L'environnement vu par ses documents : utilisation de techniques de fouille de textes dans un contexte de description linguistique[1]

Dominic Forest, Hélène Brousseau, Patrick Drouin et Gabriel Bernier-Colborne

Université de Montréal
C. P. 6128, succursale Centre-ville, Montréal (Québec) H3C 3J7, Canada
{dominic.forest, helene.brousseau.1, patrick.drouin, gabriel.bernier-colborne}@umontreal.ca

Abstract

We present a text mining approach used in a project bringing together researchers from the information sciences and linguistics. Its aim is to describe the field of the environment from its documents. We accomplish this by studying a massive corpus acquired by Web crawling including 23 514 records from 1969 websites totalling 47 364 125 tokens. In this communication we account for the text mining portion of this larger project. Using a sample of the corpus, our objective is to extract its thematic structure by applying a non-supervised clustering algorithm (topic model) to identify the hierarchical structure of documents with common themes. Furthermore, our approach is innovative given it allows us the possibility to process complex corpora composed of different textual genres including documents from multiple domains, as well as several types ranging from expert reports, newspaper articles, ideological pamphlets and vulgarization work. From our results, we are able to supply linguists with two specific outputs. Through an interactive dendrogram, we offer a thematic structure of the corpus and using the identification of themes, we are able to supply a series of specialized sub-corpora. Our results prove that using text mining to apprehend large, noisy, Web corpora improves the precision for the latter steps involved in creating linguistic descriptions.

Résumé

Nous présentons l'utilisation d'une approche de fouille de textes dans le cadre d'un projet regroupant des chercheurs en sciences de l'information et en linguistique. L'objectif du projet est la description du domaine de l'environnement à partir d'un important corpus Web de 23 514 documents moissonnés à partir de 1 969 sites Web totalisant 47 364 125 occurrences. L'étape dont nous rendons compte dans cette communication est une première étape de fouille de textes visant à extraire la structure thématique d'un échantillon du corpus en appliquant de façon itérative un algorithme de classification non supervisée afin d'identifier une structure hiérarchique de documents partageant des thématiques communes. Notre approche est novatrice, puisqu'elle permet de traiter un corpus complexe composé de différents genres textuels, dont des documents de multiples domaines et de plusieurs types incluant des rapports d'experts, des articles de journaux, des pamphlets idéologiques et des travaux de vulgarisation. À partir de nos résultats, nous sommes à même d'alimenter le travail de linguistes par la possibilité de visualiser les principaux thèmes sous la forme d'un dendrogramme de thèmes et par la création de sous-corpus spécialisés. Nos résultats démontrent que l'utilisation de cette approche basée sur la fouille de textes comme première étape exploratoire pour appréhender les corpus massifs et bruités du Web améliore la précision des étapes subséquente menant à la description linguistique d'un domaine.

Mots-clés : fouille de textes, classification non supervisée, *topic model*, environnement, linguistique

[1] Ce projet a été réalisée dans le cadre du projet « Comprendre le domaine de l'environnement textuellement et linguistiquement » financé par le Conseil de recherches en sciences humaines du Canada (Programme Savoir, 2013-2018).

1. Introduction

L'environnement est étudié par plusieurs disciplines, telles que les sciences naturelles, les sciences humaines, le droit et la santé. Il s'agit d'un domaine qui évolue rapidement, au gré des avancées technologiques et de l'apparition de phénomènes naturels extrêmes. De plus, les sphères politiques et économiques exercent une forte influence sur ce domaine. Chaque discipline aborde l'environnement d'une façon particulière, teintée par un langage et un champ d'intérêt qui lui sont spécifiques. Par exemple, l'économiste décrit les enjeux environnementaux avec un angle différent de celui du climatologue. On remarque même que certains mots et expressions peuvent avoir différents sens selon la spécialisation de son locuteur.

Dans ce contexte pluriel, nous avons décidé de laisser les documents parler d'eux-mêmes. Alors que certains ont étudié des corpus sur l'environnement par le biais d'outils issus de la lexicométrie et de la socio-informatique (Scotto d'Apollonia et al., 2014), nous avons choisi d'appréhender le nôtre en appliquant des techniques de fouille de textes, tant à des fins d'extraction d'informations qu'à des fins de modélisation thématique pour assister la description linguistique de ce domaine.

Notre démarche s'inscrit dans le cadre d'un projet visant à comprendre le domaine de l'environnement textuellement et linguistiquement. Nous cherchons à assister l'extraction d'une structure thématique à partir d'un corpus de documents afin de contribuer à la description linguistique - fondée sur la sémantique lexicale - du domaine de l'environnement. Nous émettons l'hypothèse que les thèmes identifiés au moyen des méthodes de fouille de textes constituent des noyaux conceptuels importants dans le domaine. Les thèmes sont ensuite utilisés pour amorcer la description linguistique du lexique utilisé dans les textes environnementaux. Les linguistes veulent ainsi créer des descriptions qui s'appuient sur le modèle des *Frame Semantics* (Fillmore, 1982), « des scénarios conceptuels qui fédèrent les unités du lexique qui peuvent les évoquer selon des perspectives différentes » (OLST, 2016).

L'objectif de cette étude est donc de vérifier la pertinence de l'utilisation de techniques de fouille de textes pour informer le travail de description linguistique d'un domaine multidisciplinaire, évolutif et en émergence.

Dans cet article, nous faisons d'abord état des travaux d'extraction des thèmes utilisant des techniques de fouille non supervisées. Nous décrivons ensuite le corpus à l'étude et notre méthodologie. Finalement nous présentons nos résultats, illustrés de quelques exemples de visualisation. Nous concluons en proposant quelques perspectives sur la suite de nos travaux.

2. État des connaissances et positionnement

La découverte de thèmes de façon non supervisée repose principalement sur des méthodes statistiques par lesquelles on mesure la cohésion lexicale entre les segments analysés (Eisenstein & Barzilay, 2008). Plusieurs méthodes sont utilisées, notamment la classification hiérarchique (Anaya-Sánchez et al., 2008; Gil-García & Pons-Porrata, 2010; Huang, Peng, Niu, & Wang, 2011; Pons-Porrata et al., 2007; Zeng et al., 2010), les approches sémantiques (Capasso et al., 2006; Ginter et al., 2009) ou encore les *topic models* (Blei et Lafferty, 2009; Blei, 2012).

L'hypothèse de départ de ces différentes méthodes est que les documents qui ont une proximité thématique partagent un registre lexical similaire (Gärdenfors, 2014; Schwartz et al., 2001). C'est la mesure de la similarité entre les différents documents représentés sous forme de vecteurs de mots

qui permet de faire des regroupements thématiques dans le corpus. À cet égard, plusieurs schémas de pondération peuvent être utilisés préalablement à l'opération de classification automatique. Dans certains cas, c'est simplement la fréquence d'une forme dans le document qui détermine son importance (Bracewell et al., 2009). Dans d'autres cas, on s'appuie sur la valeur de TF-IDF qui vient pondérer la fréquence des termes à l'intérieur d'un document, en mettant en relation cette fréquence avec la fréquence de la même forme dans l'ensemble du corpus (Capasso et al., 2006; Huang et al., 2011; Pons-Porrata et al., 2007).

Selon plusieurs auteurs, une problématique importante lors de l'identification et de la description de thèmes dans un corpus est que la structure thématique peut difficilement être une construction hiérarchique stricte. Certains voient davantage une relation plusieurs à plusieurs, où chaque thème peut se diviser en plusieurs sous-thèmes et où les sous-thèmes peuvent être en lien avec plusieurs thèmes (Bracewell et al., 2009; Schwartz et al., 2001). D'autres imaginent plutôt des catégories souples qui peuvent s'entrecouper (Gil-García & Pons-Porrata, 2010).

L'identification thématique dans la littérature vise principalement le traitement de corpus de nouvelles. En effet, dans le recensement des écrits, plus de la moitié des articles repérés utilisent des corpus de nouvelles, comme Reuters, pour valider les méthodes d'identification thématique proposées. Un autre volet important est celui des corpus de flux d'information en langage naturel, rétrospectif et dynamique, comme les transcriptions d'appel, les notes de rencontres ou les publications sur des microblogues, comme Twitter. Ainsi, on voit un grand intérêt pour le traitement assisté par ordinateur pour des corpus où l'information est abondante, parfois évolutive et souvent constituée à partir de flux d'information en temps réel. L'explosion des contenus publiés sur le Web pousse les chercheurs en fouille de textes à imaginer des outils d'analyse automatique pour pallier la masse documentaire disponible. Étant donné la place importante de ces travaux, une question se pose, à savoir si les méthodes utilisées pour les corpus de nouvelles sont applicables pour un corpus spécialisé dans un domaine tel que celui de l'environnement.

Peu de travaux en identification thématique utilisent des corpus spécialisés. Certains ont traité des corpus dans le domaine de la philosophie (Forest, 2006; Forest & Meunier, 2005; Meunier & Forest, 2008). D'autres corpus spécialisés, tels que le *Fisher Iris*, le *Forensic Glass* et le *Gene Expression*, ou encore, le corpus *Brown*, qui à l'origine a été constitué pour suivre l'évolution de la langue anglaise dans une perspective linguistique (Reza & Matin, 2013), sont utilisés pour mettre à l'épreuve les solutions d'identification thématique (Ghiassi et al., 2012). Les corpus spécialisés ont tendance à être plus homogènes dans leur sujet. Ainsi, l'identification et l'analyse thématique sont souvent plus fines par rapport aux articles de presse. Pour les corpus spécialisés, l'identification et l'analyse thématique sont faites par paragraphes (Reza & Matin, 2013) ou par segments (Forest & Meunier, 2005) plutôt que par documents.

À la lumière des méthodes rapportées dans la littérature, on constate que l'identification et l'analyse thématique en utilisant des algorithmes de fouille de textes sont peu utilisées dans l'analyse linguistique d'un domaine spécifique. En effet, on fait généralement appel à la fouille de textes dans ce type de contexte pour l'organisation et le repérage de l'information.

La particularité de notre approche réside dans le recours à des techniques éprouvées de fouille de textes afin de procéder à l'analyse linguistique d'un corpus issu d'un domaine spécifique.

Le corpus spécialisé sur l'environnement rejoint plusieurs aspects relevés dans la littérature. Comme les corpus de nouvelles, les documents présents dans le corpus sur l'environnement sont multidisciplinaires. Par contre, les multiples champs disciplinaires sont parfois difficiles à discerner puisqu'ils ont tous recours au champ lexical de l'environnement.

De plus, dans l'analyse du corpus sur l'environnement, il est envisageable que certains paramètres spéciaux, comme la provenance géographique et l'aspect temporel des documents, teintent l'extraction des thématiques. Par exemple, un corpus moissonné après la catastrophe nucléaire de Fukushima fera état de façon plus importante de thèmes comme les retombées radioactives et le milieu océanique qu'un corpus qui aurait été moissonné avant cette catastrophe.

3. Méthodologie

3.1. Le corpus

Le corpus analysé dans ce projet a été constitué en 2011 dans le cadre du projet PANACEA (*Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Ressources for Human Language Technologies*) du septième programme-cadre de la Commission européenne (Prokopidis et al, 2012). C'est un corpus unilingue en français composé de documents acquis sur le Web. Il a été construit de façon automatisée. Les documents détectés sont réputés être en français et avoir une appartenance au domaine de l'environnement. Ont été exclus les mots des paragraphes courts, c'est-à-dire, contenant moins de dix mots, les mots des paragraphes ayant été identifiés comme « *boilerplate* » et les mots des paragraphes écrits dans une langue autre que le français. Au final, 23 514 documents sont moissonnés à partir de 1 969 sites Web. Pour des raisons d'efficacité, mais aussi d'évaluation, notre projet a été réalisé sur un échantillon du corpus. Le tableau 1 présente les statistiques descriptives du corpus entier et le tableau 2 présente les statistiques descriptives pour l'échantillon utilisé dans la présente étude.

Tableau 1. Statistiques du corpus entier.

Nombre de documents	23 514
Nombre de sites Web	1 969
Corpus entier sans prétraitement linguistique	
Nombre de mots	47 364 125

Tableau 2. Statistiques de l'échantillon du corpus.

Nombre de documents	2 200
Sans prétraitement linguistique	
Nombre de mots	4 368 664
Nombre de mots différents	85 747
Moyenne de mots par article	1986
Avec prétraitement linguistique	
Nombre de mots	4 353 760
Nombre de mots différents	56 342
Moyenne de mots par article	1979

3.2. Traitement des données

3.2.1. Randomisation et segmentation du corpus

L'ordre des 23 514 documents du corpus a été randomisé pour offrir une représentation distribuée des différents sites Web du corpus. Ensuite, un échantillon de 2 200 documents (environ 10% du corpus) a été retenu et soumis à une opération de classification non supervisée. Il s'agit d'un échantillon aléatoire simple avec une marge d'erreur de 1,99 % et un niveau de confiance de 95 %.

3.2.2. Prétraitement linguistique et options de filtrage des données

Le corpus a été soumis à des traitements linguistiques simples et classiques : normalisation de la casse, lemmatisation simple[2] et utilisation d'un antidictionnaire permettant de retirer les mots usuels de la langue ainsi que les chiffres et les symboles. Nous avons enrichi l'antidictionnaire avec tous les noms de pays et de lieux présents à l'exception de *Kyoto*, *Rio* et *Copenhague*, qui renvoient aux conférences internationales sur le climat. De plus, certaines formes non significatives comme *wikiref* et *URL* ont été enlevées manuellement dans le but d'éviter que le regroupement des documents soit fait en tenant compte du contenant plutôt que du contenu. Après plusieurs essais, nous avons retenu 2 500 formes pour décrire les documents basées sur le facteur de la fréquence. Ce grand nombre de formes assure une représentation de tous les documents. Nous obtenons des termes fortement représentés comme le mot *Environnement*, présent dans 1574 documents sur 2 200, mais aussi des mots moins fréquents, comme *Enchère*, présent dans seulement 16 documents. Lors de l'analyse des sous-corpus, les mêmes paramètres de prétraitement linguistique ont été appliqués. Par contre seulement 300 formes sont sélectionnées en fonction de la fréquence.

3.2.3. Clustering des données

Les documents sont par la suite segmentés par paragraphe (le paragraphe étant ici défini par un simple retour de chariot) puis soumis à un algorithme de classification (*clustering*) hiérarchique puis à une analyse factorielle permettant de décrire chaque regroupement par une suite de termes discriminants. Nous avons soumis à l'algorithme de classification une matrice de 195 566 paragraphes décrits par 2 500 formes. L'objectif de notre démarche étant de vérifier la pertinence d'utiliser des techniques de fouille de textes pour informer le travail de description linguistique, nous avons opté pour une approche hiérarchique à deux niveaux qui permet d'identifier des noyaux conceptuels importants du domaine de l'environnement. Ce choix offre ainsi plusieurs points d'entrées pour effectuer la description linguistique. Chaque niveau est représenté par *n* facteurs, ou *n* thèmes. Le nombre de thèmes est établi selon l'homogénéité des thèmes et des sous-thèmes ainsi que selon la distance entre les vecteurs. Chaque thème est décrit à l'aide de mots clés sélectionnés de façon automatique. Chaque noyau obtenu lors du premier niveau de modélisation thématique crée un nouveau corpus spécialisé, composé de l'ensemble des paragraphes du thème. C'est ce nouveau corpus qui est extrait et soumis à nouveau à l'algorithme lors du second niveau de modélisation thématique (pour un schéma complet voir figure 1).

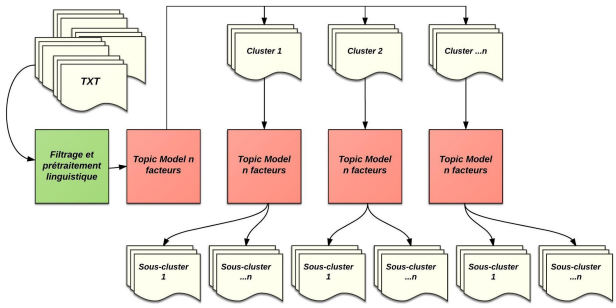

Figure 1. Schéma de l'opération de clustering.

[2] Chaque verbe est rapporté à sa racine et les mots sont rapportés au masculin singulier.

3.2.4. Évaluation des résultats

L'évaluation des résultats de l'opération de classification non supervisée demeure un défi de taille particulièrement pour les corpus n'ayant pas de données de référence (Kessler et al, 2014). En effet, n'ayant pas accès à un corpus annoté entièrement ou même en partie, il n'est pas possible d'évaluer les résultats avec des mesures classiques de précision ou de rappel.

Nous avons choisi d'évaluer les thématiques en fonction de la qualité des résultats obtenus à la seconde étape de description linguistique. Dans le cadre de notre projet, lors de l'étape de description linguistique, les linguistes travaillent à partir de corpus pour faire l'extraction de termes candidats grâce à l'outil TermoStat (Drouin, 2003); TermoStat étant « un outil d'acquisition automatique de termes qui exploite une méthode de mise en opposition de corpus spécialisés et non spécialisés en vue de l'identification des termes » (OLST, 2016). La qualité des résultats est évaluée en appliquant la méthode d'identification de termes à partir d'un corpus composé de paragraphes identifiés comme ayant une appartenance à un thème en opposition au corpus entier. Les résultats sont ensuite comparés aux entrées du DiCoEnviro comme liste de référence. La mesure d'évaluation utilisée est la précision au rang k, c'est-à-dire la précision moyenne de la liste triée de candidats-termes à chacun des rangs où on trouve un terme de la liste de référence.

4. Résultats et évaluation

Au premier niveau, nous avons jugé qu'une configuration à vingt thèmes est optimale à ce stade-ci du travail, car à dix-neuf et moins nous perdions la dimension politique du thème *Relations internationales* tandis qu'à plus de vingt thèmes il y avait trop de sous-divisions thématiques. Chacun des vingt thèmes est redivisé en six à quinze sous-thèmes. L'objectif de la présente démarche étant de nourrir le travail de linguistes, nous utilisons une bibliothèque D3.js pour créer une représentation dynamique des résultats de la classification automatique (figure 2). Cette représentation permet au linguiste de naviguer du général au spécifique à travers les thèmes du corpus dans le but d'identifier des sections thématiques intéressantes pour la description linguistique.

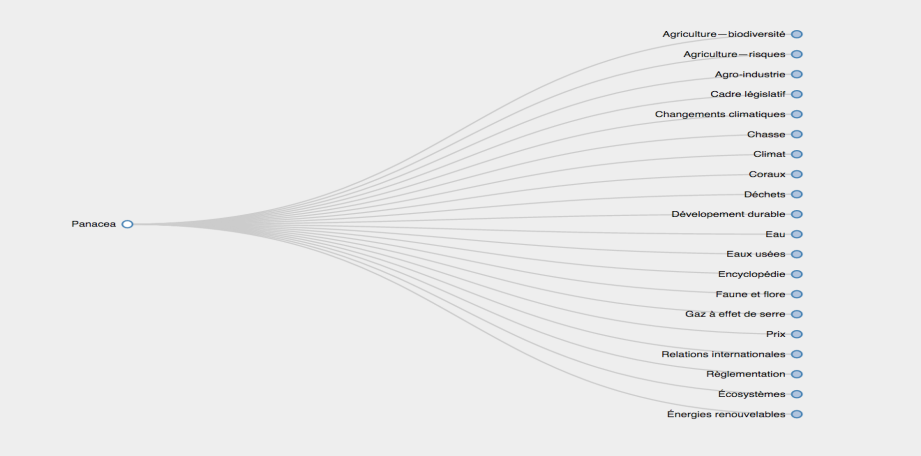

Figure 2. Dendrogramme des thématiques, premier niveau[3].

[3] Une version dynamique du dendrogramme est disponible à l'adresse www.dominicforest.me/environnement/.

Les thèmes obtenus par le moyen de la fouille de textes démontrent bien l'interdisciplinarité inhérente au domaine de l'environnement. En effet, on y trouve par exemple la dimension politique avec *Relations internationales*; les sciences naturelles présentent dans de nombreux thèmes (dont *Faune et flore*, *Coraux* et *Écosystèmes*) des thèmes qui lient l'environnement à l'humain comme *Déchets*, *Eaux usées* et *Agro-industrie*; ou encore, la technologie et l'innovation caractérisées par des thèmes comme *Développement durable* et *Énergies renouvelables.*

Chaque nœud thématique est décrit par une liste de mots qui lui sont caractéristiques et à partir desquels le *topic model* est construit (tableau 3). Les mots clés obtenus par les méthodes de fouille de textes sont utiles pour établir et confirmer les relations entre les mots. En plus d'offrir une description, chaque nœud est un sous-corpus disponible pour la description linguistique.

Tableau 3. mots clés des thématiques de premier niveau.

Nom du thème construit manuellement	**Mots clés retenus par le *topic model***
Agriculture-biodiversité	Cultiver, Diversité, Dériver, Génétique, Hybride, Migration, Mutation, Pur, Semence, Sélection, Variété, Évolutif
Agriculture-risques	Légume, Fruit, Alarmant, Test, Résider, Baser
Agro-industrie	Huile, Palme, Soja, Produit, Sucre
Cadre législatif	Alinéa, Application, Arrêter, Article, Autorisation, Conformément, Disposition, Décret, Loi, Ordonnance, Présent, Relatif, Règlement, Vertu, Visée
Changements climatiques	Climatique, Changement, Réchauffement, Adaptation, Lutte
Chasse	Chasseur, Fédération, Chasse, Départementale, Accident, Déchet
Climat	Intergouvernemental, Expert, Groupe, Climat, Évolution, Nord
Coraux	Récifs, Corallien
Déchets	Dioxide, Déchet, Oxyde, Carbone, Azote, Méthane, Combustion, Métal, Ozone, Combustible, Lourd, Radioactif, Stockage
Développement durable	Développement, Durable, Mise, Oeuvre, Gestion, Collectivité, Territorial, Aménagement, Place, Environnement, Public, Local
Eau	Nappe, Phréatique, Eau, Bassin, Rivière, Souterrain, Lacs, Versant, Superficiel, Hydrographique
Eaux usées	Épuration, Station, Égout, Eau, Boue, Traitement, User
Écosystèmes	Caribou, Appréhender, Oeuf, Course, Tortue, Progressif, Survivre, Plaine, Tenter, Centaine, Millier, Saison, Disparition
Encyclopédie	Encyclopédie, Atlas, Métier, Faune, Détail, Ménage, Photo, Flore
Énergies renouvelables	Énergie, Renouvelable, Solaire, Éolienne, Hydraulique, Électricité, Photovoltaïque
Faune et flore	Espèce, Faune, Flore, Sauvage, Animal, Milieu, Menacer, Protéger, Diversité, Oiseau, Zone, Végétal, Naturel, Envahissant, Exotique
Gaz à effet de serre	Serre, Gaz, Effet, Émissions, Réduction, Co2, Réduire, Carbone, Dioxide
Prix	Nobel, Médecine, Conjointement, Prix
Réglementation	Quota, Membre, Aéronef, Enchère, Exploitant, État, Communautaire, Conformément, Allocation, Délivrer, Compétent, Système, Gratuit, Commission
Relations internationales	Protocole, Kyoto, Ratifier, Convention, Pays, Accord, Rio

Les mots clés permettent d'établir certaines relations entre le thème et les termes qui s'y rapportent. Prenons le thème *Énergies renouvelables :* les mots clés *Solaire* et *Photovoltaïque* se rapportent à l'énergie produite à partir du soleil et le mot clé *Éolienne* à l'énergie produite par le vent. Par contre, le mot clé *Hydraulique* désigne une branche de la physique qui s'intéresse à la circulation des liquides et donc possiblement à l'énergie produite par la fracturation hydraulique. Ceci révèle que notre approche peut aussi mettre en lumière certaines oppositions quand elles sont très présentes dans les documents à l'étude. Ainsi, par l'extraction des thèmes du corpus avec des méthodes de fouille comme étape préliminaire à la description linguistique, il est possible d'établir certaines relations sémantiques dans le lexique de l'environnement.

Notons que certains thèmes absents du premier niveau apparaissent parfois au deuxième niveau. Par exemple, le thème *Transport* apparaît seulement comme un sous-thème de *Développement durable*. Cela étant, plusieurs problématiques liées à ce sous-thème comme les gaz à effet de serre apparaissent au premier niveau.

Comme le démontre la figure 3, les sous-thèmes obtenus au deuxième niveau d'extraction ont parfois des intitulés identiques aux thèmes de premier niveau. Par exemple, dans le thème *Faune et flore* (premier niveau), on y trouve les sous-thèmes *Changements climatiques* et *Développement durable* (deuxième niveau) dont l'intitulé est également utilisé pour un thème au premier niveau. Loin d'être redondants, ces sous-corpus permettent aux linguistes d'explorer un thème dans une perspective plus générale ainsi que dans une dimension spécifique, permettant d'enrichir les descriptions produites. Par exemple, les mots extraits du sous-corpus *Développement durable* niché sous le thème *Déchet* sont différents que ceux obtenus à partir du sous-corpus *Développement durable* sous le thème *Faune et flore.*

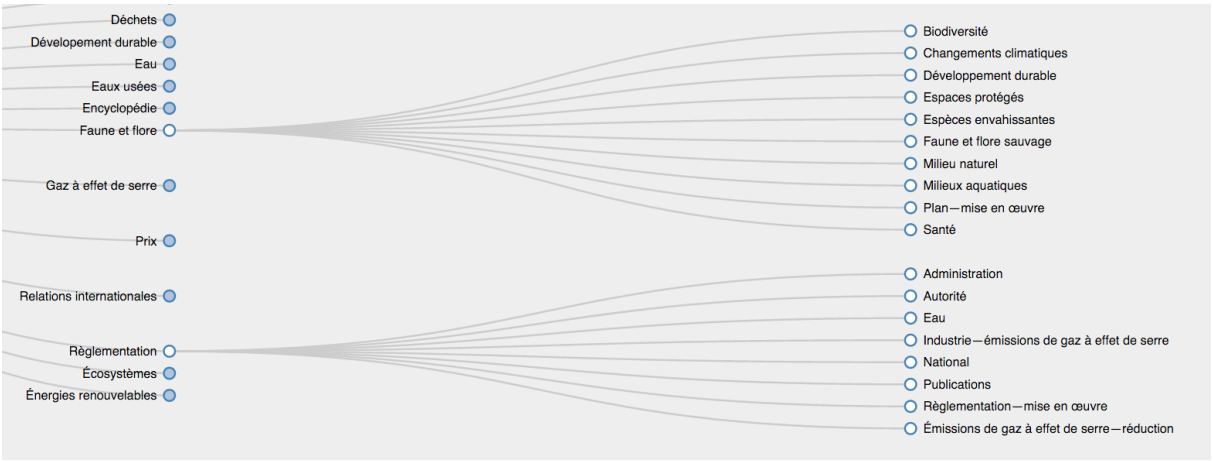

Figure 3. Dendrogramme des thématiques, exemple du deuxième niveau.

En somme, ces résultats démontrent que les techniques de fouille sont utiles pour informer le travail de description linguistique, car ces techniques offrent de nombreux points d'entrée à un corpus très volumineux. Certains thèmes comme *Effet de serre* ou *Développement durable* confirment nos attentes sur les thèmes qui devraient être présents dans un corpus sur l'environnement, tandis que d'autres comme *Chasse* et *Encyclopédie* sont plus surprenants et permettent d'enrichir la description linguistique qui suit l'étape de la fouille.

L'évaluation des résultats est faite sur un échantillon de données à partir du thème *Changements climatique*, un thème présent dans la version actuelle du DiCoEnviro (OLST, 2016), une ressource qui est en cours de développement et qui est construite à partir du même corpus utilisé dans le présent travail. La validation des résultats a été effectuée à partir de 641 unités lexicales associées aux cadres sémantiques des scénarios liés à la thématique des changements climatiques. Cette validation est entièrement automatique et n'a pas fait l'objet d'une validation humaine. Le DiCoEnviro est une ressource en constante évolution et les résultats obtenus sont donc directement liés au contenu de cette ressource lexicale au moment de la rédaction de cet article. Le DiCoEnviro comprend présentement 363 lexies différentes, dont 36 lexies qui se rapportent aux changements climatiques. En d'autres termes, notre processus d'évaluation consiste donc à évaluer la précision de l'extraction de termes sur des corpus ciblés ou non à partir d'un dictionnaire de termes.

Tableau 4. Statistiques du sous-corpus Changements climatiques.

Nombre de documents	1 344
Nombre de paragraphes	7171
Nombre de mots	242 492
Nombre de mots uniques	15 434

La première étape d'évaluation consiste à comparer la précision obtenue sur le sous-corpus sur les changements climatiques (CC) (sous-corpus obtenus grâce au processus de fouille décrit précédemment) à ceux obtenus sur le corpus entier (corpus traitant de l'environnement, mais ne traitant pas exclusivement des changements climatiques). Afin de procéder à une évaluation la plus neutre possible, la comparaison est effectuée sur des corpus de taille comparable à celle du sous-corpus des changements climatiques. Ces sous-corpus de comparaison (GEN1, GEN2, GEN3, GEN4 et GEN5) sont construits aléatoirement et automatiquement à partir du corpus entier afin de faciliter la comparaison avec le corpus spécifique et de minimiser l'effet de la taille du corpus dans les paramètres d'évaluation. Le tableau 5 présente le nombre de mots pour l'ensemble des sous-corpus. Un échantillon des résultats obtenus à partir de chaque corpus est présenté dans le tableau 6.

Tableau 5. Taille des sous-corpus généraux.

GEN1	**GEN2**	**GEN3**	**GEN4**	**GEN5**
239 584	229 452	233 442	237 038	237 919

Tableau 6. Échantillon des résultats obtenus à partir de chaque corpus.

CC	climatique, changement, réchauffement, biodiversité, *lutte, désertification, *adaptation, *serre, écosystème, carbone, pollution, forêt, *europe, *canada, atténuation, *arctique, *kyoto, impact, atmosphérique, gaz
GEN1	*france, forêt, eau, *canada, *québec, *europe, écosystème, environnement, espèce, *déchet, carbone, *etat, *etats, *chap, conservation, *ong, *serre, ressource, *fao
GEN2	biodiversité, *canada, *france, eau, écosystème, forêt, *québec, environnement, *europe, espèce, carbone, *afrique, *etat, *déchet, *chap, *ong, faune, pesticide, *ue, *etats
GEN3	*canada, *france, forêt, eau, *europe, écosystème, *québec, *chap, environnement, espèce, *déchet,

	carbone, *etats, conservation, *etat, *afrique, pesticide, *ii, ressource
GEN4	biodiversité, *france, *canada, eau, *québec, forêt, *europe, écosystème, espèce, environnement, *etat, carbone, *déchet, *etats, *serre, conservation, pesticide, *afrique, *chap, ressource
GEN5	biodiversité, *france, *canada, eau, forêt, *europe, *québec, écosystème, *etat, environnement, espèce, *chap, carbone, *etats, *ii, *déchet, conservation, pesticide, déforestation, pollution

La figure 4 illustre la précision au rang k pour le sous-corpus thématique et les 5 sous-corpus construits automatiquement. On constate rapidement que dans tous les cas, à rang égal dans la liste des candidats termes, la précision obtenue sur le sous-corpus obtenu suite à l'opération de *clustering* est supérieure à celle obtenue à partir du corpus général. L'intérêt de la démarche semble donc confirmé, surtout lorsque le terminologue s'intéresse au début de la liste des candidats puisque c'est à ce moment que le gain de précision est le plus important. Cet écart important se maintient d'ailleurs sur les 200 premiers candidats de la liste pour ensuite diminuer légèrement.

La technique permet donc de vérifier qu'à taille égale, le corpus ciblé thématiquement - identifié grâce au processus de fouille - donne de meilleurs résultats. Le graphique suivant illustre l'évaluation de la précision en fonction du rang dans la liste des candidats termes obtenus lors de l'extraction automatique. Dans tous les cas, la qualité de la liste du corpus ciblé est plus élevée.

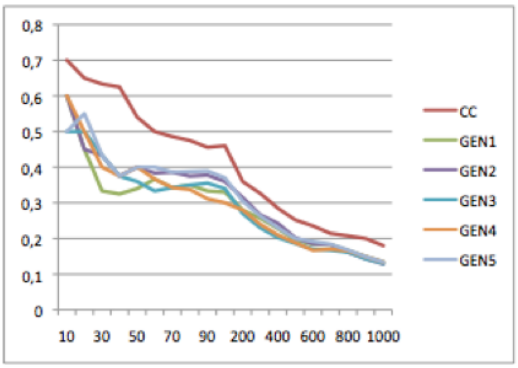

Figure 4. Précision au rang k des listes de candidats termes (corpus de taille comparable).

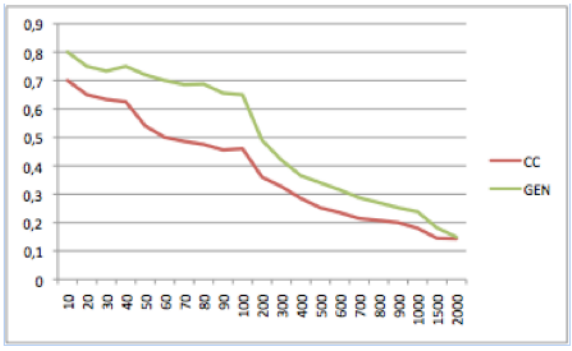

Figure 5. Précision au rang k des listes de candidats termes (corpus entiers).

Cette performance intéressante de l'extraction de termes sur le corpus thématique est encourageante. Elle nécessite tout de même une comparaison à une extraction sur le corpus entier, sans cibler *a priori* une thématique. La figure 5 présente les résultats de cette expérimentation.

Cette seconde comparaison démontre que les résultats validés automatiquement sont moins intéressants sur le corpus thématique que sur le corpus général pris en entier (donc contenant plus de données). Ces résultats peuvent sembler surprenants puisque le corpus thématique devrait conduire à candidats termes qui sont plus directement liés à la thématique recherchée. La liste des candidats termes générés à partir du corpus général contient 7 861 entrées alors que la liste du corpus sur les changements climatiques en comporte 2 096. Sur les 100 premiers candidats, la précision obtenue sur le corpus général intégral est beaucoup plus élevée que sur le corpus thématique; les courbes de précision convergent autour du rang 2 000. Il semble donc que l'approche statistique utilisée au sein de TermoStat (le calcul des spécificités, Lafon 1980) qui bénéficie d'un corpus plus gros permet de faire émerger une terminologie pertinente malgré que le corpus porte sur diverses thématiques.

Les résultats de cette deuxième expérimentation démontrent donc que la constitution d'un sous-corpus ciblé thématiquement n'est peut-être pas l'option la plus intéressante pour le travail terminologique quand le corpus général peut être analysé dans son ensemble. Cependant, étant donné que la taille des corpus ne cesse d'augmenter, il est fort probable qu'un logiciel d'extraction de termes ne puisse prendre en charge le corpus en entier. Dans un tel cas, la première expérimentation démontre qu'il vaut mieux cibler une thématique à l'aide de techniques de fouille de textes en vue de mettre sur pied un sous-corpus lié au domaine d'intérêt.

5. Conclusion et perspectives

Cet article fait état de l'utilisation de techniques de fouille de textes dans le cadre d'un projet interdisciplinaire visant la création de descriptions linguistiques du domaine de l'environnement à partir d'un grand corpus Web. Nos résultats indiquent que les techniques de fouille permettent non seulement de cibler des noyaux conceptuels importants, mais aussi d'établir certaines relations sémantiques entre les thèmes, les sous-thèmes et les descripteurs. D'autre part, notre approche basée sur des algorithmes de fouille de textes permet la création de sous-corpus thématiques ciblés à partir desquels l'extraction automatique de termes génère de meilleurs résultats que ceux obtenus à partir d'un corpus général. Par ailleurs, les techniques de fouille de textes utilisés dans ce contexte d'un grand corpus Web bruité permettent également d'éliminer les thématiques qui ne sont pas pertinentes pour la description linguistique du domaine. Suite à nos expérimentations, plusieurs pistes sont envisagées. Premièrement, nous entendons déployer la méthodologie à l'ensemble des 23 514 documents du corpus *Panacea* et procéder à une évaluation qualitative des résultats.

6. Références

Anaya-Sánchez, H. et al. (2008). A new document clustering algorithm for topic discovering and labeling. In Lazo, M. et Sanfeliu, A. éditeurs, Progress in Pattern Recog, Image Analysis and App. Springer.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4) : 77-84.

Blei, D. M. et Lafferty, J. D. (2009). Topic Models. In Srivastava, A. et Sahami, M. éditeurs, Text mining: classification, clustering, and applications. CRC Press.

Bracewell, D. et al. (2009). Category Classification and Topic Discovery of Japanese and English News Articles. Electronic Notes in Theoretical Computer Science, 225(0) : 51-65.

Capasso, P. et al. (2006). A Semantic Topic Identification System for Document Retrieval on the Web. In Proc. of Web-Age Information Management Workshop 2006, page 22.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. Terminology, 9(1) : 99-115.

Eisenstein, J. et Barzilay, R. (2008, octobre). Bayesian unsupervised topic segmentation. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 334-343.

Fillmore, C. (1982). Frame semantics. In Linguistics in the morning calm, 111-137.

Forest, D. (2006). Application de techniques de forage de textes de nature prédictive et exploratoire a des fins de gestion et d'analyse thématique de documents textuels non structurés. Thèse de doctorat. UQÀM.

Forest, D. et Meunier, J.-G. (2005). NUMEXCO: A Text Mining Approach to Thematic Analysis of a Philosophical Corpus. CH Working Papers, 1(1).

Gärdenfors, P. (2014). Conceptual spaces. Cambridge (Mass.) : MIT Press.

Ghiassi, M. et al. (2012). Automated text classification using a dynamic artificial neural network model. Expert Systems with Applications, 39(12) : 10967-10976.

Gil-García, R. et Pons-Porrata, A. (2010). Dynamic hierarchical algorithms for document clustering. Pattern Recognition Letters, 31(6) : 469-477.

Ginter, F. et al. (2009). Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. International Journal of Medical Informatics, 78(12) : e1-e6.

Huang, S. et al. (2011). News topic detection based on hierarchical clustering and named entity. Proc. of NLP-KE, pages 280-284.

Kessler R., Forest D. et Laplante A. (2014). Encore des mots, toujours des mots : fouille de textes et visualisation de l'information pour l'exploration et l'analyse d'une collection de chansons en français. In Née E., Daube J., Valette M. et Fleury S., editeurs, Actes des 12es Journées internationales d'Analyse Statistique des Données Textuelles, pages 311-322.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, Mots, 1, 127-165.

Meunier, J. G. et Forest, D. (2008). Computer assisted conceptual analysis (CACAT): the concept of Mind in the Collected Papers of C.S. Peirce. Proc. of DH 2008, pages 74-80.

Observatoire de linguistique Sens-Texte (OLST). (2016). http://olst.ling.umontreal.ca

Prokopidis, P. et al. (2012). Final report on the corpus acquisition & annotation subsystem and its components. Rapport interne WP-4.5, PANACEA Project.

Pons-Porrata, A. et al. (2007). Topic discovery based on text mining techniques. Information Processing & Management, 43(3) : 752-768.

Reza, M. F. et Matin, R. (2013). Application of data mining for identifying topics at the document level. Proc. of Informatics, Electronics & Vision (ICIEV) 2013 International Conference.

Schwartz, R. M. et al. (2001). Unsupervised topic discovery. Proc. of Lang. Models for Info. Retrieval Workshop.

Scotto d'Apollonia L. et al. (2014). Approche lexicométrique des controverses climatiques. In Née E., Daube J., Valette M. et Fleury S., éditeurs, Actes des 12es Journées internationales d'Analyse Statistique des Données Textuelles, pages 605-616.

Zeng, J. et al. (2010). Multi-grain hierarchical topic extraction algorithm for text mining. Expert Systems with Applications, 37(4) : 3202-3208.