyste n'est pas ce qui a été dit. […] ce ne sont pas des relations d'identité et de textualité qui unissent les paroles des commentateurs s'exprimant dans l'espace public et le journal qui les fait parler sous son nom, mais des rapports de sélection et de transformation* » (Krieg, 2000). Il est important de préciser aussi que le rôle de l'analyste n'est pas d'interpréter avec son imagination. Aussi, « *ce qui est écrit dans le journal n'est pas ce qui a été dit* » (Krieg, 2000). En d'autres termes, les textes ne reflètent absolument pas des paroles et doit être pris de façon autonome. L'analyse de corpus de presse ou de discours de presse est une spécificité dans l'analyse de contenu. Elle possède ses propres méthodes et techniques scientifiques.

En plus de la spécificité qu'est l'analyse de corpus de presse, il y a une dimension automatique à ajouter. En effet, dans la mission de stage, le but est d'analyse de façon automatisée un corpus de presse. En ce sens, certains outils numériques existent, qu'ils soient gratuits ou payants, et sont dédiés à ce champ disciplinaire. Ces outils permettent de faire ressortir très rapidement les mots-clés d'un discours. Le choix des outils numériques d'analyse sémantique diffèrent en fonction des besoins de l'analyste, la nature des discours à analyser ainsi que leur format. Cependant, ces outils comportent tous leurs propres limites dans leur utilisation. C'est pourquoi, une vue d'ensemble sur les outils existants et leur fonctionnalité va nous être utile pour déterminer le ou les outil(s) qui nous sera(ont) indispensable(s) et qui correspondra(ont) au mieux à nos attentes et nos besoins.

Nous allons voir quels sont les différents outils qui peuvent être utilisés dans le cadre d'une analyse de corpus de presse et pourquoi nous avons choisi Tropes. Nous continuerons sur la méthodologie à adopter, les critères d'analyse et la confection d'une grille d'analyse. Puis nous terminerons par aborder la notion de l'objectivité et sa place dans l'analyse de contenu.

2.1 Choix des outils

Dans le cadre du projet ANR MémoMines, qui vise à établir une liste des acteurs et des lieux qui ont joué un rôle pendant ou après la période de l'exploitation du charbon, une question s'est posé sur le meilleur outil à utiliser pour une analyse de données textuelles informatisée. Ils sont nombreux à exister mais il a fallut trouver celui qui répondait le mieux à nos besoins. Pour ce faire, nous allons commencer par présenter certains outils qui pourraient nous intéresser dans le cadre d'une analyse de corpus d'articles de presse automatisée, puis nous allons nous focaliser sur l'outil Tropes, qui est l'outil qui a été choisi pour cette mission de stage.

2.1.1 Présentation des différents outils

Avant de commencer quelconque présentation, il est important d'expliquer les données que nous devons analyser. Les données à analyser sont sous forme de corpus textuel. Ce corpus représente l'ensemble des articles du journal La Voix du Nord publiés sur le domaine de la mine entre 2004 et 2021. Il est constitué de plus de 1 300 articles de presse. Ce qu'il faut comprendre c'est que nous avions besoin d'un outil capable de traiter les milliers de pages d'un corpus constitué au préalable et importé par l'analyste dans cet outil. Maintenant que le sujet d'analyse et expliqué, nous pouvons commencer la présentation des cinq outils qui pourraient être utiles dans ces circonstances.

Le premier outil qui sera présenter est Iramuteq. Iramuteq est un logiciel, créer en 2010, qui « *permet de faire des analyses statistiques sur des corpus texte et sur des tableaux. Il repose sur le logiciel R et le langage Python* » (Ratinaud, 2021)[1]. C'est un logiciel disponible sous Windows, Linux et MacOs. Il permet de générer en sortie des corpus sous format CSV et ne prend qu'en entrée uniquement des corpus CSV. L'analyse de données textuelles qu'il propose repose sur la méthode de classification hiérarchique descendante de Reinert (1983, 1991), c'est-à-dire qu'elle est composée de trois modalités : la première étant la classification simple sur texte qui traite les textes dans leur intégralité et qui permet de regrouper les plus proches ; la seconde est une classification qui porte sur les segments de textes ; et la dernière est une classification sur deux tableaux, ici il n'est plus question de segments mais de regroupements de segments. L'un de ses points fort est qu'il propose une richesse informationnelle et une diversité des visualisation puisqu'il associe les différentes classes thématiques à une couleur. De plus, « *IRaMuTeQ s'articule avec TXM : tout corpus qui a été importé dans l'un des logiciels est importable dans l'autre* » (Pincemin, 2018).

Le second outil est TXM qui est un logiciel de textométrie gratuit et open-source, crée en 2009. Ce logiciel est disponible sous Windows, MacOS X et Linux. Il prend en entrée plus d'une dizaine de formats de fichiers et il peut en générer autant en sortie. Il a pour objectif de repenser les calculs textométriques pour des corpus annotés et structurés. Il est utilisé pour l'analyse de données textuelles, du discours, et du contenu ce qui lui permet d'avoir une exploitation plutôt complète des informations contenues dans un corpus. Il propose aussi une personnalisation par des scripts, des sorties graphiques interactifs et les résultats donnés peuvent être exploités par d'autres outils. C'est un logiciel très apprécié dans le domaine des sciences humaines et sociales et plus particulièrement dans les humanités numériques.

Les troisièmes logiciels sont les logiciels Hyperbase et Hyperbase Web Edition qui servent à l'exploration documentaire et statistique des textes. Il a été crée en 1989 mais il a, depuis, évolué et une version 10 en est sortie cette année, en 2021. Il est disponible uniquement sur Windows. Il y a deux fonctions qui sont combinées dans ces logiciels, une fonction documentaire et une fonction statistique. Parmi les fonctions documentaires, on retrouve une lecture naturelle du corpus, une possibilité de navigation par mots-clés, il recherche et tri les contextes et propose un index, des dictionnaires des formes, des lemmes, des codes et des fréquences. Quant à la fonction statistique, elle propose des graphes des unités, des représentations factorielles ou arborées, une extraction des segments du corpus et des représentations des cooccurences

1 Selon le site : http://www.iramuteq.org

et réseaux thématiques[2]. La différence entre Hyperbase et Hyperbase Web Edition est que ce dernier est une refonte du premier dans le but de rendre l'interface plus intuitive et disponible en ligne.

Le quatrième logiciel est Analec. Analec est un logiciel d'analyse et d'annotation de corpus écrits. Il est téléchargeable en ligne gratuitement et disponible sous Windows, Mac et Unix. Il permet la gestion, le regroupement, la visualisation et l'impression de corpus annoté et proposent des calculs de fréquences, de recherche de corrélations et une générations de tableaux qu'il est tout à fait possible de définir soi-même. Il génère en sortie des corpus sous les formats TXT, CSV et GLOZZ.

Le cinquième et dernier logiciel de cette liste se nomme Glozz. C'est un logiciel d'annotation manuelle et d'exploration de corpus textuels. Il est disponible sous OS X, Linux et Windows. Il prend en entrée uniquement des corpus en texte brut et génère en sortie du texte brut ou du CSV. Il permet une annotation manuelle de texte qui peut être déclarée en XML et que l'on peut explorer grâce à des requêtes en langage GlozzQL ou SQL.

Cette liste de logiciels est une liste non exhaustive des logiciels qui pourraient répondre le plus à nos besoins. Ces logiciels sont très intéressants lorsque l'on cherche à analyser du texte et le marquer et pour certains sont très complets en proposant des tableaux et des schémas. Cependant, il a une dimension à prendre en compte dans le choix d'un tel outil, la dimension de proximité.

2.1.2 Choix de Tropes

Le logiciel qui a été choisi pour mener à bien la mission d'analyse de corpus d'articles de presse est le logiciel Tropes. Nous allons, dans cette partie, expliquer le fonctionnement de Tropes et pourquoi nous avons opté pour celui-ci.

Tropes est un logiciel d'analyse sémantique. Il a été crée par Pierre Molette et Agnès Landré en 1994. Il est capable d'éditer des ontologies, de proposer une arborescence des références, une analyse chronologique des textes, catégoriser des termes, analyser les acteurs, permettre une extraction terminologique et un diagnostic du style du texte. Il est disponible uniquement sous Windows et c'est un logiciel gratuit. Il est utilisé en science de l'information afin de générer et de vérifier la pertinence des ontologies et des thesaurus. Le but de ce logiciel est de montrer à son utilisateur le squelette du texte analysé.

2.1.2.1 Fonctionnement de Tropes

Comment fonctionne le logiciel Tropes ? Avant de procéder à quelconques manipulations du logiciel, il est important de constituer son corpus, si c'est un corpus qu'il faut analyser, ou son texte. Tropes accepte les entrées sous format texte ANSI, HTML, Microsoft Word, etc. Dans son analyse, Tropes ne prend en compte uniquement les mots et la ponctuation. Les caractères spéciaux ne sont donc pas un problème et n'altère en aucun cas l'analyse sémantique. Cependant, les termes constitués de plusieurs mots doivent être lié avec le caractère « _ » pour compter comme un seul et même mot.

Pour analyser notre corpus et notre texte, il va falloir lancer le logiciel et ouvrir le ou les fichier(s) dans le menu [Fichier] et sélectionner les textes à analyser. Il est tout à fait possible, dans un corpus, de faire en sorte que le logiciel nous montre les résultats d'un seul fichier à la fois. Pour ce faire, il faudra aller sur l'onglet [Fichier] et sélectionner dans la liste des fichiers constituant le corpus, le fichier que l'on souhaite voir uniquement. Il est aussi possible d'en ajouter, d'en effacer ou de les trier.

Les résultats de l'analyse apparaissent sur le côté gauche de la fenêtre principale du logiciel. Afin de mieux comprendre les termes d'analyse, nous allons proposer quelques définitions de ceux-ci. Le premier terme d'analyse est le style. Il précise le style général du texte et sa mise en scène, le nombre de fichiers examinés, le nombre de propositions remarquables qui résument les parties les plus caractéristiques du texte, et le nombre d'épisodes qu'il a pu détecter. Les épisodes sont les parties représentatives du textes et sont constitués selon la chronologie du discours. Dans ces épisodes, on retrouve la notion de rafales, qui sont des mots que l'on retrouve plusieurs fois dans le discours dans un épisode donné.

---

2 Selon la page Wikipédia « Hyperbase », URL : https://fr.wikipedia.org/wiki/Hyperbase

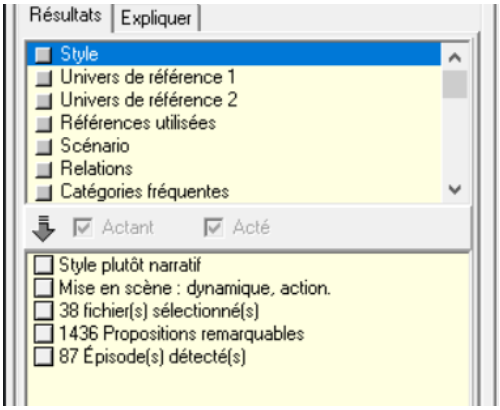

Figure 1: Résultats des styles sur le logiciel Tropes

Dans un second temps, ce sont les références qui seront analysées par le logiciel. Les références sont des objets, des concepts, des acteurs, des idées qui représentent le contexte. Elles sont regroupées selon des classes d'équivalents et selon trois niveaux : l'univers de référence 1 pour le premier niveau, l'univers de référence 2 comme second niveau et les références utilisées qui constitue le troisième et dernier niveau. Elles apparaissent par importance décroissante et sont accompagnées de leur nombre d'occurrences. Il est tout à fait possible de personnaliser la classification de ces références en ajoutant des éléments dans le scénario. Le scénario est une classification hiérarchique que le logiciel nous retourne en résultat. Il permet de personnaliser les dictionnaires du logiciels et structurer l'information comme nous le souhaitons. Pour créer ou modifier un scénario, il faudra utiliser l'outil scénario que l'on retrouve dans le menu [Outil]. Le logiciel ouvrira alors une nouvelle fenêtre destiné uniquement à ce scénario. Cet outil permet de construire notre propres classifications personnalisées. Celle-ci peut se faire manuellement à partir d'un scénario vide, automatiquement à partir du texte analysé ou nous pouvons ajouter un scénario déjà constitué au préalable. Dans ce dernier cas, il faudra sélectionner le fichier scénario qui doit avoir comme extension « .scn ». Dans un troisième temps, le logiciel est capable de nous retourner les relations qui peuvent exister entre deux termes et leur nombre d'occurrences. Les relations sont les mots qui sont retrouvés ensemble, dans le même ordre et dans la proposition du texte. Elles sont orientées suivant l'ordre d'apparition. Le dernier terme d'analyse à définir est celui des catégories. Le logiciel Tropes nous rapporte les catégories de mots du texte analysé. Il y a deux niveaux : les catégories fréquentes donc celles qui sont les plus utilisées, et toutes les catégories qui comportent les plus fréquentes et les moins fréquentes.

Figure 3: Résultats de l'univers de référence 1 sur le logiciel Tropes

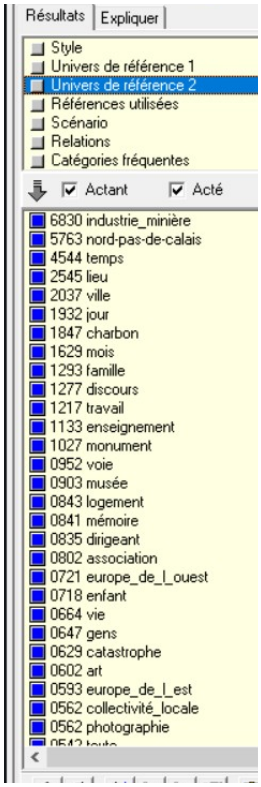

Figure 2: Résultats de l'univers de référence 2 sur le logiciel Tropes

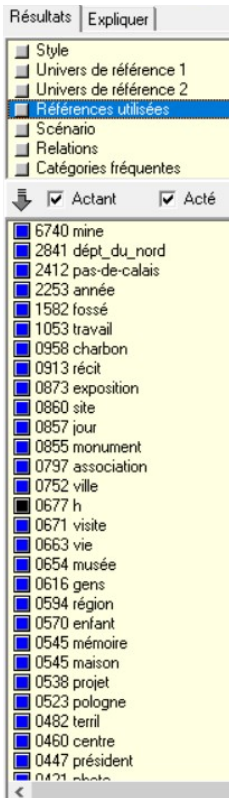

Figure 4: Résultats des références utilisées sur le logiciel Tropes

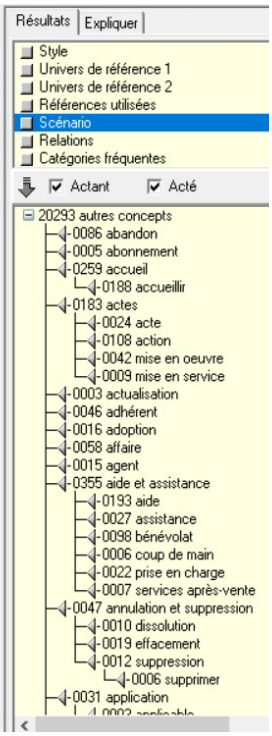

Figure 6: Résultats du scénario sur le logiciel Tropes

Figure 5: Résultats des relations sur le logiciel Tropes

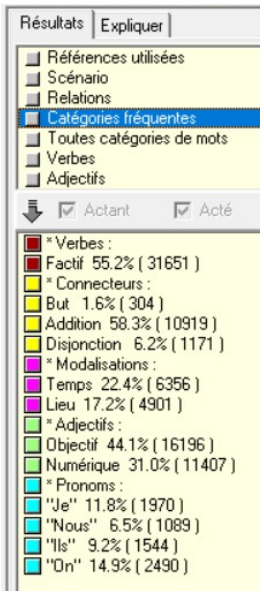

Figure 7: Résultats des catégories fréquentes sur le logiciel Tropes

Avec ces résultats sémantique plutôt complets, Tropes retourne aussi des graphes. Les graphes sont hypertextes, nous pouvons donc cliquer sur un élément pour voir le graphe qui lui est attribué. L'affichage est personnalisable, il suffit de déplacer son curseur pour faire varier le nombre de concepts affichés. Il y a cinq sortes de graphe :

- Le graphe-acteur qui est un graphe en fonction des acteurs sous forme donc de petites sphères nommées par un terme du texte et reliées entres elles si le graphe est une représentation des relations, par exemple. Sur ce graphe, ne sont représentées que les références qui présentent un nombre important de relation.
- Le graphe-aires qui est un graphe sous formes d'aires, c'est-à-dire que ces mêmes sphères sont disposés sur un système d'aires circulaires, comme notre système solaire, et leur surface varie en fonction du nombre de mots qu'elles contiennent. Ce graphe permet d'analyser l'environnement d'une référence ou d'une catégories. Les références sont orientés en fonction de leur place selon le terme central : à gauche il y aura tous les prédécesseurs du terme central et à droite tous les successeurs.
- Les graphes-étoilés, qui représentent un terme ou une relation dans le texte et tous les autres mots auxquels le terme central peut être relié. En d'autres termes, il affiche les relations entre les références ou entre une catégorie de mots et des références. Il affiche aussi la quantité de relation qui existe entre les références. Comme le graphe-aires, ce graphe analyse l'environnement d'une référence ou d'une catégorie et ce, de la même façon.
- Le graphe-répartition qui en fait, un histogramme qui montre la répartition d'une référence, d'une relation ou d'une catégorie de mot. Il est formé par la division du texte en secteurs contenant tous le même nombre de mot. Il représente la fréquence d'apparition du sujet à l'intérieur de chaque secteurs. Une barre représente un secteur et ils sont affichés par ordre chronologique. Ce graphe montre aussi une moyenne des barres grâce à une ligne en pointillés.
- Le graphe-épisodes affiche chaque rafale sous forme d'une ligne de pointillés qui indique son étendue et sa position. Les rafales sont affichées de haut en bas et de gauche à droite. Elles se succèdent selon leur ordre d'arrivée dans le texte. Ce graphe affiche aussi les épisodes sous forme de grands cadres pointillés.

Comme nous pouvons le constater, le logiciel permet une analyse très complète des données textuelles et diverses visualisations des résultats. C'est une des raisons pour laquelle nous nous sommes orientés vers celui. Mais ce n'est pas la seule.

2.1.2.2 Pourquoi Tropes ?

Bien qu'il a été crée en 1994, et qu'il a donc 27 ans, Tropes reste un logiciel très intéressant pour effectuer une analyse de corpus d'articles de presse. Il a montré qu'il été capable d'analyser plus d'une trentaines de fichiers, plutôt conséquents, en même temps sans perdre de sa rapidité. Mais ce n'est pas non plus un des raisons principales du choix de Tropes.

La première raison de ce choix que nous pouvons expliquer est que Tropes est un logiciel français, créer par des français sur la base des travaux du psychologue français Rodolphe Ghiglione. Tout porte donc à croire qu'il a été créer dans le but premier de faire avancer les méthodes et les outils scientifiques de la recherche en France. Il a évidemment été développer dans plusieurs langues, comme l'anglais, l'espagnol, le portugais ou le brésilien, il n'est pas moins qu'il est une garantie pour analyser des textes écrits en langue française, ce qui n'est pas le cas pour certains outils développés dans d'autres pays.

La seconde raison est une question de proximité. En effet, malgré son âge et sa très petite notoriété, le laboratoire GERiiCO a une certaine proximité avec ce logiciel puisque certains de ses chercheurs l'utilisent toujours et il est enseigné aux étudiants des formations en Science de l'Information et du document, notamment dans le Master Veille et Communication de l'Information Stratégique (VeCIS) dans le cadre d'un cours sur l'information et l'aide à la décision où il est question de collecter et d'analyser l'information grâce à la pratique de l'analyse automatique avec le logiciel Tropes puis de la construction de scénario et

d'analyse de corpus. Le but de ce cours étant d'apprendre aux étudiants à utiliser un outil d'analyse de l'information.

Nous arrivons donc à la dernière raison qui constitue ce choix qui est la possibilité de construire un scénario. Comme expliqué dans la partie précédente, la logiciel permet d'utiliser une classification hiérarchique des références pour l'analyse des textes. Nous avons le choix de créer notre propre classification directement avec le logiciel, en choisir un établi au préalable ou de laisser le logiciel le créer lui même à partir des termes des textes. C'est une fonctionnalité qui représente un atout par rapport à d'autres logiciels comme Iramuteq qui propose une unique méthode de classification hiérarchique, pour ne prendre que cet exemple.

Finalement, bien des outils d'analyse sémantique étaient disponibles pour nous aider dans l'analyse de corpus de presse et la création d'une grille d'analyse mais un seul validait tous les critères de sélection pour cette tâche. Sachant aussi que comme il est étudié dans un des Master du département des Sciences de l'Information et du Document, il constituait une opportunité, pour le stagiaire choisi, d'avancer dans l'apprentissage de l'utilisation de cet outil, qu'il aura l'occasion de rencontrer quelques mois plus tard, ainsi que de connaître la bonne méthodologie à utiliser.

2.2 Méthodologie utilisée

Il existe une méthodologie générale de l'analyse de contenu et elle consiste en deux étapes selon André Désiré Robert et Annick Bouillaguet dans le chapitre II de leur ouvrage « L'analyse de contenu », publié en 2007. La première étape consiste à une travail en amont de l'analyse qui est de définir précisément le sujet à traiter et ainsi d'élaborer une problématique autour de ce sujet. Le but de la problématique est d'essayer de traiter sous un nouvel angle l'objet de recherche et d'apporter de nouvelles réponses. La seconde étape de cette méthodologie générale est donc l'analyse de contenu en elle-même qui comporte à son tour plusieurs phases. La première phase de la seconde étape est la pré-analyse qui est déterminer sur quel support établir l'analyse. La seconde phase est celle de la catégorisation. Elle a comme objectif de traiter une première fois les documents à analyser en catégorisant les éléments les plus pertinents du corpus. Ces catégories doivent être pertinentes, exhaustives, exclusives et objectives. La troisième phase est le codage et le comptage des unités et la dernière phase mène l'interprétation des résultats.

Afin de mieux comprendre ce qu'est la méthodologie à adopter dans le cadre d'une analyse de corpus de presse, nous allons définir la méthodologie utilisée pour le cas de l'analyse de corpus de presse pour le projet ANR MémoMines. Nous rentrerons ensuite plus en détails dans cette méthodologie et présenter quels sont les critères qui permettront d'analyser ce corpus de presse, puis terminer sur le sujet de la grille d'analyse qui est la finalité de cette analyse de contenu et, évidemment, du stage.

A mon arrivée en tant que stagiaire sur le projet ANR MémoMines, une partie du corpus à analyser avait déjà été sélectionné et récupéré via le serveur Europresse qui permet aux professionnels de l'information et de la communication de consulter un répertoire régulièrement mis à jour d'articles de presse. Ce serveur est donc un répertoire des sources en ligne de plus de 55 000 sources reconnus. Il a été crée par la société Cision qui est leader mondial des technologies de l'information et de la veille. En plus de la récupération du corpus qui avait déjà commencé, des fichiers indiquant des listes de lieux et d'acteurs, ayant joués un rôle pendant et après les siècles d'exploitation du charbon, avaient aussi déjà été rédigés. Le but de cette analyse était de confirmer ces listes établies, et de les modifier si un nouvel acteur ou un nouveau lieu était retrouvé dans le corpus d'articles de presse. Aussi, un thesaurus a été créée dans le but d'avoir une représentation terminologique du patrimoine minier.

La première tâche qui m'a été attribué dans mon stage a été de prendre connaissance du projet et ses enjeux. Une documentation constituée de documents scientifiques m'a été fourni afin de connaître le contexte dans lequel ma mission de stage s'inscrit. Ensuite, il a fallut compléter le corpus. En effet, les articles de la seconde partie de l'année 2020 et des premiers mois de l'année 2021 n'avait pas encore été collecté. Pour ce faire, j'ai du suivre la méthode de construction du corpus qui m'a été indiqué. Cette méthode consiste à interroger le répertoire Europresse selon trois requêtes. Pourquoi trois requêtes ? Le serveur Europresse ne donne pas la possibilité d'exporter un nombre illimité d'articles c'est pourquoi il a fallut scinder les exportations en trois lots, selon la longueur des articles, c'est-à-dire : les articles courts comprenant entre 100 et 297 mots, les articles moyens entre 304 et 695 mots puis les articles longs. Dans ces premières données, on constate que les intervalles ne sont pas très précises car Europresse ne donne pas de définition stricte de la longueur des articles. Voici donc les requêtes permettant l'exportation des articles :

1. **Articles courts** : TIT_HEAD= (charbon | mines | mine | minier | mineur | fosse) & LG= court & LEAD= ((charbon | mines | mine | minier | mineur | fosse)! ("le fossé" | "au fossé")) ! TEXT= ("école des mines" | antipersonnel | "mine de rien" | "grise mine" | noeux | douchy | bully | auchy | halluin | chine | marles | enquin | carmaux | nomades | monoxyde)
2. **Articles moyens** : TIT_HEAD= (charbon | mines | mine | minier | mineur | fosse) & LG=moyen & LEAD= ((charbon | mines | mine | minier | mineur | fosse)! ("le fossé" | "au fossé"))! TEXT= ("école des mines" | antipersonnel | "mine de rien" | "grise mine" | noeux | douchy | bully | auchy | halluin | chine | marles | enquin | carmaux | nomades | monoxyde)
3. **Articles longs** : TIT_HEAD= (charbon | mines | mine | minier | mineur | fosse) & LG=long & LEAD= ((charbon | mines | mine | minier | mineur | fosse)! ("le fossé" | "au fossé")) ! TEXT= ("école des mines" | antipersonnel | "mine de rien" | "grise mine" | noeux | douchy | bully | auchy | halluin | chine | marles | enquin | carmaux | nomades | monoxyde)

Les mots compris dans les requêtes sont les termes qui devraient figurer dans les articles de presse afin qu'ils soient pertinents. Elles sont larges et ont été créées volontairement dans ce sens. Elles entraînent donc du bruit qu'il faudra réduire en sélectionnant manuellement les articles en effectuant une lecture des titres et des premières lignes des articles. Aussi, il faudra ajouter des filtres dans la recherche, notamment pour définir la source qui est ici La Voix du Nord et faire une recherche selon les années afin d'avoir un document pour une année d'articles.

Suite à la récupération du corpus, tout d'abord sous le format PDF, il a fallut nettoyer les documents. Europresse propose de télécharger les titres sous le format PDF ou le format HTML. Le format PDF occasionnent du bruit, notamment une page de garde avec des notes de la Société Cision, une ou plusieurs pages de sommaires en fonction du nombre d'articles sélectionnés pour une année, certaines métadonnées qui ne sont pas utiles à l'analyse notamment des certificats d'utilisations et autres métadonnées juridiques. C'est là le premier problème que nous avons rencontré. Tout d'abord j'avais personnellement un problème à faire entrer les fichiers PDF sur le logiciel Tropes. Il donc fallut trouver un convertisseurs qui pouvaient supporter une bonne centaine de fichiers à convertir en même temps et convertir correctement. Nous n'avons pas trouvé de convertisseurs qui convenait à nos attentes. Aussi, le nettoyage d'un corpus en format PDF nécessitait de nombreuses manipulations qui nous auraient fait perdre beaucoup de temps car elles n'étaient pas encore totalement connues. Un choix s'est donc posé à nous, soit continuer à pratiquer plusieurs manipulations sur du PDF, soit récupérer tout le corpus en HTML et pratiquer des transformations plus précises et plus rapides sur les fichiers en utilisant le langage XSLT. Nous avons donc opté pour la seconde option. Une fois le corpus récupérer totalement en HTML, nous avons transformés ces fichiers en ne gardant que les éléments pertinents pour l'analyse et dans les formats souhaités. Une première transformation du corpus a été effectuée qui est une transformation des documents HTML en format texte brut ne comprenant uniquement le titre des articles et leur contenu pur.

xsl:stylesheet

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
3      xmlns:xs="http://www.w3.org/2001/XMLSchema"
4      exclude-result-prefixes="xs"
5      version="2.0">
6      <xsl:output method="text" encoding="iso-8859-15"/>
7      <xsl:template match="body/article" name="regle1">
8          ------------------------------------
9          <xsl:for-each select="header/div/p">
10             <xsl:if test="@class='titreArticleVisu rdp__articletitle'">
11                 <xsl:value-of select="."/>
12             </xsl:if>
13         </xsl:for-each>
14         >&#x09;
15         <xsl:for-each select="section/div/div/p">
16             <xsl:value-of select="."/>
17         </xsl:for-each>
18     </xsl:template>
19 </xsl:stylesheet>
```

Figure 8: Script de la transformation en texte brut

Résultats

```
1   html,body{font-family:Arial,'tahoma';font-size:11px;background-color:#fff;margin:5p
2       ------------------------------------
3       Rafraîchissement autour du Mémorial du  mineur... avant son déménagement ?
4       >
5       par Anne-Claire Guilain bethune@lavoixdunord.fr Bruay-La-Buissière. Vous avez s
6       ------------------------------------
7       À l'époque où les pompiers des  mines intervenaient en torpédo
8       >
9       Auberchicourt. Que de visages, aujourd'hui disparus. Le service d'incendie comp
10      ------------------------------------
11      Le bassin  minier et ses artistes inspirent le collectif  Mines de rien
12      >
13      par christophe lecouteux lens@lavoixdunord.fr Hénin-Beaumont. L'un est spéciali
14      ------------------------------------
15      Comment le plateau de « Germinal » est arrivé au Centre historique  minier
16      >
17      par Julien Gilman douai@lavoixdunord.fr lewarde. Le Covid ne réserve pas que de
18      ------------------------------------
19      Accusé de délaisser les affiliés au régime  minier, Filieris répond
20      >
21      Par Élise Forestier lens@lavoixdunord.fr Liévin. Début février, Laurent Duporge,
22      ------------------------------------
23      Grand défenseur du bassin  minier, Gilbert Rolos, ancien maire, est décédé
24      >
25      Par Céline Debette lens@lavoixdunord.fr Sallaumines. Le 30 juin 2012, à Saint-P
26      ------------------------------------
27      Terrils avec vue sur site  minier UNESCO cherchent vaches pour pâturer
28      >
29      Par Anna Morello lens@lavoixdunord.fr Oignies. Cherche six à huit occupants pou
30      ------------------------------------
```

XSL - xml - transfo_txt.xsl (transfo_txt1)

Figure 9: Résultats de la transformation en texte brut

Deux
autres transformations ont été effectuées. L'une en gardant le format HTML et l'autre en basculant sur le format XML. Dans ces deux nouvelles versions, le corpus contient le titre et le contenu des articles, mais aussi le nom de la source et la ville, la date de publication et le numéro de l'article.

```
    version="2.0">
    <xsl:output method="html" encoding="UTF-8"/>
    <xsl:template match="body" name="regle1">
        <html>
            <head>
                <title>Resultats de la transformation</title>
            </head>
            <body>
                <header>
                </header>
                <section>
                    <xsl:for-each select="article">
                        <article>
                            <xsl:for-each select="header/div/p">
                                <xsl:if test="@class='titreArticleVisu rdp__articletitle'">
                                    <h1><xsl:value-of select="."/></h1>
                                </xsl:if>
                            </xsl:for-each>
                            <xsl:for-each select="header/div/span">
                                <xsl:if test="@class='DocPublicationName'">
                                    <h2><xsl:value-of select="."/></h2>
                                </xsl:if>
                                <xsl:if test="@class='DocHeader'">
                                    <h3><xsl:value-of select="."/></h3>
                                </xsl:if>
                            </xsl:for-each>
                            <xsl:for-each select="section/div/div/p">
                                <p><xsl:value-of select="."/></p>
                            </xsl:for-each>
                            <xsl:for-each select="footer/div/div">
                                <xsl:if test="@class='publiC-lblNodoc'">
                                    <p><xsl:value-of select="."/></p>
                                </xsl:if>
                            </xsl:for-each>
                        </article>
                    </xsl:for-each>
                </section>
            </body>
        </html>
    </xsl:template>
</xsl:stylesheet>
```

Texte Grille Auteur

Figure 10: Script de la transformation en HTML

Résultats

```
    <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
    <title>Resultats de la transformation</title>
</head>
<body>
    <header></header>
    <section>
        <article>
            <h1>L'odyssée du  charbon</h1>
            <h2> La Voix du Nord  DOUAI </h2>
            <h3>  dimanche 19 septembre 2004 182 mots </h3>
            <p>Au Centre historique minier de Lewarde, les Journées du Patrimoine, sont consacrées
               cette année au thème des sciences techniques. Ce dimanche, les visiteurs pourront
               accéder gratuitement aux différentes expositions permanentes et temporaires du musée,
               ainsi qu'à deux animations,  Histoire de la fosse Delloye  et  L'Odyssée de la vie
               sur Terre  présentées au cours de cette journée par les médiateurs culturels du Centre.
               </p>
            <p> L'animation  Histoire de la fosse Delloye  entraînera le public dans une promenade
               guidée du carreau où seront notamment évoquées les caractéristiques architecturales
               et les infrastructures d'une fosse des années 1930...Quant à  L'odyssée de la vie
               sur Terre , elle permettra de découvrir les différentes phases de l'évolution de la
               vie sur Terre en parcourant les 455 mètres de l'échelle des temps géologiques géante.
               Les visiteurs pourront apprécier en particulier certaines étapes importantes de cette
               odyssée, comme l'apparition de la vie il y a 3,8milliards d'années, puis les premiers
               vertébrés... pour terminer avec l'hominisation. Par ailleurs, la visite des galeries
               guidée par un ancien mineur sera proposée à demi-tarif soit 5,30Euro(s)E pour les
               adultes et 3,70Euro(s)E pour les enfants, de 9 heures à 19h30 avec la fermeture de
               la billetterie à 17h30. Enfin, les deux expositions temporaires présentées lors de
               ces Journées du Patrimoine sont liées à l'immigration dans le Bassin minier du Nord-Pas-de-Calais:
               Ahmed, Wladislaw, Dario... tous «gueules noires»</p>
            <p>et  tout un spectacle: collection d'affiches polonaises d'Henri Juskowiak.</p>
            <n>news·20040919·VN·20040919571</n>
```

Figure 11: Résultats de la transformation en HTML

Figure 12: Script de la transformation en XML

Après avoir nettoyé le corpus et extrait tous les éléments jugés pertinents pour l'analyse, nous avons entrer le nouveau corpus en texte brut sur Tropes afin de faire une première analyse des résultats retournés par le logiciel. Un nouveau problème est survenu, celui du scénario renvoyé par Tropes. Le scénario proposé par Tropes ne correspondait pas à ce que nous souhaitions retirer de l'analyse. En effet, le premier scénario que nous retourne Tropes est un scénario qu'il a lui-même créer grâce à ses dictionnaires et en fonction des éléments qu'il retrouve dans le corpus. C'est une fonctionnalité très utile lorsque l'on cherche à trouver des éléments inconnus mais dans notre cas, rappelons-le, nous cherchons à confirmer les listes préétablies par les collègues avant le début du stage. Il a donc fallut créer des nouveaux scénarios à partir de scénarios vides pour les lieux et les acteurs, et inscrire directement des codes de structuration dans le fichier du thesaurus. La création de nouveaux scénarios à partir d'un fichier vide nécessite une réflexion sur la structuration. Nous avons choisi pour le scénario des lieux de classer les villes de la région Hauts-de-France en fonction de l'arrondissement dans lequel elles font parties, et du département dans lequel les arrondissements font partis.

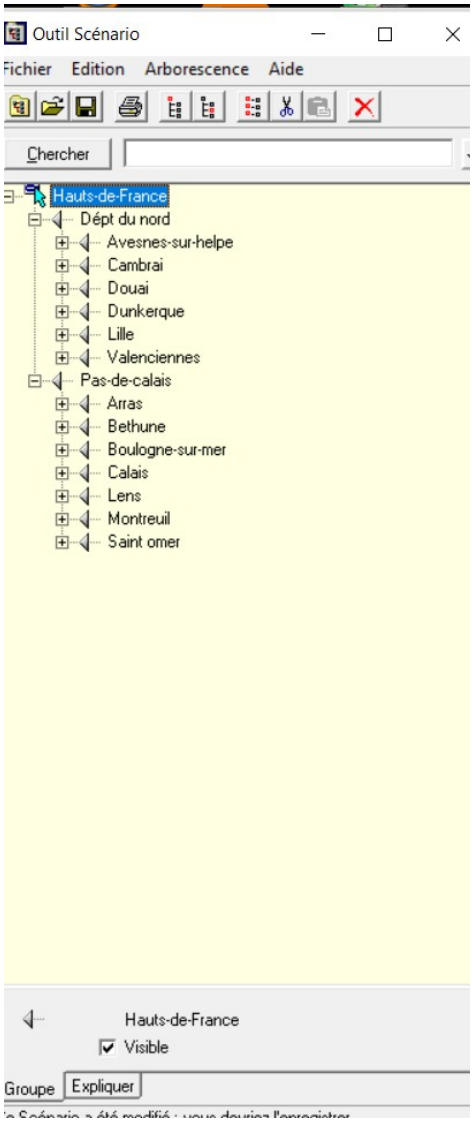

Figure 13: Scénario des lieux 1

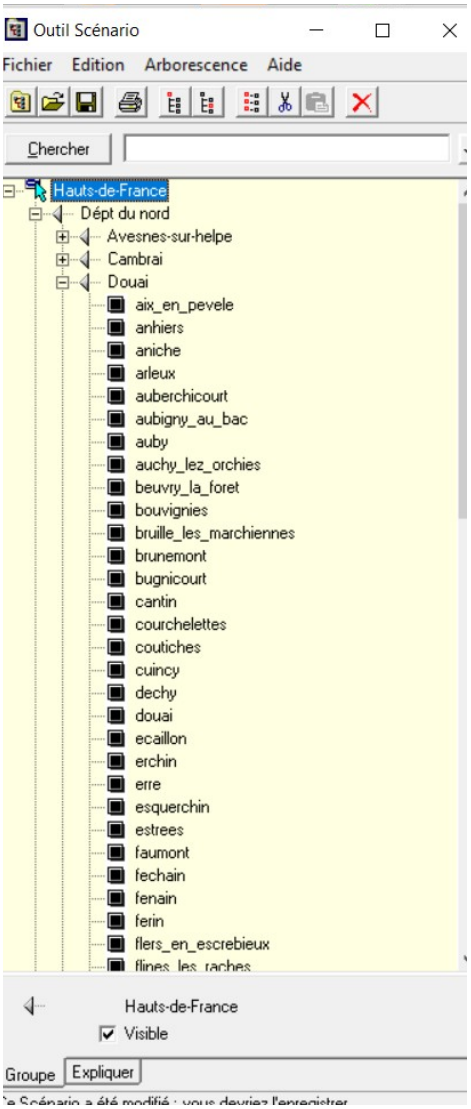

Figure 14: Scénario des lieux avec détail sur les villes

Pour les acteurs, nous avons choisi de structurer le scénario en fonction de si ils sont des associations, des musées ou des sites miniers et pour chacun, en fonction de leur région et de leur département.

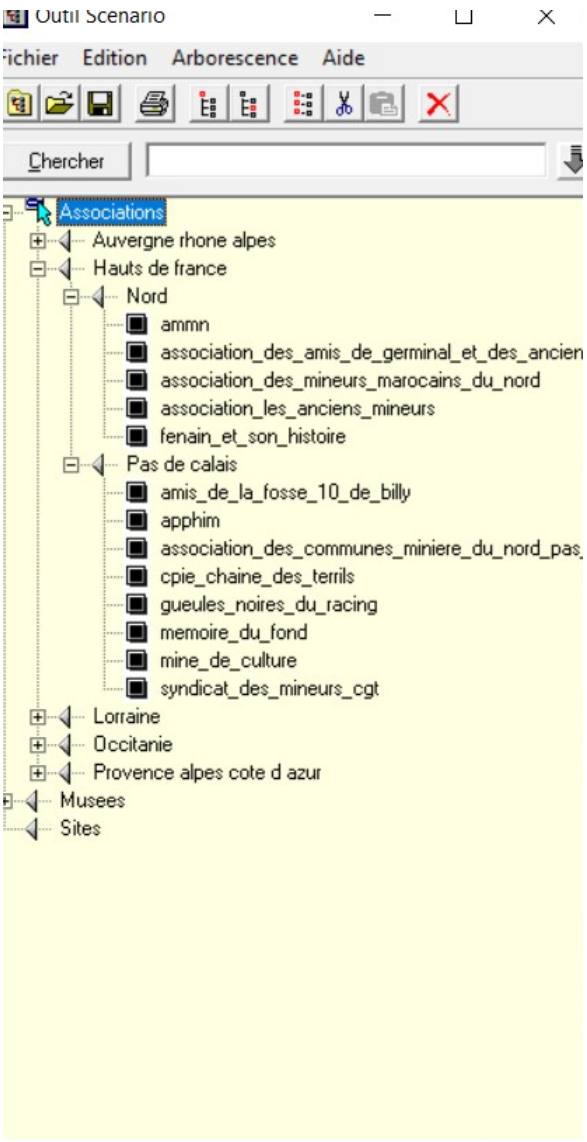

Figure 15: Scénario des acteurs

Le thesaurus a été recopié entièrement sur un fichier.scn. Il était possible aussi d'ouvrir le fichier du thesaurus sur un bloc-note et y ajouter des codes. Les codes à ajouter dans le fichier du thesaurus correspondent à position dans le thesaurus de l'élément codé. Il existe des codes en fonction des niveaux et en fonction de leur nature, soit si un élément est un groupe qui contient d'autres éléments, ou si un élément est un terme uniquement.

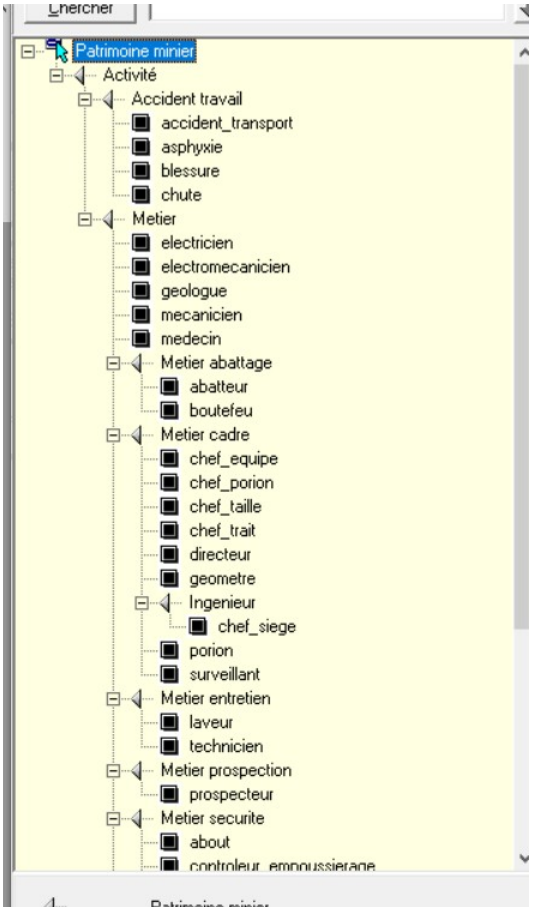

Figure 16: Scénario du thesaurus du patrimoine minier "ThesoMines"

Les scénarios se définissent ici comme des grilles d'analyse du contenu dans le sens où nous avons établit une liste des grandes catégories et des unités correspondantes, qui peut aussi bien servir à un corpus d'articles de presse qu'à d'autres corpus textuels. Ces scénarios, qu'ils soient utilisés seuls ou bien combinés, permettent à l'analyste d'évaluer le corpus de la façon la plus objective possible, bien qu'ils ne suppriment pas totalement la subjectivité. Cela dit, bien que les scénarios permettent de confirmer les listes d'acteurs et de lieux déjà trouvés dans des analyses d'autres corpus, ils ne permettent pas d'en définir de nouveaux. Ils marquent effectivement les unités déjà déterminés au cours du projet mais d'autres unités peuvent encore se cacher dans le corpus.

Pour conclure cette partie sur la méthodologie, nous justifierons le choix de l'utilisation de l'outil Tropes pour l'analyse sémantique par le fait que le logiciel nous donnait accès aussi bien aux résultats que le logiciel obtenait de lui-même qu'à la possibilité de créer nos propres catégories mais surtout nos propres grilles d'analyse. Bien que le logiciel est un dictionnaire plutôt large, les propositions de références ne convenaient pas totalement aux références que nous possédions au préalable, et ne permettaient pas de produire une analyse qui tendait vers ce que nous cherchions. C'est un logiciel très utile qu'il faut continuer d'améliorer notamment peut-être au niveau de la création des scénarios, dans le sens où cela serait plus pratique de pouvoir intégrer un scénario à partir d'un fichier autre qu'avec l'extension « .scn », qu'il est délicat de créer en dehors du logiciel. Cependant, c'est un outil qui favorise énormément l'objectivité dans les analyses, étant donné que l'humain n'est pas spécialement obligé de faire de nombreuses manipulations pour avoir des résultats satisfaisants.