Applied Semiotics / Sémiotique appliquée, nº 26 (2018)

UN ALGORITHME POUR EXTRAIRE LES SEGMENTS QUI EXPRIMENT UN CONCEPT – PREMIÈRES EXPÉRIMENTATIONS[1]

Louis Chartrand, Jean-Guy Meunier, Davide Pulizzotto et Francis Lareau
Université du Québec à Montréal
lochartrand@gmail.com

Résumé
L'analyse conceptuelle, en philosophie comme ailleurs, pourrait profiter de techniques d'assistance par ordinateur ; cependant, l'absence de modélisation du concept dans le texte ralentit le développement de ces outils. De surcroît, l'heuristique la plus utilisée, qui définit l'expression d'un concept à l'expression d'un ou de quelques mots, pose des problèmes d'ambiguïté et ne rappelle pas tous les segments textuels intéressants pour l'analyse. Dans cet article, les auteurs introduisent l'hypothèse que le thème agit comme lieu d'expression du concept, et proposent CoFiH, un algorithme qui rappelle des segments pertinents à un concept spécifié dans une requête. À l'aide de ce dernier est effectuée l'analyse du concept de LANGAGE dans *Gender Trouble*, œuvre écrite par Judith Butler, et les résultats sont évalués quantitativement et qualitativement. Comparé à l'heuristique de base, CoFiH produit un bien meilleur rappel et des résultats généralement supérieurs, laissant entrevoir de nombreuses applications.

Introduction

Dans le texte qui suit, nous nous intéressons au problème de la détection des concepts aux fins de l'analyse conceptuelle philosophique. Malgré quelques développements prometteurs, l'absence de modèle concret et exhaustif pour décrire l'expression d'un concept dans le texte est un frein au développement de l'assistance informatique à l'analyse conceptuelle philosophique basée sur des corpus textuels.

Nous présentons ici une voie prometteuse pour aborder ce problème. D'une part, nous proposons quatre hypothèses pour décrire l'expression des concepts dans le texte, et d'autre part, nous proposons une méthode qui exploite ces hypothèses afin de détecter automatiquement la présence d'un concept dans un segment de texte.

Dans la section 1, nous introduisons le contexte philosophique dans lequel s'inscrit la lecture et l'analyse conceptuelle assistée par ordinateur (LATAO) et formulons le problème faisant l'objet de cet article. Dans la section 2, nous formulons les hypothèses sur lesquelles s'appuie notre méthode, introduisant au passage la notion d'aspect qui nous permettra de lier

[1] Les auteurs tiennent à remercier Audrey Rousseau pour sa révision et ses commentaires judicieux.

le concept à son expression dans le texte. Dans la section 3, nous introduisons un algorithme exploitant ces éléments théoriques pour rappeler les segments textuels pertinents à l'analyse du concept (CoFiH). Dans la section 4, l'algorithme CoFiH est appliqué sur des données textuelles et les résultats sont présentés dans la section 5. Finalement, dans la section 6, nous discutons de la portée de ces résultats et de leurs limites.

1. Analyse conceptuelle et assistance informatique

L'analyse conceptuelle joue un rôle fondamental en philosophie, que ce soit dans le cadre des analyses génératives des causes de divers phénomènes que l'on retrouve chez les Anciens, dans l'interprétation de concepts fondamentaux comme la justice ou la vérité, qu'on ne possède souvent qu'intuitivement sans pouvoir donner de définition précise, ou dans la décomposition de notions en leurs composantes essentielles. De cette longue tradition, l'analyse conceptuelle a hérité une grande diversité de méthodes qui varient tant par les moyens que par les buts poursuivis. Ainsi, Beaney (2015), dans un long survol historique des méthodes d'analyse, les regroupe dans les grandes catégories des approches génératives (retrouver les causes), interprétatives et décompositionnelles. Haslanger (2012, chap. 13) s'inspire plutôt de Quine (1951) pour les regrouper selon le type de questions qu'elles posent : on retrouve ainsi les analyses descriptives (« Qu'est-ce que X auquel réfère notre concept de X ? »), conceptuelles (« Quel est notre concept de X ? ») et amélioratives (« Pour remplir son rôle, comment devrait être notre concept de X ? »). À chacune de ces approches, on peut ajouter une dimension temporelle qui caractérise les approches généalogiques (enquêtes sur la genèse et le développement du concept). De surcroît, le concept peut être manifeste (le concept tel qu'on en parle explicitement), opérationnel (le concept tel qu'il se manifeste implicitement à travers nos pratiques) ou objectif (target concept : dans le cadre d'une analyse améliorative, le concept qui remplirait le mieux son rôle), avec les conséquences méthodologiques qui s'en suivent.

Ces différents types d'analyse conceptuelle relèvent souvent de projets forts différents. Ainsi, une analyse proprement conceptuelle du concept de PARENTALITÉ (pour reprendre l'exemple de Haslanger, 2006) n'invalide pas une analyse améliorative du même concept : le fait que l'on conçoive généralement la parentalité comme essentiellement biologique ou génétique n'en fait pas nécessairement le meilleur concept pour rendre compte des familles dans nos sociétés ou pour orienter les institutions qui s'adressent à elles. Cependant, ils emploient des outils semblables pour parvenir à leurs fins : expériences de pensées, pompes à intuitions, historiographies philosophiques, interprétation de textes, etc.

Un nouveau venu parmi ces outils est la lecture et analyse de textes assistée par ordinateur (LATAO, Meunier et coll., 2005). Celle-ci vise à exploiter les techniques de fouille de texte dans le but, d'une part, de permettre l'étude de corpus trop volumineux pour

être lus de façon traditionnelle et, d'autre part, de permettre de jeter un regard neuf sur ceux-ci, puisque l'ordinateur nous révèle souvent des aspects du texte qui passent inaperçus à la lecture (p. ex., Chartrand et coll., 2015). À travers l'étude du lexique, qui révèle les thèmes étudiés dans un corpus (Forest, 2006) et leurs interrelations, la LATAO peut assister la chercheuse ou le chercheur à cibler les passages-clés et à indiquer, par exemple, les types de propos qui sont associés avec un concept. En ce sens, elle a le potentiel de faciliter grandement la recherche documentaire : par exemple, elle peut assister la modélisation de l'évolution d'un concept à travers le temps pour faciliter une étude généalogique, ou donner un instantané des thèmes associés à un concept pour une recherche proprement conceptuelle.

Cependant, pour ce faire, il manque encore à la LATAO un modèle satisfaisant de l'expression du concept dans le texte. L'occurrence d'une lexie canonique (p. ex., si le mot « esprit » est présent, on en induit que le concept d'esprit est exprimé), qui est l'heuristique la plus commune pour ce faire, souffre de plusieurs défauts. Premièrement, un mot ou un lexème peut posséder plusieurs significations (polysémie). Un mot comme « esprit » peut avoir un sens proche de « âme » ou « substance pensante », mais peut aussi référer à un fantôme ou être employé dans une locution comme « dans l'esprit de la fête de Noël ». Inversement, on peut parler de l'esprit sans jamais en employer le mot, usant à la place de périphrases comme « le sujet qui pense » ou de synonymes comme « entendement » (synonymie). Deuxièmement, dans le cadre d'une analyse conceptuelle, ce ne sont pas seulement les phrases contenant un lexème pouvant référer au concept étudié qui sont intéressantes, mais aussi celles qui, sans l'impliquer explicitement, peuvent contribuer à son analyse et fournir les clés de son interprétation. On cherchera donc non seulement les segments où le concept ciblé est exprimé implicitement ou explicitement, mais aussi ceux où il est joue un rôle de façon indirecte. Par exemple, dans le cas où le concept manifeste diverge du concept opérationnel, il faut savoir extraire les segments où le concept opérationnel est exprimé, même s'il n'est pas directement mis en scène. En ce sens, il faut un modèle du concept qui permette d'explorer ces voies.

Notre problème se dessine donc ainsi : nous cherchons à automatiser le rappel des segments textuels qui contiennent l'information nécessaire à l'analyse d'un concept donné. Autrement dit, notre algorithme devrait être en mesure, à partir d'une requête désignant le concept à rechercher, de rappeler les segments textuels qu'une personne experte désignerait comme utile à l'analyse conceptuelle de ce concept. Conséquemment, notre hypothèse est qu'il est possible d'automatiser la collecte de segments textuels en vue d'une analyse conceptuelle. Nous entendons le démontrer à l'aide d'une preuve de concept, en décrivant et en appliquant une méthode sur un texte philosophique, soit *Gender Trouble* de Judith Butler. Dans la recherche que nous présentons ici, nous proposons une approche pour rappeler les segments de textes pertinents à un concept en exploitant les régularités lexicales dans le texte.

2. Hypothèses de modélisation

Dans cette section, nous clarifions les termes contenus dans la problématique qui, bien qu'ils puissent sembler clairs dans un contexte de pratique, ne le sont pas dans un contexte d'implémentation. Le mot « concept », par exemple, est très polysémique, puisqu'il peut référer à différents types de représentations jouant différents rôles dans la cognition (Machery, 2009 ; Harnad, 2009). De la même façon, il existe une grande variété d'approches pour rendre compte du contenu d'une phrase (Beaney, 2015), et en l'absence de contexte fournissant des critères, il semble impossible de déterminer avec assurance de la pertinence ou de la non-pertinence d'un contenu pour l'analyse d'un concept.

Dans une large mesure, il n'y a pas lieu de définir ces termes : il importe peu, par exemple, de savoir quel rôle le concept joue dans l'apprentissage ou dans le comportement en général, pourvu qu'on puisse s'exprimer sur sa façon de se manifester dans le texte. Aussi notre approche s'appuie-t-elle plutôt sur les quatre hypothèses suivantes, qui nous permettent de concevoir un certain rapport entre le texte et ses concepts au travers duquel on peut étudier les seconds à partir du premier.

Hypothèse 1. Est pertinent à un concept ce qui participe aux thèmes où intervient ledit concept.

Le thème, qui se traduit autant par *topic* que par *theme* dans la littérature anglophone, apparaît dans la littérature scientifique comme étant ce à propos de quoi est un ensemble de segments textuels qui ont cet « à propos » en commun. Cependant, au-delà de cette caractérisation très générale, autant « thème » que « *topic* » prennent des sens différents selon la discipline et l'auteur.e qui les invoque (Rimmon-Kenan, 1995). Bien que le thème soit parfois associé à la phrase ou la proposition, et qu'il soit souvent contrasté au rhème comme rassemblant les informations déjà connues de la personne qui lit, ce n'est pas dans ce sens où nous l'entendons ici. Pour nos fins, le thème est plutôt une macrostructure (Van Dijk, 1977) qui rassemble un ensemble de phrases et plus généralement de segments de texte qui partagent des similarités au niveau du contenu. Ainsi, dans des segments d'un même thème interviendront des concepts communs ou semblables, lesquels seront en relations entre eux de façon cohérente au travers des différentes manifestations du thème dans le texte.

Le thème peut nous aider à produire un critère de pertinence d'un segment de texte pour l'étude d'un concept. Comme on suppose que le thème constitue une macrostructure discursive qui implique des concepts, et que tous les segments qui participent d'un thème nous informent sur celle-ci, on peut supposer qu'ils nous informent aussi indirectement sur les concepts qu'elle rassemble. Autrement dit, toute information sur le thème nous renseigne sur le concept qui y participe. Aussi, tous les segments textuels qui participent d'un thème apparaissent-ils ainsi comme pertinents pour l'étude d'un concept.

Hypothèse 2. Le thème est une variable sous-jacente du texte

En d'autres mots, la probabilité qu'un lexème apparaisse dans un segment dépend de la présence ou de l'absence des différents thèmes d'un corpus.

Un texte est fait de lexèmes – on ne peut y observer directement de concepts ou de thèmes. Pour accéder à ceux-ci à partir des données textuelles, on suppose donc qu'il existe une relation indéterminée entre, d'une part, les unités structurantes du discours comme le thème et, d'autre part, les observables du texte. En supposant que ces derniers dépendent conditionnellement des premiers, on peut alors, étant donné les lexèmes que l'on observe, estimer la probabilité que chaque thème soit présent.

Dire que des thèmes conditionnent le texte comme variables latentes signifie qu'il est possible de décrire la structuration du texte en employant des variables latentes qui conditionnent l'apparition des mots dans le texte, sans pour autant que l'on ait à décrire les mécanismes par lesquelles on en vient à observer une relation de dépendance conditionnelle. L'hypothèse 2 ne se prononce donc pas sur l'interprétation à faire de la nature de la dépendance conditionnelle.

Cependant, nous supposons que, lorsqu'un corpus est relativement uniforme au niveau pragmatique (les personnes impliquées dans la communication sont les mêmes, le jeu de langage reste à peu près le même), la structuration du texte que décrivent ces variables latentes permet de rassembler les segments de textes qui sont à propos de la même chose, et donc que ces regroupements peuvent être interprétés comme des thèmes. Nous rejoignons en cela la tradition des modèles topiques en informatique (Blei et coll., 2003 ; Landauer et coll., 1998).

> **Hypothèse 3**. L'empreinte lexicale d'un segment qui exprime un thème en particulier obéit, *ceteris paribus*, à une distribution normale ayant pour moyenne un vecteur prototypique dudit thème.

Il est possible de représenter un corpus sous la forme d'un modèle vectoriel de façon à permettre de calculer facilement la similarité entre paires de segments de texte. Bien qu'il existe des alternatives qui représentent aussi le texte sous forme vectorielle (Mikolov et coll., 2013 ; Pennington et coll., 2014), les méthodes basées sur le comptage de mots, popularisées notamment par Salton (1989 : 131-9), restent populaires et actuelles (Levy et coll., 2015) : soit T un corpus textuel en langue naturelle écrite, composé de segments $t_i \in T$. Soit W l'ensemble des mots de contenu $w_j \in W$ qui se retrouvent au moins une fois dans le corpus T. Alors on peut construire pour chaque segment t_i un vecteur $d_i = (a_{i1}, a_{i2} \dots a_{in})$, où $n = |W|$, et où $a_{ij} = 1$ si le mot w_j apparaît dans le segment t_i et $a_{ij} = 0$ autrement. Les vecteurs d_i constituent ensemble les colonnes de la matrice D, et partagent un même espace vectoriel, de sorte qu'on peut mesurer leur similarité de diverses façons (distance euclidienne, cosine, PPMI, etc.).

On suppose que, sur cet espace vectoriel, non seulement les segments qui appartiennent au même thème tendent à se regrouper, mais que ces regroupements ont un point central (le vecteur prototype du thème) autour duquel les segments se distribuent en

suivant une distribution normale (gaussienne). Autrement dit, on considère que les facteurs qui font qu'un segment s'éloigne du vecteur prototype thème sont des facteurs indépendants des propriétés de ce prototype que l'on peut approximer aléatoirement.

> **Hypothèse 4.** L'ensemble des segments qui contiennent l'expression canonique d'un concept constitue, en termes d'expression des thèmes, un échantillon représentatif de l'ensemble des segments pertinents audit concept.

Comme on l'a dit précédemment, l'expression canonique d'un concept (p. ex., le mot « arbre » pour parler du concept ARBRE) n'est pas un signe certain de la présence d'un concept dans le contenu d'un segment de texte : l'expression canonique peut être utilisée sans évoquer le sens voulu (homonyme, métaphorique, expression consacrée, etc.) ou, au contraire, le concept peut être évoqué sans recourir à celle-ci (notamment à l'aide d'une périphrase). Cependant, mesurer la présence d'un concept à la présence de cette expression reste une heuristique éprouvée en analyse de texte : en témoigne la popularité constante des concordanciers (Pincemin, 2006) et les développements plus récents de techniques d'analyse conceptuelle qui continuent de l'employer (p. ex., Meunier et coll., 2005 ; Sainte-Marie et coll., 2011).

De fait, si on peut s'attendre à ce que cette heuristique ne permette pas de trouver tous les segments exprimant un concept donné, elle peut quand même fournir à l'analyste des segments qui expriment presque tous les aspects d'un concept donné. Le contexte de l'analyse de texte présente souvent des dynamiques qui favorisent que les segments contenant le mot canonique associé à un concept X expriment presque toujours effectivement ce concept X. Premièrement, on choisit rarement de faire une analyse conceptuelle sur un concept banal – il s'agit presque toujours d'un concept qui soulève des enjeux d'interprétation, et donc dispose d'une expression consacrée. De plus, l'analyste ayant la liberté d'adapter la requête au concept, ille la modulera le plus souvent de façon à exclure des lexèmes ambigus. Deuxièmement, les corpus d'étude ont souvent une certaine homogénéité qui rend également l'expression des concepts plus homogène, et encore une fois, la sélection se fait parfois de façon à exclure des emplois divergents de l'expression canonique. Troisièmement, les auteur.e.s mêmes, lorsque confronté.e.s à un concept qu'illes jugent important, ont tendance à éviter d'employer une expression canonique dans un sens différent, afin d'éviter d'être mal compris.es.

Si l'usage de mots-clés permet de rappeler seulement une fraction des segments où s'exprime le concept donné, mais très peu de segments où il ne s'exprime pas, on peut supposer, par approximation, qu'il s'agit d'un échantillon de l'ensemble des segments où l'on peut retrouver le concept d'intérêt. On suppose également que cet échantillon est représentatif des thèmes auxquels participe le concept, c'est-à-dire que les segments exprimant chaque thème se retrouveront dans des proportions semblables dans l'échantillon et dans l'ensemble des segments où le concept se trouve exprimé.

3. Méthode

Dans cette section, nous présentons notre approche, qui consiste à employer le thème et son extension comme une passerelle vers l'extension des segments pertinents à l'analyse du concept (suivant l'hypothèse 1).

Sommairement, notre stratégie consiste d'abord à extraire automatiquement les thèmes du sous-corpus des segments contenant l'expression canonique du concept étudié (le fait de se restreindre à ce sous-corpus permettant de mieux distinguer les différents thèmes qui s'y retrouvent et rend le calcul plus rapide). L'hypothèse 4 nous dit que ces thèmes sont ceux où est exprimé le concept d'intérêt. Ensuite, pour chaque thème, on emploie l'hypothèse 3 pour modéliser la distribution des segments de texte participant d'un thème, et on rappelle tous les segments faisant partie de la zone à l'intérieur de laquelle, selon le modèle, 95 % des segments participant du thème devraient se retrouver. Le résultat du rappel d'information est l'ensemble de tous les segments rappelés.

En termes plus précis, la méthode que nous employons ici se présente sous la forme d'une chaîne de traitement (3.1) qui implémente l'algorithme CoFiH (algorithme 1), à quelques ajustements près. Elle part d'une requête qui exprime le concept à analyser et qui peut être exprimé sous forme d'un ensemble de mots qui désignent typiquement un concept Q (p. ex., le concept de CORPS donnerait : {« corps », « corporel », ...}), et peut être représentée sous forme d'un vecteur $q = \{a_{q1}, \dots, a_{qN}\}$ sur l'espace vectoriel de la matrice D. Elle vise à extraire l'ensemble $E \subset D$ des segments textuels qui sont pertinents à l'analyse du concept Q.

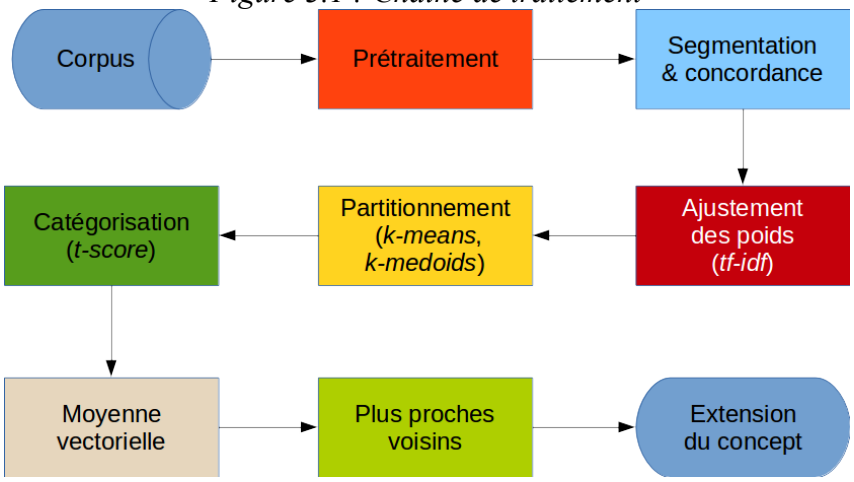
D'abord, le texte est prétraité, ce qui comprend l'exclusion des caractères non alphabétiques, l'extraction des mots, l'exclusion des mots fonctionnels et la lemmatisation. On procède ensuite à la segmentation et à la construction de la matrice D, puis on produit une concordance – c'est-à-dire qu'on fait une deuxième matrice qui ne conserve que les segments de texte qui contiennent les mots indiqués dans la requête q (ligne 2, algorithme 1). Suit l'ajustement des poids de la matrice B à l'aide d'une pondération TF-IDF, après quoi on effectue un partitionnement non supervisé sur B (ligne 3). Ensuite, pour chaque classe c_k obtenue, on calcule la statistique t (t-score) de chaque mot, et on construit une matrice D'_k à partir de D en n'incluant que les colonnes des mots qui ont les l meilleures statistiques t et en ne gardant que les lignes qui appartiennent à la classe c_k (lignes 5 à 7). Ici, on choisit $l = 200$, mais il peut être commode de définir l comme une fraction α de N (où $\alpha \in [0; 1]$). Dans l'espace formé par $D_k{}'$, on fait la moyenne μ_k (ligne 8) et l'écart-type σ_k (ligne 9). On suppose que l'extension d'un thème, modélisé par la classe c_k, suit une distribution normale autour d'un point central, et donc que les segments qui expriment ce thème s'agglutinent autour de son centre. On construit donc pour chaque classe c_k l'extension J_k du thème correspondant qui contient tous les segments dont la distance euclidienne est inférieure ou égale à $2\sigma_k$ (ligne 10). L'extension E du concept est donc l'union des extensions J_k (ligne 11), à l'exclusion des thèmes que l'analyste aura jugés impertinents à son analyse.

Algorithme 1 : CoFiH

Données : D, q, α
Résultat : E

$l \leftarrow \alpha N$;
$B \leftarrow \{d_i | d_i \cdot q > 0 \, d_i \in D\}$;
$C \leftarrow \text{Partition}(B)$;
pour chaque $c_k \in C$ **faire**
  $T_k \leftarrow \{t_i | \text{tScore}_i \, (D, c_k) \text{ parmis les } l \text{ plus hauts}\}$;
  $D' \leftarrow (\{b_i | b_i \in D^T, i \in T_k\})^T$;
  $D'_k \leftarrow \{d'_i | d'_i \in D', i \in c_k\}$;
  $\mu_k \leftarrow \frac{1}{|c_k|} \sum_{d'_i \in D'_k} d'_i$;
  $\sigma_k \leftarrow \sqrt{\frac{1}{|c_k|} \sum_{d'_i \in D'_k} \|d_i - \mu_k\|^2}$;
  $J_k \leftarrow \{d_i | \, |d'_i - \mu_k| \leq 2\sigma_k\}$;
$E \leftarrow \bigcap_k J_k$

Figure 3.1 : Chaîne de traitement



4. Expérimentation

En guise d'expérimentation préliminaire, notre chaîne de traitement est appliquée au livre Gender Trouble de Judith Butler, qui a été préparé et segmenté en 295 segments de textes qui correspondent aux paragraphes tels que l'auteure les a déterminés. Le concept choisi pour l'analyse conceptuelle est celui de LANGAGE, lequel est réputé structurant dans ce livre puisqu'il aborde la question de la normativité à travers les formes langagières. Une autre raison pour le choix de ce concept est qu'il ne pose pas *a priori* de difficulté particulière – par exemple, le mot canonique correspondant au concept étudié n'est pas particulièrement polysémique dans ce corpus : lorsque Butler emploie le mot « *language* », elle réfère constamment au même concept spécifique de LANGAGE. Le corpus étant lemmatisé et donc exempt de formes plurielles, le vecteur q ne contient une valeur que pour le mot « *language* ».

Algorithme 2 : Partitionnement médiatisé par une matrice de consensus

Données : Matrice D
Résultat : Partition C de dimension $m \times n$
initialiser D' de dimension $m \times m$;
pour $k = 2$ **à** $m - 1$ **faire**
    $p \leftarrow \text{kmeans}(D, k)$;
    pour $i = 1$ **à** m **faire**
        pour $j = i + 1$ **à** m **faire**
            si $p_i = p_j$ **alors**
                $D'_i j \leftarrow D_i j + 1$;
                $D'_j i \leftarrow D_j i + 1$;
$D' \leftarrow \text{Normaliser}(D')$;
$k \leftarrow 2$;
On paramètre kmedoids pour rendre la meilleure partition après 100 000 itérations ;
répéter
    pour $i \leftarrow 1$ **à** 5 **faire**
        $p_i \leftarrow \text{kmedoids}(D', k)$;
    $k \leftarrow k + 1$;
jusqu'à *ce que* $p_1 ... p_5$ *soient tous différents*;
$C \leftarrow \text{kmedoids}(D', k - 1)$;

On suppose que la fonction k-means a pour sortie un vecteur de taille m contenant pour chaque vecteur d_i de A l'indice de la classe à laquelle il a été assigné.

En guise de méthode de partitionnement, on emploie une méthode basée sur un consensus de partitions *k-means* décrite par Chartrand et coll. (2015) et illustrée par l'algorithme 2. D'autres méthodes de partitionnement non supervisées auraient pu être employées.

5. Résultats préliminaires

L'application de notre méthode permet de rappeler un total de 181 segments sur 295 et issus de six thèmes différents. En guise de comparaison, seuls 79 segments dans tout le corpus contiennent le mot « *language* », au singulier ou au pluriel. Les dix mots ayant les plus hautes statistiques *t* pour chaque thème sont illustrés au tableau 5.1.

5.1 Thématiques observées

Tableau 5.1 : Mots ayant les plus hautes statistiques t pour chaque thème de LANGAGE

$J_1, N = 65$		$J_2, N = 99$		$J_3, N = 38$	
language	2.14572	gender	2.88358	sex	2.23217
law	1.5134	sex	2.70939	gender	1.66755
within	1.47986	language	2.19573	category	1.42242
subject	1.4615	body	2.11899	wittig	1.38842
symbolic	1.19562	within	2.01178	woman	1.37262
kristeva	1.19289	category	1.82115	within	1.19541
gender	1.16644	wittig	1.78457	body	1.12398
drive	1.15816	identity	1.75503	one	1.03498
body	1.15169	cultural	1.60528	cultural	1.02591
terms	1.13246	power	1.50021	subject	0.973431

$J_4, N = 2$		$J_5, N = 34$		$J_6, N = 52$		$J_7, N = 154$	
drive	0.735424	language	1.7627	subject	1.97031	gender	3.01173
language	0.661291	kristeva	1.60412	language	1.82851	body	2.75759
poetic	0.638354	law	1.58619	gender	1.59314	sex	2.71506
linguistic	0.571098	maternal	1.37928	wittig	1.52947	within	2.45032
univocal	0.493756	paternal	1.36366	woman	1.32914	identity	2.39456
symbolic	0.467318	semiotic	1.25474	sex	1.32467	cultural	1.9093
meaning	0.452355	drives	1.21694	system	1.21063	can	1.90626
manifest	0.447802	symbolic	1.1544	power	1.15767	language	1.89055
kristeva	0.422382	body	1.15128	theory	1.11439	one	1.88027
domain	0.408491	within	1.1395	within	1.07304	woman	1.79879

À partir des mots du tableau 5.1, on peut produire une interprétation de chaque classe qui les emploie et les met en relation les uns avec les autres. Ainsi, on peut interpréter la classe J_1 comme parlant des normativités (*laws*) à l'égard du corps (*body*, *gender*) à travers le langage (*language*, *symbolic*, *within*), les classes J_2, J_3 et J_7 comme prenant pour objet principal le duo genre/sexe, la classe J_5 comme parlant du langage en rapport aux rôles familiaux (duo paternel/maternel), et la classe J_6 comme évoquant la construction de la subjectivité féminine à l'intérieur d'un système de pouvoir. On note aussi une certaine association de vocabulaire avec des auteur.e.s : « *symbolic*, *semiotic*, *drive* », avec Kristeva et « *gender*, *sex*, *woman* », avec Wittig.

Ces interprétations semblent se confirmer dans certains passages qui font partie des classes en question. Par exemple, dans J_1, dont le tableau 5.1 suggère qu'elle traite des normativités à l'égard du corps, on retrouve des remarques comme la suivante, portant sur le langage en rapport aux normes sociales :

> Kristeva challenges the Lacanian narrative which assumes cultural meaning requires the repression of that primary relationship to the maternal body. She argues that the semiotic is a dimension of language occasioned by that primary maternal body, which not only refutes Lacan's primary premise, but serves as a perpetual source of subversion within the Symbolic. For Kristeva, the semiotic expresses that original libidinal multiplicity within the very terms of culture, more precisely, within poetic language in which multiple meanings and semantic nonclosure prevail. In effect, poetic language is the recovery of the maternal body within the terms of language, one that has the potential to disrupt, subvert, and displace the paternal law.

Dans J_5, qui semble traiter du langage en rapport aux rôles familiaux, on retrouve des passages comme celui-ci, liant les pulsions (*drives*) à un renversement de l'ordre familial/paternel :

> The alleged psychosis of homosexuality, then, consists in its thorough break with the paternal law and with the grounding of the female ego, tenuous though it may be, in the melancholic response to separation from the maternal body. Hence, according to Kristeva, female homosexuality is the emergence of psychosis into culture: The homosexual-maternal facet is a whirl of words, a complete absence of meaning and seeing; it is feeling, displacement, rhythm, sound, flashes, and fantasied clinging to the maternal body as a screen against the plunge... for woman, a paradise lost but seemingly close at hand. For women, however, this homosexuality is manifest in poetic language which becomes, in fact, the only form of the semiotic, besides childbirth, which can be sustained within the terms of the Symbolic.

et dans J_6, dont le tableau 5.1 nous dit qu'elle évoque la construction de la subjectivité féminine à l'intérieur d'un système de pouvoir, on retrouve des réflexions comme celle-ci, sur la (re)construction de l'identité féminine :

> Precisely because female no longer appears to be a stable notion, its meaning is as troubled and unfixed as woman, and because both terms gain their troubled significations only as relational terms, this inquiry takes as its focus gender and the relational analysis it suggests. Further, it is no longer clear that feminist theory ought to try to settle the questions of primary identity in order

> to get on with the task of politics. Instead, we ought to ask, what political possibilities are the consequence of a radical critique of the categories of identity. What new shape of politics emerges when identity as a common ground no longer constrains the discourse on feminist politics? And to what extent does the effort to locate a common identity as the foundation for a feminist politics preclude a radical inquiry into the political construction and regulation of identity itself?

Il semble donc que parmi les segments que rappelle CoFiH, il y en a qui correspondent aux interprétations que l'on peut produire en lisant les listes de mots du tableau 5.1.

5.2 Comparaison avec une classification humaine

Afin de pouvoir quantifier l'efficacité de notre méthode, nous avons tenté de mesurer les résultats obtenus par CoFiH en relation avec un étalon de référence produit par un annotateur humain. Celui-ci (l'auteur principal) a classifié un sous-ensemble du corpus composé de 50 segments choisis au hasard selon la pertinence ou la non-pertinence pour l'étude du concept de **langage** chez Butler. L'étalon a ensuite été employé pour évaluer les résultats obtenus par CoFiH sur les mêmes segments, de même que ceux qu'on obtient en employant une heuristique de base consistant à rappeler les segments qui contiennent le mot « *language* » au singulier ou au pluriel.

Pour ce faire, nous employons les mesures de rappel, de précision et la mesure F1, qui sont couramment employées pour évaluer les algorithmes de recherche d'information. Le *rappel* est la part des documents retrouvés par l'heuristique qui sont véritablement pertinents selon l'étalon de référence. La *précision* est la part des documents pertinents selon l'étalon de référence qui ont été retrouvés par l'heuristique. La *mesure F1*, quant à elle, est une moyenne harmonique du rappel et de la précision, et équivaut au double du produit de ces mesures divisé par leur somme. Au contraire de la précision et du rappel, sur lesquels on peut tricher respectivement en diminuant ou en augmentant le nombre de segments retrouvés, la mesure F1 est souvent utilisée comme indicateur général de la performance d'un algorithme de rappel d'information, les deux autres mesures servant plutôt à l'analyse. Les résultats sont illustrés au tableau 5.2.

Tableau 5.2 : Efficacité sur le corpus de test

	Rappel	Précision	Mesure F1
CoFiH	**0,79**	0,71	**0,75**
Heuristique de base	0,36	**0,91**	0,51

6. Discussion

Notre méthode se démarque notamment parce qu'elle introduit la notion de thème comme médium pour retrouver les contextes pertinents au concept étudié, ce qui permet aussi de représenter les concepts sous des formes qui peuvent être d'assistance à l'analyste. Ainsi, comme preuve de concept, nous avons illustré comment les résultats reproduits dans le tableau 5.1 peuvent être interprétés comme reflétant certaines thématiques, lesquelles peuvent ensuite se retrouver à la lecture du texte.

Ces résultats sont cependant anecdotiques, et la forme des représentations des thèmes (à l'aide de mots-clés) peut apparaître comme n'étant pas particulièrement informative. En effet, cette représentation est bruitée, en ce sens que l'ordre des termes ne reflète pas parfaitement la thématique qu'on peut y lire, et on peut supposer que ce bruit reflète sans doute d'autres dynamiques dans le texte qui sont encore inexpliquées. De plus, la simple succession de mots-clés ne rend pas compte de l'ambiguïté des termes ni des relations que les lexèmes ou les concepts qu'ils expriment entretiennent les uns avec les autres. Enfin, sans les outils de validation pour ce genre de tâches qui restent encore à développer, il est difficile de se faire une idée de leur réelle utilité.

Ceci dit, notre principal objectif reste d'identifier des segments qui sont pertinents à l'analyse parce qu'ils expriment le concept à analyser. À ce titre, en termes absolus, CoFiH réussit beaucoup mieux que la simple concordance, augmentant le score F1 de presque 50 % (tableau 5.2). Bien que l'on n'ait pas de données ici sur l'accord entre personnes expertes, l'exercice d'annotation étant assez subjectif, il est probable que leurs jugements varieraient dans une bonne mesure, à côté de laquelle un score F1 de 0,75 ferait relativement bonne figure.

Il convient cependant d'admettre que l'heuristique de base semble connaître sa part de succès, avec une précision très forte (91 %). Le sens commun en recherche d'information veut que le succès dans la précision se fasse souvent aux dépens du rappel : si on choisit un seul segment dont on est certain qu'il possède la qualité voulue, on obtient une précision de 100 %, mais un rappel très bas puisqu'on n'a qu'un seul des segments parmi l'ensemble de ceux que l'on veut rappeler. Suivant cette logique, on peut supposer que l'heuristique de base parvienne à avoir une forte précision parce qu'elle rappelle très peu de segments. Par ailleurs, on peut croire que cette haute précision s'explique aussi par les facteurs mentionnés à la section 2 dans la discussion en marge de l'hypothèse 4. De fait, il semble que Butler n'emploie que rarement le terme « *language* » dans les cas où il n'exprime pas le concept qu'elle développe dans le cadre de *Gender Trouble*. Or le danger d'une telle stratégie, c'est que ces facteurs ne peuvent jouer pour toutes les analyses conceptuelles. Par exemple, dans des corpus où on étudie un même mot qui peut prendre plusieurs sens, on peut s'attendre à ce que l'heuristique de base soit beaucoup moins efficace, puisque la précision s'effondrerait.

En ce sens, non seulement CoFiH paraît-il supérieur en termes de rappel, mais on peut croire que ses performances en termes de précision resteraient bonnes dans le cadre de corpus différents. Certes, l'algorithme CoFiH est également susceptible de perdre en précision, puisqu'il emploie lui aussi une stratégie de rappel des segments à l'aide de mots-clés (ligne 2, algorithme 1). Cependant, cette vulnérabilité est mitigée d'au moins deux façons. Premièrement, CoFiH ne nécessite pas nécessairement que l'échantillon formé par les

segments qui contiennent l'expression canonique du concept étudié soit parfaitement représentatif de la proportion des thèmes : même si les proportions varient, il suffit que les thèmes restent suffisamment présents pour être détectés par l'algorithme de partitionnement. Autrement dit, l'hypothèse 4 peut toujours être soutenue pour des raisons méthodologiques même si elle ne se vérifie pas empiriquement, dans la mesure où cela n'empêche pas l'algorithme de détecter tous les thèmes pertinents. Deuxièmement, si une plus grande polysémie des mots canoniques vient introduire des thèmes qui ne sont manifestement pas pertinents à l'analyse que la personne experte souhaite effectuer, par exemple parce qu'ils impliquent des homonymes du concept étudié, on peut facilement soustraire ces thèmes de l'extension finale.

Malgré ses résultats encourageants, l'évaluation que nous avons présentée en section 5.2 a cependant ses limites. D'une part, le fait de n'avoir qu'un seul annotateur pose problème, et il faudrait refaire l'exercice avec d'autres expert.e.s pour avoir des résultats fiables. D'autre part, comme on l'a mentionné plus haut, l'exercice d'annotation est difficile : même une personne experte ne peut savoir hors de tout doute si un segment est pertinent à l'analyse conceptuelle, car c'est souvent en faisant l'analyse conceptuelle que l'on découvre ce qui nous est utile. C'est d'autant plus vrai que différentes analyses avec différents angles d'approches et différents objectifs ne se pencheront probablement pas sur les mêmes segments. Le travail de la personne qui fait l'annotation est de produire un indicateur d'utilité d'un segment pour une analyse conceptuelle sur un concept X, mais cette utilité est une notion très floue, et il est très difficile d'instruire l'annotation à cet effet. Dès lors, non seulement l'étalon de référence peut-il être inexact et refléter des effets subjectifs, mais si la consigne est mal comprise, il pourrait être carrément trompeur, et il serait difficile de s'en rendre compte si la confusion affecte la plupart des gens qui font l'annotation. En somme, d'une part, il nous faut plus de données d'annotation et trouver des façons de valider les validations, et d'autre part, il faut s'assurer que la performance telle qu'indiquée par un outil comme la mesure F1 d'une comparaison à un étalon de référence est garante de la pratique.

Ceci dit, les résultats rapportés ici semblent néanmoins prometteurs en termes de pratique. S'il est clair que les méthodes de représentations des thèmes pourraient être développées, il n'en reste pas moins que l'on peut immédiatement voir deux usages très concrets à CoFiH dans l'analyse conceptuelle : (1) réduire le corpus d'étude à un ensemble plus limité de contextes textuels susceptibles d'exprimer le concept étudié, permettant ainsi de réduire la charge de lecture, et (2) exploiter la classification par thème pour produire des pistes d'analyses qui peuvent être employées pour diriger l'analyse conceptuelle.

Conclusion

Dans cet article, nous avons, d'une part, introduit l'idée de passer par le thème comme passerelle dans le contexte de la recherche des segments textuels pertinents pour une analyse

conceptuelle et, d'autre part, présenté une méthode et un algorithme opérationnalisant cette idée, le tout dans une perspective d'assistance à l'expert.e.

Tout ce qui est présenté ici, que ce soit l'articulation de la notion de thème, l'algorithme CoFiH ou la validation, constitue un premier effort et mérite d'être développé. En effet, notre articulation du thème constitue une première articulation qui pourrait être peaufinée, l'algorithme CoFiH constitue une heuristique relativement rudimentaire qui ne modélise pas dans le détail l'expression du thème ou du concept dans le texte, et la validation est assez limitée, étant basée sur un seul texte, un seul annotateur, et un ensemble de tests somme toute limité.

Nos résultats semblent cependant indiquer qu'il s'agit d'un horizon prometteur. D'une part, l'ampleur de la supériorité de CoFiH sur l'heuristique de base dans notre évaluation et l'absence de méthode alternative (à notre connaissance) pour la détection de concept dans un cadre comme le nôtre suggère que CoFiH peut d'ores et déjà constituer une option raisonnable pour l'assistance à l'analyse conceptuelle. D'autre part, on peut espérer que de nouveaux développements dans la modélisation du thème pourront produire des résultats encore meilleurs sur des tâches de rappel de segments, et qu'ils permettront de développer des techniques s'appliquant à d'autres thèmes de l'assistance à l'analyse conceptuelle, voire à d'autres domaines du traitement automatique des langues ou de la linguistique computationnelle.

Notices biobibliographiques

Titulaire d'une maîtrise en philosophie, **Louis Chartrand** est candidat au doctorat en informatique cognitive à l'Université du Québec à Montréal, et détenteur d'une bourse doctorale du CRSH. Il est membre du Laboratoire d'analyse cognitive de l'information (LANCI) depuis 2010, où il travaille au sein de l'unité de Lecture et analyse de texte assistée par ordinateur (LATAO), qu'il a coordonnée de 2013 à 2015. Ses intérêts de recherche se situent principalement en philosophie de l'esprit et des sciences cognitives, en analyse conceptuelle, en fouille de texte et en analyse des réseaux sociaux et sociosémantiques.

Jean Guy Meunier est professeur titulaire au département de philosophie de l'Université du Québec à Montréal. Son enseignement s'étend sur plusieurs domaines, passant par l'informatique cognitive, la philosophie et la sémiotique. Co-directeur du Laboratoire d'analyse cognitive de l'information (LANCI), il est également associé à l'Institut des sciences cognitives de l'UQAM et membre de l'Académie de philosophie des sciences. En 2007, il a reçu le prix Réalisation exceptionnelle dans le domaine des arts et des lettres en mode numérique de la Société canadienne des humanités numériques, en reconnaissance de son travail pionnier dans la lecture et l'analyse de texte assistée par ordinateur. Il a publié plus d'une centaine d'articles sur plusieurs disciplines allant de la philosophie à l'informatique en passant par la linguistique computationnelle. Il dirige présentement une subvention majeure en Forage conceptuel de texte assisté par ordinateur.

Détenteur d'une maîtrise en communications, **Davide Pulizzotto** est étudiant au doctorat en sémiotique à l'Université du Québec à Montréal. Il est membre du Laboratoire d'analyse cognitive de l'information (LANCI) depuis 2010, et y coordonne l'unité de Lecture et analyse de texte assistée par ordinateur (LATAO) depuis 2015. Son projet doctoral vise à intégrer les outils

sémiotiques fondés sur les structures narratives aux méthodologies de l'analyse de texte assistée par ordinateur.

Francis Lareau est titulaire d'une maîtrise en philosophie. Il est membre du Laboratoire d'analyse cognitive de l'information (LANCI) depuis 2013. Ses intérêts de recherche portent sur le forage conceptuel, la philosophie de l'esprit et la fouille d'argument.

Ouvrages cités

BEANEY, M. (2015), « Analysis », dans E. N. ZALTA (dir.), *The Stanford Encyclopedia of Philosophy (Spring 2015 edition)*, http://plato.stanford.edu/archives/spr2015/entries/analysis/, consulté le 10 décembre 2015.

BLEI, D. M., A. Y. NG, et M. I. JORDAN (2003), « Latent Dirichlet Allocation », *Journal of Machine Learning Research* 3, p. 993-1022.

CHARTRAND, L. et J. G. MEUNIER (2015), « Peindre Magritte avec des mots : analyse conceptuelle dans l'œuvre de Magritte à l'aide d'un corpus de descripteurs sémiotiques », *Les cahiers de l'ISC*, nº 4.

FOREST, D. (2006), « Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés », Thèse de doctorat, Montréal, Université du Québec à Montréal.

HARNAD, S. (2009), « Concepts: The Very Idea », *Canadian Philosophical Association Symposium on Machery on Doing without Concept*, Ottawa/Ontario, http://eprints.soton.ac.uk/268029/, consulté le 19 novembre 2015.

HASLANGER, S. (2006), « What Good Are Our Intuitions?: Philosophical Analysis and Social Kinds », *Aristotelian Society Supplementary*, vol. 80, nº 1, p. 89-118.

HASLANGER, S. (2012), *Resisting Reality: Social Construction and Social Critique*, New York/NY, Oxford University Press.

LANDAUER, T. K., P. W. Foltz et D. LAHAM (1998), « An Introduction to Latent Semantic Analysis », *Discourse processes*, 25, nº 2-3, p. 259-284.

LEVY, O., Y. GOLDBERG et I. DAGAN (2015), « Improving Distributional Similarity with Lessons Learned from Word Embeddings », *Transactions of the Association for Computational Linguistics*, nº 3, p. 211-225.

MACHERY, E. (2009), *Doing Without Concepts*, New York, Oxford University Press.

MCNAMARA, T. P. (2005), *Semantic Priming. Perspectives from Memory and Word Recognition,* New York/NY et Hove/East Sussex, Psychology Press.

MEUNIER, J. G., I. BISKRI et D. FOREST (2005), « Classification and Categorization in Computer Assisted Reading and Analysis of Texts » dans H. Cohen, & C. Lefebvre (dir.), *Handbook of Categorization in Cognitive Science*, Amsterdam, Elsevier, p. 955-978.

MIKOLOV, T., I. SUTSKEVER, K. CHEN, G. CORRADO et D. JEFFREY (2013), « Distributed Representations of Words and Phrases and their Compositionality », *Advances in neural information processing systems*, nº 26, p. 3111-3119.

PENNINGTON, J., R. SOCHER, et C. D. MANNING (2014), « Glove: Global Vectors for Word Representation », *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12, p. 1532-1543.

PINCEMIN, B. (2006), « Concordances et concordanciers : de l'art du bon KWAC », dans F. RASTIER, M. BALLABRIGA, C. DUTEIL-MOUGEL et B. FOUQUIÉ (dir.), *XVII^e colloque d'Albi Langages et signification – Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation,* Albi, CALS-CPST, p. 33-42.

QUINE, W. V. O. (1951), « Two Dogmas of Empiricism », *Philosophical Review*, vol. 60, nº 1, p. 20-43.

RIMMON-KENAN, S. (1995), « What is a Theme and How Do We Get at it? », dans C. BREMOND, J. LANDY et T. G. PAVEL (dir.), *Thematics: New Approaches*, New York, State University of New York Press, p. 9-19.

SAINTE-MARIE, M. B., J. G. MEUNIER, N. PAYETTE et J.-F. CHARTIER (2011), « The Concept of Evolution in the *Origin of Species*: a Computer-Assisted Analysis », *Literary and Linguistic Computing*, vol. 26, nº 3, p. 329-334.

SALTON, G. (1989), *Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer*, Reading, Addison-Wesley.

VAN DIJK, T. A., (1977) *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*, New York/NY, Addison-Wesley Longman.