Chronique d'un échec : identification des métaphores dans les écrits des géographes

Suzanne Mpouli*

* *HTL, Université de Paris, CNRS, F-75013 Paris, France*
suzanne.mpouli@u-paris.fr

RÉSUMÉ. *La métaphore présente un intérêt indiscutable pour étudier en diachronie l'évolution des idées dans les textes scientifiques relevant des sciences humaines et sociales. Cependant, malgré les différentes méthodes proposées en pour détecter automatiquement les métaphores, très peu de travaux de recherche ont essayé de les appliquer à ce genre de textes. Dans cet article, nous présentons une tentative d'identification des métaphores conceptuelles dans des textes de géographie en français et en anglais qui utilise une méthode reposant sur LDA (Heintz* et al.*, 2013). Si la méthode testée s'avère, à l'issue de nos expérimentations, inadéquate pour notre objectif final, elle nous a cependant permis de cibler les difficultés inhérentes à ce type de projet ainsi que de futures perspectives de recherche.*

ABSTRACT. *Metaphors are often perceived as being essential to the diachronic study of ideas formulated in scientific texts pertaining to social sciences. However, very few research endeavours have tried to apply any existing NLP metaphor detection method to such texts. The present article describes an attempt to identify conceptual metaphors in geography texts written in English and in French using an LDA-based method (Heintz* et al.*, 2013). Although that particular method was ultimately unsuitable for our final goal, it enabled us to circumscribe the specific challenges inherent to this type of project as well as future research perspectives.*

MOTS-CLÉS : *métaphore, géographie, topic modelling (LDA), fouille de textes, humanités numériques.*

KEYWORDS: *metaphor, geography, topic modelling (LDA), text mining, digital humanities.*

1. Introduction

Aucune figure de style n'a suscité tant de fascination ni fait couler autant d'encre que la métaphore, qualifiée à juste titre de « figure des figures » (Deguy, 1969). Plus qu'un simple ornement du langage, elle devient avec Lakoff et Johnson (1980), le fondement même de notre système de pensée : la métaphore nous permet de mieux décrire et appréhender les multiples phénomènes abstraits qui nous entourent à l'instar des idées, des mouvements et du temps. On peut ainsi, grâce à une métaphore, projeter de manière sélective certains traits d'un *domaine source*, typiquement assez concret sur un *domaine cible*, généralement plus abstrait.

Si la rhétorique étudie principalement la métaphore comme un ornement du langage, elle la différencie en premier lieu des autres figures de style par sa capacité à doter temporairement un terme d'une nouvelle signification. Ainsi, Aristote (1922) la définit comme une figure où il s'opère un « transfert par analogie » d'un mot à un autre tandis que Dumarsais (1818) précise qu'elle s'effectue « en vertu d'une comparaison qui est dans l'esprit ». La métaphore reposant sur une comparaison, elle est généralement, dans les ouvrages de rhétorique, mise en parallèle avec une autre figure de style reposant sur le même procédé : la comparaison figurative. Celle-ci, cependant, au contraire de la métaphore établit explicitement une comparaison entre des unités lexicales au moyen d'un terme de comparaison, traditionnellement « comme » en français. À titre d'illustration, Aristote (1922) oppose la métaphore « **Ce lion** s'élança » à la comparaison « [Achille] s'élança comme **un lion** » et conclut que cette dernière est moins agréable car plus longue. Du point de vue structurel, on a dans les deux phrases un comparé « Achille » et un comparant « lion » ; le comparé étant absent dans la métaphore, on parlera d'une métaphore *in absentia*. Le domaine cible mentionné étant « l'homme » et le domaine source « l'animal », la théorie de la métaphore conceptuelle (Lakoff et Johnson, 1980) classera ces deux phrases sous la métaphore conventionnelle UN HOMME EST UN ANIMAL.

Comme l'exemple de métaphore donné ci-dessus attribué à Homère, bon nombre d'exemples que proposent les rhétoriciens afin de discuter de l'esthétique du langage sont empruntés à des auteurs de textes littéraires, déjà reconnus pour leur maniement décrié ou admiré mais néanmoins toujours singulier de la langue. Commentant l'emploi des mots chez les poètes grecs anciens, Aristote (1922) considère l'usage juste des métaphores comme la qualité primordiale qu'un auteur doit posséder et la seule véritable marque d'un talent indéniable. De par leur pouvoir évocateur et le lien qu'elles entretiennent avec l'imaginaire, les métaphores constituent des outils de choix sur lesquels les auteurs peuvent non seulement innover du point de vue linguistique, marquer l'esprit de leurs lecteurs, mais aussi mettre l'accent sur des thèmes, des émotions ou des événements particuliers de leur récit. Si la présence des métaphores est attendue dans un texte littéraire, leur emploi dans d'autres types de discours, notamment le discours scientifique, a parfois été critiqué.

Cette condamnation de l'utilisation des métaphores dans le discours scientifique découle vraisemblablement à la fois de leur faculté d'ouvrir les portes de l'imaginaire

et de leur statut d'accessoire du langage. En effet, pour Bachelard (1967), par exemple, l'histoire de chaque science se caractérise par un dépouillement de la langue de toute métaphore et analogie, obstacles à l'accès à la vraie connaissance. Néanmoins, dans la pratique, les métaphores restent omniprésentes dans les textes scientifiques dans lesquels elles assurent non seulement un rôle didactique mais aussi créent des cadres de référence quasi universels (Molino, 1979). Ceci se vérifie clairement dans les trois types de fonctions qu'Ascher (2005) distingue dans les métaphores utilisées par les géographes :

– la fonction pédagogique, qui a surtout une valeur illustrative, se borne à constater une ressemblance et n'établit aucun développement analogique plus poussé.
Exemple : L'homme est **un loup** pour l'homme ;

– la fonction heuristique où l'analogie est plus exploitée et facilite l'analyse de phénomènes abstraits en mettant en exergue leurs similitudes avec d'autres phénomènes.
Exemple : Le concept de « modernité **liquide** » chez Bauman qui souligne les changements constants de l'ère actuelle dominée par des technologies à courte durée de vie et toujours en évolution ;

– la fonction modélisatrice où la métaphore pose les bases d'un modèle qui traduit les réflexions théoriques de l'auteur et qui étaye son argumentation.
Exemple : La métaphore textile du réseau social dans laquelle la société est assimilée à un filet constitué de liens.

Hormis les différents rôles qu'elles servent, les métaphores en géographie revêtent un intérêt majeur du point de vue épistémologique : en effet, depuis l'Antiquité, le courant qualitativiste copie l'écriture littéraire, produisant ainsi des textes dans lesquels la métaphore joue un rôle essentiel (Lévy, 2006). Cependant, les divers travaux qui se sont attelés à analyser et catégoriser les métaphores dans le discours des géographie se sont parfois limités à un type spécifique de métaphore, comme la métaphore organiciste (Bachimon, 1979 ; Berdoulay, 1982 ; Archer, 1993), ou alors ont dressé un panorama assez large des métaphores utilisées dans un sous-domaine de la géographie (Daniels et Cosgrove, 1993), souvent sans trop s'attacher à la dimension chronologique et sa signification. De fait, si l'on peut dire avec certitude que le vivant, le théâtre, le textile, la physique, les mathématiques, la mécanique ou encore l'écologie ont servi de domaines sources à des métaphores géographiques, il apparaît plus compliqué de recenser tous les auteurs qui y font référence et tous les termes qu'ils convoquent. Le projet GÉONUM se propose de combler ce vide, d'une part en s'appuyant sur la théorie de la métaphore conceptuelle ainsi que sur des méthodes de traitement automatique des langues et, d'autre part, en se focalisant sur des métaphores dans lesquelles la géographie est utilisée comme domaine cible en conjonction avec un domaine source prédéfini.

Au vu de la nécessité de ne pas se limiter à une structure syntaxique particulière de métaphores (par exemple, métaphores adjectivales ou verbales) et de pouvoir identifier les domaines sources et cible, la méthode de Heintz *et al.* (2013) nous est apparue comme celle qui correspondait le mieux aux besoins de ce projet. Dans la section

suivante, nous présentons en détail les particularités et les différentes étapes de cette méthode. Le code utilisé pour les expériences rapportées dans l'article de référence n'étant pas disponible, nous avons dû entièrement réimplémenter le système quasi à l'identique pour le tester sur des données en anglais. Dans la section 3, nous revenons sur les résultats de cette expérimentation et procédons à une analyse des principales erreurs que nous avons relevées. La section 4, quant à elle, porte sur les modifications apportées à la méthode initiale suite à nos premières expérimentations et sur les résultats obtenus sur des données en français. Enfin, dans la section 5, nous jetons un regard critique sur notre approche avant de conclure.

2. Description de la méthode de détection automatique des métaphores choisie

De manière générale, la métaphore du point de vue computationnel est assimilée à un écart sémantique créé par l'association de termes partageant peu, voire aucun sème. Elle peut donc se fonder sur la violation des préférences sémantiques (Fass, 1991 ; Kintsch, 2000 ; Krishnakumaran et Zhu, 2007), l'opposition concret abstrait (Turney *et al.*, 2011 ; Tsvetkov *et al.*, 2014) ou l'utilisation de termes appartenant à des domaines sémantiques distincts (Schulder et Hovy, 2014 ; Shutova *et al.*, 2017). C'est dans cette dernière famille de méthodes que rentre la méthode de Heintz *et al.* (2013) retenue pour nos expérimentations qui, en plus d'être applicable à plusieurs langues, a été conçue dans une optique sensiblement identique à la nôtre : détecter dans des textes bruts des métaphores ayant un domaine cible spécifique GOVERNANCE combiné à des domaines sources prédéfinis.

Contrairement à d'autres méthodes de détection de métaphores qui utilisent les thématiques extraites d'un corpus au moyen d'algorithmes de *topic modelling* comme un paramètre pour l'apprentissage automatique (Klebanov *et al.*, 2009 ; Bethard *et al.*, 2009 ; Klebanov *et al.*, 2014 ; Jang *et al.*, 2015), cette méthode s'en sert plutôt pour retrouver les mots appartenant à chaque domaine présélectionné. Plus concrètement, cette méthode présuppose qu'une phrase est métaphorique si un domaine source et un domaine cible prédéterminés font partie de ses thématiques prédominantes. Ainsi, dans une phrase telle que « *Moderates, we all hear, are an* ***endangered species*** », prévalent le domaine source, *Animaux* et le domaine cible, POLITIQUE auxquels sont respectivement rattachés les termes *endangered species* et *Moderates*.

Côté ressources, cette méthode non supervisée requiert une liste de domaines sources et cible, des articles de Wikipédia dans la langue de travail et une implémentation de *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003). Ce dernier point est sans aucun doute l'atout majeur de cette méthode. Très utilisés en recherche d'information et en fouille de textes, les modèles thématiques constituent une famille de modèles probabilistes génératifs qui permettent non seulement de faire émerger les mots appartenant aux thématiques présentes dans une collection de documents mais aussi de prédire quelles thématiques cette collection partage avec d'autres.

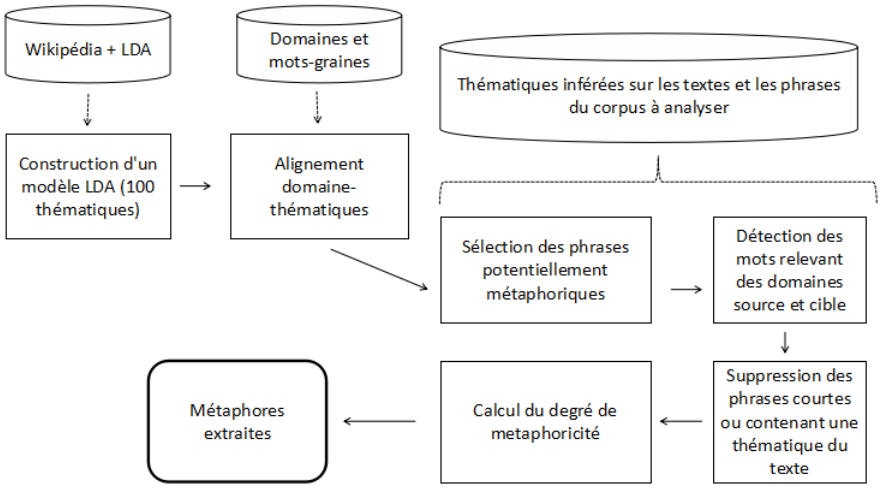

Figure 1. *Schématisation de la méthode de Heintz* et al. *(2013)*

Initialement, la méthode de Heintz *et al.* (2013) a été testée sur un ensemble de textes en anglais et en espagnol avec des résultats satisfaisants : les auteurs rapportent une F-mesure de 59 % pour les expérimentations en anglais. Dans la suite de la section, nous détaillerons les principales phases de cette méthode schématisées dans la figure 1.

2.1. *Phase manuelle autour des domaines cible et sources*

Il s'agit ici d'établir une liste des domaines cible et sources que l'on veut détecter, à raison d'un seul et unique domaine cible et de plus d'un domaine source. Heintz *et al.* (2013) ont défini 62 domaines sources ; nous en avons retenu 42 en concertation avec des spécialistes de l'épistémologie de l'écriture des géographes. Ensuite, pour chaque domaine, nous avons proposé des mots-graines, c'est-à-dire des termes qui, à notre sens, font typiquement partie du champ lexical de ce domaine. Heintz *et al.* (2013) proposent au maximum 4 mots-graines par domaine. Ainsi, un domaine comme *Art* devrait contenir des termes comme « musée », « art », « peinture » et « culture ». Cependant, pour prendre en compte la diversité des sous-disciplines existant en géographie, nous avons suggéré 9 mots-graines pour notre domaine cible. Le tableau 1 recense l'ensemble des domaines ainsi que les mots-graines choisis pour chacun d'eux dans les deux langues de travail ; nous avons indiqué entre parenthèses l'équivalent du mot-graine en anglais lorsque celui-ci diffère du français.

Domaines	Mots-graines
Art	peinture (*painting*), art, musée (*museum*), culture
Astronomie	astronomie (*astronomy*), étoiles (*stars*), planètes (*planet*), télescope (*telescope*)
Barrière	barrière (*barrier*), grillage (*fence*), clôture (*closing*), frontière (*frontier*)
Biologie	biologie (*biology*), espèces (*species*), nature, gène (*gene*)
Chimie	chimie (*chemistry*), acide (*acid*), enzyme, oxygène (*oxygen*)
Compétition	champion, gagner (*win*), course (*race*), médaille (*medal*)
Construction	bâtiment (*building*), pièce (*room*), mur (*wall*), édifice (*house*)
Corps	corps (*body*), tête (*head*), mains (*hands*), os (*bones*)
Créatures	titan, dragon, monstre (*monster*), sorcière (*witch*)
Désordre	fouillis (*jumble*), désordre (*disorder*), enchevêtrement (*tangle*), chaos
Dynamique	mouvement (*going*), bouger (*move*), avancer (*forward*), progrès (*progress*)
Économie	économie (*economy*), argent (*money*), financier (*financial*), banque (*bank*)
Éducation	éducation (*education*), école (*school*), étudiant (*student*), université (*university*)
Émotion	joie (*joy*), colère (*anger*), peur (*fear*), amour (*love*)
Expansion	fertilité (*fertility*), descendant (*offspring*), reproduire (*breed*), développer (*expand*)
GÉOGRAPHIE	territoire (*territory*), terre (*land*), environnement (*environment*), montagne (*mountain*), rivière (*river*), désert (*desert*), climat (*climate*), population, société (*society*)
Guerre	guerre (*war*), bataille (*battle*), armée (*army*), assaut (*attack*)
Langage	langage (*language*), verbe (*verb*), mot (*word*), alphabet
Livre	écrit (*writing*), livre (*book*), page, imprimer (*print*)
Maladie	cancer, guérison (*healing*), douleur (*pain*), maladie (*disease*)
Mathématiques	fonction (*function*), centre (*center*), matrice (*matrix*), cercle (*circle*)
Mécanique	rouage (*gear*), machine, chaîne (*chain*), panne (*breakdown*)
Météorologie	vent (*wind*), tempête (*tempest*), rafale (*blow*), pluie (*rain*)
Meuble	meuble (*furniture*), table, tapis (*carpet*), lit (*bed*)
Monarchie	roi (*king*), reine (*queen*), royaume (*kingdom*), empereur (*emperor*)
Mort	mourir (*die*), enterrement (*burial*), tombeau (*grave*), disparu (*extinct*)
Musique	rythme (*rhythm*), musique (*music*), concert, jazz
Nourriture	nourriture (*food*), viande (*meat*), manger (*eat*), boire (*drink*)
Organe	organe (*organ*), poumons (*lungs*), cœur (*heart*), cerveau (*brain*)
Physique	gravité (*gravity*), énergie (*energy*), nasa, satellite
Politique	gouvernement (*government*), président (*president*), élections (*elections*), politique (*politics*)
Prison	prison, crime, coupable (*culprit*), meurtre (*murder*)
Religion	paradis (*paradise*), saint, dieu (*god*), église (*church*)
Richesse	richesse (*wealth*), bijoux (*jewels*), trésor (*treasure*), riche (*rich*)
Sens	ouïe (*hearing*), vue (*seeing*), doux (*sweet*), amer (*bitter*)
Sport	tennis, football, athlétisme (*athletics*), champion
Statique	arrêté (*stopped*), figé (*rooted*), fixe (*fix*), rester (*stay*)
Technologie	technologie (*technology*), ordinateur (*computer*), internet, software
Théâtre	acteurs (*actors*), jouer (*play*), scène (*stage*), rôle (*role*)
Tribunal	tribunal (*court*), droit (*law*), avocat (*lawyer*), légal (*legal*)
Véhicule	véhicule (*vehicle*), essence (*gas*), voiture (*car*), train
Vêtement	veste (*coat*), plier (*fold*), porter (*wear*), déchirer (*tear*)
Voyage	voyage (*trip*), séjour (*journey*), voyager (*travel*), touristes (*tourists*)

Tableau 1. *Domaines et mots-graines sélectionnés pour nos expérimentations*

2.2. *Extraction des thématiques*

On commence par dresser une liste de mots vides, définis comme étant les 500 mots les plus fréquents du corpus d'articles issus de Wikipédia, afin de pouvoir retrancher ces mots dans la suite du processus. Puis, LDA est utilisée sur ce corpus pour extraire 1 000 thématiques en itérant 1 000 fois et en optimisant les hyperparamètres toutes les 100 itérations. Pour cette étape, la méthode d'origine utilise MALLET (McCallum, 2002)[1], librairie dont nous nous sommes également servis ensuite pour inférer les thématiques de chaque texte et de chaque phrase une fois le modèle LDA construit.

2.3. *Alignement des domaines aux thématiques extraites*

LDA identifiant les différentes thématiques par des nombres aléatoires, le but de cette étape est de pouvoir associer de manière pertinente une ou plusieurs thématiques à un unique domaine source ou cible. Pour ce faire, on considère comme les thématiques liées à un domaine, les n thématiques dans lesquelles la somme de probabilités LDA des mots-graines d'un domaine particulier est maximale pour une thématique et supérieure à la valeur d'un seuil prédéfini z_{align}. Plus formellement, soient t une thématique, c un domaine à aligner et $K(c)$ l'ensemble des mots-graines w associés à ce domaine, t et c sont alignés si : $\sum_{w \in K(c)} p(w|t) > z_{\text{align}}$

Dans l'article de Heintz *et al.* (2013), au maximum 3 thématiques sont alignées avec un domaine source et au maximum 5 avec le domaine cible tandis que le seuil z_{align} est fixé à 0,01. De plus, grâce à cet alignement, on peut également lier les thématiques identifiées précédemment dans chaque phrase et texte aux domaines sources et cible correspondants.

2.4. *Sélection des phrases potentiellement métaphoriques*

En accord avec l'hypothèse de départ, on cherche à déterminer ici les phrases qui parlent à la fois du domaine cible et de l'un des domaines sources prédéfinis. On va donc considérer les 10 thématiques avec la plus forte probabilité inférées par le modèle LDA construit à partir des articles de Wikipédia. Il est impératif d'une part qu'au moins une thématique alignée avec le domaine source et au moins une thématique alignée avec le domaine cible figurent parmi ces 10 premières thématiques inférées par le modèle LDA. D'autre part, on considère séparément pour les thématiques associées au domaine source et au domaine cible la plus forte probabilité en s'assurant que celle-ci est supérieure à un seul prédéfini, z_{relC} respectivement de 0,06 pour le domaine source et de 0,1 pour le domaine cible en anglais chez Heintz *et al.* (2013). Dans le cas où toutes ces conditions ne sont pas respectées, la phrase est ignorée.

1. `http://mallet.cs.umass.edu/`

Soient t une thématique, C un domaine aligné et $\Lambda(C)$ l'ensemble des thématiques dominantes de la phrase x à analyser, ce domaine est jugé être une thématique prédominante de cette phrase si : $\sum_{t \in \Lambda(C)} p(t|x) > z_{\mathrm{relC}}$.

2.5. *Identifications des mots utilisés dans la métaphore*

Dans le lot de phrases restantes, on va détecter précisément quels mots rattachés aux domaines cible et sources résultant de la phase précédente sont utilisés pour construire la métaphore. Cet ensemble de mots A'_C correspond aux mots qui appartiennent simultanément à un domaine prédominant C et à la phrase à analyser x tel que : $\sum_{t \in C} p(t|x) > z_{\mathrm{word}}$

Dans l'article de référence, z_{word} est fixé à 0,1. Par ailleurs, tout mot ne peut être associé qu'à un et un seul domaine. Ainsi, soient A_T, l'ensemble des mots du domaine cible et A_S, celui des mots du domaine source :

$$A_S = A'_S - A'_T$$

et inversement,

$$A_T = A'_T - A'_S$$

2.6. *Application des filtres*

Heintz *et al.* (2013) recensent 2 filtres qui doivent être mis en place pour diminuer le bruit et éliminer des phrases non pertinentes. Premièrement, l'article propose de ne garder que les phrases qui comptent au minimum 4 mots qui ne font pas partie de la liste des mots vides établis à partir des fréquences de mots dans le corpus Wikipédia. Deuxièmement, il suggère d'exclure toutes les phrases dont l'une des 10 thématiques prédominantes est aussi une thématique principale du texte car si le domaine source se retrouve dans tout le texte, il y a de fortes chances qu'il soit employé de manière littérale. À titre d'illustration, dans un texte qui parle de construction d'autoroutes, une phrase telle que « *Congress needs to pass a new highway bill* » serait ignorée pour la détection de métaphores liées au domaine GOVERNANCE même si elle combine bien ce domaine cible et un domaine source prédéfini : *highway* étant fréquemment utilisé dans le texte, son sens est forcément littéral.

2.7. *Calcul du score final*

Enfin, on classe une phrase qui a passé les deux filtres comme étant métaphorique si son degré de métaphoricité est supérieur à un seuil z_{final} manuellement fixé dans l'article à – 10 pour l'anglais. Définissons pour une thématique t, un domaine C, une phrase x et un mot w :

– $\lambda_C = \sum_{w \in K(C)} p(w|t)$

– $\rho_C(x) = \sum_{t \in \Lambda(C)} p(t|x)$
– $R_C(x) = \arg\max_{t \in \Lambda(C)} p(t|x)$
– $\omega_C(w) = \max_{w \in x} \sum_{t \in C} p(t|x)$

Toute phrase x contient une métaphore mettant en relation un domaine cible T et un domaine source S préalablement alignés à une thématique LDA si :

$$\ln\left(\lambda_S \times R_S(x) \times \rho_S(x) \times \omega_S(w) \times \lambda_T \times R_T(x) \times \rho_T(x) \times \omega_T(w)\right) > z_{\text{final}}$$

3. Galop d'essai : expérimentations sur des données en anglais

Cette expérimentation ayant déjà fait l'objet d'une première publication (Beligné *et al.*, 2017), nous allons décrire brièvement l'implémentation du système, le corpus utilisé et les résultats obtenus avant de passer à l'analyse des erreurs et ce que nous en avons retiré pour la suite du projet.

3.1. *Présentation du dispositif expérimental*

Pour cette première expérimentation, la méthode de Heintz *et al.* (2013) a été entièrement réimplémentée en Python. Si la librairie MALLET (McCallum, 2002) a été utilisée pour la création du modèle LDA ainsi que l'inférence des topiques, il est important de signaler que nous n'avons exploité que la moitié des articles de Wikipédia sélectionnés aléatoirement. Nous nous sommes servis de NLTK (Loper et Bird, 2002) pour segmenter les textes en phrases et effectuer la tokénisation. De plus, les seuils inhérents à la méthode étant fixés manuellement, après avoir testé différentes valeurs, nous en avons modifié certaines. Ainsi, nous avons fixé les valeurs respectives de z_{align}, z_{relT} (prédominance du domaine source dans la phrase) et z_{relS} (prédominance du domaine cible dans la phrase) à 0,0075, 0,007 et 0,3. Le corpus d'étude a été constitué à partir de 17 articles écrits entre 2012 et 2016 et comportant entre 30 et 256 phrases pour une taille totale de 41 620 mots. Le tableau 2 décrit plus en détail le corpus utilisé et donne plus d'informations sur l'origine de ces articles ainsi que sur les sujets dont ils traitent.

Afin d'évaluer la performance de la méthode choisie, ce corpus a été annoté par deux annotateurs suivant le protocole MIPVU (Steen *et al.*, 2010) qui consiste de manière très succincte, premièrement à lire le texte pour s'en imprégner puis à faire une deuxième lecture plus poussée en s'arrêtant sur chaque unité lexicale pour déterminer si elle est utilisée métaphoriquement ou non. Sur les 1 527 phrases que contenait au total le corpus, 365 ont été classées comme renfermant une métaphore portant sur un terme géographique. Pour cette tâche d'annotation, l'accord interannotateur calculé au moyen du Kappa de Cohen (Cohen, 1960) est de 0,63, ce qui est de loin supérieur au taux de 0,48 rapporté par Heintz *et al.* (2013) qui avaient également fait appel à deux annotateurs.

Support	Type	Nombre d'articles	Sujets
Atlantic Geology[2]	Revue scientifique	4	Roches, sédiments, marais salant
Nature Climate Change[3]	Revue scientifique	1	Niveau de l'eau
Hypergeo[4]	Encyclopédie en ligne	2	Érosion, océan
Geocurrents[5]	Forum d'échanges d'idées	2	Caucase, Australie
123helpme[6]	Base de devoirs en ligne	3	Hongrie, Papouasie, région arctique
The Guardian[7]	Journal	2	Population mondiale, région de Douro
Carbonbrief[8]	Site d'informations sur le changement climatique	2	Croissance des arbres, dégradation des forêts
American Geophysical Union [9]	Site d'informations sur la géophysique	1	El Niño

Tableau 2. *Description du corpus utilisé pour les expérimentations en anglais*

3.2. *Analyse des erreurs*

Dans l'ensemble, la performance du système s'est avérée plus que décevante. Seuls 27 des 43 domaines prédéfinis ont pu être automatiquement alignés avec une thématique LDA. En outre, comme le montre la matrice de confusion (tableau 3), les résultats finaux sont inexploitables pour une future automatisation : 58,9 % de précision et 18,9 % de rappel. Il faut préciser que ces résultats ne concernent que la classification d'une phrase comme étant métaphorique ou non. En regardant de plus près, au niveau de l'attribution des domaines sources, nous nous sommes rendus compte que sur les

2. https://journals.lib.unb.ca/index.php/ag
3. https://www.nature.com/nclimate/
4. http://www.hypergeo.eu/
5. http://www.geocurrents.info/
6. https://www.123helpme.com/
7. https://www.theguardian.com
8. https://www.carbonbrief.org/
9. https://news.agu.org

69 phrases correctement détectées, le bon domaine source n'a été trouvé que dans une seule phrase : « *The county of* ***hungry*** *is inhabited by roughly — 9,919,128 as of July 2014 [. . .]* ». Cependant, au vu du contexte d'utilisation, on s'aperçoit vite que cette phrase comporte deux coquilles : *county* au lieu de *country* et *hungry* à la place de « Hungary ». Cette première découverte soulève la question de la qualité non seulement de l'OCR des textes choisis, mais aussi de leur annotation. En effet, on pourrait supposer que des locuteurs natifs auraient remarqué de telles erreurs et les auraient corrigées.

		Méthode automatique		
		Métaphores	**Non-métaphores**	**Total**
Annotation manuelle	**Métaphores**	69	48	117
	Non-métaphores	296	1114	1410
	Total	365	1162	

Tableau 3. *Matrice de confusion pour les expérimentations en anglais*

En ce qui concerne les autres phrases correctement identifiées comme étant métaphoriques, nous avons décelé trois types d'erreurs liées à la détection des domaines sources :

– un domaine source aligné assigné à la place d'un autre domaine source aligné (26 phrases).
Exemple : *The Saint John River* ***bisects*** *the city at a location down river of the point of tidal influence that extends up-river to Mactaquac Da*m [*Politique* au lieu de *Mathématiques*] ;

– un domaine source aligné assigné à la place d'un domaine source non aligné (14 phrases).
Exemple : *To start we will look at the river that* ***cuts*** *through the grand city* [. . .] [*Théâtre* au lieu de *Mort*] ;

– un domaine aligné assigné à la place d'un domaine hors liste ou inexistant (28 phrases), ce qui relève à nouveau d'un problème d'annotation.
Exemple : *Freeze-thaw exploitation is dependent upon microclimate, which* ***controls*** *the number of cycles. . .* [attribué à *Tribunal* ; pas de domaine source précisé].

Pour mieux comprendre pourquoi l'assignation des domaines sources et l'identification des mots utilisés dans la métaphore fonctionnent aussi mal, nous sommes inté-

ressés dans un premier temps à l'alignement obtenu à partir du modèle LDA construit. Nous en avons retiré d'une part que contrairement aux substantifs, les verbes et les adjectifs ont tendance à avoir des probabilités significatives dans plusieurs domaines. Par exemple, *called* se retrouve à la fois dans les domaines *Mathématiques* et *Nourriture* tandis que *high* est présent dans les domaines GÉOGRAPHIE et *Éducation*. De plus, la probabilité attribuée à chaque mot est assez basse, la probabilité la plus haute dans un domaine aligné étant très souvent de l'ordre de 0,05. En classant ces probabilités par ordre décroissant et en considérant pour chacun des 27 domaines alignés les cent premières probabilités, on constate que seuls 81 mots ont une probabilité comprise entre 0,05 et 0,02 tandis que 204 mots ont une probabilité de l'ordre de 0,01. En comparant les scores d'alignement avec certains exemples donnés par Heintz *et al.* (2013), on note tout de suite une grande différence. Par exemple, pour le domaine *Véhicule*, leurs scores d'alignement sont de 0,035, 0,29 et 0,022 alors que nous obtenons pour le même domaine des valeurs deux fois ou plus inférieures : 0,011, 0,014 et 0,009. De plus, avec notre modèle LDA, en utilisant les mots-graines fournis par Heintz *et al.* (2013), nous ne parvenons pas à aligner avec succès comme eux le domaine *Animals*. L'une des hypothèses plausibles qu'on pourrait avancer pour expliquer cet échec serait que la partie de Wikipédia utilisée pour notre expérimentation couvre très mal ou peu certains domaines.

Ensuite, nous avons analysé la détection des thématiques prédominantes dans les textes et les phrases. Très souvent, le domaine inféré par le modèle LDA est erroné parce que le domaine source de la métaphore est souvent utilisé ponctuellement et n'est pas dominant dans toute la phrase.
Exemple : *The relief changes from the floodplain at an average of 10 m elevation to greater than 100 m in elevation at the top of the steep valley* ***wall*** *to the south to be related to the emplacement of the thrust slice, but is directed toward the west*.
Ici, le domaine source attendu, *Construction*, est surtout représenté par le mot *wall* même si on pourrait argumenter que *elevation* est utilisé dans le domaine de l'architecture et qu'*emplacement* peut désigner l'endroit où se situe un bâtiment. Dans d'autres cas, le domaine source est choisi par défaut car il figure parmi les 10 thématiques dominantes sans pour autant forcément être le domaine prédominant de la phrase.
Exemple : *The Artic is one of the few* ***unspoilt*** *wilderness areas in the world and must be conserved*.
Le domaine source attribué ici automatiquement *Éducation* est en fait la quatrième thématique mais devient par défaut la première thématique puisque les trois premières thématiques inférées ne sont pas alignées. Le même problème se rencontre lorsque l'on regarde de plus près la reconnaissance des domaines dans les faux positifs. À titre d'illustration, dans la phrase « *The typical aspect of the contact can be recognized as a gently westward-dipping plane* », deux raisons justifient que cette phrase soit considérée à tort comme contenant une métaphore liée au domaine *Musique* :

– la plupart des mots de la phrase se trouvent dans une thématique associée au domaine *Musique* ;

– *Musique* est la cinquième thématique prédominante dans la phrase avec une probabilité de 0,029 tandis que la première thématique a une probabilité de 0,18, et GÉOGRAPHIE, la troisième, une probabilité de 0,03.

Nous avons par la suite comparé, dans la mesure du possible, notre dispositif expérimental avec les informations que nous avons pu glaner de l'article de Heintz *et al.* (2013), les auteurs n'ayant pu nous fournir une copie du code et des données utilisés. Dans un premier temps, nous avons remarqué que presque tous les exemples donnés comme étant bien analysés par le système se rapportent à des phrases se composant d'une proposition indépendante, ce qui rend difficile toute interférence d'autres domaines sources. Il en est de même du seul exemple comptant deux propositions « *Moderates, we all hear, are an* ***endangered species*** ». Si on élimine tous les mots vides, on se retrouve avec 4 mots pleins qui limitent grandement le nombre de thématiques prédominantes possibles : *moderates*, *hear*, *endangered*, *species*. Il va sans dire que beaucoup de phrases de notre corpus sont bien plus complexes syntaxiquement ou comportent beaucoup plus de mots pleins. Pour y remédier, nous avons procédé à des tests avec des 4-grammes à la place de phrases entières mais n'avons pas relevé d'améliorations significatives.

Enfin, il est important de souligner que la méthode de Heintz *et al.* (2013) n'a jamais été évaluée sur l'ensemble de leur corpus sur lequel nous n'avons aucune précision hormis son origine (articles de journaux et de blogs). Les données avec lesquelles ils affirment avoir obtenu une F-mesure de 59 % sont composées de 600 phrases tirées pour une moitié, pour chaque domaine source et domaine cible, de phrases ayant les cinq plus hauts degrés de métaphoricité finaux auxquels s'ajoutent d'autres phrases jugées métaphoriques sélectionnées en fonction de leur degré de métaphoricité. L'autre moitié, quant à elle, contient 300 phrases classées par le système comme étant non métaphoriques et choisies au hasard. Dans la seconde évaluation dans laquelle il a été demandé à des utilisateurs d'AmazonTurk de juger de la métaphoricité des 250 phrases auxquelles leur méthode a attribué le plus haut degré de métaphoricité, la métaphoricité moyenne des métaphores conceptuelles annotées est de 0,39.

4. Nouveau protocole expérimental avec des données en français

À la suite de cette analyse d'erreurs qui a fait remonter les faiblesses du système implémenté, nous avons défini différentes modifications dans le protocole expérimental qui à notre sens devraient nous permettre d'obtenir des résultats plus pertinents.

Premièrement, nous avons résolu de guider l'extraction des thématiques LDA afin de mieux faire ressortir les domaines qui nous intéressent en sélectionnant automatiquement les articles de Wikipédia qui s'y rapportent. Par ailleurs, nous avons aussi décidé de faire varier le nombre de thématiques extraites afin de mesurer leur incidence sur l'alignement thématique domaine. Enfin, dans l'idée d'attribuer de plus fortes probabilités aux mots réellement caractéristiques d'une thématique, la probabilité qu'un mot appartienne à une thématique donnée a été calculée différemment.

Cette fois, nous avons travaillé sur des données en français, changement qui s'explique par deux principaux facteurs :

– le manque d'adéquation entre le corpus en anglais et notre objectif final : en effet, le corpus en anglais se compose en grande majorité de textes informatifs non scientifiques qui ne sont pas pour la plupart écrits par des géographes et ne correspondent donc pas au type de textes que nous souhaiterions analyser à terme ;

– les différents défauts de l'annotation déjà mentionnés plus haut : plutôt que de reprendre l'annotation sur des textes, somme toute, peu pertinents et avec des annotateurs qui maîtrisent approximativement l'anglais, travailler en français nous garantissait un accès plus facile à des locuteurs natifs formés en géographie.

Dans la suite de cette section, nous allons présenter les différentes modifications apportées à la méthode originale, puis nous nous intéresserons au corpus utilisé ainsi qu'à son annotation avant de discuter les résultats obtenus.

4.1. *Sélection sémantique des articles de Wikipédia*

L'idée de filtrer les articles en fonction de leur contenu nous a été inspirée par les travaux de Phan *et al.* (2008) qui utilisent des mots-clés pour choisir des pages Wikipédia parlant des thématiques qui les intéressent. Si l'extraction de thématiques LDA qui en résulte donne de bons résultats, malheureusement, aucune précision n'est donnée sur le choix de mots-clés ou le seuil minimal de mots en commun requis pour qu'un texte soit considéré comme pertinent. Plutôt que de nous limiter aux mots-graines de chaque domaine, afin de couvrir plus exhaustivement les domaines sources et cible, nous avons choisi d'utiliser le *Dictionnaire électronique des mots* (Dubois et Dubois-Charlier, 2010) qui associe chaque lemme d'une entrée à un domaine sémantique prototypique[10].

Au total, le *Dictionnaire électronique des mots* compte 32 des 43 domaines prédéfinis pour cette tâche, le présupposé étant que les 11 domaines absents (*Statique*, *Dynamique*, *Mort*, *Expansion*, *Sens*, *Désordre*, *Richesse*, *Prison*, *Monarchie*, *Compétition*, *Barrière*) sont soit inclus intégralement ou partiellement dans d'autres domaines, soit latents, dans les textes présélectionnés. Par exemple, le domaine *Monarchie* fait partie du domaine *Politique*, *Compétition* de celui de *Sports*, *Prison* de *Tribunal*, etc.

Une fois cette liste de mots-clés construite, nous avons utilisé Morphalou3[11] pour ajouter automatiquement les formes fléchies de chaque mot. Puis, nous avons gardé tout article de Wikipédia qui avait au moins 10 mots en commun avec l'un des domaines à aligner. Ce seuil a été choisi en tenant compte du fait que plusieurs des métadonnées des pages de Wikipédia telles que « Portail », « Catégorie », « Notes et références », « Articles connexes » sont susceptibles de contenir des termes qui peuvent appartenir au lexique de mots-clés. Il va donc de soi que les pages de Wikipédia sont

10. http://rali.iro.umontreal.ca/DEM/domaines/index.html
11. https://repository.ortolang.fr/api/content/morphalou/2/LISEZ_MOI.html

uniquement nettoyées *a posteriori*. Enfin, nous avons réduit le corpus aux articles les plus longs, c'est-à-dire, ceux qui dépassent 9,90 ko afin d'avoir suffisamment de contexte pour chaque mot.

4.2. *Augmentation du nombre de thématiques LDA extraites*

Le nombre de 100 thématiques fixé par Heintz *et al.* (2013) semble relativement modeste, surtout lorsque l'on pense d'une part, au nombre de thématiques que Wikipédia peut couvrir et d'autre part, qu'on aligne un domaine source au maximum avec 3 thématiques. Navarro Colorado et Tomás (2015) ayant montré qu'une sortie de 1 000 ou 2 500 thématiques sur l'ensemble de Wikipédia permet d'obtenir une meilleure granularité de thématiques et de mieux circonscrire dans une même thématique les mots appartenant au même champ lexical, nous avons également construit toujours à partir de notre corpus Wikipédia en français, des modèles LDA avec 1 000 et 2 500 thématiques.

4.3. *Identification des mots discriminants pour chaque thématique*

La métaphore se produisant surtout au niveau lexical, et prenant en compte la dimension probabiliste de LDA, il nous a semblé primordial de pouvoir définir si un mot est caractéristique d'une thématique particulière, en calculant différemment la probabilité qu'un mot relève d'une thématique donnée. Ainsi, soit z une thématique extraite, w un mot du corpus Wikipédia, et d un document de ce corpus :

$$p(z/w) = \frac{p(w/z) \times p(z)}{p(w)}$$

où $p(z) = \sum p(z/d) \times p(d)$ et p$(w) = \sum p(w/z) \times p(z)$

Une fois, ces nouvelles probabilités générées, l'entropie de Shannon a servi à filtrer les mots passe-partout qui ne sont discriminants pour aucune thématique, c'est-à-dire ceux dont aucune probabilité dans une thématique ne dépasse l'entropie de toutes leurs probabilités :

$$H(p(z/w)) = -\sum p(z/w) \log p(z/w)$$

Enfin, nous avons éliminé dans chaque phrase tous les mots non discriminants avant de procéder à l'inférence des thématiques.

4.4. *Présentation du corpus et de la méthode d'annotation*

Une dizaine d'articles parus entre 1972 et 1999 dans *L'Espace Géographique* a été sélectionnée au hasard pour constituer notre corpus d'évaluation. Les trois décennies

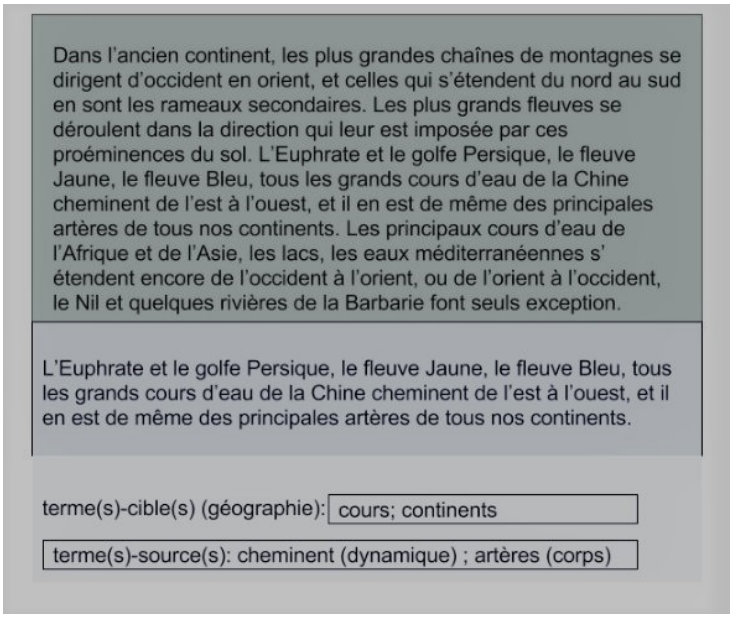

Figure 2. *Vue d'une phrase à annoter*

que recouvre ce corpus sont représentées (au moins 3 articles par décennie). Ce corpus rentre dans diverses sous-disciplines géographiques telles que la méthodologie de la recherche en géographie, la géographie économique ou encore la géographie humaine. Tous les textes sont issus du portail Persée[12].

Pour annoter le corpus, nous avons choisi de faire appel à des étudiants en master de géographie. À cet effet, nous avons mis en place une plate-forme en ligne dans laquelle chaque texte s'affiche phrase par phrase. Afin de la situer dans son contexte, chaque phrase est toujours affichée en dessous du paragraphe dont elle est tirée. Une fois la phrase lue, les annotateurs doivent indiquer si elle contient une métaphore ou non. Dans le premier cas, ils doivent alors rentrer les termes cibles et sources de la métaphore et préciser uniquement le ou les domaines auxquels le ou les termes sources appartiennent (figure 2). Avant le début de la phase d'annotation proprement dite, une séance d'explication a été organisée durant laquelle les grandes lignes du projet ont été exposées aux étudiants. Nous avons également conçu un guide d'annotation qui recensait tous les domaines sources, présentait des exemples d'annotation et décrivait les différentes étapes à suivre pour identifier les métaphores selon le protocole MIPVU (Steen *et al.*, 2010).

12. `https://www.persee.fr/`

Nous disposions au total de 9 annotateurs qui ont annoté partiellement 4 textes, soit 435 phrases dont 51 ont été considérées comme étant métaphoriques par plus de la moitié des annotateurs. On constate que le taux d'accord interannotateur est plutôt bas en ce qui concerne les phrases métaphoriques (0,30) par rapport à celui de l'étiquetage des domaines sources (0,52).

4.5. *Résultats et discussion*

Afin de mieux cibler les failles du système d'identification des métaphores, en plus de mesurer sa précision et son rappel pour la détection des phrases annotées comme étant métaphoriques, nous avons également évalué sa précision en ce qui concerne la détection des domaines sources.

Au terme de la phase de filtrage sémantique de Wikipédia, nous avons obtenu un corpus final de Wikipédia de 68 404 articles pour un total de 1 570 718 mots uniques[13]. Pour l'ensemble des tâches relevant de l'extraction des thématiques, nous avons utilisé Gensim (Řehůřek et Sojka, 2010) en itérant chaque fois 1 000 fois et en ne gardant pour chaque thématique que les 2 000 premiers mots avec la plus forte probabilité. Le tableau 4 montre qu'effectivement, déjà en réduisant le nombre de pages Wikipédia, on améliore légèrement l'alignement des domaines, cependant on reste dans le même ordre de résultats que ce que nous avions déjà obtenu pour l'expérimentation en anglais.

Corpus	Nombre de thématiques	Nombre de domaines alignés
Wikipédia - corpus entier	100	20
Wikipédia intermédiaire (873 807 articles)	100	22
Wikipédia final (68 404 articles)	100	23

Tableau 4. *Domaines alignés en fonction des corpus Wikipédia*

De même, comme on peut le voir dans le tableau 5, sur le modèle LDA construit à partir du corpus Wikipédia final, en ne considérant que les cinquante premiers mots et 2 500 thématiques, on améliore sensiblement l'alignement des domaines de sorte que 100 % des concepts sont alignés. Néanmoins, la comparaison des différentes performances du système listées dans le tableau 6 laisse à penser qu'un alignement optimal

13. La version de Wikipédia utilisée est celle de 20/07/2017, qui comporte environ 1 900 000 pages (articles, pages de discussion, pages de désambiguïsation…). Le fichier XML brut obtenu a été nettoyé de toutes les balises superflues avec WikiExtractor (`https://github.com/attardi/wikiextractor`).

n'a aucun impact majeur, au contraire. En effet, les meilleurs résultats sont obtenus avec 100 thématiques conformément à ce que proposent Heintz *et al.* (2013). En outre, plus on augmente le nombre de thématiques extraites, plus on fait baisser le rappel du système sans pour autant améliorer la détection des domaines sources.

Nombre de thématiques	Nombre de mots considérés avec la plus forte probabilité	Nombre de domaines alignés
100	2 000	23
1 000	20	42
2 5000	50	43

Tableau 5. *Impact du nombre de thématiques extraites sur l'alignement thématiques domaine*

Paradigmes	Précision système	Rappel système	Précision domaines sources
100 thématiques	16,60 %	49,00 %	0,04 %
100 thématiques + $p(z/w)$	16,20 %	43,10 %	0 %
100 thématiques + $p(z/w)$ + entropie	18,10 %	7,80 %	0 %
1 000 thématiques	18,40 %	23,50 %	0,10 %
1 000 thématiques + $p(z/w)$	15,20 %	25,40 %	0 %
2 500 thématiques	75,00 %	5,80 %	0 %
2 500 thématiques + $p(z/w)$	18,20 %	8,80 %	0 %

Tableau 6. *Performances du système*

Comme à l'étape précédente, nous avons procédé à une analyse des erreurs. D'abord, en ce qui concerne l'alignement des domaines, nous avons pu cibler quatre principaux problèmes :

– certains domaines sont ambigus, ce qui complique leur alignement. Le domaine *Barrière*, par exemple, est associé à des mots du type « sydney »,« bactérie »,« infectieuse »,« sauveteurs », et donc se réfère plutôt vraisemblablement à la Grande barrière de corail ;

– de même, certains des mots-graines proposés sont polysémiques et faussent l'alignement thématique domaine. C'est ainsi que la présence du mot « étoiles » lie le domaine *Astronomie* à une thématique qui fait plus penser au monde du cinéma : « passionné », « herbert », « julia », « marcello »« akkad », « réfuter »… ;

– certains domaines alignés ne sont pas suffisamment distincts et semblent contenir des mots se rapportant à d'autres domaines. Dans les thématiques alignées aux domaines *Art* et *Chimie*, on retrouve des noms d'autres domaines ; « religion » dans le premier domaine et dans le second, « biologie » et « astronomie ». Dans ce dernier exemple, il serait plus plausible que cette thématique relève des la science en général ou des sujets scolaires. De même, toujours dans une des thématiques associées au domaine *Art*, on retrouve plutôt des mots qui relèvent clairement de la géographie : (« peuples », « population », « régions », « rurales ») ;

– le grand nombre d'entités nommées surtout en ce qui concerne les lieux dans des domaines autres que GÉOGRAPHIE, par exemple « Paris » et « Tokyo » dans le domaine *Art*.

Au niveau de l'attribution des domaines sources, nous avons noté que la détection des mots impliqués dans la métaphore pose toujours problème. Par exemple, dans la phrase « La porte d'un parc est, à l'origine, un ensemble de réalisations, situées au fond d'une vallée, au plus près des limites de la zone centrale », le système conclut que la métaphore met en relation « origine » du domaine GÉOGRAPHIE et « ensemble » du domaine *Construction* respectivement à la place de « parc » et « porte ».

5. Bilan et nouvelles perspectives de travail

Au vu de ces nouveaux résultats négatifs, nous ne pouvons que conclure que la méthode testée est totalement inadaptée à notre objet de recherche ; avons-nous pour autant perdu notre temps ? Loin de là. D'une part, cet échec nous a permis de nous interroger sur le rôle que joue le type de textes dans la détection automatique des métaphores et d'autre part, nous a poussés à nous poser d'importantes questions méthodologiques touchant aussi bien la formulation de notre objet de recherche que notre protocole d'annotation.

La question de la portabilité de la méthode choisie rejoint une problématique récurrente en humanités numériques où les algorithmes de traitement automatique des langues disponibles ont souvent été uniquement testés sur un seul type de textes, typiquement des articles de journaux. En ce qui concerne la détection automatique des métaphores dans les textes, l'utilisation d'articles journalistiques s'explique d'autant plus qu'à cause du nombre limité, voire inexistant, de corpus pour plusieurs langues et de la perception subjective de ce que constitue une métaphore chez des annotateurs non experts, les recherches menées se sont surtout focalisées sur les types de textes les plus susceptibles de nous renseigner sur ce qui se passe au quotidien et comment nous le ressentons. Afin d'adopter la stratégie la plus appropriée face à un genre de textes jusque-là inexploré en détection automatique des métaphores, il nous semble opportun de nous demander dans un premier temps, quelle fonction joue la métaphore dans le discours des géographes et si toutes les métaphores sont équivalentes. En d'autres termes, des syntagmes du type « la porte du parc » ou encore « le flanc de la colline » traduisent-ils une idéologie particulière de l'auteur ou un simple choix stylistique ?

Exactement, à partir de quel moment les métaphores des géographes commencent-elles à devenir singulières et signifiantes ?

Pour répondre à ces questions, nous avons examiné des métaphores de géographes citées en exemple dans différents articles. On constate ainsi qu'en général, ces métaphores expriment une façon de penser et de voir le monde, mais aussi que très souvent, elles traversent plusieurs écrits du même auteur (par exemple Vidal de La Blache et Jean Brunhes pour la métaphore organiciste). Par conséquent, ces métaphores idéologiques pourraient se concevoir comme la surprésence dans un texte ou chez un auteur de termes appartenant à un ou plusieurs domaines distincts de la géographie. Comment donc déceler cette surprésence ?

À notre sens, cela passe en premier lieu par la construction d'un lexique exhaustif pour chacun des domaines prédéfinis. Cependant, en regardant de plus près notre liste initiale de domaines, leur hétérogénéité saute immédiatement aux yeux. En effet, si certains d'entre eux appartiennent à des champs d'étude bien définis (*Biologie*, *Art*...), d'autres au contraire se réfèrent à un concept bien précis (*Mort*, *Barrière*, *Expansion*...). Par conséquent, deux stratégies distinctes doivent être mises en place. En ce qui concerne le premier groupe de termes, pour les domaines qui y sont répertoriés, le plus simple nous paraît de tirer parti de la liste de termes par domaine extraite du *Dictionnaire électronique des mots* (Dubois et Dubois-Charlier, 2010) et des métadonnées des articles de Wikipédia liant un terme à un portail spécifique. Pour le second cas de figure, il serait plus approprié d'utiliser des dictionnaires de synonymes, des bases lexicales listant les mots de la même famille ou encore des modèles de plongements de mots préconstruits pour construire automatiquement le champ lexical de ces concepts. Une vérification manuelle pourrait être envisagée pour éliminer les termes les moins pertinents.

Ensuite, afin de pouvoir détecter la surutilisation des lexiques construits, il nous faudra un plus large corpus diachronique composé d'une part d'articles de géographie subdivisés en sous-disciplines et étiquetés par auteur et d'autre part, de textes plus génériques de la même période portant sur une toute autre discipline qui nous serviront de référence. Ainsi, nous pourrons inférer si un domaine est surreprésenté et régulièrement convoqué chez un auteur non seulement par rapport à sa présence relative dans les textes de référence, mais aussi dans les écrits des géographes contemporains à l'auteur en question. Il est important de souligner ici que certains domaines sont étroitement liés entre eux et ne devraient pas être impérativement considérés comme des domaines complètement séparés, par exemple *Organe*, *Mort* et *Maladie*.

Enfin, une fois les phrases intéressantes ainsi isolées, nous pourrons nous fonder sur différents scénarios syntaxiques de constructions métaphoriques tels que ceux définis par Tamine (1979), Krishnakumaran et Zhu (2007) et Dodge *et al.* (2015) pour identifier les termes relevant du domaine cible impliqués dans la métaphore. Mise à part la possibilité de détecter les métaphores *in absentia* ainsi que celles dans lesquelles le terme du domaine source est remplacé par un pronom personnel, cette nouvelle approche présente différents autres avantages. D'abord, elle nous évitera de construire une liste de termes pour le domaine GÉOGRAPHIE, tâche particulièrement

compliquée au vu des différents sous-domaines et thématiques que la géographie peut couvrir. Ensuite, l'annotation manuelle ne se concentrera, dans un premier temps, que sur les phrases présélectionnées. Enfin, elle nous permettra de construire de façon semi-automatique un corpus annoté qui pourra nous permettre ultérieurement d'utiliser des méthodes d'apprentissage supervisé pour découvrir les métaphores qui nous auraient échappé.

6. Conclusion

La métaphore joue un rôle primordial dans les textes de sciences humaines et sociales non seulement en tant qu'élément du langage, mais aussi en tant qu'agent de la circulation des idées. À cet effet, nous avons présenté dans cet article un projet en humanités numériques qui a pour but de détecter des métaphores dans les écrits des géographes et utilise principalement la méthode de Heintz *et al.* (2013) testée en anglais et en espagnol pour identifier des métaphores conceptuelles ayant pour domaine cible GOVERNANCE. Bien que nos expérimentations en français et en anglais soient loin d'avoir généré des résultats exploitables pour la suite, cette étude non seulement pose d'importantes questions sur l'adéquation de certaines méthodes de traitement automatique des langues pour des textes de sciences humaines et sociales, mais aussi nous a permis de radicalement redéfinir notre approche afin de mieux répondre à nos objectifs de départ.

Remerciements

Ce travail a été financé par l'Institut universitaire de France (IUF) et s'est effectué en collaboration avec Sabine Loudcher (ERIC, Université Lumière Lyon 2), Julien Velcin (ERIC, Université Lumière Lyon 2), Max Béligné (EVS, Université Lumière Lyon 2) et Isabelle Lefort (EVS, Université Lumière Lyon 2).

7. Bibliographie

Archer K., « Regions as social organisms : The Lamarckian characteristics of Vidal de la Blache's regional geography », *Annals of the Association of American Geographers*, vol. 83, n° 3, p. 498-514, 1993.

Aristote, *Poétique et rhétorique*, Garnier Frères, 1922.

Ascher F., « La métaphore est un transport. », *Cahiers internationaux de sociologie*, n° 1, p. 37-54, 2005.

Bachelard G., *La Formation de l'esprit scientifique : Contribution à une psychanalyse de la connaissance*, Vrin, 1967.

Bachimon P., « Physiologie d'un langage. L'organicisme aux débuts de la géographie humaine », *Espace Temps*, vol. 13, n° 1, p. 75-103, 1979.

Beligné M., Campar A., Chauchat J.-H., Lefeuvre M., Lefort I., Loudcher S., Velcin J., « Détection automatique de métaphores dans des textes de Géographie : une étude prospective », *Actes de la Conférence sur le traitement automatique des langues naturelles (TALN), 2017*, vol. 2, p. 86-93, 2017.

Berdoulay V., « La métaphore organiciste : Contribution à l'étude du langage des géographes », *Annales de géographie*, p. 573-586, 1982.

Bethard S., Lai V. T., Martin J. H., « Topic model analysis of metaphor frequency for psycholinguistic stimuli », *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, Association for Computational Linguistics, p. 9-16, 2009.

Blei D. M., Ng A. Y., Jordan M. I., « Latent dirichlet allocation », *Journal of machine Learning research*, vol. 3, , p. 993-1022, 2003.

Cohen J., « A coefficient of agreement for nominal scales », *Educational and psychological measurement*, vol. 20, n° 1, p. 37-46, 1960.

Daniels S., Cosgrove D., « Landscape metaphors in cultural geography », *Place/culture/representation*, 1993.

Deguy M., « Vers une théorie de la figure généralisée », *Critique*, vol. XXV, n° 269, p. 841-861, 1969.

Dodge E., Hong J., Stickles E., « MetaNet : Deep semantic automatic metaphor analysis », *Proceedings of the Third Workshop on Metaphor in NLP*, p. 40-49, 2015.

Dubois J., Dubois-Charlier F., « La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration », *Langages*, n° 3, p. 31-56, 2010.

Dumarsais C., *Les Tropes, éd*, Fontanier, Paris, 1818.

Fass D., « met* : A method for discriminating metonymy and metaphor by computer », *Computational Linguistics*, vol. 17, n° 1, p. 49-90, 1991.

Heintz I., Gabbard R., Srivastava M., Barner D., Black D., Friedman M., Weischedel R., « Automatic extraction of linguistic metaphors with lda topic modeling », *Proceedings of the First Workshop on Metaphor in NLP*, p. 58-66, 2013.

Jang H., Moon S., Jo Y., Rose C., « Metaphor detection in discourse », *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 384-392, 2015.

Kintsch W., « Metaphor comprehension : A computational theory », *Psychonomic bulletin & review*, vol. 7, n° 2, p. 257-266, 2000.

Klebanov B. B., Beigman E., Diermeier D., « Discourse topics and metaphors », *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, p. 1-8, 2009.

Klebanov B. B., Leong B., Heilman M., Flor M., « Different texts, same metaphors : Unigrams and beyond », *Proceedings of the Second Workshop on Metaphor in NLP*, p. 11-17, 2014.

Krishnakumaran S., Zhu X., « Hunting Elusive Metaphors Using Lexical Resources. », *Proceedings of the Workshop on Computational approaches to Figurative Language*, p. 13-20, 2007.

Lakoff G., Johnson M., *Metaphors we live by*, University of Chicago Press, 1980.

Lévy B., « Géographie et littérature. Une synthèse historique », *Le globe*, vol. 146, p. 25-52, 2006.

Loper E., Bird S., « NLTK : The Natural Language Toolkit », *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and*

Computational Linguistics - Volume 1, ETMTNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 63-70, 2002.

McCallum A. K., « MALLET : A Machine Learning for Language Toolkit », 2002, http ://mallet.cs.umass.edu.

Molino J., « Métaphores, modèles et analogies dans les sciences », *Langages*, n° 54, p. 83-102, 1979.

Navarro Colorado B., Tomás D., « A fully unsupervised Topic Modeling approach to metaphor identification », *Actas del XXXI Congreso de la Sociedad Espanola para el Procesamiento del Lenguaje Natural (SEPLN 2015). http ://www. dlsi. ua. es/˜ borja/NavarroTomas_PosterSEPLN2015. pdf*, 2015.

Phan X.-H., Nguyen L.-M., Horiguchi S., « Learning to classify short and sparse text & web with hidden topics from large-scale data collections », *Proceedings of the 17th international conference on World Wide Web*, ACM, p. 91-100, 2008.

Řehůřek R., Sojka P., « Software Framework for Topic Modelling with Large Corpora », *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45-50, 2010. `http://is.muni.cz/publication/884893/en`.

Schulder M., Hovy E., « Metaphor detection through term relevance », *Proceedings of the Second Workshop on Metaphor in NLP*, p. 18-26, 2014.

Shutova E., Sun L., Gutiérrez E. D., Lichtenstein P., Narayanan S., « Multilingual metaphor processing : Experiments with semi-supervised and unsupervised learning », *Computational Linguistics*, vol. 43, n° 1, p. 71-123, 2017.

Steen G. J., Dorst A. G., Herrmann J. B., Kaal A., Krennmayr T., Pasma T., *A Method for Linguistic Metaphor Identification : From MIP to MIPVU*, vol. 14, John Benjamins Publishing, 2010.

Tamine J., « Métaphore et syntaxe », *Langages*, n° 54, p. 65-81, 1979.

Tsvetkov Y., Boytsov L., Gershman A., Nyberg E., Dyer C., « Metaphor detection with cross-lingual model transfer », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, vol. 1, p. 248-258, 2014.

Turney P. D., Neuman Y., Assaf D., Cohen Y., « Literal and metaphorical sense identification through concrete and abstract context », *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 680-690, 2011.