UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CHAÎNES DE TRAITEMENT POUR LA DÉTECTION DE CONCEPTS DANS LE CONTEXTE DE L'ANALYSE CONCEPTUELLE PHILOSOPHIQUE BASÉE SUR DES DONNÉES TEXTUELLES

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

LOUIS CHARTRAND

JUIN 2019

REMERCIEMENTS

Comme le dit Machery (2017), un·e auteur·e accumule des dettes envers ses proches et ses collègues qu'ille ne pourra jamais espérer pouvoir rembourser. La vérité, dans mon cas, c'est que je ne pourrai probablement même pas en rendre compte adéquatement. J'espère que celles et ceux que j'aurai oublié m'en feront part, afin que je puisse me faire pardonner.

En premier lieu, je désire mentionner mon directeur de thèse, Jean-Guy Meunier, mon guide et mon mentor qui, depuis huit ans, m'assure un soutien indéfectible. Je retiens ses nombreux enseignements sur le métier, le leadership, son ouverture et sa vision du futur. Qu'il reçoive ma plus vive gratitude et mes mercis les plus chaleureux.

Je remercie chaleureusement mon co-directeur, Mohamed Bouguessa pour toute son aide, sa compréhension et sa présence. Mon projet revêtant un caractère particulier, M. Bourguessa a cru en celui-ci, en a bien saisi le sens et la portée et m'a accompagné dans toutes les étapes touchant l'informatique.

Dans cette aventure, j'ai également eu la chance de côtoyer d'autres mentors qui m'ont prodigué une aide inestimable. Jackie Cheung m'a accueilli pour huit mois dans son groupe de travail en linguistique computationnelle à McGill, où j'y ai vécu une expérience extraordinaire avec lui et avec ses étudiant·es. Ayant un pied en linguistique et l'autre en informatique, Jackie m'a énormément aidé à réaliser mes ambitions. Élias Rizkallah m'a également beaucoup aidé à comprendre les problèmes auxquels je faisait face avec l'annotation. Johanne Saint-Charles et Pierre Mongeau m'ont également accueilli dans leur groupe de travail, où j'ai

pu bénéficier de leurs rétroactions et de leurs idées. J'aimerais particulièrement les remercier tous les quatre pour leurs encouragements, leur gentillesse et leur générosité tout au long du doctorat.

Mes collègues, en particulier Davide Pulizzotto, Jean-François Chartier, Maxime Sainte-Marie, Francis Lareau, Marie-Noëlle Bayle, Jean Danis, Alaidine Ben Ayed, Tan Ngoc Le et José López, avec qui j'ai travaillé sur la LACTAO, ont une grosse part dans cette thèse. Je les remercie pour leur générosité, leur énergie et leur confiance, et bien davantage : on a grandi ensemble, et j'espère que l'on continuera à le faire à travers d'autres projets. Je remercie également mes autres collègues du LANCI, qui méritent bien plus qu'une ligne de remerciement : Louise Caroline Bergeron, Alice Livadaru, Anne Brel Cloutier, Simon Brien, Julien Ouellette-Michaud, Janie Brisson, William Beauchemin, Julián Trujillo et Mylène Legault. Chacune et chacun d'entre eulles m'ont aidé à un point ou un autre de ma thèse, me prêtant leur oreille, m'aidant à clarifier mes idées, alimentant ma réflexion, répondant à mes questions, etc. Pour les mêmes raison, merci à Nathalie Voarino, dont l'amitié est pour moi un véritable pilier de résilience. Un gros merci aussi à Geneviève Dick, qui entre autre choses a relu le premier article de la thèse, ainsi qu'à Brooke Struck et à Benoit Potvin, qui ont répondu avec enthousiasme à mes questions et à mes invitations à discuter ceci ou cela. Merci surtout à mon amoureuse, Audrey Rousseau, qui m'a sensibilisé à la dimension du discours et qui m'en apprend chaque jour. Enfin, merci à Luce Vermette, René Chartrand et Eamon Leonard pour les relectures en fin de parcours.

Comme cette thèse repose en grande partie sur un travail d'annotation, j'ai dû recruter des juristes volontaires pour faire ces annotations. Je n'y serais jamais arrivé si je n'avais pu compter sur la générosité de gens de mon entourage, que je remercie très chaleureusement.

Je voudrais également remercier les étudiant·es que j'ai côtoyé·es dans le groupe de linguistique computationnelle de McGill et dans le Groupe-Réseaux de l'UQAM de m'avoir si bien reçu, et d'avoir si généreusement écouté et commenté mes idées.

Un gros merci également à Mylène Dagenais, à Hakim Lounis et à Elisabeth Lindsay qui m'ont appuyé et guidé dans les aspects plus techniques du parcours du doctorat, et ce avec énormément d'intelligence et de générosité. J'ai aussi eu la chance de pouvoir calculer les modèles topiques et d'enrobages de mots sur un bel ordinateur très puissant, mais qui a été le centre de toute une aventure bureaucratique. Je remercie chaleureusement Christophe Malaterre et Serge Robert pour avoir administré et défendu ce projet.

Finalement, ce qui est peut-être ma plus grosse dette de doctorat : j'aimerais remercier du fond du cœur tous les gens qui m'ont soutenu mentalement et émotionnellement pendant mon doctorat. Je pense en particulier à mes parents, mon amoureuse, Audrey, qui a vu de loin le pire, mais également à mes amies proches, Anne et Nathalie, avec qui j'ai eu des longues et très précieuses discussions, à Benoit, Eamon et Élise, qui m'ont donné leur soutien moral. Je pense également à celles et ceux pour qui j'ai été moins présent à cause de cette thèse. Merci aussi à ma famille pour le constant soutien.

Enfin, cette thèse a été financée de plusieurs façons par le CRSH (bourse de doctorat, bourse Armand-Bombardier, subvention Savoir) et le FRQSC (bourse de doctorat). J'ai également profité d'une subvention de rayonnement de la part de l'Institut des Sciences Cognitives, et j'ai utilisé du matériel qui a été financé par une subvention du FCI. J'aimerais remercier les organismes qui m'ont financé, et qui m'ont permis de travailler dans des conditions optimales.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

Tableau Page

LISTE DES FIGURES

Figure Page

RÉSUMÉ

Dans les dernières années, l'analyse conceptuelle en philosophie a pris un tournant empirique, notamment avec l'essor de la philosophie expérimentale et, plus localement, de la Lecture et analyse conceptuelle de texte assistée par ordinateur (LACTAO), ouvrant ici la porte au développement d'un type d'analyse conceptuelle basée sur l'étude des corpus de données textuelles. Cependant, certains défis techniques viennent encore freiner l'essor de ce type de méthode. En particulier, l'heuristique couramment employée pour détecter la présence d'un concept dans le texte, l'heuristique du mot-clé, tend à exclure systématiquement certains contextes où le concept est employé implicitement, et à inclure des contextes où le ou les mots que l'on associe habituellement à un concept sont employés dans un sens très différent. La présente thèse attaque ce problème en deux étapes, qui sont présentées dans trois articles. Dans une première étape, on discute les notions principales de cette question – ANALYSE CONCEPTUELLE et CONCEPT – afin d'interpréter le problème de la détection de la présence du concept dans le texte. Un portrait du type d'analyse conceptuelle philosophique susceptible de prendre en compte des données empiriques est avancé, et sur la base de celui-ci, on énonce un problème pour le concept de CONCEPT. Une solution est alors proposée en puisant dans la téléosémantique de Millikan (1984), et on montre comment son application permet à la fois de faire un protocole d'annotation pour la détection de la présence du concept dans le texte, et de proposer des avenues d'automatisation pour la même tâche. Dans une deuxième étape, des chaînes de traitement exploitant des modèles topiques sont conçues et sont évaluées. Pour l'évaluation de celles-ci, un protocole d'annotation est conçu et soumis à des participant·es. Deux ensembles de chaînes de traitement sont ensuite testées, l'un reposant sur l'allocation de Dirichlet latente (LDA) de Blei *et al.* (2003) et l'autre reposant sur le Latent Concept Topic Model de Hu et Tsujii (2016). Les résultats des chaînes de traitement des deux ensembles s'avèrent mieux corrélées avec les jugements humains que l'heuristique du mot-clé, mais les meilleurs résultats viennent de chaînes construites à partir de la LCTM, dont certaines sont également plus flexibles dans la formulation du concept ciblé qu'elles permettent.

MOT-CLÉS : philosophie expérimentale, analyse conceptuelle, LACTAO, modèle topique, enrobages de mots, sémantique distributionnelle

INTRODUCTION

Cette thèse puise sa motivation de deux traditions distinctes : d'une part, la Lecture et analyse conceptuelle de texte assistée par ordinateur (LACTAO), qui a été développée à l'UQAM par Jean-Guy Meunier et son équipe (Chartier *et al.*, 2008 ; Meunier *et al.*, 2005 ; Meunier et Forest, 2009 ; Sainte-Marie *et al.*, 2011), et, d'autre part, la philosophie expérimentale (Knobe, 2003 ; Knobe et Nichols, 2007 ; Weinberg *et al.*, 2001).

La LACTAO

La LACTAO est née de la volonté d'appliquer des méthodes algorithmiques afin de contribuer à l'interprétation de textes, en particulier dans l'optique de la recherche en sciences humaines. Son objectif n'est donc pas d'extraire certaines informations, de modéliser ou de répliquer le processus de compréhension, mais de faciliter la compréhension chez les lectrices et lecteurs expert·es humain·es en proposant de l'assistance sous diverses formes (Meunier *et al.*, 2005).

Pour ce faire, la LACTAO se concentre principalement sur des méthodes de découverte. Par exemple, on tentera de réduire la tâche de la personne qui lit en synthétisant les informations et en identifiant les passages les plus pertinents (Le *et al.*, 2016 ; Pulizzotto *et al.*, 2016). On établira un portrait des liens entre les concepts (Sainte-Marie *et al.*, 2011). On rassemblera et annotera les documents qui sont similaires (Meunier *et al.*, 2005). Etc.

Ce faisant, le soin d'interpréter les résultats et de les valider revient à l'interprète, c'est-à-dire la personne qui lit. C'est donc dire, par exemple, que la LACTAO

ne tente le plus souvent non pas de produire des preuves ou de l'évidence, mais plutôt d'orienter le regard de l'interprète dans une direction. Ce dernier peut ensuite tâcher de corroborer la lecture qu'ille en tire en revenant au texte, ne serait-ce que pour regarder des extraits.

Le texte, et plus précisément le corpus qu'il constitue, se trouve alors à être l'objet d'investigation. Si, comme le suggère Rastier (2005), un corpus doit être "aimé", c'est non seulement parce que cet amour est garant de la compréhension de la situation dans laquelle s'insèrent les discours qu'il contient, mais également parce que les questions posées par l'interprète portent d'abord sur le corpus.

En se mettant au service de celles et ceux qui aiment le texte, la LACTAO fait un pari épistémologique qui paraît audacieux dans un contexte nord-américain : celui de privilégier l'objet observé et de ne pas se pencher sur les généralisations que l'on peut tirer de son observation. Lorsqu'on étudie un thème ou un concept dans le texte, les observations valent pour le contexte particulier du corpus étudié, et nous éclairent d'abord sur le corpus. S'il y a généralisation à faire par la suite concernant une communauté, une culture, ou l'espèce humaine, elle demande des étapes supplémentaires (comme l'observation d'autres corpus) et exige des garanties additionnelles conséquentes. Mais la LACTAO, qui vise une compréhension de phénomènes propres au corpus étudié, n'en fait pas son obligation.

Philosophie expérimentale

Alors que la LACTAO part d'un amour du corpus, la philosophie expérimentale se propose au contraire de confronter les hypothèses des philosophes, dont la portée se veut souvent universelle, au test de la vérification empirique. Par exemple, une stratégie d'argumentation courante en philosophie consiste à faire émerger des intuitions en employant des scénarios fictifs – des expériences de pensée. Celles-ci

peuvent ensuite être traitées comme évidences en faveur ou en défaveur de théories philosophiques (Bealer, 1998) – c'est là l'essence de ce qu'on appelle la "méthode des cas" (*method of cases*). La philosophie expérimentale porte un regard sceptique sur ce genre de méthode, employant l'expérimentation auprès de participant·es humain·es comme un moyen de tester ses présupposés. Elle s'inspire notamment dans cette démarche des méthodes développées en psychologie cognitive et en psychologie sociale. Ainsi, un de ses projets phares a été de tenter de vérifier que les scénarios proposés faisaient bien émerger les intuitions que leurs auteur·es supposaient (e.g. Weinberg *et al.*, 2001).

La philosophie expérimentale considère donc les données qu'elle produit par expérimentation comme des données empiriques susceptibles de jouer le rôle d'évidence. Si certaines branches de la philosophie, en particulier la philosophie de l'esprit, ont intégré dans leur pratique le traitement des données empiriques venant d'autres sciences, cette intégration n'avait pas donné lieu à des discussions métaphilosophiques très approfondies. L'introduction de ces méthodes en philosophie devient alors l'occasion d'une réflexion sur le rapport entre cette discipline et les sources de données empiriques qui pourraient l'alimenter. Cependant, ce rapport est encore presque exclusivement pensé dans des termes falsificationnistes : comme le note Pohlhaus (2015), il n'est pas question ici de donner la voix à la personne ordinaire pour qu'elle nous dise comment elle emploie les concepts philosophiques, mais bien de produire de l'évidence qui peut être confrontée aux théorie philosophiques.

Par ailleurs, la philosophie expérimentale hérite des traditions dont elle questionne (mais aussi de la psychologie cogntive, dont elle reprend les méthodes) l'intérêt pour des énoncés universaux. Ainsi, il n'y a pas d'intérêt pour des portraits d'une forme de conceptualisation ou de façon de penser qui soit située dans une communauté ou dans un espace-temps. Tout au plus retrouve-t-on des travaux qui

reproduisent une expérience sur plusieurs groupes dont le positionnement ethnoculturel est très vaguement défini : on parlera par exemple d'étudiant·es universitaires à Hong Kong (Machery *et al.*, 2004) sans investiguer leur ancrage culturel, ou de participant·es recruté·es selon leur état-nation de résidence, sans égard au fait que ces états sont souvent pluri-ethniques (Machery *et al.*, 2017). De plus, même quand la variable culturelle rentre en ligne de compte, l'hypothèse à démontrer ou à corroborer porte généralement sur l'universalité ou l'invariance d'un concept ou d'une manière de penser à travers les cultures.

Bien qu'ils aient en commun de mettre la philosophie en rapport avec l'empirie, la LACTAO et la philosophie expérimentale peuvent sembler être opposées dans leurs approches. Il n'y a cependant pas lieu de choisir entre ces deux traditions : cette opposition peut être comprise comme une complémentarité.

Certes, on peut, suivant Pohlhaus (2015), déplorer que l'universalisme de la philosophie expérimentale contribue à reconstruire un sujet universel, dont le récit tend à occulter la diversité des récits particuliers, en particulier des personnes qui sont en position de minorité ou qui sont victimes d'oppression. En ce sens, la LACTAO, avec son emphase sur une connaissance intime du corpus, peut amener une alternative salutaire. On peut également se lamenter du manque d'attention au processus de découverte en philosophie expérimentale, et au péril de la pauvreté des hypothèses qui en résulte. En particulier, on peut douter des vertus de décrire les modes de pensée de non-occidentaux exclusivement en termes de théories occidentales. L'approche de la LACTAO, qui part des données pour produire des hypothèses, peut être vue encore une fois comme une meilleure alternative.

Ceci dit, il y a peut-être lieu de développer une position mitoyenne entre le développement d'une pensée presque exclusivement axée sur des énoncés universels et une autre axée presque exclusivement sur des études de cas. Il y a peut-être

lieu de parler de ce que les cultures ont en commun de par le partage des mêmes contingences (e.g. Hannon, 2015), ou de faire un portrait général des variations.

Cette thèse se situe donc au croisement de ces deux traditions, se proposant de répondre à un problème commun.

Du point de vue de la LACTAO, les dernières années ont montré qu'il y avait un besoin pour une meilleure caractérisation du concept. D'une part, comme l'ont illustré Sainte-Marie *et al.* (2011), l'expression des concepts dans un texte va bien au-delà de ce qui est apparent si on se limite aux endroits où un concept est nommément mentionné. Le concept ÉVOLUTION n'apparaît qu'un petit nombre de fois dans l'*Origine des espèces* de Darwin (le nombre exact varie selon l'édition), mais il n'en est pas moins central pour comprendre l'argumentation et la structuration du texte. D'autre part, la polysémie pose aussi une difficulté : le mot "évolution", chez Darwin, est souvent employé non pas pour parler des changements qui s'opèrent de génération en génération d'individus dans une espèce, mais de l'évolution des embryons vers une forme adulte. Bref, si l'on tente d'étudier le concept d'ÉVOLUTION chez Darwin, on ne peut s'en tenir au mot qui semble le mieux correspondre au concept : non seulement cette heuristique nous cacherait-elle la réelle importance de l'évolution chez Darwin, mais elle menacerait d'incorporer au tableau des dimensions sémantiques ou de discours qui appartiennent à un autre concept. Comme beaucoup d'études dans la LACTAO portent sur un concept (e.g. McKinnon, 1973 ; Chartier *et al.*, 2008 ; Estève, 2008 ; Meunier *et al.*, 2005 ; Sainte-Marie *et al.*, 2011), il s'agit là d'un problème important.

Du point de vue de la philosophie expérimentale, on pourrait croire que le problème est beaucoup moins important, étant donné que l'expérimentateur·trice a un grand contrôle sur les conditions de l'expérimentation, et qu'ille peut utiliser ce contrôle pour réduire l'ambiguïté dans les communications. Cependant, comme le

note Machery (2017), les données produites par la philosophie expérimentale sont bruitées, et ce même lorsqu'elles produisent des résultats robustes. Prenons l'effet Knobe (Knobe, 2003), selon lequel les gens ont plus tendance à tenir pour moralement responsables celles et ceux qui ont sciemment causé un effet secondaire si celui-ci est néfaste. Initialement, Knobe voit dans la plus grande propension des participant·es à attribuer la responsibilité pour un effet néfaste une asymétrie qui se situe dans le concept lui-même : il suggère donc que l'on jette un coup d'œil à la façon dont les gens appliquent effectivement le concept de RESPONSABILITÉ, et appelle les philosophes à considérer l'impact négatif ou positif comme un critère potentiel dans son application. Cependant, bien que l'expérience ait fait l'objet de plusieurs réplications concluantes, de nombreuses hypothèses surgissent rapidement proposant que l'effet Knobe n'est pas causé par une asymétrie dans le concept, mais dans divers effets psychologiques, ou découlant de la perspective que les participant·es peuvent prendre en jugeant du cas qui leur est soumis (McCann, 2005 ; Nichols et Ulatowski, 2007 ; Wright et Bengson, 2009 ; Young *et al.*, 2006). Knobe lui-même croit maintenant qu'il s'agit d'un effet plus général du jugement moral qui aurait un impact sur un large éventail de comportements (Pettit et Knobe, 2009), mais le débat est loin d'être clos : chaque année apporte sa moisson de nouvelles études (Kneer, 2018 ; Michael et Szigeti, 2018) et de nouvelles hypothèses (Egré, s. d. ; Mizumoto, 2018) sur l'effet Knobe.

Malgré l'engouement autour de l'effet Knobe, il peut sembler que la philosophie expérimentale peine à nous renseigner sur le concept qui était initialement son enjeu. Étant donnés les moyens déployés dans le cadre de ce projet de recherche, on peut se demander, si la philosophie expérimentale dispose vraiment des outils pour rendre compte d'un concept. Dans le cas de l'effet Knobe et du concept de RESPONSABILITÉ lorsqu'appliqué à des effets secondaires, la philosophie expérimentale a adopté une approche incrémentaliste, où chaque nouvelle étude

vient, à l'aide de variations sur le protocole original, affiner le portrait que l'on a de l'effet Knobe. Cependant, s'il est vrai que l'on s'approche sans doute de la vérité à chaque étude, il n'en reste pas moins que le chemin à parcourir peut être très long. D'où ceci qu'après 15 ans et des dizaines d'études, cet effet – qui n'est qu'une partie de l'application du concept de RESPONSABILITÉ – semble encore nous échapper.

Comme le cas de l'effet Knobe en témoigne, le coût d'une approche incrémentaliste la rend inaccessible pour la plupart des projets. Des approches qualitatives, basées sur des entrevues par exemple, peuvent nous donner une meilleure idée de la cartographie générale d'un concept. Cependant, la façon la moins coûteuse et, possiblement, la plus exhaustive, est probablement d'exploiter des corpus de données textuelles très volumineux, comme le font déjà la linguistique de corpus et la linguistique computationnelle pour mieux comprendre les mécanismes du langage. Bref : la solution à la difficulté qu'a la philosophie expérimentale à faire un portrait adéquat des concepts qu'elle étudie pourrait se trouver dans la LACTAO, si on avait la solution au même problème pour la LACTAO.

La détection de présence du concept

En somme, entre la LACTAO et la philosophie expérimentale, il y a donc un problème général commun : comment faire le portrait ou l'analyse d'un concept dans un corpus de données textuelles ? Cette question demande d'abord qu'on ait répondu à un problème plus particulier : comment détecte-t-on si un concept est présent dans un texte ? Enfin, si l'on veut préciser davantage, on doit noter que dans le cas de la LACTAO, on cherche à automatiser l'assistance à la lecture, et que du côté de la philosophie expérimentale, on a besoin d'un portrait du concept étudié relativement exhaustif, et qui doit donc être produit à partir d'un corpus

trop volumineux pour être lu par des humains. Dès lors, notre question est la suivante : comment, par des moyens automatiques, détecte-t-on si un concept est présent dans un texte ?

Cette dernière question sera la question centrale de cette thèse. Et l'hypothèse qui sera défendue est que l'on peut effectivement détecter automatiquement la présence de concepts en retrouvant la présence des topiques (au sens de Blei *et al.*, 2003) où le concept recherché joue un rôle constitutif.

Pour désambiguïser autant la question que la réponse que je propose, il convient de clarifier certains aspects.

Premièrement, il convient de noter que l'on conçoit ici la détection de concept comme une *tâche* au sens de la linguistique computationnelle – c'est-à-dire qu'elle ne relève pas de l'application ou de l'opérationnalisation, mais qu'elle constitue plutôt un problème général à résoudre pour le développement de la discipline et pour faciliter la résolution de problèmes plus concrets[1]. En l'occurrence, son objectif est de faciliter le développement de l'analyse conceptuelle de texte basée sur des données textuelles en général, et non pas dans un cas particulier. Cette

[1] Le concept de tâche en informatique est rarement théorisé ou explicité, mais structure néanmoins le développement de plusieurs branches, dont l'apprentissage automatique et la linguistique computationnelle. Il semble répondre à au moins deux contraintes. D'une part, étant donné le foisonnement de méthodes et d'approches dans ces disciplines, il n'est pas toujours praticable de tester toutes les méthodes possibles pour résoudre un problème donné, d'autant plus qu'elles demandent souvent de maîtriser des concepts et des techniques mathématiques très avancé·es qui peuvent demander un investissement de temps considérable. Les tâches permettent d'identifier des méthodes robustes, qui fonctionnent dans de très nombreux cas, et qui sont susceptibles d'être utiles sur un très vaste ensemble de contextes d'application. D'autre part, les paramètres pertinents d'un problème dans un contexte précis ne sont pas toujours connus ou saillants d'emblée. Par exemple, dans le cas de la détection de concept, il est difficile de juger à partir de quel degré de rappel on possède toute l'information pertinente pour donner un portrait complet du concept qui nous intéresse, à moins de savoir à l'avance ce que l'on cherche, de la même façon qu'on ne peut savoir si un résumé est fiable et contient toute l'information pertinente si l'on n'a pas préalablement lu la version complète du texte. En ce sens, la recherche d'une méthode robuste permet de mitiger les risques inhérents à des techniques qui ont recours à des heuristiques pour répondre à des problèmes complexes.

tâche consiste simplement à localiser dans un corpus un concept préalablement identifié (et exprimé sous forme d'une représentation concrète, comme une chaîne de caractère ou un vecteur). En conséquence, elle entretient une relation avec, à tout le moins, une certaine conception de l'analyse conceptuelle, mais pas avec une opérationnalisation particulière de cette analyse conceptuelle. Par exemple, une fois que l'on a identifié qu'un concept est présent dans un ensemble de segments, on peut traiter cette information de différentes façons : on peut tenter de représenter le concept présent à l'aide de cooccurrents, ou on peut employer ces derniers pour produire une liste de concepts et de mots similaires, qui peuvent jouer le rôle de synonymes ; on peut tenter de rendre compte du discours que contiennent ces segments, par exemple en les lisant, si c'est possible, ou en en lisant un échantillon ou un résumé généré automatiquement ; on peut également tenter d'analyser ces discours en extrayant les concepts importants et en tentant de comprendre comment ils sont liés, de façon à se faire une idée du contexte dans lequel évolue le concept qui nous intéresse ; on peut croiser cet ensemble avec des données géographiques, chronologiques et/ou concernant les réseaux sociaux des auteur·es dans le but d'étudier la généalogie d'un concept ; etc. En ce sens, l'importance de la détection de concept ne vient pas de son rôle dans une chaîne de traitement particulière, mais plutôt du fait que la détection de concept rend possible un ensemble de chaînes de traitement qui contribueraient à l'analyse conceptuelle.

Deuxièmement, il faut souligner que cette question, qui est la question de recherche de cette thèse, relève essentiellement de l'informatique. Le rôle du travail philosophique qui est présent dans les pages qui suivent est, pour les fins de cette thèse[2], d'une part de justifier la pertinence de nouvelle tâche (et l'importance

[2]Les chapitres qui suivent étant des articles, ils doivent simultanément répondre aux exigences découlant du rôle qu'ils ont dans l'argumentation de la thèse et aux exigences découlant

de la distinguer de tâches similaires en informatique, comme la désambiguïsation des sens des mots et le rappel d'information), et, d'autre part, d'informer la formulation précise de cette tâche[3] de façon à s'assurer qu'elle remplisse un besoin. Partir de la philosophie, dans ce cas-ci, ne procède pas d'une association privilégiée entre les méthodes d'analyse conceptuelle qui y fleurissent et l'analyse des données textuelles. En effet, on peut être certain que la détection de concept puisse profiter à d'autres méthodes similaires en sciences humaines, par exemple dans le domaine de l'analyse du discours, de la lexicographie, des sciences infirmières, des consultations publiques, etc. On peut voir en l'analyse conceptuelle philosophique une première application de la détection de concept, au sein de laquelle elle doit d'abord être adéquatement formulée avant d'essayer de voir comment elle peut être adaptée à d'autres méthodes et à d'autres cadres théoriques.

Troisièmement, comment doit-on interpréter "concept" ici ? Le concept de CONCEPT est notoirement polysémique (Harnad, 2009 ; Machery, 2009 ; Murphy, 2004). Cette polysémie vient en partie du fait que l'on veut faire accomplir au concept différents rôles : expliquer la façon que l'on a de catégoriser, organiser l'information, rendre possible les inférences, expliquer la façon dont on communique, etc. Aussi, pour choisir le bon concept de CONCEPT pour le genre de tâche que l'on veut accomplir – l'analyse conceptuelle philosophique basée sur un corpus textuel – il faut d'abord bien comprendre cette tâche.

Quatrièmement, comment peut-on déterminer si on a répondu à la question qui est posée ? Ici, par exemple, une preuve de concept suivie d'une interprétation des résultats, comme on en trouve souvent dans les travaux de la LACTAO (e.g. Char-

---

de leur question propre.

[3]Un réviseur a noté qu'on aurait très bien pu partir d'un autre contexte, par exemple en analyse du discours. Effectivement, le choix de l'analyse conceptuelle en philosophie est arbitraire–on aurait pu partir d'un autre cadre.

tier *et al.*, 2008 ; Danis, 2012 ; Sainte-Marie *et al.*, 2011), ne saurait suffire. Dans le cas qui nous intéresse, l'utilité des résultats obtenus est un mauvais indicateur : un sous-ensemble des segments où un concept est présent peut nous donner une représentation utile de celui-ci, mais ce n'est pas nécessairement une représentation complète ou équilibrée. Il faut donc ici un autre indicateur. Or, comme les humains savent faire cette tâche, on peut donc employer des annotations par des humains pour évaluer les performances d'une méthode automatique. Cependant, pour ce faire, il faut pouvoir instruire l'humain à accomplir cette tâche (et pas une autre tâche similaire).

À la question de la thèse – comment détecter automatiquement un concept dans le texte – s'ajoutent trois autres questions subsidiaires : (1) Qu'est-ce que l'analyse conceptuelle philosophique basée sur un corpus ? (2) Quel concept de CONCEPT rend compte de l'objet de l'analyse conceptuelle philosophique basée sur un corpus ? et (3) Comment peut-on instruire un humain à détecter la présence d'un concept dans des données textuelles ?[4]

Ces trois questions diffèrent de la question de la thèse en termes du type de réponse qu'elles demandent. Dans tous les cas, ces questions sont l'énonciation de problèmes qui en appellent à des solutions : on peut solutionner la question de la thèse avec un algorithme qui détecte les concept, alors que les questions subsidiaires demandent une théorie de l'analyse conceptuelle basée sur des données

[4]Il convient de noter ici que ces questions subsidiaires ne correspondent pas aux objectifs premiers de la thèse, mais interviennent simplement comme des questions auxquelles il faut répondre pour formuler la tâche de la détection de concept. Autrement dit, la problématique fondamentale de la thèse correspond à la question de la thèse mentionnée au début de cette section : "Comment, par des moyens automatiques, détecte-t-on si un concept est présent dans un texte ?" Les objectifs de la thèse sont donc principalement de formuler la tâche de la détection de concept, de proposer des méthodes qui l'accomplissent et d'évaluer ces méthodes. Les questions subsidaires dont il est question ici ne doivent être abordées que dans la mesure où l'on veut formuler la détection de concept d'une façon qui réponde aux besoins et aux particularités de l'analyse conceptuelle basée sur des données textuelles.

textuelles, une théorie du concept appropriée, et un protocole d'annotation. Cependant, seule la première question peut être évaluée expérimentalement, parce qu'elle seule tente de répliquer une fonction qui est déjà remplie par des être humains. En contraste, le défi auquel nous invitent les questions subsidiaire est un défi d'innovation conceptuelle, puisqu'il n'existe pas, dans le paradigme de la philosophie analytique, de théorie d'analyse conceptuelle adaptée pour tirer profit des données textuelles. Le travail demandé par ces questions est donc davantage un travail théorique : il s'agit d'adapter et de modifier des théories existantes pour leur faire faire le travail qu'on veut leur faire faire. Les garanties qui le soutiennent se trouvent dans l'argumentation, et donc il ne saurait être évalué que par le regard minutieux d'expert·es, et indirectement par le test de leur application à travers des questions plus concrètes comme celle de la détection de la présence d'un concept dans le texte.

Ceci suggère une séparation du travail de la thèse selon les deux disciplines qu'elle mobilise et auxquelles elle se réclame. Le rôle de la partie philosophique de la thèse est de clarifier l'interprétation que l'on peut donner (et l'évaluation que l'on peut faire) à la question de la thèse, et elle accomplit ce rôle en donnant des réponses aux questions subsidiaires. En cela, la philosophie joue le rôle de clarification conceptuelle qu' on lui attribue couramment (e.g. Wittgenstein *et al.*, 2001, paragr. 109), mais également un rôle d'ingénierie conceptuelle (Brun, 2016 ; Cappelen, 2018 ; Haslanger, 2012). Le rôle de la partie informatique de la thèse est de produire des solutions au problème de la détection de la présence du concept dans les données textuelles, et ainsi de répondre à la question de la thèse telle qu'interprétée à la lumière de la partie philosophique.

Aussi cette thèse se divisera-t-elle en deux parties : l'une philosophique et l'autre informatique.

Approches théoriques

La partie philosophique est constituée du premier article, qui est conçu pour jouer son rôle de clarification conceptuelle dans le cadre de cette thèse tout en considérant une question philosophique qui puisse lui permettre de s'inscrire dans les débats philosophiques contemporains. Cet article s'inscrit dans le cadre de la philosophie de tradition analytique, et ce à deux niveaux. Au niveau de l'article, la question qu'il adresse, soit la question de l'interprétation de la similarité pour la comparaison de concepts en regard de leur identité ou de leur différence, est conçue pour s'inscrire dans certains débats de la philosophie analytique. Au niveau de la thèse, pour pouvoir répondre à une question comme "qu'est-ce que l'analyse conceptuelle ?"[5] il faut pouvoir identifier une communauté de pratique. Dans ce cas, il s'agit de la communauté des philosophes de tradition analytique. Ce choix est un peu atypique d'un projet issu de la LACTAO – ceux-ci trouvent souvent des ancrages dans les traditions européennes de philosophie et de sciences sociales (e.g. Forest, 2002 ; Danis, 2012). Il s'explique par la volonté de pouvoir dialoguer avec la philosophie expérimentale et d'aborder le genre de questions qu'elle aborde.

Par ailleurs, comme pour beaucoup de travaux en philosophie, il est difficile de situer l'approche poursuivie dans un cadre théorique précis : l'argumentation de la partie philosophique emprunte des éléments de plusieurs penseur·ses. Cependant, la conception de l'analyse conceptuelle adoptée tire davantage de l'explication carnapienne que d'autres influences, et le concept de CONCEPT est résolument ancré dans le paradigme téléosémantique de Millikan (1984). Ces choix sont expliqués dans le premier article : dans le cas de l'approche carnapienne, il s'explique

[5]Conçue comme pratique. En philosophie, l'analyse conceptuelle est souvent associée à une forme de représentation (cf. King, 1998) ; ici, il est question de méthodes pour produire une représentation d'un concept.

par ceci qu'elle est encore la plus développée (en particulier chez Brun, 2016), les autres approches se concentrant seulement sur certains aspect de l'analyse conceptuelle. Le choix de la téléosémantique millikanienne s'explique quant à lui par le fait que, d'une part, il s'agit, à ma connaissance, à la fois d'une articulation très complète d'un concept évolutionniste de FONCTION et de l'articulation la plus complète d'une philosophie du langage basée sur une conception évolutionniste de la fonction. Ce faisant, elle permet à la fois de répondre à la question de la similarité qui est posée dans le premier article, et de proposer des applications concrètes pour des tâches comme la détection de concept dans le texte.

La partie informatique, quant à elle, se situe au croisement de la linguistique computationnelle et de l'apprentissage automatique. Au niveau de la théorie linguistique, elle est basée sur la sémantique distributionnelle (Lenci, 2008 ; Sahlgren, 2008). Ce choix s'impose pour un certain nombre de raisons : ses nombreux succès en traitement automatique des langues naturelles dans les dernières décennies, particulièrement pour les approches non-supervisées ; sa contribution essentielle aux fondements de la LACTAO ; et le fait qu'elle permet un raccord théorique avec l'approche téléosémantique.

Au niveau de la modélisation, les solutions informatiques proposées relèvent de la modélisation probabiliste graphique (Koller et Friedman, 2009) et, plus particulièrement, elles se situent dans la lignée de travaux sur les modèles topiques inspirés de l'allocation Dirichlet latente (LDA) de Blei *et al.* (2003). L'approche de la modélisation probabiliste graphique permet de représenter des contraintes bayésiennes, ce qui lui permet, dans le cas du texte, de représenter des processus génératifs qui permettent de révéler des variables cachées, comme les topiques ou les concepts. Même si on peut arguer avec Bengio *et al.* (2013) que des approches de Deep Learning pourront peut-être ultimement mieux rendre compte d'autres suppositions concernant la structure des données, l'approche des modèles proba-

bilistes graphiques bénéficie encore d'un avantage en termes de flexibilité pour la représentation d'hypothèses sur la génération du texte. La revue de littérature sur les modèles topiques qui mobilisent les enrobages de mots (*word embeddings*) en témoigne d'ailleurs de façon éloquente : sur 21 études recensées, 17 emploient des modèles probabilistes graphiques inspirés de la LDA, contre seulement deux qui emploient des modèles connexionnistes.

Plan

Pour répondre à la première question subsidiaire, qui porte sur la nature de l'analyse conceptuelle, le premier article, intitulé "Similarity in conceptual analysis and concept as proper function", fait une revue des débats actuels sur la question. Dans celui-ci, j'identifie les trois principaux paradigmes au sein desquels la question est discutée en philosophie analytique : l'explication carnapienne (Carnap, 1950), l'analyse haslangerienne (Haslanger, 2012) et la méthode des cas (notamment Bealer, 1998 ; Machery, 2017 ; Sosa, 2007), et je fais un survol de chacun. Constatant que ces paradigmes sont largement complémentaires en ceci qu'ils se penchent sur des aspects différents de l'analyse conceptuelle, une synthèse est produite pour caractériser la conception de l'analyse conceptuelle qui sera employée dans le reste de la thèse.

La question du concept de CONCEPT est ensuite abordée indirectement, à travers la question du comment évaluer la similarité entre deux concepts. Après avoir argué que de comparer les concepts par une représentation de leur intension ou de leur extension ne saurait convenir au contexte d'analyse conceptuelle qui nous intéresse, j'argue que la meilleure façon reste de comparer leurs fonctions propres, au sens de Millikan (1984)[6]. Je propose ensuite un concept de CONCEPT pour

[6]Dans cette conception de la fonction, qui vient de la biologie et qui se rapporte à

lequel la similarité est effectivement mesurable en termes de similarité de fonctions propres, en adaptant le concept millikanien de MOT.

Enfin, la dernière partie du premier article sert, du point de vue de l'article, à illustrer comment le concept de CONCEPT développé dans la partie précédente peut être employé dans l'analyse conceptuelle basée sur des données textuelles. Du point de vue de la thèse, elle illustre comment le concept de CONCEPT et la conception de l'analyse conceptuelle développée dans les parties précédentes peuvent se traduire dans l'opérationnalisation de la détection de la présence de concept dans le texte, que cette détection soit faite par un humain ou par un algorithme informatique.

Les deux articles suivants proposent des solutions informatiques au problème de la détection du concept. Pour ce faire, ils proposent des chaînes de traitement conçues à partir de modèles topiques appris sur un large corpus spécialisé (décisions de la Cour d'appel) et partiellement annoté.

Le deuxième article, intitulé "Detecting Large Concept Extensions for Conceptual Analysis", modélise les topiques à l'aide de la LDA, laquelle est apprise avec deux différents algorithmes d'apprentissage (Griffiths et Steyvers, 2004 ; Hoffman *et al.*, 2010). Les liens entre topiques et les concepts qui les constituent sont modélisés par les liens entre topiques et mots tels que décrits par la probabilité apprise qu'un mot soit écrit si un concept est activé. La méthode est un succès, au sens où certaines heuristiques démontrent une grande amélioration sur l'heuristique de base, et l'expérimentation démontre l'avantage de modéliser un topique comme constitué de plusieurs concepts.

Cependant, représenter un concept comme un mot est problématique, puisqu'au-

---

l'évolution, la fonction correspond davantage au rôle que, par exemple, la fonction mathématique ou informatique.

tant la théorie développée dans le premier article que les succès du deuxième article suggèrent que le concept s'exprime très souvent sans le ou les mots ou expressions auquel il est le plus associé. Aussi le dernier article, intitulé "Mixing syntagmatic and paradigmatic information for concept detection" tente-t-il de donner une meilleure modélisation en incluant les concepts dans la modélisation et en permettant de formuler le concept dont on cherche l'extension sous une forme autre que lexicale. Pour ce faire, il emploie le modèle LCTM (Hu et Tsujii, 2016), qui modélise les concepts dans l'espace paradigmatique formé par un modèle d'enrobages de mots, afin de construire différentes chaînes de traitement. Les résultats démontrent que plusieurs de ces chaînes de traitement offrent de meilleures performances que toutes les heuristiques testées dans le deuxième article.

CHAPTER I

SIMILARITY IN CONCEPTUAL ANALYSIS AND CONCEPT AS PROPER FUNCTION[1]

Mise en contexte

Il vise à présenter les base théoriques à partir desquelles se comprend la détection de la présence du concept dans le texte. Premièrement, il puise dans la tradition de la philosophie analytique afin de situer et de décrire l'analyse conceptuelle basée sur des données textuelles. Deuxièmement, il tire de cette contextualisation le critère de similarité, et montre qu'il doit être interprété de façon fonctionaliste (au sens de Millikan, 1984). Il propose par le fait même une caractérisation du concept compatible avec cette interprétation. Troisièmement, il fait les liens entre les notions développées et une éventuelle opérationnalisation en illustrant comment ces notions peuvent informer des applications pratiques – en particulier la détection de concept, que ce soit par un humain ou par un algorithme.

En ce sens, cet article pose les bases théoriques d'une analyse conceptuelle philosophique reposant sur des données textuelles et illustre comment ces bases peuvent être mobilisées dans la détection de concepts dans le texte.

---

[1] A previous version of this article, which I authored alone, was submitted to the journal *Ergo*.

Abstract

In the last decades, experimental philosophers have introduced the notion that conceptual analysis could use empirical evidence to back some of its claims. This opens up the possibility for the development of a corpus-based conceptual analysis. However, progress in this direction is contingent on the development of a proper account of concepts and corpus-based conceptual analysis itself that can be leveraged on textual data. In this essay, I address this problem through the question of similarity: how do we evaluate similarity between two concepts, as similarity relates to identity? After a survey of prominent conceptual analysis methods, I propose a cursory account of corpus-based conceptual analysis. Then I formulate the question of similarity, and argue for an account that is functionalist in Millikan's (1984) sense. In this process, I propose a new account of concept that bases itself on millikanian teleosemantics in order to account for concepts' contribution in discourse. I then illustrate its fruitfulness by showing how it enables accounts of concept presence detection in textual data, both automatically and by a human judge.

Say a philosopher, named Alice, wants to study a given concept—in particular, she wants a picture of how it is being used. She gathers a very large corpus, large enough that for most concepts, she will have enough instances in the text so that she can observe the full variety in kinds of sentences, narratives, arguments and contexts in which it is used. In other words, her corpus is large enough to assume that it is representative of the kinds of discourses that run within the context where it was collected. As a very competent reader, Alice can intuitively pick up concepts when they are used.

However, she can hardly translate this "picking up concept" into a set of procedures[2]. Indeed, this "picking up concepts" should not be assimilated with, say,

[2] A reviewer brought to my attention an argument that can be brought up against any empirically based conceptual analysis that resorts to studying folks' understanding of a concept. Say we want to study ordinary people's knowledge of a concept, then we have to grant that the participants of our study understand the concept in question. But then, if it is folk knowledge, admittedly it is shared by all the community, and unless the researcher is from a different

picking up words that stand for it: concepts are often implicit, and they may exert their influence on a text's content and structure without there being a word or set of words that make reference to it. When Alice picks up a concept's presence in the text, she is not merely recognizing material shapes, but recognizing the concept's role in the discourse's structure, at least as she understands it. This kind of operation is opaque even from Alice's point of view: while we can learn to better read and understand from others, we cannot tell exactly which operations take us from ink shapes to a certain concept.

In this scenario, if Alice has read the corpus, she probably has strong grounds for trusting herself with the various judgments that she makes as part of her interpretative activity. So long as she trusts her competence as a reader, she could go through the texts, identify the sentences which mobilize the concept she is interested in, and make an inventory of what the corpus tells us about it. The problem is that she probably lacks the time and resources to read the whole corpus by herself. As she needs to outsource parts of her reading and interpretative process in order to be able to treat massive amounts of data, she might not be able to trust the results of this outsourcing, even if she knows which operations are being performed. Given the opacity of her competence as a reader, even a

---

culture, then she should also understand the concept as well as her subject. Why, then, doesn't she simply reflect on her concept, and produce an account of it from her armchair?

We can appease this worry by noting that the researcher might be more interested in the concept than simply her personal account of it, which might be skewed by her social positioning and personal history. However, there is more to it, as this argument fails if we distinguish, as we have done here, between the capacity to use a concept in everyday uses, such as discourse structuring or comprehension or for producing statements about the world, and the capacity to represent that concept as an object for discussion (or, in other words, to make it explicit). We might call the first kind of capacity "operative knowledge" of the concept being inquired, and the second kind might be called the "theoretical knowledge" of the concept. At the beginning of the study, we might grant the researcher an operative knowledge of the concept she wishes to study, but what she is looking for is a theoretical knowledge of it—she is certainly not looking for information she already has.

Cf. also footnote 11 on page 36.

simple operation such as detecting a concept in the text becomes a challenge.

An algorithm that detects the presence of concepts in textual data, such as the ones developed by researchers of the LANCI in the last decade (e.g. Chartrand *et al.*, 2016, 2017; Pulizzotto *et al.*, 2016)[3], might go a long way towards helping Alice. Indeed, given the importance of concepts in philosophical practice, we have speculated that the lack of computational tools to detect concepts in textual data is one of the reasons why philosophy is lagging behind other disciplines of social science and of the humanities with regards to the penetration of natural language processing and text mining in the research practice (Chartrand *et al.*, 2016). While there is some opacity in the way these algorithms make their interpretative decisions, computer scientists will usually lean on our faith in human judgment in order to validate their algorithms: they engineer and evaluate them by comparing them to what humans would do when they perform the same operation. For example, Chartrand *et al.* (2017) had participants annotate text segments for the presence of a concept, and evaluated their method against this metric.

One might argue that this strategy of relying on the trust we have on human competence merely displaces the problem. Indeed, in practice, even getting annotators to make the right calls require that the annotation protocol be well thought through—short of which they might be fulfilling a different task. This, in turn, requires that we have a good understanding of what it is to detect concept presence in text.

In post-war analytic philosophy, the association of conceptual analysis with *a priori* (non-empirical) knowledge (e.g. Jackson, 1998) has meant that questions pertaining to topics such as the observability of concepts in empirical data have remained underdeveloped. While recent discussions around experimental philoso-

[3]See also Chartrand (n.d.).

phy and its methods have led to some progress on this topic, the focus on adapting methods from cognitive and social psychology on one side and on the role of intuitions on the other has meant that little has been developed to characterize the role of concepts in natural language. On the other hand, concepts have been discussed as instantiated in language, such as in the notion of "lexical concept", but it is often in a very limited role, where the concept is viewed as attached to a particular expression or lexical pattern, which typically brings up the concept in question by referring to it (e.g. Fodor, 1998; Evans, 2006). But concepts in discourse are often implicit; they may have an important role in structuring narratives or discourses without being attached to specific expressions. Conceptual analysis would be incomplete if it failed to account for the roles a concept plays when it is not directly expressed through reference.

Therefore, while it is probably true that the lack of algorithmic tools is an obstacle to the development of corpus-based conceptual analysis, it also seems that it is in need of a proper account of concepts (as it plays a role both in formulating a question in conceptual analysis and in concept detection) and corpus-based conceptual analysis itself[4].

In this essay, I address this problem through the question of similarity: how do we evaluate similarity between two concepts, as similarity relates to identity? Concepts are public entities, and they achieve their roles by being repeated from an instance to another. However, individual humans likely don't internalize concepts exactly the same way, which is to say that we likely have slightly different accounts

---

[4] Not to be confused with, for example, Meunier *et al.*'s (2005) *Computer-Assisted Conceptual Analysis of Text.* Meunier *et al.*'s aim is to unearth associations of a concept (as it is explicitly employed in text) in order to contribute new knowledge to an interpretation. Corpus-based conceptual analysis, in contrast, shares a similar objective with experimental philosophy as it is employed for the sake of conceptual analysis: namely, the idea is to give an account of a concept as it is employed in relevant linguistic behaviour.

of the same concepts. Furthermore, as we keep learning and updating these accounts, it is likely that those also change across time—I probably don't have the same account of the concept of CAT (the animal) as when I was five years old. Still, we say of my current concept of CAT that it is the same (in the relevant sense) as the concept of CAT that I had as a five-years-old. How do we judge this identity?

As we shall see, this question poses itself slightly differently in the context of conceptual analysis and in the context of concept detection. However, I will argue that there can be a single answer to these two varieties of the similarity problem.

This essay is divided in two broad sections. In section 1.1, I compare various ways of understanding conceptual analysis: the method of cases (and Machery's (2017) understanding of it in particular), Haslanger's (2012) three types of conceptual analysis, and Carnap's (1950) explication. I propose that these accounts are mostly complementary, and offer a synthesis. In section 1.2, I formulate and address the problem of similarity. I assess three ways of understanding Carnap's similarity criterion: intension, extension and function; and I argue that similarity by function is superior to its alternatives. To replace Carnap's vague account of function, I offer a millikanian account of it, and I show how it translates into an account of the concept of CONCEPT and into a heuristics to measure similarity between concepts. Finally, in section 1.3, I illustrate how the millikanian framework, and in particular similarity as function, plays out in corpus-based analysis and in concept presence detection in particular.

1.1 Varieties of conceptual analysis

When talking about conceptual analysis in philosophy, two different ideas come to mind.

Firstly, in the mind of most analytic philosophers, the term "conceptual analysis" conjures a specific type of proposition, with the concept that is being analyzed (the *analysandum*) on one side, its deconstruction into other concepts on the other (the *analysans*), and an operator that asserts some form of identity between the two terms. Usually, this proposition expresses the *analysans* in the form of necessary and sufficient conditions: for instance, "a brother is a male sibling" expresses that the concept BROTHER can be analyzed into the concepts MALE and SIBLING, with both being necessary conditions for BROTHER, and being jointly sufficient. In this sense, a conceptual analysis is a form of representation. It does not tell much about how we can arrive to propositions of this type, but it does tell us about the constraints coming from the form and the properties and paradoxes that come from it (Cf. King, 1998; Jackson, 2013).

The second sense that is associated with the term "conceptual analysis", on the other hand, speaks of method rather than form. To some philosophers (Chalmers and Jackson, 2001; Jackson, 1998; Lewis, 1970), it evokes a method to produce a proposition that would be a conceptual analysis in the first sense. Traditionally, conceptual analysis has been mostly about unravelling "our" concept of something, which a philosopher can often study through her own account of this concept, in a *a priori* manner—viz. without inquiring outside of the realm of her own mind. But it need not be that way, and indeed, many (e.g. Haslanger, 2012; Machery, 2017) use this term to refer to explicitly empirical methods.

This section's aim is to make a short review of current accounts of empirically informed conceptual analysis. The motivation for this boils down to this: conceptual analysis is the context within which we shall understand both concept similarity and concepts themselves. In other words, our accounts of concepts and concept similarity will be those that serve the account of conceptual analysis that we shall adopt. Therefore, this section can be thought of as a clarification of the

main concerns of this article, those of concept similarity (how is it measured?) and concept detection (how can it be theorized for operationalization?). In a first subsection, I propose a historical perspective on the roots of conceptual analysis in the naturalist/rationalist debate, with an eye for the *a priori*/*a posteriori* debate, which has been polarizing the way we understand conceptual analysis and engineering in philosophy, especially in the second half of last century. Then I go on to describe the main frameworks through which the relation between conceptual analysis and empirical data have been theorized in the last few years – the method of cases, haslangerian analysis, and carnapian explication[5]. Finally, I show how those different accounts fit together in the context of corpus-based conceptual analysis.

1.1.1 Historical roots of conceptual analysis

While analysis has been a prominent part of the philosopher's toolset for millennia, we often trace back contemporary analysis to Kant and his analytic/synthetic distinction. Kant is interested in statements as subject-predicate pairs, and calls "analytic" those in which the predicate is contained in the subject and "synthetic" for which it is not the case. For instance, the idea of having three sides is present within the concept of TRIANGLE, thus "All triangles have three sides" is an analytic statement. This dichotomy is closely associated to another, which deals with the means of acquiring truth values for a statement: if we need experience

[5]The reader could probably point out to other frameworks that could fit the bill. In particular, one might argue that debates around natural kinds, for instance, should be addressed. However, these accounts address a very limited subset of concepts: those whose main goals are to refer to natural phenomena in ways that enable descriptions of the world that are as accurate as possible. Not all language is scientific language, and for good reasons; most concepts are adapted to other activities and fulfill other objectives which are not less commendable (Cf. Haslanger, 2012; Carus, 2008). Furthermore, I have not addressed other historically significant accounts of conceptual analysis, as I felt I should prioritize on current accounts of conceptual analysis. Machery (2017) has addressed the same three frameworks, admittedly for similar reasons.

of the world to determine such truth values, then a statement is *a posteriori*, if it can be determined without experiencing the world, then it is *a priori*. Kant thought that there were no such things as analytic *a posteriori* statements, and his successors mostly rejected the possibility of synthetic *a priori* statements, such that, for most purposes, these two dichotomies are usually addressed as a single one, with the analytic *a priori* being opposed to synthetic *a posteriori*.

For the better part of the 20th century, analysis has thus been presented as a polar opposite to empirical inquiry. This said, the place occupied by this dichotomy in philosophy of science goes beyond the mere separation of analytic from empirical truths, as a defining research question has been to deterine how analytic truths are to be integrated in the body of scientific knowledge (Rey, 2018). For instance, at least for the early Carnap of the *Logische Aufbau der Welt* (Carnap, 1928), in order to have content, the theoretical terms with which scientific theories and claims are formulated ought to be reducible to observation terms. Analysis, or "rational reconstruction", is thus the production of a form of definition, whereby a scientific term is related through rules to observation terms. Such definitions, however, have different conditions of adequacy than sentences about empirical terms: whereas the latter gets a truth value when confronted with observation, definitions are adequate if they reflect a convention (Rey, 2018). As such, one must distinguish between the language in which empirical statements are produced and the language of reconstruction, with the former reflecting observation and experimentation, and the latter reflecting convention. To a degree, one can thus see the project of the *Aufbau* as attempting to draw a line between analysis and empirical inquiry and between the corresponding languages and epistemologies, and attributing them their roles in the production of scientific knowledge.

With Quine's "Epistemology Naturalized" (Quine, 1971), the debate eventually becomes polarized between reformers of the project of the *Aufbau*—soon joined by

defenders of *a priori* methods of analysis—and defenders of naturalized epistemology—often called respectively "rationalists" and "naturalists." Rather than reconstructing the meaning of empirical concepts through analysis, Quine's suggestion is that we study how the construction of these terms actually proceeds. Knowledge is thus seen as a natural phenomenon, and the project of epistemology should be to study where and how it emerges. The same goes for empirical concepts, whose meanings are not to be determined by an elaborate definition leading us all the way to a primordial empirical language, but by a function of the processes of categorization they enable.

One of the central friction points is around the question of *a priori* statements. A naturalized epistemology would seek to replace *a priori* analysis of empirical terms with scientific accounts of those terms as they are reliably employed (Rysiew, 2017). On the surface, it might seem like it is just another, perhaps more scientific, way of determining what our concepts are. However, rationalists would argue that naturalists who think that they are turning their back on *a priori* intuitions are in fact presuming or assuming them (e.g. Bealer and Strawson, 1992). For example, as Bealer suggests, we need to use intuitions to determine what counts as empirical evidence rather than, say, *a priori* intuitions, imaginations or memory. Alternatively, if our starting-point intuition about mundane concepts were wildly unreliable, we might not be able to bootstrap them to acceptable concepts.

Prominent responses to this challenge often choose to concede Bealer's point, to the extent that they concede that empirical inquiries need a starting point. But those starting-point judgments need not be interpreted as *a priori*. To Kornblith and others (2002, p. 13), "the extent to which naive investigators agree in their classifications is not evidence that these judgments somehow bypass background empirical belief, but rather that background theory may be widely shared." Even judgments which seem to rely on information that we share from the moment we

are born are likely informed by lessons learned through our species' evolution. From Kornblith's perspective, *a priori* judgments, or at least the judgments that are referred to with this expression, exist and are relevant in epistemology, but they are best explained as natural abilities that draw from experience, including the experience of our ancestors.

It is unclear that this response really addresses the qualms of traditional epistemologists, as explaining intuitions as natural empirically-informed abilities relies ultimately on intuitions, and this explanation isn't available to the epistemic agent at its starting point. On the other hand, Kornblith suggests that we might not be more justified in trusting *a priori* intuitions whose legitimacy seems somewhat supernatural. Thus the debate over the *a priori* takes a sort of chicken-or-the-egg flavour: it seems to depend on which perspective—e.g. the natural or the phenomenal—one is starting from.

The distance between naturalists and rationalists should not be overstated. On the one hand, of course, Kornblith's arguments does not target the practice of using *a priori* intuitions, but rather suggests that the source of their legitimacy do not lie where rationalists think it is. On the other hand, rationalists are not necessarily opposed to the project of revising our account of knowledge in light of discoveries in cognitive science (BonJour, 2006), and neither do they take *a priori* intuitions to be unrevisable in light of empirical knowledge (Bealer and Strawson, 1992).

Furthermore, the middle way between a "pure" naturalism and a "pure" traditional aprioristic epistemology is actually well-travelled. For instance, Goldman (Goldman, 2005, 1986) has argued consistently that intuition-based conceptual investigation must be the starting point of epistemological inquiries (Rysiew, 2017). On his account, intuitions can be interpreted as a window to our internal concepts,

and methods to elicit them can be seen as ways to gather evidence for conceptual analysis. On the rationalist side of things, Canberra planners have gone so far as to reclaim the "naturalist" label, in part because of their general commitment to physicalism, and their lack of commitment to the primacy of the *a priori* over the *a posteriori* (Braddon-Mitchell, 2009). Moreover, the rationalist's armchair often looks suspiciously susceptible to empirical inquiry: for instance, the Canberra planners' "two-step" method for conceptual analysis begins by collecting all the platitudes about this concept[6] (Nolan, 2009).

It would also be a mistake to associate empirical inquiry with the naturalists to the exclusion of the rationalists on account of their positions with regards to the *a priori*/*a posteriori* dichotomy. After all, the initial positivist project, as it is developed in the *Aufbau*, far from developing a discipline disconnected from empirical inquiry, portrays philosophy as "the handmaiden of science" (Braddon-Mitchell, 2009). Furthermore, this separation did not necessarily imply that the analysis should stick to the armchair. It is explicitly in this spirit that Arne Naess pioneered experimental methods strikingly similar to modern experimental philosophy during his years attending the Vienna Circle, and while the project has not been well received by all of the Vienna Circle regulars, Carnap himself saw this as a positive development (Murphy, 2014; Naess, 1938).

More recent attempts at informing philosophers' accounts of concepts are also hard to split along the rationalist/naturalist lines, but a generalization can perhaps be made: while naturalists analyze concepts to ensure that they capture the right phenomena or objects, rationalists put more emphasis on capturing *our* concept of something. This is not unexpected, as the naturalist project is more about

---

[6]There is some controversy around what should count as a platitude. Generally speaking, these would be claims that reflect commonplace uses of the concept.

building concepts from observation and experimentation[7], whereas the rationalist project begins with an assessment of the concepts we have before diving into data. A typical naturalist project would be, for instance, to determine whether the physical extension of the concept MIND should be limited to the brain or diffuse into a creature's environment (cf. Clark and Chalmers, 1998; Clark, 2008; Hurley, 1998; Rupert, 2009), and would draw heavily on research in psychology, anthropology, neuroscience, etc. to argue for its case. Conversely, more typical of a rationalist project would be to probe laypeople's intuitions about a concept in order to determine how they understand it.

Therefore, experimental philosophy, which will provide part of the framework for contextualizing corpus-based concept analysis, is probably more rooted in the rationalist tradition, and might be thought as the rationalist response to the naturalists' use of cognitive science research for their own projects. This is evidenced by its focus on thought experiments and its methodological reliance on intuitions. However, it is worth noting that not all of this focus is an endorsement: in fact, while it is far from forming the bulk of the research in experimental philosophy (cf. Knobe, 2016), much of it is devoted to what has been dubbed *the negative program*, viz. a critique of the reliance on intuitions in philosophy. As a result, it is probably best to think of experimental philosophy having its roots in both traditions.

---

[7]While both are ways of capturing empirical data, experimental studies and observational studies differ in the degree of control being exerted by the researcher. In experiments, the phenomenon being studied is is provoked, typically in controlled conditions, such that causes and effects can be isolated. In observational studies, the researcher has no control over the phenomenon she is observing, she might make for more realistic environments, but makes it more difficult to ascertain causality and to control for unwanted interactions, among other things. While most of experimental philosophy has indeed been experimental, corpus analysis would rather qualify as observation.

1.1.2 The method of cases

The method of cases is, at its core, a sort of narrative that goes as follows. We have a concept which we suspect to have a certain attribute. For instance, we might imagine that in a categorization task—when judging whether a certain limit-case object is a representative of the said concept, or not—we think that having a certain feature is important in determining where it belongs. So we think up cases or scenarios where the said feature can be isolated, and test our judgment on it to see where it leads us. For example, Knobe (2003) suspects that whether a side-effect is positive or negative can have an impact on whether the person who brought it about is responsible for it or not. So he concocts this scenario:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
>
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
>
> They started the new program. Sure enough, the environment was harmed.

He tours around Central Park submitting either this case, or a similar one where every instance of the verb "harm" is replaced by "help", and asks whether the chairman is responsible. As people are twice as likely to say he is in the "harm" cases, Knobe concludes that the valence of the side-effect (its being good or bad) is important in the folk concept of RESPONSIBILITY[8].

---

[8]At least, it was the case in 2003. Since then, Knobe has adopted the view that this

Alternatively, one can do this kind of experiment esoterically, between the author and its readers. The Gettier cases (Gettier, 1963) are often understood like this: Gettier thinks that there is more to knowledge than its common analysis—according to which knowledge is justified and true belief. Therefore, he proposes this case (p.122):

> Suppose that Smith and Jones have applied for a certain job. And suppose that Smith has strong evidence for the following conjunctive proposition:
>
> (d) Jones is the man who will get the job, and Jones has ten coins in his pocket.
>
> Smith's evidence for (d) might be that the president of the company assured him that Jones would in the end be selected, and that he, Smith, had counted the coins in Jones's pocket ten minutes ago. Proposition (d) entails:
>
> (e) The man who will get the job has ten coins in his pocket.
>
> Let us suppose that Smith sees the entailment from (d) to (e), and accepts (e) on the grounds of (d), for which he has strong evidence. In this case, Smith is clearly justified in believing that (e) is true.
>
> But imagine, further, that unknown to Smith, he himself, not Jones, will get the job. And, also, unknown to Smith, he himself has ten coins in his pocket. Proposition (e) is then true, though proposition (d), from which Smith inferred (e), is false. In our example, then, all

---

effect is probably more of a widespread cognitive effect than a feature of the concept of RESPONSIBILITY (Pettit and Knobe, 2009).

> of the following are true: (i) (e) is true, (ii) Smith believes that (e) is true, and (iii) Smith is justified in believing that (e) is true. But it is equally clear that Smith does not know that (e) is true; for (e) is true in virtue of the number of coins in Smith's pocket, while Smith does not know how many coins are in Smith's pocket, and bases his belief in (e) on a count of the coins in Jones's pocket, whom he falsely believes to be the man who will get the job.

Gettier thus concludes that the "justified true belief" account of knowledge is inadequate.

As Sosa (2007) points out, it is not necessary to associate the method of cases to conceptual analysis. For instance, he suggests, cases can be used to argue for or against any philosophical theory, including those that are not about concepts.

For Sosa and others (among which Bealer, 1998; Chalmers, 2014; Goldman, 2007; Gopnik and Schwitzgebel, 1998; Ludwig, 2007), intuitions are what drives us to make the relevant judgments on the cases. Their value comes from our competence in making judgments—for instance, when making judgments about concepts, these judgments would derive from our competence in using those concepts. We might assume, in turn, that we would have acquired this competence from living in a society that uses those concepts, or from our experience in using this concept. As a result, the method of case can be seen as a way to highlight those intuitions and make them explicit.

However, the proponents of an intuition-based method of cases have struggled, over the years, to establish intuitions as sources of evidence or other epistemic guarantees for the method of cases. Formulations of the concept INTUITION (understood in the context of philosophical method) are numerous, although they rarely have clear boundaries, and all the most prominent formulations have been

the target of numerous critiques. Variety in accounts of a concept is not in itself a problem—we lack consensual accounts of many important concepts, such as COGNITION, LIFE and DEATH, and this does not count as a failing for the theories that rely on those concepts. But intuition-mongerers have had to defend against charges that intuitions are too volatile to fulfill their epistemic role (Alexander and Weinberg, 2007; Machery *et al.*, 2004; Swain *et al.*, 2008) on one side and arguments that they are not central in the practice or logical structure of philosophical argumentation (Cappelen, 2012; Deutsch, 2015; Williamson, 2008; cf. also Nado, 2016; Pohlhaus, 2015) on the other[9].

Following Machery (2017), defenders of intuitions can be broadly divided into two broad camps[10]. On the one hand, there are those who would have intuitions be a special kind of mental state or competence, whom Machery calls "particularists" and "exceptionalists": for them, intuitions are not just any opinion or felt state, they are distinct in virtue of things like type of content (e.g. abstract or modal), psychological or phenomenological properties (automaticity, speed, "being drawn to", etc.), etiology (e.g. came from experience, accepted competence), epistemic status (e.g. reliable opinion), etc. The reason why we would want intuitions to be particular or exceptional lies primarily in the mechanics of the subject/object dichotomy as it functions in the dispatch of epistemic work: we expect one side of the dichotomy to do one part of the epistemic work (provide evidence) and the other to do the other part (evaluate the information, synthesize it, draw conclusions, etc.). Mixing those responsibilities can yield paradoxes and fallacies (Cf. Williamson, 2008, 2013; Ichikawa, 2009). For instance, say I grant evidential sta-

---

[9]There has been others charges against intuitions. For example, Machery (2017) argues that they are a bit of a nomological dangler and Pohlhaus (2015) argues (among other things) that, in the way they are formulated and employed, intuitions rely on an assumption of universality which is epistemically noxious.

[10]Machery talks about three camps, but one delineation is more important that the other.

tus to intuitions, but I think of intuitions simply as being no different to other judgments. Then using my intuition of P as evidence for my judgment that P would translate into justifying my judgment that P as evidence for my judgment that P. Even if we manage to work around this paradox, there is a legitimate concern that we could be tainting our intuitions with our opinions and motivations[11]. Shielding intuitions from the rest of the mental lore by affirming their distinctness serves to avoid this kind of difficulties.

On the other hand, there are the minimalists (among whom Machery, 2017; Ichikawa, 2009; Williamson, 2008) who hold that the method of cases does not require that there be a special epistemic status for intuitions. For the minimalist, intuitions have no special phenomenology, they have no special epistemic or semantic status, they are neither necessarily analytic or justified *a priori*, and they have no special etiology. They certainly don't come from a different faculty, they have no special psychological status (they can be fast and automatic, but also slow and deliberate), and they don't need to be obvious or conscious.

The motivation for this view largely comes from the perceived failure to pin down specific properties that stand up to scrutiny and successfully manage to map

---

[11]This argument also offers an answer to the question posed by a reviewer (cf. also footnote 4): "Why doesn't the researcher simply reflect on her concept, and produce an account of it from her armchair?" From this point of view, armchair reflection runs the risk of contaminating the data. As Machery (2017: 234-5) argues, probing others' intuitions provides protection against this risk.

A partisan of armchair methods might counter by arguing that armchair methods are rarely confined to a single armchair, but that such conceptual analyses are actually developed in the interaction with colleagues and graduate students. It is not completely clear that this would solve the problem, as a researcher's colleagues might be as motivated as her towards a conclusion. However, even if this difficulty were circumvented, Machery argues that it might not be reasonable to expect that the general population will share the intuitions of a small group of philosophers, as it has often turned out not to be the case (234-5). This is why experimental philosophers typically avoid probing philosophers' intuitions when they want to know what is the "ordinary" account of a concept, or the account of a wider population. This is also why, in the following chapters of this thesis, we avoid using a corpus authored by philosophers.

onto what and only what we would want to consider as intuitions. For instance, Williamson (2008) argues that judgments obtained after some reflection should not be less eligible to the status of intuition than spontaneous "seemings". Indeed, it seems to conform with what philosophers do: we don't need a lot of deliberation to agree that the Searle in Searle's Chinese room does not understand the Chinese characters he's manipulating, but people will take a pause before answering a trolley problem or the violinist dilemma. In a different kind of argument, Williamson shows that restricting what we call intuitions by invoking particularities can lead to undergeneration. His example is that we should count it as intuitive that there are mountains, but these restrictions often lead to categorize this statement as not intuitive. Intuitions are thus behaviourally and phenomenologically diverse, and, as Nado (2016) notes, this almost certainly means that we are also facing psychologically diverse phenomena. To account for intuitions, it would seem, we need to be liberal, and accept any judgment or opinion.

But then, minimalism (or liberalism, as Ichikawa, 2009 calls it) could be facing the same problem that particularists and exceptionalists were trying to avoid in the first place. Opinions are exactly the kind of things that an argument is meant to sway, so using them as evidence and as ground for said argument seems like begging the question. Machery's defence is to embrace what he calls "sociological psychologism" and consider intuitions as indicators for judgments that happen to be widely shared, as opposed to being the ultimate support of philosophical arguments. So, while opinions from a reader or from the author might indeed have been corrupted, we can survey the opinions of those outside of the ivory tower, and get a good idea of what the widely shared judgment is from participants untainted by philosophical debates.

Thus, by expanding the domain of what counts as intuition, Machery significantly expands the domain of what can count as evidence for a certain account

of a certain concept. As he notes, the judgments provoked by common thought experiments such as Kripke's Gödel case and Foot's trolley case lack the features that mainstream particularist theories of intuition deem they should have, hence the significance of minimalism for case-based experimental philosophy.

Given the importance that intuitions have historically had in conceptual analysis as its source of evidence, what counts as intuition is relevant for both case-based experimental philosophy and corpus-based conceptual analysis. However, if it turned out that particularists were right, the method of case would simply have to restrict the judgments it elicits, and perhaps philosophers would try to rewrite the Gödel and trolley cases so as to elicit truly intuitive judgments. As such, particularists do not pose much of a threat to experimental philosophy. On the other hand, it is not impossible that particularism about intuitions could restrict our ability to produce data from textual corpora. Indeed, it is not clear that, given our current level of understanding of writing, we could discriminate or control for the linguistic behaviours observed in a text that are not the product of intuition, understood the way a successful particularist theory would understand it. Because minimalism advocates for an inclusive account of INTUITION, it dispenses us with the need to determine which linguistic behaviour reflects intuitive judgments, and which behaviour reflects non-intuitive judgments. Minimalism makes it possible to look at all the linguistic behaviour involving a concept that can be found in a corpus.

1.1.3 Haslangerian analysis

While the method of cases gives us some tools to formulate the nature and role of empirical data in philosophical analysis, it isn't sufficient for analyzing a concept. What makes a proper portrait of a concept depends on what we are trying to do

with it, therefore we need tools to formulate our philosophical research projects and derive research and evaluation strategies.

When philosophers are tasked to study thinking tools, they can be involved in three broad types of projects. Firstly, they can study the thinking tool as it is used. For example, they might try to give an adequate portrayal of a concept as they believe it is used from their own understanding of the concept. They might also try to study a corpus (as Von Eckardt, 1995 did when studying the concept of COGNITIVE). Or, as experimental philosophers have been doing, they might design and run experiments in order to understand through behaviour how participants conceive and employ these thinking tools.

Secondly, philosophers can study thinking tools as they work within their own systems. This might take the form of studying a concept within a formal system. For instance, a philosopher might study how removing the law of the excluded middle gives rise to different regimes of logic and logical thinking. This work has often been characterized as being *a priori*, but one might also think that it is analog to building reduced models of a new plane in order to study its aerodynamics: the idea is to play with the object to understand how it behaves in various conditions that can be expected to arise.

Thirdly, philosophers might study a thinking tool with the express objective to improve it. They might appreciate how the tool works in a certain context, and wish to adapt it to another or they may believe it has a certain failing, and wish to modify it in order to correct it. They might even think that current tools are not getting the job done in a fundamental way, and try to build up new thinking tools from the ground up in order to replace them. Typically, philosophers who engage in such works have a definite idea of the function of the thinking tools they target, and they will take measures to ensure this function is properly fulfilled.

Concepts are a kind of thinking tool. When we take it to be a constituent of propositions, a concept enables the representation of a set of thoughts that mobilize it. When we take it as holding knowledge about regularities in the world, it enables inferences about its objects (Millikan, 1998). As a result, philosophers who study concepts can also adopt three kinds of strategies, which mirror the three ways of approaching thinking tools in general. They are broadly those described by Sally Haslanger (2012, chp. 13). The first one is the *conceptualist* or *internalist* approach: it corresponds to the question, "what is *our* concept of X?"—"our" corresponding to "our community", as in the community that is involved in using and perpetuating a concept, be it philosophers, members of a Western society, etc. Here one might argue that once a philosopher goes about the world asking others if they share their intuitions about a concept, "*internalist*" becomes a bit of a misnomer. Nevertheless, the gist of the question is the same: it is about achieving an understanding of concepts as they are actually used as human thinking tools.

The second is the *descriptive* approach to analysis. Haslanger describes it as being involved in understanding "what objective types (if any) our epistemic vocabulary tracks" (p. 386). For example, a descriptive analyst might wonder if our concept of DOLPHIN actually corresponds to what dolphins really are. Or a descriptive analysis might inquire if what falls under the concept of DOLPHIN should really fall under this concept. Realizing that there is no principled reason to lump together oceanic dolphins and river dolphins, a philosopher involved in such a project might propose that this concept should only refer to oceanic dolphins. Thus, there is a normative aspect to this approach: if the concept being studied is not as efficient as we might like in tracking its objective kind, then the analyst will suggest adjustments. However, the suggestion of changing the concept DOLPHIN here is not driven by any concern for our understanding of dolphins or our interactions with them, but rather on the apparent disjointedness of the category.

Nor is it concerned with whether current use of the concept DOLPHIN would require this revision. It is concerned with the workings of the concept within its conceptual system—in other words, it studies concepts as thinking tools as they work within their own system.

With this in mind, we might want to think that the descriptive approach should apply not only for the referential function of concepts, but also for other functions. Indeed, concepts don't always refer to something, they only do when it is their role to do so. For instance, when concepts are expressed in verbs like "apologize" or operators like "and", they often function to position the participants of a conversation, or to determine how other concepts fit together in the context of a proposition. Some concepts might even work within a formal system that lacks semantics, and then, studying what the concept does—and whether it does it well—isn't about object tracking but about whether it performs the relevant operations in the relevant contexts. As a result, perhaps "*functionalist*" is a better term to describe this approach than "*descriptive*"[12].

Finally, projects of *ameliorative analysis*[13] put this function in question: "What is the point of having the concept in question [...]? What concept (if any) would do the work best?" (p.386) In such works, we might find a philosopher introducing a new fauna of concepts in order to achieve a theoretical goal (e.g. Millikan, 1984: chp. 1-3), or arguing for a redefinition of a concept in order for it to fulfill a new,

---

[12]Haslanger's views are strongly realist, in that, for her, concepts like KNOWLEDGE and JUSTICE have a referent that is an objective kind—not a natural kind, but a social kind. As a result, from her point of view, all concepts that may be of interest for conceptual analysis have referents. I'm not willing to commit to such a view, and I feel this work can be of use to those who are also disinclined to adopt it, hence my redefinition of descriptive analysis to one that is more tolerant of non-realist views.

[13]Haslanger (2012) also uses the term "analytical approach" (p. 352) in reference to a tradition in contemporary feminist theory. But authors who reference her work mostly use "ameliorative analysis".

or a modified, role.

These three approaches are connected in various ways. For instance, there is a sense in which all projects are ameliorative: whether someone is making explicit how we use a concept, or what its role and operation within a system is, this person is always making a new concept to represent what she found. We do not encode the concept of APPLE in the same way when we use it to recognize apples in visual stimuli or when we use it to talk and think about them. Likewise, it is different to encode a concept simply to perform its role in a formal system or to describe to a colleague the role it has been given in it. Inevitably, there is change, which involves gain in the fact that the concept can be used in new contexts (e.g. explanation, reasoning, etc.) and often some losses. In any case, any successful conceptual analysis, whether conceptualist, functionalist or ameliorative in its approach, will yield a new concept that is identical to its initial target in a way, but different in that it is tailored for new contexts. There is also a sense in which conceptual analysis always involves a descriptive or functionalist analysis: one could not forego a careful examination of a concept's function and functioning before proposing adjustments, and without attending to function, a conceptualist analysis would be a mere description of use cases. And finally, any functionalist or ameliorative analysis ought to have a grounding in actual use, which involves a conceptualist analysis.

This is not to say that conceptualist, functionalist and descriptive approaches are all the same. The difference lies in the purported contribution: a conceptualist analysis' contribution lies in a better understanding of actual use, a functionalist analysis must give us a better understanding of the function and operation of the concept and an ameliorative analysis should improve on the function itself. However, no single analysis stands by itself: a good understanding of a concept involves a good understanding of its use, of how it operates and of its role.

This being said, to Haslanger, differences in types of analyses actually boil down to different projects with different objects. Among them, she makes a distinction between *manifest* and *operative* concepts. The former is typically the one that transpires from an explicit description, like a law or a rule (e.g. "a kilogram shall be defined as whatever weighs as much as the standard in Paris"), while the latter is more like the often implicit, but effective characterization that transpires from use. The object of a conceptualist analysis is typically the manifest concept, while that of descriptive or functionalist analysis usually is the operative concept. Meanwhile, ameliorative analysis' object is the concept that it tries to create: the target concept.

The problem here is that Haslanger does not provide us with a clear-cut way to distinguish between manifest and operative concepts. While she does suggest some criteria, they are sometimes in tension with the examples she produces. For instance, on page 388:

> Consider again my requests to Zina (my daughter) that she lower the volume of her music. Suppose I don't want to listen to music with misogynistic lyrics. I have a concept of misogynistic lyrics and I also have a rough-and-ready responsiveness to what she is listening to. When Zina complains about my interventions into her listening, I may come to find that my responses are not tracking misogynistic lyrics after all, even though that's the concept I was attempting to use to guide my interventions. Let's call the concept I thought I was guided by and saw myself as attempting to apply, the *manifest concept.*

In this passage, the manifest concept seems to be the conscious one, the one that she perceives to be applied, while the operative concept is the one that actually reflects her interventions. Indeed, as she explains a little later (p. 398):

> The manifest concept is the concept I take myself to be applying or attempting to apply in the cases in question. The operative concept is the concept that best captures the distinction as I draw it in practice.

However, in another example, she speaks of the concept of PARENT, as in the institution of parent-teacher conferences. A parent for a human being is usually understood to be one of the two persons that are the immediate progenitors of a person. However, when the school invites parents to a parent-teacher conference, they actually mean to invite the primary caregivers of the children attending classes. Here, there is no disconnect between the concept the school authorities take themselves to be applying and the one that is actually applied: if asked who counts as a parent, they would describe the caregiver, not the progenitor. Therefore, in this example, Haslanger describes the manifest concept as "the concept that speakers generally associate with the term", and the operative concept is "the concept that captures how the term works in practice"[14] (p. 390). In a further example (p. 368-370), she speaks of how her son and his school have different definitions for the tardiness: for the school, following the official rulebook, a student is tardy if she arrives in class past 8:25, but in practice, teachers have different policies concerning tardiness, so a student can arrive past 9 on Wednesday and still be on time, because the Wednesday teacher will not mark her down as tardy. Once again, both of these conceptions of tardiness are conscious and explicit (perhaps the second one is a little less so, as it would not do to make it explicit it in some contexts, but students can certainly do so), so the difference here rather seems to be that the manifest concept is the one that is associated with an authority, while the operative concept is not.

---

[14]By this, Haslanger probably has in mind something along the lines of "social and institutional practice" rather than "linguistic practice".

If the boundary between the manifest and the operative seems to be changing, it might because rather than being a problem, it is a feature of the dichotomy that it is dependant on the context. In *Resisting Reality*, this dichotomy first comes when she addresses theories of the concept RACE[15]. These theories take conceptual analysis to be mainly descriptivist: concepts (at least the referential kind) must map onto essences. Concerning races (as applied to humans), there is no identifiable essence. Therefore, there are no races, and the concept RACE is superfluous—that is, if we take the descriptivist analysis to be the only legitimate one. Distinguishing between manifest and operative concepts enables Haslanger to argue that there is more than one aspect of the concept, which in turn suggests that there should be more than one possible type of analysis. This opens up the possibility for a complementary account of RACE, that is based on the operative concept: in such an account, race might be seen as a concept that plays a variety of social roles, branding those who are identified with a race for discrimination and special treatments.

Given the context in which it is introduced, it seems that the manifest/operative distinction serves to open door rather than circumscribe the concept in a dichotomy. It is meant to distinguish her projects from other types of conceptual analyses which have wrongly assumed to be the only game in town. Branding their object as the manifest concept enables Haslanger to simultaneously frame their contribution to a larger research endeavour, while opening up the space for new types of conceptual analyses. But once those projects lose their claim to monopoly or higher authority, it isn't clear at all that the types of conceptual variations they are interested in is systematically different to the ones they are not interested in.

---

[15]This is Haslanger's example (2012: 383-385).

Therefore, we may read Haslanger's contribution as being mostly critical: conceptual analysis cannot be understood as transparent representation of an object (the concept) into a descriptive language. Rather it is dependent on the kind of purposes that we have for this description, and on how we have carved the object (which contexts are to be included in the study). Furthermore, how concepts function, be it in scientific or naturalistic project or in a social context, is fundamental: conceptualist and descriptive or functionalist projects will ultimately aim to faithfully represent how the manifest or operative concept functions in the contexts they wish to represent, while ameliorative concepts will try to reforge their object into a target concept that performs the function that we want them to perform.

But there is perhaps one fundamental dimension where Haslanger's three types of concept actually differ, and it is in what we might call their empirical grounding (what Haslanger calls "subject matter"). While a conceptualist analysis might produce a concept that fails to capture a natural kind and that fails to have any noticeable effect within the discourse community for which it was crafted, it will be judged much more harshly if it fails to capture what our concept actually is, because it will have failed on its own terms. One might say that, by virtue of how Haslanger's framework divides conceptual analysis labour, conceptualist analyses (even if they are done from the armchair) will have their empirical grounds in phenomena that indicate how we conceive of a concept (e.g. linguistic or categorization behaviour) and descriptivist analyses will have their empirical grounds in phenomena that indicate something about the objects themselves. Even ameliorative analyses will be grounded in an understanding of the context in which concepts are meant to play as well as the role they play in it (in particular, their purposes, the constraints that act on them, the mechanism in which they participate, etc.). Haslanger, of course, frames this as knowledge of the why: "why do

we have a concept or belief?" And it is from the answer to this question that the analyst can move to the central question of ameliorative analysis: "what concept (if any) would do the work best?" It should be obvious that without knowledge on the workings of a concept in a linguistic context and/or in a community, any speculation on this matter would be moot.

One takeaway from this is that, firstly, all conceptual analyses have an empirical ground, and therefore, all conceptual analyses have the potential to profit from empirical data[16]. This is why even the rationalist tradition, which has been symbolically associated with armchair speculation, has seen its proponents attracted to various attempts to incorporate empirical data in the debate, as we have seen in section 1.1.1. Secondly, analyses of every type actually have some stake in the empirical grounds of other types of analyses, because despite the division of labour, a successful concept has to be successful in working with the concepts that we actually have (as opposed to the ones a philosopher might think we have), in fulfilling its function and in being suited for the context in which it is meant to play. This is why, for example, in the context of carnapian explication, experimental philosophy has been proposed to play a role in informing us on those three aspects of the concept (Koch, 2019; Pinder, 2017; Shepherd and Justus, 2015). Thirdly, there is nevertheless value in distinguishing among these different types of conceptual analysis, if only to apply the right standards of evaluation.

[16]Furthermore, because experimental philosophy works by provoking linguistic behaviour, this statement also applies in principle to textual data found in corpora. Indeed, if in a study experimental philosophers elicited linguistic behaviour that could not be found in any possible corpus constructed from speech and writing taking from natural settings, then this would suggest that the experiment is not ecologically valid.

1.1.4 Carnapian explication

Carnap's project lies somewhere between scientific ambitions of the conceptualist and descriptive projects and the revisionary ambitions of the ameliorative projects. In *The Logical Foundations of Probability* (Carnap, 1950), for example, Carnap is involved in clarifying concepts that scientists already commonly use: DEGREE OF CONFIRMATION, INDUCTION, PROBABILITY. These concepts are "usually sufficiently well understood for simple, practical purposes" (2), but Carnap gives himself the task of reaching a more precise understanding of them through the method of *explication*. Put this way, his project does not seem revisionary: we expect those concepts to keep functioning the same way, at least in the simple and practical purposes for which they are already commonly employed. However, Carnap recognizes that clarifying a concept necessarily produces a new concept: clarifying ambiguity involves determining features of a concept, and thus modifying it. Hence, explication has both a conservative and a revisionary aspect.

Carnapian explication describes a process by which we form a new, more precise concept (the *explicatum*) from a typically less precise and relatively unscientific concept (the *explicandum*). Much like Haslanger's ameliorative analysis, the process is described, at least by Carnap himself (1950: chp. 1), as a two-step process[17]. Since the *explicandum* is relatively imprecise, the problem that an individual explication is meant to solve is, at the outset, never very clear. This is why we ought to clarify the *explicandum*. Carnap suggests that this can be achieved by giving examples of contexts where we use the concept we wish to explicate, and examples of contexts where we might think it is being used, but where the

[17]Brun (2016) further analyses explication into a four-step process, which he obtains by making the role of evaluation in explication more explicit. Thus, after clarifying the *explanandum*, one should clarify how the various criteria should be interpreted in the context of the explication, and the fourth and final step is simply a critical appraisal using those criteria.

concept mobilized is actually a distinct concept. For example: "I mean by the explicandum 'salt', not its sense which it has in chemistry but its narrow sense in which it is used in the household language" (Carnap, 1950: 4-5). Once we have clarified the *explicandum* as such, we can go on to provide an explication, which "may be given, for instance, by the compound expression 'sodium chloride' or the synonymous symbol 'NaCl' of the language of chemistry" (*idem*).

Furthermore, explication is usually about taking a concept from a conceptual paradigm (or "system of concepts", as Brun, 2016 calls it) and making it available to another. Explicating table salt with a chemical formula enables us to insert this concept into chemical discourse. For scientific purposes, explication can thus serve as a bridge between various models (Meunier, 2017). For instance, if we want to know whether crows are more intelligent that finches, we might need to explicate the folk concept of intelligence into a relevant ethological framework. From this, we might want to formalize this ethological concept of intelligence, so that crow behaviour and finch behaviour can be made comparable. Then, this formalized concept can be explicated into a physical/experimental concept that can be used to construct experimental protocols and applied to observed behaviour. In this string of operations, the explicator must have guarantees that, at each transition, what has been learned about a concept from the *explicata* can permeate back to the *explicanda*. For example, if crows do better than finches in a set of experiments, it is through a well-constructed structure of explications that we can convert statements about behaviours into general statements like "Crows are more intelligent than finches" in everyday settings.

Carnap does not spell out many constraints on what an explication may be, but he gives us the means to guide and evaluate it in the form of four criteria. The first is *similarity*: the *explicatum* must be reasonably similar to the *explicandum*, which is to say that in the contexts that count, the *explicatum* has to be able to do the

job of the *explicandum.* Then comes *exactness*: since this is the whole idea of the project of explication, it stands to reason that the *explicatum* has to be more exact, "so as to introduce the explicatum into a well-connected of scientific concepts." (Carnap, 1950: 7). Of particular importance is the criterion of *fruitfulness*, which he describes as usefulness for the formulation of universal laws in *The Logical Foundations of Probability* (Carnap, 1950). As Dutilh Novaes and Reck (2017) put it, the *explicatum* ought to be conducive to the production of new knowledge. Finally, the weakest criterion is *simplicity*—a criterion that Carnap presents as a tie-breaker between equally good candidates for an *explicatum.* In other words, an explication that is simpler (more parsimonious) than another is *ceteris paribus* also better.

While the context of scientific discovery is central to Carnap's project, it might be useful to generalize explication beyond this restricted endeavour. This would mean, for instance, that a concept being fruitful might mean more than just enabling the creation or formulation of new knowledge, but that it also could lead to some social improvements. In fact, even when restricting ourselves to the scientific project, we can see how fruitfulness comes to represent very different things in different cases. For example, Carnap discusses the explication of the everyday concept FISH (which would include such things as whales and cuttlefishes) with the scientific concept PISCES (cold-blooded gill-bearing vertebrates[18]). There are things that can be said of pisces that cannot be said of fishes (in the old sense): that they evolved from an amphioxus-like creature, that they are chordates, etc. Because it carves nature at the right joints, this new conceptualization enables new generalizations. Now, compare this with the kind of explication we employ to make experimentation possible: for example, when we explicate CONSCIOUS-

[18]This is Carnap's definition, but it is not perfect, as it would include the axolotl, but exclude some lungfishes. Wikipedia's *Fish* entry defines them as "gill-bearing aquatic craniate animals that lack limbs with digits."

NESS as "the content of the participants' experience as she or he is able to convey it". In such a case, the *explicatum* does not shed any light on the subject matter by itself, but only because it enables the construction of an experimental protocol. It appears that fruitfulness is not tied to any specific way of contributing to knowledge—all that is required is that it enables knowledge to progress further. Therefore, employing explication for other purposes is by no means a big stretch[19].

As a referee of Dutilh Novaes (2018) notes, it is easy at this point to imagine that ameliorative analysis and explication might, to a certain degree, be fruitfully thought of as being approximately the same. Dutilh Novaes initially counters by noting that explication projects can be pursued within any approach of conceptual analysis that Haslanger describes, including conceptualist or descriptivist/functionalist approaches. She also notes that in exactness, explication possesses a criterion that is absent from Haslangerian analysis, and that, in turn, ameliorative analysis employs tools, such as ideology critique[20], that are absent from explication. However, on the one hand, if some descriptivist projects can also be explications, perhaps it is that descriptive analyses could also be described in terms of an ameliorative analysis, with the goal of the amelioration being of a scientific nature (perhaps "carving nature at its joints", or promote a goal that would

---

[19]Some might argue that the move would affect the exactness criterion. Indeed, the association between exactness and scientificity might seem like a natural one, but it isn't clear that it actually applies more specifically to scientific discourse. Brun (2016: 1222) argues that exactness is about such various objectives as reducing ambiguity (including reducing the amount of cases where we can doubt whether a concept applies or not), not leading to paradoxes and allowing for finer and more precise descriptions. Unlike Dutilh Novaes (2018), I don't think we should read the exactness criterion as a call for formalization. Beyond the connection with Enlightment ideals (Carus, 2008), the conceptual hygiene that is evoked through the criterion of exactness does seem like a practical necessity for entertaining the desideratum of fruitfulness: how could we affirm that a new concept is fruitful, if we ignore whether it will lead to paradoxes or if we do not know how it will behave in limit cases? Furthermore, these are all important preoccupations in Haslanger's politically motivated ameliorative analyses as well.

[20]Ideology critique is an analysis that is focused on the thinking tools like concepts and narratives that we employ to navigate the world. Cf. Haslanger (2012: 17-22).

lead to new knowledge down the line). On the other hand, Dutilh Novaes argues that explication and Haslangerian ameliorative analysis should take inspiration on each other, which is testament to the fact that their respective virtues are not restricted to the domains of issues from which each type of analysis originates. Perhaps, down the line, we will see explication projects that include ideology critiques, and ameliorative analyses that make attempts at formalization, such that they will become indistinguishable. Furthermore, neither Haslanger nor Carnap mean to define their respective forms of analysis. As we mentioned, Haslanger's trichotomy seems to be aimed at opening the stage for new types of analysis, rather than closing down the possibilities to three rigid types of projects. Carnap, on his side, formulated his criteria in ways that could afford a variety of interpretations. Therefore, while we can agree with Dutihl Novaes that explication and Haslangerian analysis are still different things, it is not clear that it should remain so.

This said, there is at least one sense in which explication goes farther than haslangerian analysis, and it is in its capacity to link concepts from different systems of concepts in a relation of identity. Haslanger takes Appiah's descriptivist analysis of RACE and her own ameliorative analysis of it to be about the same concept. However, there is little substance about this identity, which goes from having its extension in things like genes and phenotypes to things like social representations and dynamics. Through Carnap's account of similarity, which will be unpacked in section section 1.2, we can explain why conceptualist, descriptivist and ameliorative analyses can give different but equally valid accounts of the same concept.

Unsurprisingly, explication's binding powers also come in handy with experimental philosophy and corpus-based conceptual analysis, where it connects the language of the hypothesis to the language of experimentation. As a conceptual analysis project mobilizes a corpus in order to answer its questions, it must go from the

concept as it lives in the context that has motivated the conceptual analysis to a concept that lives in its empirical domain. Most importantly, the connection between the origin concept and the target concept must be articulated in such a way that discoveries on the empirical side can translate into insights for conceptual analysis. Articulating an explication means articulating the conditions under which what can be said about the *explicandum* can also be said about the *explicatum*, and vice-versa.

1.1.5 Conceptual analysis for Alice

One interesting recent development in empirically-informed methods of conceptual analysis is a novel interest in articulating methods that have mostly evolved separately. Dutilh Novaes (2018) has investigated the intersection of explication and haslangerian amelioration analysis and has argued for convergence, and Machery (2017: 215-7) draws a similar parallel. Meanwhile, Shepherd and Justus (2015), Pinder (2017) and Koch (2019) have explored the possibility of using experimental philosophy to inform explication—though the main prize might have been to provide experimental philosophy with a method that avoids intuitions and its pitfalls.

The consensus, so far, has been that syncretism is probably a winning strategy for the development of conceptual analysis. After all, in articulating explication and haslangerian analysis, we have seen that we gain from an extension of the available tools on both sides (Dutilh Novaes, 2018); we also gain Haslanger's insight into the division of labour, and carnapian explications capacity to articulate the different accounts of a single concept that are conceived through conceptualist, functionalist and ameliorative analyses. Similarly, experimental philosophy contributes empirical grounding to explication, while explication brings in a way to connect

concepts from philosophical discussions into the experimental (or observational) realm, thus foregoing the need for intuitions.

I find no reason to doubt this consensus. As noted in the previous section, haslangerian analysis and explication are formally the same. As such, adopting Dutilh Novaes' strategy of recuperating insights and tools of inquiry from both sides makes sense. Thus, a syncretic analysis would probably employ the elaborate methods devised for carnapian explication (Cf. Brun, 2016), but it would also have to position itself with regards to the division of labour, and use this position to leverage the information from other types of analyses of the same concept in order to improve upon itself.

However, syncretism might be a bit more difficult to achieve when it comes to determining how empirical data can inform analysis. For Shepherd and Justus (2015) and Koch (2019), experimental philosophy (and, we can assume, textual data analysis as well) can inform the explication preparation phase, where one has to get a clearer account of the *explicandum*. For Pinder (2017), on the other hand, the contribution experimental philosophy can make to explication preparation is too small to make it worthwhile, and the way should rather be to use experimental philosophy to probe the conceptual environment to which the *explicatum* is destined in order to predict if it will be successful enough to be adopted by the community. Koch (2019) disagrees, finding uptake to be a poor indicator of success, and judging Pinder's plan difficult to materialize.

That being said, taking into account the division of labour in conceptual analysis, it seems that the most productive contribution of experimental philosophy for an explication should depend on the kind of explication. If we are in a conceptualist project, then the explication preparation is certainly the most important step, as the goal of the *explicatum* is to enlighten us on the *explicandum*. On the

other hand, while we might agree that uptake might not be a good indicator of a concept's quality, an empirical study of the *explicatum*'s conceptual environment is crucial for an ameliorative project, as we need to predict how the modified concept will play in it. As such, the debate between Sheppherd, Justus, Pinder and Koch is probably misguided.

Finally, one might wonder where the method of cases fits in this picture. Machery's (2017) take is that an overly aprioristic method of case, whereby philosophers only investigate their own intuitions and that of their friends, is empirically underpowered and might beg the question. His solution is to study intuitions on more representative samples, which then enables him to adopt a minimalist view of intuitions. As such, because it enables the full breadth of textual data from corpora to be taken into account, this solution also does a lot towards making corpus-based conceptual analysis practicable.

Thus, we might describe a syncretic method of conceptual analysis as follows.

To Alice, a conceptual analysis might start with an inquiry into the problem she's facing[21]. Firstly, she needs to clarify her problem—in particular, as Carnap suggests, she needs to clarify which concept she wishes to analyze (we will call it the *original concept*). She might also want to determine whether her project is ameliorative, conceptual or descriptive, and determine how her conceptual analysis might play in the philosophical debates in which she wishes to engage. From this, she will have an idea of the discursive and pragmatic space that the concept that will be constructed in the analysis (let's call it the *target concept*) will have to inhabit. Hence, she can determine what purpose the target concept is meant to fulfill. From the knowledge of the target concept's purpose and purported context, she can infer constraints: form of representation, contexts where it needs to play

[21]Here I take inspiration from Brun's (2016: §3) "recipe" for explication.

the same function as the original concept, degree of exactness, etc. There will likely also be constraints that are specific to the function the target concept is meant to fulfill. For instance, if the target concept is meant to represent in a theoretical discussion a concept as it is used in a corpus, it will only be adequate if it efficiently reflects the original concept, and if Alice can be justified in thinking so. If, on the other hand, she wants to improve on a concept as it is used in a corpus, she will need to highlight how the original concept performs its function, and where it could be improved.

From this, Alice can plot a path getting to the target concept. In her case, this means that she will first need to determine what kind of corpus she needs to construct. On the one hand, from the way her project is formulated, she will be able to draw conclusions as to which set of assertions are relevant. For instance, she might want to have a corpus that is representative of the linguistic and discursive behaviour of the community that uses the original concept. Then, she needs to determine what kind of contexts are contexts where the original concept is present—that is, performing the discursive functions that are relevant to the question at hand—and which contexts are actually good indicators for the concept being analyzed. While Machery argued that there is no reason to demand that she discriminates according to, say, the mental faculty that is involved in applying the concept, there might be cases where a concept would appear to be associated to another concept only for discursive reasons that fall outside of its function: for example, if our corpus is collected during the 2018 World Cup, the concept STADIUM might seem strongly associated with the concept RUSSIA, but this does not reflect on the function of any of those concepts.

Finally, from observations and experiments on the corpus[], Alice can propose a target concept that fulfills the criteria as previously stated and interpreted, and evaluate how well the new concept fulfills its objectives.

1.2 Similarity

Between the vagueness of the terms employed to theorize conceptual analysis and the difficulties that arise in operationalization, theoretical and methodological difficulties abound—which is why we can only rejoice in the increasing interest philosophers have been putting into method and metaphilosophy. In the rest of this article, we will be concerned with a pair of related problems with regards to similarity.

The first problem is about identification between the original concept (in Carnap's words, the *explicandum*) and the target concept (the *explicatum*): how should we understand this relation? Which requirement comes with it? It might seem that this problem is sometimes treated a bit lightly. For instance, in the *Logical Foundations of Probability*, Carnap (1950) insists that the requirement of similarity should be flexible (although he doesn't give any limits to this flexibility) and in "Replies and systematic expositions" (Carnap, 1963), he proposes that if we interpret similarity as synonymy, we should allow at least three different senses of synonymy to be employed, depending on the context. Haslanger, on her part, does not explicitly address it, and Machery sees it as merely embodying a form of conservatism: "Concepts should not be modified without reason, and when they are modified they should be modified as little as possible" (Machery, 2017: 215).

Surely, however, similarity is about more than just conservatism. Take cases where the target concept is meant to play a role for the original concept: for example, cases where the target concept is an *explicatum* that enables us to do experiments. Here, the objective is to learn new things about the target concept that will also apply to the original concept. Such a transfer from target to original concept supposes that the two concepts are similar enough that, barring some constraints, properties of one concept can be justifiably applied to the other. Inversely, in

order to fulfill their role, the new concepts need some of the information that is contained in the old one. Indeed, it is hard to imagine a case where we don't want to transfer knowledge or functions from an original concept to the target concept. There is more at stake here than, say, prevent costly and unnecessary change: we need to justify the transfers between original and target concepts.

The second problem hits closer to home for philosophers who (like Alice) practice conceptual analysis with empirical data, and perhaps observational data in particular: how exactly are we justified in identifying two instances of the use of a concept as two instances of the use of the same concept? For example, how can we feel secure in thinking that two cases of thought experiments mobilize the same concept, or that the concept they mobilize is the concept we wish to inquire? Alternatively, in corpus-based conceptual analysis, how can we feel justified in thinking that two segments exhibit traces of the presence of the same concept?

These two questions could have demanded two distinct answers. Indeed, the causal threads which link concepts in those two questions might be of different nature: in a deliberate conceptual analysis like an explication or an ameliorative analysis, the target concept is constructed from the original one, whereas the dynamics of concept diffusion, drift and repeated reinterpretation that occurs naturally in a community are much less deliberate and likely are the result of a very different, natural evolution. As a result, we might expect that what unites and distances concepts in those two contexts would turn out to be very different. However, as we shall see, there is a single answer to these two problems.

In the rest of this section, we will assess various propositions for establishing how similar or dissimilar concepts can be: firstly, by similarity of intension, then by similarity of extension, and then by similarity in function. "Intension" and "extension" are terms that tend to take different definitions depending who you are

speaking to. However, generally speaking, intensions correspond to the internal content or the essential properties of a term or a concept. For instance, if the concept is determined by a definition, it would be that definition; if it is a cluster concept[22], then it might be a list of properties, perhaps along with weights representing the importance of the property for a object instantiating the concept. For our purpose, intension shall be a representation of the properties germane to a concept. Extension is more straightforward: it is the objects that the concept is meant to represent. For our purpose, the extension of a concept is the set of possible objects that would be its instances if they are/were real[23].

I will argue that there are major issues with similarity of intension, and extension that make them poor candidates for the similarity criterion. On the other hand, not only is function more apt to account for the similarity in a diversity of contexts, but it comes with a perk: it afford natural cutoff points for judgments of identity.

1.2.1 Similarity by intension

The question of similarity touches on the question of what is fundamental in a concept. When we say that humans are similar to chimpanzees, it often comes with some kind of evidence: sometimes, it is about DNA ("we share 98% of our DNA"), sometimes it is about ancestry ("they are our closest relatives"), sometimes it is about phenotype, behaviours like problem-solving or social mores or cultural transmission, etc. Whatever is mentioned, it usually is deemed fundamental, at least in the discursive context, of what it is to be a human or a chimpanzee as

---

[22]An object instantiate a cluster concept if it possesses a certain number of the attributes that are associated to this concept, while none of these attributes is necessary or sufficient for instantiating the concept. Cf. Searle (1958).

[23]I do not assume here that all concepts have an extension, as we will be clear in section 1.2.2.

species. Of course, if we believe that chimpanzee or human essence (as species, of course) lies in DNA, phenotypic comparisons are not out of question, as genotype is a huge factor in determining phenotype, and, *ceteris paribus*, individuals with similar genotypes also have similar phenotypes. But if we can, we might as well hear it from the horses mouth, and check genotypic similarities. As such, we can assume that philosophers who think that a concept is its intension will also think that concepts which are similar to each other are concepts whose intensions are similar.

Those who see concepts as being intensions typically think of a concept as being its essential properties or predicates (with what is essential being largely dependant on what one believes to be the essential role of a concept in an organism's cognitive economy). Essentially intensional concepts can be found in a large variety of philosophical and scientific traditions. In traditional conceptual analysis (e.g. King, 1998), a concept just is its decomposition into necessary and sufficient conditions—a brother just is something that has both the property "sibling" and the property "male". In cognitive psychology, concepts take different logical forms (Harnad, 2009; Machery, 2009; Murphy, 2004), but they are also characterized by a form of subject-predicate association, even if the predicate is often fuzzy and neither necessary nor sufficient for categorization. Intensional criteria are also common in computer science. Proponents of the method of case are also typically fond of the intensional concept. For instance, Machery (2017) suggests that a concept is a set of belief-like states ("bliefs") about the substance.

To evaluate similarity between intensions, we can encode them into digital representations. A standard way to do this is to code properties as variables, while a concept can be coded as a data point. In such a case, it might seem that geometrical measures such as the euclidean or the cosine distance would be an obvious choice to give us a good idea of how similar or dissimilar two concepts

can be. However, while it would work well for cases like prototype or exemplar concepts, it would not work with concepts that need to be represented using more complex forms of representation, like schemata (Minsky, 1975), and that cannot be represented as a point in a high-dimensional space without loss of information.

This said, there are perhaps other ways of measuring similarity and difference between digital representations that could perhaps bridge the gap between representations of very different forms. For instance, a promising avenue might be to think of two concepts as being a few modifications away from each other. By computing the minimum amount of modifications needed to go from one string representation to another, measures like the Levenshtein distance (Levenshtein, 1966) and its derivatives can give us an estimation of this drift, and effectively tell us how similar these representations are. These measures can (and often are) easily adapted to measure differences between objects that have different logical forms; they could therefore be adapted to measure differences between representations of the intensions of two concepts. Furthermore, as these measures can be adapted to measure representations in various forms, they could possibly be applied to any concept intension, no matter its form.

However, it isn't clear that similar intensional traces actually mean similar concepts (let alone identical). Take the FISH/PISCES example: one might describe a fish as an aquatic animal, whereas the PISCES intension also includes other properties, like having a skull, a notochord and gills and lacking digits on the limbs. Intensionally, it would seem that concepts like SIRENS (a family of gilled limbless aquatic salamander) are a lot more similar to PISCES than FISH is to PISCES: sirens also have notochords, skulls, gills and no digits on the limbs, but only the last property can be expected of all fish in the old prescientific meaning of FISH. This problem also arise in more natural settings, as we commonly describe the same things in various ways. Berenice might think that, essentially, water is H_2O,

while Charles might think of it as a transparent liquid with the ability to quench thirst. Their intensional pictures of water have nothing in common, yet they have little difficulty agreeing that they are talking about the same thing. There might be a sense of CONCEPT in which it is relevant to say that this person and I have different concepts of WATER. But if we are trying to understand how the community in which they both live understands the concept of WATER, it seems like both their voices should be included. In other words, for our purposes, intensional similarity does not seem to do the work we would want it to do.

1.2.2 Similarity by extension

The obvious next step after putting intension aside is to take a look at its externalistic twin, similarity by extension. Carnap (1950: 7), reports that this criterion is employed by Karl Menger (1943) with definitions: "A good definition of a word must include all entities which are always denoted and must exclude all entities which are never denoted by the word." It is worth noting that Carnap, however, does not endorse this view for concepts (cf. Brun, 2016).

Nevertheless, similarity by extension has some things going for it. In particular, it would work a little better in practice, at least with the example that we just mentioned. Whether you think of water as H_2O or as a liquid that quenches thirst, the extension remains the same. Indeed, it would seem sensible to think that it is because these two intensions refer to the same thing that Berenice and Charles are talking about the same thing. This said, in the FISH/PISCES case, Carnap notes that the latter is much narrower in extension than the former, and thus, that "they do not even approximately coincide". However, thanks to work pioneered by Rosch (1973), we now know that, at least in people's minds, not every instance of a concept counts equally: there is a sense in which a carp is more of a fish than

a seahorse, or in which an apple is more of a fruit than a pineapple. Perhaps the proper way of measuring similarity is through a weighted metric that gives more importance to co-extension in instances that are more emblematic of the concept. In such regards, PISCES and FISH are certainly similar, as they conjure the same exemplars of carps and trouts.

However, there are grounds to doubt that it is always fruitful to think of a concept as referential (even sometimes going through great lengths to find a domain where it can be instantiated). Quite often, it is more useful to think of a concept as a tool, say, to structure discourse, knowledge and behaviours. For instance, some concepts are used in ways that suggest that their main or only function is to position an assertion pragmatically or rhetorically (e.g. APOLOGIZE in "We apologize to our readers."), or to convey mood or attract attention (e.g. IMPORTANT in "This package is very important."). If we adopted similarity by extension, then we might be unable to use the similarity criterion on those concepts, which might be a problem.

A way out would be to find strategies to assign an extension to every concept. Perhaps we should force ourselves to think of APOLOGIZE as a verbal form referring to acts of positioning oneself in discourse, and of IMPORTANT as having for extension the set of all things that are deemed important by someone. One worry with this solution is that this might actually change with the meaning of a concept. It does seem that there is something performative in calling something important that goes beyond asserting that something belongs to the set of things that are important. Much like explaining a joke will ruin it, explaining why we think that something is important will not have the same effect as calling it important.

More importantly, when concepts are abstract, we may be tempted to draw their

extensions in more than one place, all of which might be equally adequate. Brun (2016) gives us an example of this from Stalnaker (1976):

> "the proposition [a sentence expresses] will be a function taking possible worlds into truth values. Equivalently, a proposition may be thought of as a set of possible worlds [...]". (Stalnaker, 1976: 80)

Here, PROPOSITION_1 extends on functions, while PROPOSITION_2 extends on sets. Therefore the extensions for PROPOSITION_1 and PROPOSITION_2 are disjointed. While we should expect those two *explicata* to be very similar, judging only by extensions would tell us that they are very different. It seems that, at least for concepts which lack extension in the physical world, extensions are not an appropriate way of judging similarity.

1.2.3 Similarity by function

Arguably, one of the best interpretations of Carnap's own understanding of his criterion of similarity is by way of comparing concept functions. Indeed:

> The explicatum is to be similar to the explicandum in such a way that in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted. (Carnap, 1950: 7)

This "most cases" ought to be interpreted as "relevant cases" (Brun, 2016; Dutilh Novaes and Reck, 2017)—i.e. relevant for the problem the explication is meant to solve. Thus, the idea here is that in those relevant cases, the *explicatum* and the

explicandum are interchangeable in use, which is to say that they perform about the same function[24].

Similarity by function has much going for itself. Firstly, it avoids the problems that we ran into with intension and extension. Stalnaker's explications for PROPOSITION are equivalent because, even though they have different intensions and extensions, they can still perform the same function (at least in most contexts). While it is possible to have explications where intensions and extensions are completely different in *explicandum* and *explicatum*, there still needs to be functional overlap: for a piece of knowledge to be applicable to both, there needs to be a sentence embodying that knowledge where they play the same function.

Secondly, even though these functions are not the traditional functionalist's kind[24], they can be realized in multiple ways. Thus, variations between or within individuals are not an issue. It doesn't matter if different authors in a corpus think of a concept in different ways: if they are using it in a similar way, we can be confident that it is the same (or about the same) concept. This also means that we can encode a concept in different ways (as Machery, 2009 suggests) and still be talking about the same thing, so long as there is some functional overlap.

Still, Carnap's account of similarity by function is quite thin, and if we're to apply it systematically, we need more details. Functionalism about concepts can find a more elaborate account if we mobilize Ruth Millikan's (1984, 1998, 2017) works, as it provides a more precise notion of "function" and a system of concepts to go with it. To understand it, we must first understand that Millikan's account is tightly dependent on a peculiar understanding of the ecosystem that houses cultural artefacts such as words and concepts.

---

[24]Here, function has a sense similar to "purpose" or "role", and is not closely related to the mathematical function or the function in computer science. Therefore, we should not think of this functionalism in the traditional way. Cf. Millikan (1984: 18).

Millikan starts from the realization that words, concepts and other cultural artefacts are subject to evolutionary pressures of sorts. On the one hand, they are subject to replication: effectively, we use the words, concepts and other linguistic devices that we have encountered before, so we are replicating the communicative behaviour we have seen in others. On the other, replication is not perfect. If Debra teaches a concept to Eleanor, Eleanor might retain a slightly modified version of the concept, or might even use it in slightly different ways. Debra herself might use a concept in a standard way on most days, but might get creative in certain circumstances, and count on the intelligence and culture of her audience to play with that concept's meaning. Therefore, there is also a space for innovation and semantic drift.

Furthermore, replication does not happen randomly: if we replicate a concept, a word or another language device (as Millikan calls them), it is because we wish to accomplish something, and the language device helps us accomplish that something. Another way of saying it is that we replicate them for a certain function. The function in virtue of which linguistic device is replicated is what Millikan calls this device's "proper function." That is, the proper function is not necessarily the function in virtue of which a specific instance of a linguistic device has been used, but rather the function in virtue of which the linguistic device is being replicated in general in a linguistic community. In other words: the function that ensures that a linguistic devices remains alive in its community. So, to take an example: what identifies a word is its proper function—the function that ensures its being replicated across time and contexts. Depending on the context, "happy" can express a lasting contentment or an ephemeral joy: these two "happy" express different things, have different functions, and therefore are actually different words despite their being associated with the same morpheme.

The nature of linguistic communication, with its requirements for some sort of

alignment between speakers and hearers, creates what Millikan calls "stabilizing proper functions" or "standardizing proper functions." Indeed, under normal conditions, the function of a language device ought to be of value for both the hearer and the speaker, or else the exchange would collapse for lack of cooperation. This implies a certain uniformity in function: if the speaker wants a certain reaction from the hearer, she better stay conservative and employ language devices as they are most employed. Inversely, if the hearer wants to extract the right information from the exchange, she will want to use conservative interpretations.

Now, how should we account for concepts and their functions in such an ecosystem? For simplicity's sake, let us first consider words once again. To Millikan, a word alone has no proper meaning of its own; rather, meaning is imparted to sentences (1984: 80), and words have meanings in the context of sentences. Sentences are themselves constructed by replicating syntactic forms—that is, patterns of word arrangements that serve specific rhetorical purposes. For instance, "Long live the revolution!" and "Down with the tyrant!" share a simple syntactic form, where the first slot serves to express a sentiment towards the object that is in the second slot. Other syntactic forms, like "Would you ... ?", "Could you ... ?", "I would like ... please." need to be adapted to (and with) other syntactic forms in order to construct a proper sentence. Thus, the role of word in a sentence is mediated by the syntactic form that inserts themselves in the sentence.

While it may happen that a conceptual analysis actually analyzes a word (in Millikan's terminology) rather than a concept (indeed, Brun, 2016 argues that Carnap explicitly accounts for this possibility in the context of explications), prototypical conceptual analyses from Carnap and Haslanger portrait concepts as accomplishing a lot more than just sentences. Millikan's concept of WORD is tightly associated with lexical forms on the one hand and sentences on the other. It is close to what computational linguists call "sense": a single semantic unit

associated with a word or expression. On the other hand, Carnap and Haslanger see concepts as structuring discursive, scientific and political practices in general. For instance, PISCES organizes entities in such a way that makes it possible to make new true statements about them, and Haslanger's concept of PARENT acts as a sort of gatekeeper for some social institutions and practices.

What about concepts? Millikan does have a concept of CONCEPT, but it is not quite what Carnap and Haslanger are talking about when they are talking about concepts. For Millikan, most of the time, "concept" is short for what she calls "empirical concept" (her writing implies that there might be non-empirical concepts, but to my knowledge, she never really develops this notion). Empirical concepts are public ways of referring to kinds or properties: they are shared like words, and they can be used to identify the entity they refer to as well as hold information about it (Millikan, 2017: 47). As such, at its core, such a concept is the ability to categorize between instances and non-instances (Millikan, 1984: 253), but because this ability is dependent on what I know about the concept, it is also the set of information that we have about its object(s).

However, on the one hand, Millikan herself has recently rejected her concept of CONCEPT (Millikan, 2017: 49: "my claim is that there are no such things as empirical concepts"). On the other hand, even if we were to argue against her that there really were such things as empirical concepts, much like Millikan's concept of WORD, this concept of CONCEPT does not fulfill the role that we need it to fulfill. Not only does it differ with the concept of CONCEPT that is employed in conceptual analysis, but it fails to account for many of the concepts we may wish to study. Indeed, if there are such things as concepts that do not refer, we would like to be able to account for them. Thus, Millikan's concept of CONCEPT might be too restrictive for our purposes.

This does not mean that we should abandon the idea of a concept of CONCEPT that is functionalist in a millikanian sense. Millikan provides us with an elaborate system of concepts to talk about cultural artifacts in terms of proper functions and selection; there is no reason why we should not be able to construct a concept of CONCEPT out of it that satisfies our needs.

Millikan (1984) puts a strong focus on language, but she addresses questions that are narrowly focused on relatively esoteric topics of philosophy of language using concepts that have a much wider applicability. In the first three chapters, she constructs a set of concepts meant to talk about biological and cultural artifacts in terms of what accounts for their pervasiveness—viz. their systematic reproduction and selection. This is where she defines such concepts as proper functions and stabilizing proper functions. This part then serves as a theoretical ground for the rest of the book, which is more narrowly about language, and in particular the topics of language that drew the interest of the analytic philosophers at the time. Millikan's interest in, say, proper function, is thus narrow, but her system of concepts have been applied elsewhere—including in philosophy of biology (e.g. Schwartz, 1999), epistemology (e.g. Plantinga, 1993; Nolfi, 2016), meta-ethics (e.g. Wisdom, 2017) and semiotics (Menary, 2007).

Concepts, as they interest conceptual analysts, touch on a large and discontinuous domain. They are employed in language, but not only at the level of sentences: they are used to construct narratives, stories, tropes, arguments, and other linguistic constructions that structure discourse and lie above the level of sentences. They are also manifested in social, scientific and political practices, rituals and institutions, such as laws, parent-teacher conferences and experimentation best practices. These are all things that are reproduced in a similar way that sentences and language devices are reproduced.

For instance, parent-teacher conferences are events that get replicated regularly in schools across the world. To use Millikan's terminology, they are members of reproductively established families: because they successfully fulfill a certain proper function, they are allowed to be reproduced, and thus form a family of similar instances. Much like sentences (Millikan, 1984: 22), parent-teacher conferences are formed from a variety of model: they retain forms that are also prevalent in other meetings (like plans for orders of business, presentation rounds, etc.) as well as forms that are relevant to the kinds of discourses or social context which is specific to these kinds of meetings. Furthermore, they articulate various cultural devices (parents, teachers, children, learning, child development, speech, etc.) in a coherent whole that promotes the proper function of the parent-teacher conference. In the same way that the concept PARENT acquires a meaning in a sentence, the concept PARENT also acquires a meaning in a parent-teacher conference. Indeed, concepts relate to higher-level cultural artefacts like parent-teacher conferences or narratives in the same way words relate to sentences: they have no proper function of themselves, but rather, they have derived proper function from their association with other devices to form these higher-level entities.

Therefore, we might think of a concept as the thing that composes higher-level entities, in accordance with the role that is imparted to them by the discursive forms[25] that model the higher-level entity. It is an analog to the millikanian word in the context of the sentence—indeed, in a sense, the millikanian word is a special type of concept for which the higher-level entity is the sentences. If grasping a word is grasping its proper function as manifested in the syntactic forms that bind it to sentences, then grasping a concept is grasping its proper function as manifested in the discursive forms (for lack of a better word) that bind it to

[25]I think of the discursive forms as higher-level entities' analog to the syntactic forms for sentences.

higher-level entities.

Now that we have placed our concept of CONCEPT in a millikanian system of concepts, we can wonder how it helps us with similarity. If a concept is grasped by way of its proper function, it is also individuated by way of its proper function. Given that it is always a derived proper function—derived from the proper function of the higher-level entities that the concept composes—we can get a sense of a concept's proper function through the contexts in which the concept is used. Hence perhaps Carnap's suggestion: a good way of getting a sense of a concept is to enumerate relevant contexts where it is mobilized; and a good way of comparing two concepts that are suspected to be similar is to test whether they could replace one another in these contexts.

Another way of framing it is to consider a community where two concepts are both used. Let us assume we have access to all the linguistic discourses they produced (we can also assume this corpus to be indefinitely large), and know in each instance which concept was used. Then, we could compare two concepts by comparing their contexts—that is, the higher-level entities in which they were involved—and the roles played within those contexts. The concepts would be similar insofar as they play the same roles as assigned by the same discursive forms, and insofar as they are put in the same relationship with other concepts that are mobilized in their respective contexts.

While we lack such capacities as that of collecting every instance of every use of every concept in a community, or that of automatically identifying an instance of a concept to its type in every situation, this idealization does suggest some indicators. Carnap's heuristic of producing assertions is one of them, but it only seems appropriate in cases like explication or ameliorative analysis, where the target concept or the *explicatum* has as of yet no real existence. Producing assertions

can serve as a form of simulation for the sake of predicting how the new concept will function in its new environment. When we have data on how a concept was actually used, as is the case in Alice's scenario, we can use heuristics to identify concept occurrences in the corpus and construct representations of their contexts. These representations can be compared to produce an index of similarity. This is the principle behind paradigmatic relations in distributional semantics (Sahlgren, 2008) and word embeddings in computer sciences (Mikolov and Chen *et al.*, 2013), which have proved very efficient at predicting word similarities and at uncovering semantic relationships between words. Thus, not only is functional similarity measurable in a corpus, but its measure is a well-established practice in corpus linguistics and natural language processing.

As we have seen, a functionalist's similarity avoids the pitfalls that similarity by intension or extension fall into. Furthermore, it explains why Carnap's heuristic for judging of similarity is a good one, and it hints at an explanation for the success of similarity indices based on paradigmatic relations in natural language processing. And as we mentioned, it also comes with an additional perk in that a millikanian framework can afford a clear-cut criterion for a concept instance to belong to a concept type.

In the spirit of flexibility, Carnap did not suggest any way of finding a cut-off point beyond which the *explicatum* is too far from the *explicandum* for there to be a sort of identity between the two. In the context of an explication, it might not be too important of a problem: the idea is that any loss in similarity ought to be offset with gains in fruitfulness and precision, so we don't necessarily need a cut-off point. But in the context of a corpus, it can be important to know where a concept begins and ends, and which contexts mobilize a concept of interest and which only mobilize a similar, but distinct, concept.

Millikan has a proposition for a cutoff point in the context of words (cf. Millikan, 1984: 72-5). Words, in her account, are individualized by reference to their genealogy—they are to be categorized with words that are reproduced from the same lineage of previous uses of the same word type. The verb "to mean" comes from a different etymology as the noun "mean", which entails that they are obviously not the same word. But "to mean" can take a variety of meanings: to convey meaning, to intend, to be important, to be sincere or even to bring about. Until the 19th century, "to mean" as "to be important" was not a standard usage of the verb "to mean". Certainly, in some context, English-speakers would have been able to make sense of a usage of "to mean" in this sense, but they would have understood by inferring from the proximity in meanings and from the context. Therefore, even in the case of an abnormal usage, reproduction of the verb "to mean" would have been driven by the proper function of "to mean" when it means "to convey meaning". When it became standard for "to mean" to mean "to be important", then what was driving this use of "to mean" was not the same proper function—communicative acts were not successful in virtue of an inference from a similar meaning, but rather because audiences were habituated to see "to mean" as meaning "to be important". Therefore, the driving force of the reproduction of "to mean" as meaning "to be important" was now a new proper function. This, to Millikan, is the birth of a new word. This kind of stabilizing proper function is what Millikan calls a *least type*: the narrowest proper function that manages to drive its reproduction. The same reasoning can be applied to concepts in general. Therefore, to determine whether two concepts are identical, one simply needs to determine if their function can be boiled down to the same least type.

To recapitulate, we have argued that similarity between concepts should be understood as similarity by function. In order to clarify this proposition, we have turned to millikanian teleosemantics and its concept of proper function, which con-

tributes a more precise account of function (as proper function) while grounding it into the dynamics of communications within a community. From this basis, we have introduced a concept of CONCEPT using Millikan's ontology, as an analog to her concept of WORD which can play the roles of a concept within the realm of conceptual analysis. As such, this concept articulates its roots in Millikan's system of concepts with the requirements of conceptual analysis, as we have come to understand it in section 1.1. Furthermore, it is haslangerian in the way we defined higher-level entities to include not only the usual descriptive structures like propositions, theories and models, but also non-descriptive linguistic structures and even structures that shape our political interactions like schools and parent-teacher conferences. But it is also carnapian inasmuch as it addresses a problem in carnapian explication.

1.3 Millikanian concepts for corpus-based conceptual analysis

Now, we might wonder if the millikanian framework described above can actually help with corpus-based conceptual analysis.

From a certain perspective, it could appear that this account of functional similarity is bad news for corpus-based conceptual analysis, because functions are not what we observe directly. In practice, we can never be certain that the apparition of a word, for example, has been driven by the reproduction of a certain proper function rather than an another. Furthermore, a function in this sense is harder to express than a subject-predicate association or a subset of an extension.

However, there are reasons to think that concept as function actually makes things easier from Alice's perspective. Firstly, a proper function is not about what caused something, but about what normally causes something in a certain environment under normal, everyday conditions. Therefore, it is not opaque: it could not be

transmitted or replicated if it were private. Furthermore, given that we are talking about communication events, function cannot lie, say, in the emitter alone; the reader, as a participant in the communication, also has in principle a privileged access to the function of its components. Secondly, while there is no straightforward way to represent a function, it might be possible to find proxies for it. For instance, in a corpus, we might expect functions to be associated with distributional patterns.

We can explain and illustrate these points by showing how they can be applied in corpus-based conceptual analysis, and in the task of detecting concepts in particular.

On the one hand, we mentioned in the introduction that computer scientists use human judgments as a way to evaluate and improve the efficiency of an algorithm, and thus to give us confidence in its judgment. In the case of concept detection, there is no material obstacle to asking humans to do exactly the kind of task that we are asking the algorithm to do, and then comparing their answers. Thus, while Alice might not have to detect concept presence by herself if an algorithm does it for her, someone at some point has to be able to make those judgments.

In normal conditions, if there is a shared body of linguistic devices between receivers and emitters, humans usually have an intuitive grasp of when a concept is present in discourse. Thus it makes sense to ask participants "Is concept C present here?" However, there are various shades and variations to this perception: not all concepts feel present in the same way or the same degree. A concept may be present in the theme—for example, it might be the very subject we are discussing about—but it can also play a supporting role in the argumentation, or be vaguely alluded to.

So what task is it exactly that people making judgments about concept presence

are making?

It is relevant here to recall that higher-level entities can only play their linguistic role properly if all of their constituents are taken into account. Concepts, as we have described them above, are constituents of higher-level entities which, in the context of a text corpus, means higher-level discursive entities, such as sentences, of course, but also narratives, arguments, and other higher-level entities that structure text. These higher-level entities are what embody the message—what is being communicated. They are thus what needs to be understood in order for the communication event to be successful. And since they are constituted by concepts, it is necessary (but not sufficient) that the receiver and emitter share a grasp of the concepts that constitute the higher-level entity. In other words, higher-level entities and their message cannot be transmitted without mobilizing their constituting concepts.

Thus, if we are faced with a message, we are faced not only with the concepts explicitly mentioned or set on the centre stage, but also with concepts that play supporting roles, without which the message would be different. It stands to reason that these supporting roles are both essential to the message, and qualitatively different from centre-stage roles. Therefore, if we wish to draw a portrait of how a concept is being used, or of its overall function in language and behaviour in general, our portrait of it should account for supporting roles as much as the more glamorous ones. Therefore, our task when detecting concept use is to get all the concepts constituting the higher-level discourse entities that structure the text, be they centre-stage or not, explicit or not.

In normal conditions, readers should be able to pick out concepts even when they are implicit and play supporting roles, because the understanding of higher-level discourse entities depends on it. However, given how language purposefully

draws our attention to centre-stage concepts, this demands that efforts be made in order to get our attention at the right place. In an annotation protocol, this means coming up with devices to force the annotator to focus on supporting role concepts. For example, Chartrand *et al.* (2017) came up with a two-step annotation process. The first step is meant to build a pool of concepts that can be used in a second step that is designed to limit the bias against supporting role concepts: a text segment is presented to an annotator, and she is tasked with providing five concepts that she deems to be present in the text segment. In the second step, on the one hand, the annotator is given a concept and a text segment, so that she cannot discriminate in favour of centre-stage concepts. On the other hand, annotators are asked for the concept's *degree* of presence—this way, they can express that a concept is not centre-stage without being tempted to express it by marking the concept as absent.

So, this is how our concept of CONCEPT translated into annotation protocols, but can we leverage it for automatic processing? As we have alluded in the previous section, millikanian linguistics offer a natural ground for distributional semantics, which can in turn be used to make indirect representations of concepts' proper functions.

Firstly, it useful to explain two fundamental concepts in distributional semantics: *syntagmatic* and *paradigmatic* relations (cf. Sahlgren, 2008).

When words are *syntagmatically* related, it usually means that they are encountered in the same documents, text segments or sentences—in other words, syntagmatically related words co-occur significantly more than syntagmatically unrelated words: they are often neighbours. This is significant because, while full sentences don't repeat themselves in a corpus, people usually use about the same words to talk about the same things, and discourse about a topic is something

that is typically repeated. Thus, co-occurring words are typically found in texts which are thematically similar (they talk about the same things). Therefore, syntagmatic similarity between two words signals unity in the themes we can express with those two words.

On the other hand, *paradigmatically* related words are words that co-occur with the same words—they have the same neighbours. Typically, paradigmatically related words can play similar roles in the same clauses: replacing one by another might change the meaning, but it will still mean something, and, often, this will form another sentence that is susceptible to be found in the corpus. Paradigmatic relations thus approximate relations of synonymy, with the caveat that antonyms are usually paradigmatically very close, given that they differ only on one dimension and that they will play similar roles in sentences.

Syntagmatic and paradigmatic relations have been leveraged in various way by researchers in natural language processing. Syntagmatic relations are often leveraged in vector representations where words are represented by the documents where they occur, or vice-versa. Through clustering, we can get groups of documents or words that are thematically related, and groups of documents, although topic models are now more commonly used to represent thematic units which are linked to both words and documents. Syntagmatic relations are also used in a large variety of tasks, including automatic summarization, information retrieval (finding a document from a keyword query), recommendation engines, etc. Paradigmatic relations are leveraged to make vector representations where words are represented through other words in terms of their propensity to co-occur, typically within a very short window. On top of finding synonyms, these representations can be used for tasks that involve word composition, for word-sense disambiguation (disambiguating different meanings or sense for a single morpheme), to enhance some topic models (it is particularly useful for inferring topics for short texts, like twit-

ter statuses), for language models (e.g. predicting which will be the next word), etc.

Coming back to millikanian linguistics, higher-level entities tend to reproduce themselves (not as exact copies, as words do, but rather in the same fashion as sentences reproduce themselves), for the same reason any linguistic device does: they successfully serve a purpose in the social and discursive landscape where they are enacted. Now, there are a large variety of factors affecting word use—simply overhearing someone use a word in a conversation nearby certainly makes us more likely to reuse this word. However, *ceteris paribus*, that we are expressing a certain narrative or story, for example, will strongly determine the words we will use to express it. This is partly because the concepts that constitute them condition lexicon by way of favouring words that can be used to express them, but also because playing a certain role in discourse is more readily achieved using some types of words rather than others. For example, while they may be argued for the same conclusions about the same themes, racist discourses from far-right extremists and from mainstream conservative politicians usually will not share the same vocabulary, because they are not staged in the same settings (Van Dijk, 1993).

This association of a recurring vocabulary to recurring higher-level entities could at least partly explain the phenomenon of syntagmatic relations in distributional semantics (Sahlgren, 2008). Two documents, two sentences, or two text segments are similar to the degree that they share the same words. This may very well be because, as higher-level entities condition vocabulary, sharing the same vocabulary indicates a common involvement in expressing the same higher-level discourse entity. Thus, *ceteris paribus*, similar vocabulary means shared involvement in the same higher-level entities.

These syntagmatic relations can then be leveraged by topic models (e.g. Blei *et al.*, 2003) and clustering algorithms (Meunier *et al.*, 2005) to find clusters of textual units that can be read as expressing the same higher-level entity. In other words, higher-level entities are traceable in the text because of the way they condition it and because of the lexical trace that they leave.

Conversely, concepts, words and other linguistic devices that participate in constructing higher-level discourse entities can be described in terms of the other linguistic devices that participate in the same higher-level entities. This is because variations in higher-level entities typically conserve the same discursive forms, and these discursive forms select their constituent by their broad functions. For instance, in "Long live the King!", "King" can easily be replaced with "Queen", as they have similar functions. To a lesser degree, the same can be said of any figure or entity that has a strong authority. Thus, association with "Long live" in a large corpus might indicate that the word "king" and "queen" can fulfill the same function of being the object of approval as an authority. Given that this form can take variations, "king" and "queen" might also be associated with "Down with", which would indicate that they can also both function as objects of disapproval as authority. And so on with other variations, and other discursive forms.

As these cooccurrences accumulate, we can have a decent portrayal of the propensity of two linguistic devices to embody the same functions: this would correspond to the paradigmatic relations in distributional semantics (Sahlgren, 2008). This explains why counting word cooccurrences in large generalist corpora is such a good indicator of synonymy, as well as various semantic properties of representations made this way (cf. Mikolov and Chen *et al.*, 2013; Mikolov and Sutskever *et al.*, 2013; Pennington *et al.*, 2014).

Now, if the repetition of higher-level entities can predict word distributions, then,

conversely, from word distributions, we can infer, at the very least, the probable presence of a higher-level entity. This is essentially how probabilistic topic model work: topics are higher-level structures that are inferred to explain word distributions. And if we assume that topics are composed of concepts, and that their influence over word distributions is a function of the concepts they are composed with, then we can use these word distributions not only to identify which topics are present in a document, but also which concepts compose these topics. This is the hypothesis that is followed in Chartrand *et al.* (2017) and Chartrand (n.d.), using different models of both concepts and topics.

Similarly, paradigmatic relations ensure that at least words can be represented in another way, through their close neighbours, which, as we mentioned, can act as a proxy for function. However, we know that not only words can fruitfully be represented on these vector spaces, but new meaningful vectors can be constructed from word vectors. Furthermore, other entities can be usefully modelled in the same vector space as word embeddings, as shown by the success of algorithms like doc2vec (Lau and Baldwin, 2016), which represents documents as vectors, sense2vec (Trask *et al.*, 2015) which deals with word sense and LCTM (Hu and Tsujii, 2016), which uses a vector model for concepts to construct topics. Therefore, there are good reasons to be optimistic concerning the modellization of paradigmatic relations for concepts using distributional semantics.

Thus, distributional semantics offers two ways of getting at concepts by observing its traces in textual corpora: through the higher-level entities that they construct, and through the modellization of functional similarity through the space of paradigmatic relations.

In this light, it appears that accepting concept similarity by function is actually quite compatible with the main insights of distributional semantics. This,

in turns, opens up the possibility of using technology built on these foundations for detecting concept presence automatically. As Chartrand *et al.* (2016) and Chartrand *et al.* (2017) have shown, we can use syntagmatic relations to find higher-level entities, which will in turn tell us where their constituting concepts are likely to be present. And as Chartrand (n.d.) suggests, recent progresses on paradigmatic relations might enable us to determine which concepts constitute these higher-level entities. Furthermore, as works in computer-assisted reading and conceptual analysis of text suggest (Chartier *et al.*, 2008; Le *et al.*, 2016; Meunier *et al.*, 2005; Sainte-Marie *et al.*, 2011) these techniques can also be leveraged for other aspects of corpus-based conceptual analysis, like for representing certain aspects of the target concept.

1.4 Conclusion

In this article, two main objectives were sought. On the one hand, I pursued the general goal of providing a theoretical framework for corpus-based conceptual analysis. On the other hand, this general, operative objective was pursued through a question: which account of similarity is best adapted to answer the challenges of corpus-based conceptual analysis?

In the first section, I sought to give a general picture of how an empirically-based conceptual analysis might be conceived and theorized. I did this through an overview of the leading accounts of conceptual analyses as method or philosophical endeavour in analytic philosophy—in particular, I reviewed Carnapian explication, Haslangerian analysis, and some perspectives on the method of cases. I noted that the contributions of these accounts are largely complementary, and used this insight to distill these contributions into a general account of corpus-based conceptual analysis as a method.

This first section acts as a sort of introduction for the second section, whereby the hermeneutical resources necessary for formulating the problem of the second section are presented and put together. The second section builds on this, first by formulating its driving question—which account of conceptual similarity is best adapted to corpus-based conceptual analysis? Given how corpus-based conceptual analysis has a foot in conceptual analysis and another in observation of concepts in corpora, this question plays on two very different contexts: the application of the similarity criterion to evaluate the target concept (or *explicatum*, to use Carnap's terminology) on the one hand, and the distinction between the concepts' representation in corpora on the other. In both contexts, I argued that concepts' similarity should be based on resemblance of their proper functions rather than comparison of extensions or intensions. To do so, I went over Millikan's (1984) system of concept and adapted her notion of WORD in order to account for and describe concepts as they are studied by empirically based conceptual analyses.

Finally, in order to illustrate how the framework developed in section 1.2 contributes to the more general objective of providing theoretical grounds for corpus-based conceptual analysis, I showed in section 1.3 how the millikanian framework and concept similarity as function can be leveraged to operationalize concept presence detection. This was done both for concept presence detection as performed by a human and by a computer algorithm.

As such, my contribution is threefold: (1) I provided an argumentation in favour of assessing concept similarity by way of comparing proper functions, (2) I provided a framework that formulates accounts of concept and conceptual analysis for corpus-based conceptual analysis, and (3), I illustrated how this framework is leveraged in concept presence detection.

On the one hand, this provides a theoretical basis that justifies both the way the

concept presence detection problem is formulated in Chartrand *et al.* (2017) and the way annotations were performed. Furthermore, it helps us better formulate the assumptions behind the algorithmic approaches defended by Chartrand *et al.* (2017) and Chartrand (n.d.), and lends some support to it.

While some work has been done to promote corpus-based conceptual analysis (Andow, 2016; Bluhm, 2013), it still represents a new way of approaching philosophical method within analytic philosophy, and, as such, disposes of very few hermeneutical resources to account for itself. One can only hope that the work presented in this paper can contribute to addressing this want—and perhaps inspire further work in this direction.

References

Alexander, J. and Weinberg, J. M. (2007). Analytic Epistemology and Experimental Philosophy. *Philosophy Compass*, *2*(1), 56–80.

Andow, J. (2016). Qualitative tools and experimental philosophy. *Philosophical Psychology*, *29*(8), 1128–1141.

Bealer, G. (1998). Intuition and the Autonomy of Philosophy. In M. DePaul and W. Ramsey (Eds.) (eds.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry* (pp. 201–240). Rowman & Littlefield.

Bealer, G. and Strawson, P. F. (1992). The incoherence of empiricism. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *66*, 99–143. retrieved JSTOR

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bluhm, R. (2013). Don't Ask, Look! Linguistic Corpora as a Tool for Concep-

tual Analysis. In M. Hoeltje, T. Spitzley, and W. Spohn (Eds.) (eds.), *Was dürfen wir glauben?: Was sollen wir tun? Sektionsbeiträge des achten internationalen Kongresses der Gesellschaft für Analytische Philosophie e.V.* (pp. 7–15). DuEPublico.

BonJour, L. (2006). Kornblith on Knowledge and Epistemology. *Philosophical Studies, 127*(2), 317–335.

Braddon-Mitchell, D. (2009). Naturalistic Analysis and the a Priori. In D. Braddon-Mitchell and R. Nola (Eds.) (eds.), *Conceptual Analysis and Philosophical Naturalism.* MIT Press.

Brun, G. (2016). Explication as a Method of Conceptual Re-Engineering. *Erkenntnis, 81*(6), 1211–1241.

Cappelen, H. (2012). *Philosophy without intuitions.* Oxford : Oxford University Press.

Carnap, R. (1928). *Der logische Aufbau der Welt.* Berlin-Schlachtensee : Weltkreis Verlag.

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago : University of Chicago Press.

Carnap, R. (1963). The Philosophy of Rudolf Carnap. In *The Philosophy of Rudolf Carnap* (pp. 859–1013). La Salle : Open Court.

Carus, A. W. (2008). *Carnap and Twentieth-Century Thought: Explication as Enlightenment.* Cambridge : Cambridge University Press.

Chalmers, D. J. (2014). Intuitions in Philosophy: A Minimal Defense. *Philosophical Studies, 171*(3), 535–544.

Chalmers, D. J. and Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review, 110*(3), 315–61.

Chartier, J. F., Meunier, J. G., Danis, J. and Jendoubi, M. (2008). Le travail conceptuel collectif: une analyse assistée par ordinateur du concept d'ACCOMMODEMENT RAISONNABLE dans les journaux québécois. *Actes des JADT 2008*, 297–307.

Chartrand, L. (n.d.). *Mixing syntagmatic and paradigmatic information for concept detection.* Manuscrit soumis pour publication.

Chartrand, L., Cheung, J. C. K. and Bouguessa, M. (2017). Detecting Large Concept Extensions for Conceptual Analysis. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 78–90). Springer, Cham.

Chartrand, L., Meunier, J.-G., Pulizzotto, D., González, J. L., Chartier, J.-F., Le, N. T., ... Amaya, J. T. (2016). CoFiH: A heuristic for concept discovery in computer-assisted conceptual analysis. In *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 1). Nice, France.

Clark, A. (2008). *Supersizing the mind.* (s. l.) : Oxford University Press.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7.

Deutsch, M. E. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method.* (s. l.) : MIT Press.

Dutilh Novaes, C. (2018). Carnapian explication and ameliorative analysis: a systematic comparison. *Synthese*, 1–24.

Dutilh Novaes, C. and Reck, E. (2017). Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization. *Synthese, 194*(1), 195–215.

Evans, V. (2006). Lexical Concepts, Cognitive Models and Meaning-Construction. *Cognitive Linguistics*, *17*(4).

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong.* Oxford : Oxford University Press.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Goldman, A. (2005). Kornblith's Naturalistic Epistemology. *Philosophy and Phenomenological Research*, *71*(2), 403–410.

Goldman, A. I. (1986). *Epistemology and Cognition.* Cambridge, MA : Harvard University Press.

Goldman, A. I. (2007). Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status. *Grazer Philosophische Studien*, *74*(1), 1–26.

Gopnik, A. and Schwitzgebel, E. (1998). Whose Concepts Are They, Anyway?: The Role of Philosophical Intuition in Empirical Psychology. In M. R. DePaul and W. Ramsey (Eds.) (eds.), *Rethinking Intuition* (pp. 75–91). Lanham: Rowman and Littlefield.

Harnad, S. (2009). Concepts: The Very Idea. Canadian Philosophical Association Symposium on Machery on Doing without Concepts. Retrieved {http://eprints.soton.ac.uk/2680

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique.* Oxford : Oxford University Press.

Hu, W. and Tsujii, J. (2016). A latent concept topic model for robust topic inference using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 380–386).

Hurley, S. L. (1998). Vehicles, Contents, Conceptual Structure and Externalism. *Analysis*, *58*(1), 1–6. retrieved Oxford University Press

Ichikawa, J. J. (2009). *Intuitions and Begging the Question.* Manuscrit soumis pour publication.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford : Oxford University Press.

Jackson, M. B. (2013). Conceptual Analysis and Epistemic Progress. *Synthese*, *190*(15), 3053–3074.

King, J. C. (1998). What is a Philosophical Analysis? *Philosophical Studies*, *90*(2), 155–179.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, *63*(3), 190–194.

Knobe, J. (2016). Experimental Philosophy is Cognitive Science. In J. Sytsma and W. Buckwalter (Eds.) (eds.), *A Companion to Experimental Philosophy.* Blackwell.

Koch, S. (2019). Carnapian Explications, Experimental Philosophy, and Fruitful Concepts. *Inquiry: An Interdisciplinary Journal of Philosophy*, *0*(0), 1–18. doi : 10.1080/0020174X.2019.1567381

Kornblith, H. and others. (2002). *Knowledge and its Place in Nature.* Oxford : Oxford University Press.

Lau, J. H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368.*

Le, N. T., Meunier, J.-G., Chartrand, L., Pulizzotto, D., Lopez, J. A., Lareau,

F. and Chartier, J.-F. (2016). Nouvelle méthode d'analyse syntactico-sémantique profonde dans la lecture et l'analyse de textes assistées par ordinateur (LATAO). In *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 2).

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Lewis, D. (1970). How to Define Theoretical Terms. *Journal of Philosophy*, *67*(13), 427–446.

Ludwig, K. (2007). The Epistemology of Thought Experiments : First Person Versus Third Person Approaches. In P. A. French and H. K. Wettstein (Eds.) (eds.), *Midwest Studies in Philosophy* (pp. 128–159). Blackwell.

Machery, E. (2009). *Doing without Concepts.* (s. l.) : Oxford University Press.

Machery, E. (2017). *Philosophy Within its Proper Bounds.* (s. l.) : Oxford University Press.

Machery, E., Mallon, R., Nichols, S. and Stich, S. P. (2004). Semantics, Cross-Cultural Style. *Cognition*, *92*(3), 1–12.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded.* (s. l.) : Palgrave-Macmillan.

Menger, K. (1943). What is dimension? *The American Mathematical Monthly*, *50*(1), 2–7.

Meunier, J.-G. (2017). Theories and Models: Realism and Objectivity in Cognitive Science: Objectivity and Truth in Science. [Theories and Models: Realism and Objectivity in Cognitive Science]. In E. Agazzi (Ed.) (ed.), (pp. 331–352).

Cham : Springer International Publishing.

Meunier, J. G., Biskri, I. and Forest, D. (2005). Classification and categorization in computer assisted reading and analysis of texts. In H. Cohen and C. Lefebvre (Eds.) (eds.), *Handbook of categorization in cognitive science* (pp. 955–978). Elsevier.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism.* (s. l.) : MIT press.

Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More Mama, more milk, and more mouse. *Behavioral and Brain Sciences, 21*(1), 55–65.

Millikan, R. G. (2017). *Beyond Concepts: Unicepts, Language, and Natural Information.* (s. l.) : Oxford University Press.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.) (ed.), *The psychology of computer vision* (pp. 211–277). New-York : McGraw-Hill.

Murphy, G. (2004). *The big book of concepts.* Cambridge, MA : MIT press.

Murphy, T. (2014). Experimental Philosophy: 1935-1965. In T. Lombrozo, J. Knobe, and S. Nichols (Eds.) (eds.), *Oxford Studies in Experimental Philosophy*

(pp. 1–325). Oxford University Press.

Nado, J. (2016). The intuition deniers. *Philosophical Studies, 173*(3), 781–800.

Naess, A. (1938). *"Truth" as Conceived by Those who are Not Professional Philosophers.* Oslo : (n. é.).

Nolan, D. (2009). Platitudes and Metaphysics. In D. Braddon-Mitchell and R. Nola (Eds.) (eds.), *Conceptual Analysis and Philosophical Naturalism.* MIT Press.

Nolfi, K. (2016). *Epistemically Flawless False Beliefs.* Montréal, Québec : Fillosophie. Manuscrit soumis pour publication.

Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Pettit, D. and Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*(5), 586–604.

Pinder, M. (2017). Does Experimental Philosophy Have a Role to Play in Carnapian Explication? *Ratio, 30*(4), 443–461.

Plantinga, A. (1993). *Warrant and Proper Function.* Oxford : Oxford University Press.

Pohlhaus, G. (2015). Different Voices, Perfect Storms, and Asking Grandma What She Thinks: Situating Experimental Philosophy in Relation to Feminist Philosophy. *Feminist Philosophy Quarterly, 1*(1).

Pulizzotto, D., Lopez, J. A., Jean-Guy, J.-F. C., Tan, M. L. C. F. L. and Ngoc, L. (2016). Recherche de «périsegments» dans un contexte d'analyse conceptuelle assistée par ordinateur: le concept d'«esprit» chez Peirce. In *JEP-TALN-RECITAL*

2016 (Vol. 2). Paris.

Quine, W. V. (1971). Epistemology naturalized. *Akten des XIV. Internationalen Kongresses für Philosophie, 6*, 87–103.

Rey, G. (2018). The Analytic/Synthetic Distinction. In E. N. Zalta (Ed.) (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.). Metaphysics Research Lab, Stanford University.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology, 4*(3), 328–350.

Rupert, R. D. (2009). *Cognitive systems and the extended mind.* (s. l.) : Oxford University Press, USA.

Rysiew, P. (2017). Naturalism in Epistemology. In E. N. Zalta (Ed.) (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.). Metaphysics Research Lab, Stanford University.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies, 20*, 33–53.

Sainte-Marie, M. B., Meunier, J.-G., Payette, N. and Chartier, J.-F. (2011). The concept of evolution in the Origin of Species: a computer-assisted analysis. *Literary and linguistic computing, 26*(3), 329–334.

Schwartz, P. H. (1999). Proper Function and Recent Selection. *Philosophy of Science, 66*(3), 210–222.

Searle, J. R. (1958). Proper names. *Mind, 67*(266), 166–173.

Shepherd, J. and Justus, J. (2015). X-Phi and Carnapian Explication. *Erkenntnis, 80*(2), 381–402. retrieved Springer Netherlands

Sosa, E. (2007). Experimental Philosophy and Philosophical Intuition. *Philosophical Studies, 132*(1), 99–107. retrieved Springer

Stalnaker, R. (1976). Propositions. *Issues in the Philosophy of Language*, 79–91.

Swain, S., Alexander, J. and Weinberg, J. (2008). The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp. *Philosophy and Phenomenological Research, 76*(1), 138–155.

Trask, A., Michalak, P. and Liu, J. (2015). sense2vec-A fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388.*

Van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & society, 4*(2), 249–283.

Von Eckardt, B. (1995). *What is cognitive science?* Cambridge, MA : MIT press.

Williamson, T. (2008). *The Philosophy of Philosophy.* Malden, MA : Wiley-Blackwell.

Williamson, T. (2013). How Deep is the Distinction Between A Priori and A Posteriori Knowledge? In A. Casullo and J. C. Thurow (Eds.) (eds.), *The A Priori in Philosophy* (pp. 291–312). Oxford University Press.

Wisdom, J. (2017). Proper Function Moral Realism. *European Journal of Philosophy, 25*(4), 1660–1674.

CHAPTER II

DETECTING LARGE CONCEPT EXTENSIONS FOR CONCEPTUAL ANALYSIS[1]

Mise en contexte

Cet article constitue une première tentative pour développer une chaîne de traitement qui réponde au problème de la détection de concept. Il produit deux contributions importantes pour la thèse. Premièrement, il présente une méthode d'annotation qui découle des considérations théoriques développées dans le premier article, et qui suit le portrait schématique d'une détection de concept par un humain décrite dans la dernière section avant la conclusion. Ce faisant, un corpus annoté de façon à permettre une évaluation de la tâche de la détection de concept est produit. Ce corpus, de nature juridique, a été choisi notamment parce que, bien qu'il soit d'une taille appréciable, il est le produit d'une communauté très réduite de juges qui sont facilement identifiables comme auteur·trices—caractéristique qui n'a pas été exploitée dans cette étude, mais qui pourrait aider à comprendre certaines observations à l'avenir. Par ailleurs, il pourra éventuellement servir à l'étude de questions de philosophie du droit.

---

[1]Co-écrit avec Jackie C. K. Cheung et Mohamed Bouguessa, et reproduit avec l'accord des auteurs. Initialement publié dans P. Perner (éd.), *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings*, Lecture Notes in Computer Science, vol. 10358, Springer, 2017.

Deuxièmement, il propose un certain nombre de chaînes de traitement basées sur un modèle d'allocation Dirichlet latente (LDA) du corpus, et montre que certaines d'entre elles peuvent être efficace – plus efficace, en tout cas, que l'heuristique du mot-clé, qui est encore la norme dans la majorité des analyse conceptuelles dans le domaine des humanités numériques. Ce faisant, il exploite les relations syntagmatiques présentes dans le corpus et décrites dans le premier article comme des moyens de parvenir à un modèle des entités de haut niveau que constituent les concepts dans le discours, mais n'exploite pas les relations paradigmatiques qui ont le potentiel de représenter les relations de similarité entre fonctions des concepts.

Abstract

When performing a conceptual analysis of a concept, philosophers are interested in all forms of expression of a concept in a text—be it direct or indirect, explicit or implicit. In this paper, we experiment with topic-based methods of automating the detection of concept expressions in order to facilitate philosophical conceptual analysis. We propose six methods based on LDA, and evaluate them on a new corpus of court decision that we had annotated by experts and non-experts. Our results indicate that these methods can yield important improvements over the keyword heuristic, which is often used as a concept detection heuristic in many contexts. While more work remains to be done, this indicates that detecting concepts through topics can serve as a general-purpose method for at least some forms of concept expression that are not captured using naive keyword approaches.

2.1 Conceptual Analysis as a Computational Linguistics Problem

Conceptual analysis in philosophy can refer, in a technical sense, to the discovery of *a priori* knowledge in the concepts we share (Chalmers and Jackson, 2001; Jackson, 1998; Laurence and Margolis, 2003). For instance, philosophers will say that "male sibling" is a proper analysis of the concept BROTHER, because it

decomposes its meaning into two other concepts: a brother is nothing more and nothing less than a male sibling. Doing so allows us to make explicit knowledge that is *a priori*, or in other words, knowledge that is not empirical, that can be acquired without observation: for instance, the knowledge that a brother is always a male. In a broader sense, it can refer to the philosophical methods we use to uncover the meaning and use of a concept in order to clarify or improve it (Dutilh Novaes and Reck, 2017; Haslanger, 2012). Given philosophy's focus on conceptual clarity, the latter has been ubiquitous in practice. These methods usually seek to make explicit key features of the concept under scrutiny, in order to construct an account of it; be it a formalized representation that can be expressed in terms of necessary and sufficient conditions, or a more intuitive and pragmatic account of it.

Among the empirical sources upon which conceptual analysis relies, textual data is one of the most important. While armchair philosophy (which relies on thought experiments and intuitions) helps one give a better account of his or her own concepts, contact with texts provides an essential perspective. As a result, philosophers often build *corpora*, i.e. databases of texts that likely use or express a particular concept that is undergoing analysis.

In philosophy as elsewhere, corpora need to be broad enough to cover all the types of usages of the concept under scrutiny, lest the analysis fails to be exhaustive. In other disciplines of social science and humanities, the necessity of grounding analysis in corpora has lead researchers to harness text mining and natural language processing to improve their interpretations of textual data. Philosophy, however, has remained untouched by those developments, save for a few projects (Braddon-Mitchell and Nola, 2009; Meunier *et al.*, 2005).

One important obstacle to the adoption of those methods in philosophy lies in

the lack of proper concept models for conceptual analysis. Keyword approaches to identifying concepts can run into ambiguity problems, like polysemy and synonymy. Furthermore, they can only detect explicit concepts, whereas passages where a concept is latent are bound to also interest the analyst. Latent concept approaches, such as latent semantic analysis (LSA) (Deerwester *et al.*, 1990) or latent Dirichlet allocation (LDA) (Blei *et al.*, 2003), can work to alleviate ambiguity problems and detect latent semantic expressions, but the dimensions they generate ("concepts" in LSA, "topics" in LDA) are thematic, not conceptual. While concepts typically refer to abstract entities or entities in the world, themes, topics and other thematic units are discursive: they can only describe features and regularities in the text.

The problem we address in this paper is that of retrieving textual segments which are relevant to philosophical conceptual analysis. Considering that conceptual analysis is interested in the entire set of textual segments where a queried concept is present in any form, the task at hand is to detect segments whose discourse expresses, implicitly or explicitly, a queried concept. Our concept detection problem distinguishes itself from traditional information retrieval problems in that the aim is to retrieve text segments where the queried concept is *present*, rather than text segments that are *relevant* to the queried concept. In the context of a relevance search, the inquirer will look for the minimum number of documents that can give the maximum amount of generic information about the queried concept; for instance, a web search for "brother" will likely return dictionary definitions and the Wikipedia entry for this word. In the context of a presence query, the inquirer will look for all of the documents where the queried concept is present, thus enabling a more subtle understanding of the concept in all its shades. A search for the presence of the concept BROTHER might thus return texts in genetics or inheritance law as well as implicit evocations of brotherly love in a play. On the

other hand, this concept detection problem also differs from entity recognition or traditional concept mining, as the concept does not need to be associated with a word or an expression. While these problems focus on one particular way a concept can be expressed, conceptual analysis will be interested in any kind of expression of a concept, be it direct or indirect, explicit or implicit.

As such, in section 2.2, we clarify what counts as concept expression for the sake of conceptual analysis, and we distinguish it from other similar notions. In section 2.3, we describe methods to detect a queried concept's expression in textual segments from a corpus. In section 2.4, we present how these methods were implemented and tested, including how an annotated corpus was built, and in section 2.5, results are laid out. Finally, in section 2.6, results are discussed, in a bid to shed light on the underlying assumptions of the methods employed.

2.2 Concept Detection

While conceptual analysis can take many forms, it can always be enhanced by taking empirical data into account. Philosophers who set out to make a concept's meaning explicit through its analysis typically already possess the said concept, and can thus rely on their own intuitions to inform their analysis. However, their analysis can be improved, both in terms of quality and in validity, by being compared with other sources. This explains, for instance, the appeal of experimental philosophy, which has developed in the last 15 years as a way of testing philosophical intuitions using the tools of cognitive and social psychology (Knobe and Nichols, 2007). However, these inquiries have their limits: the intuitions they aim to capture are restricted to a specific time and scope, as they are provoked in an artificial setting. Textual corpora give us the opportunity to study concepts in a more natural setting, and in broader populations, or in populations which are

hard to reach via conventional participant recruitment schemes (experts, authors from past centuries, etc.).

In order to use data from textual corpora, philosophers now have access to the methods and techniques of computer-assisted analysis of textual data. Those methods and techniques are both numerous and diverse, but there are some common characteristics. For instance, they typically involve various steps, which, together, form treatment chains (Fayyad *et al.*, 1996; Meunier *et al.*, 2005): textual data are preprocessed (cleaning, lemmatisation, etc.) and transformed into suitable representations (e.g. vector-space model); then, specific treatment tasks are performed, and finally their output is analysed and interpreted. Furthermore, concepts must be identified in the text, in order to extract their associations to other features that can be found in textual data, such as words, themes and other concepts.

One way of identifying a concept in the text is to identify textual segments in which it is expressed. This expression can take many forms: it can be a word that explicitly refers to a concept in a very wide variety of contexts ("moose" for MOOSE), a description ("massive North American deer"), or embedded in an anaphoric reference ("the animal that crossed the street", "*its* habitat"). It can also be expressed in such a way that it is not tied to any specific linguistic expression. For instance, it can appear in the background knowledge that is essential in understanding a sentence (for instance, in talking about property damage that only a moose could have done), or in relation to the ontological hierarchy (for instance, the concept MOOSE can be expressed when talking about a particular individual moose, or when talking about cervidae).

Our objective here is to test methods of identifying such expressions in textual segments. In other words, our goal is to detect, within a corpus, which passages

are susceptible to inform our understanding of how the concept is expressed in a corpus. As such, concept detection can be seen as a useful step in a wide variety of computer-assisted conceptual analysis methods. For instance, it can act as a way of reducing the study corpus (i.e., the corpus on which a concepual analysis is based) to make it more digestible to a human reader, or it can signal that the semantic content of the segments where the queried concept is detected is likely to be related to the concept, and thus enable new ways of representing it.

Because conceptual analysis is focused on a concept's expression in discourse, concept detection is interested in its presence in discourse. This can mean that the concept is explicitly present, and that it can be matched to a word or an expression, but this presence can also be found in other ways. It can be present in the postulates of the argumentation, without which the passage would be impossible to understand. (For instance, talk of incarceration takes on a very different meaning if we lack the concept of sentencing for a crime). It can be a hypernym to an explicit hyponym, if its properties, expressed to the hyponym, are important enough to the discourse content that we can identify the hypernym as a relevant contributor to the proposition. It can be present in the theme that's being expanded in the passage. It can be referenced using a metaphor or an anaphor. To synthesize, this criterion can be proposed: *a concept is expressed in a textual segment if and only if possession of a concept is necessary to understand the content of the segment.*

Concept detection is similar to other popular problems and projects that have been developed within NLP. However, important distinctions justify our treating it as a different kind of problem.

For instance, concept detection differs from information retrieval (IR) in that presence, rather than relevance, is what we are looking for. For instance, while IR

might be interested in giving priority to text segments where a queried concept is central, this is of little importance to a conceptual analysis, as salient and less salient expressions of a concept are likely to give different yet equally important dimensions of a concept of interest. Conversely, while IR is interested in relevance of a document to a concept even if it is absent, such a rating is meaningless if one is only looking for presence or absence.

It also differs from other tasks which are geared towards presence detection, such as named-entity recognition or coreference resolution. While a concept can be present because a word or expression directly refers to it, it is not absent because no such expression exists in a sentence or another textual unit. In other words, a concept can be present in a text segment even if no single word or expression refers to it. It can be present in virtue of being part of the necessary background knowledge that is retrieved by the reader to make sense of what she or he is reading. Concept detection, as we mean it, should detect both direct and indirect presence of concepts.

2.3 LDA Methods for Detecting Concepts

The presence of a concept as described in section 2.2 can therefore be expressed in various ways: direct explicit reference, anaphorical or metaphorical reference, implicit argumentative or narrative structures, etc. In order to detect these different types of presence, one may expect that we should fragment the task of concept detection into more specific tasks attuned to specific types of presence. In other words, we could detect concept presence by running various algorithms of named-entity recognition or extraction, coreference resolution, topic models, etc. Each of these would detect a specific way in which a concept can become present in a text, and we would rule that a concept is present in a text segment

if it has been detected with any of the methods employed. However, not only is such an approach potentially very time consuming, it makes it very hard to have a constant concept representation: these various algorithms will accept different types of representations of the queried concept, and as such, it will be hard to guarantee that they are all looking for the same concept.

One way around this problem is to hypothesize that while these various expressions of a concept are expressed in different ways, they may be conditioned in similar ways by latent variables. We suppose, in this way, that topics—i.e. underlying discursive and narrative constructions which structure a text, cf. (Blei *et al.*, 2003)—are such latent variables that condition the expression of words and concepts alike. For instance, if the topic "family dinner" is present in a text excerpt, it makes it likely for words such as "table", "mother", "brother" to be present, and unlikely for words such as "clouds" or "mitochondria" to be present; and in a similar fashion, concepts such as FOOD, BROTHER and MOTHER are likely to be expressed and concepts such as ORGANELLE and CLOUD are likely absent.

Therefore, given a concept expressed as a word that is typically associated with it, we can find topics in which it is expressed, and use those topics to find the textual segments where it is likely to be present.

We implement this approach using two different algorithms for learning an LDA model, one that is based on Hoffman's online learning algorithm (Hoffman *et al.*, 2010) and one that is based on Griffiths & Steyvers's Gibbs sampler (Griffiths and Steyvers, 2004).

2.3.1 Online Learning

Hoffman's algorithm (Hoffman *et al.*, 2010) is an online variational Bayes algorithm for the LDA. As such, it relies on the generative model that was introduced by Blei (Blei *et al.*, 2003).

Blei's model uses this generative process, which assumes a corpus D of M documents each of length N_i:

1. Choose $\theta_i \sim \text{Dirichlet}\,(\alpha)$, where

   $i \in \{1 \dots M\}$, the topic distribution for document

   i

2. Choose $\phi_k \sim \text{Dirichlet}\,(\beta)$, where

   $k \in \{1 \dots K\}$, the word distribution for topic k

3. For each of the word positions i, j, where $j \in \{1, \dots, N_i\}$

   , and $i \in \{1, \dots, M\}$:

   1. Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
   2. Choose a word

      $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$.

Here, α and β are parameters of the Dirichlet prior on the per-document topic distributions and on the per-topic word distribution respectively; θ_i is the topic distribution for document i; and ϕ_k is the word distribution for topic k.

Through online stochastic optimization, the online LDA algorithm learns θ (the topic distributions for each document) and ϕ (the word distribution for each topic).

Thus, it is possible to know which topics are likely to be found in each document, and which words are likely to be found for each topic.

Using this information and given a queried concept represented as a word, we can use ϕ to find the topics for which it is among the most important words, relatively, and then use θ to find the documents in which these topics have a non-negligible presence. We thus have a set of documents which are likely to contain the queried concept.

2.3.2 Gibbs Sampling-LDA

While it uses the same LDA model, Griffiths & Steyvers's algorithm (Griffiths and Steyvers, 2004) operates very differently. Rather than estimating θ and ϕ, it learns instead the posterior distribution over the assignments of words to topics $P(z \mid w)$, and it does so with the help of Gibbs sampling, thus assigning topics to each word. After a certain number of sampling iterations (the "burn-in"), these assignments are a good indicator of there being a relationship between word and topic, and between topic and document. From them, we can pick the topics that have been assigned to a given word in its various instanciations, and retrieve the documents to which these topics have been assigned. Furthermore, when necessary, ϕ and θ can be calculated from the assignments.

2.3.3 Concept Presence in Topics

We assume that the presence of a concept in a topic is indicated by the presence of a word typically associated with the concept in question. Therefore a topic's association with a word is indicative of its association with the corresponding concept. The LDA model explicitly links words to topics, but in a graded way: each word is associated with each topic to a certain degree. From this information,

we can use various heuristics to rule whether a concept is involved in a topic or not.

In this study, we tested these heuristics:

Most Likely: The queried concept is associated to the topic which makes its corresponding word most likely to occur.

Highest Rank: The queried concept is associated with the topic in which its corresponding word has the highest rank on the topic's list of most likely words.

Top 30 Rank: The queried concept is associated with the topics in which its corresponding word is among the top 30 words on the topic's list of most likely words.

Concrete Assignment: In the Gibbs Sampling method, individual words are assigned to topics, and word likelihood given a topic is calculated from these assignments. We can thus say that a word is involved in a topic if there is at least one assignment of this topic to this word in the corpus.

Using these heuristics and an LDA model (learned using either Hoffman's or Griffiths & Steyvers's method), we can determine for a given concept the topics in which it is involved.

Depending on the learning method, we can then determine which textual segments are associated to a given topic. On one hand, in Hoffman's method, when a topic is assigned to a segment, there will be a non-zero probability that any given word in the segment is associated with the topic in question. On the other hand, when learning the LDA model using Gibbs Sampling, we'll consider that a topic is

associated to a textual segment if there is at least one word of this segment that is associated with the topic in question.

Thus, from a given concept, we can retrieve the segments in which the concept is likely expressed by retrieving the textual segments that are associated to the topics which are associated to the queried concept.

2.4 Experimentation

2.4.1 Corpus

Algorithms were tested on a French-language corpus of 5,229 decisions from the *Cour d'appel du Québec* (Quebec Court of Appeal), the highest judicial court in Quebec. Much like philosophical discussions, arguments in juridical texts, and in decisions in particular, are well-developed, and nuances are important, so we can expect concepts to be explained thoroughly and employed with precision. However, there is much more homogeneity in style and vocabulary, and this style and vocabulary are more familiar to the broader public than in typical philosophical works, which facilitates annotation. Thus, court decisions are likely to afford complex conceptual analyses, but lack the difficulties that come with the idiosyncrasies of individual philosophical texts.

Court decisions were divided into paragraphs, yielding 198,675 textual segments, which were then broken down into words and lemmatized using TreeTagger (Schmid, 1994). Only verbs, adjectives, nouns and adverbs were kept, and stopwords were removed.

In order to provide a gold standard against which we could evaluate the perfor-

mances of the chosen algorithms, annotations were collected using CrowdFlower[2].

In a first "tagging" step, French-speaking participants were given a textual segment and were instructed to write down five concepts which are expressed in the segment—more specifically, the criterion mentioned in the instructions was that the concept must contribute to the discourse (in French: "*propos*") expressed in the segment. 25 participants annotated 105 segments in this way, yielding 405 segment annotations for a total of 3,240 segment-concept associations.

Data obtained from this first step can tell us that a concept is present in a segment, but we can never infer its absence from it, as its absence from the annotations could simply mean that the annotator chose to write down five other concepts and had no more place for another one. Therefore, it was necessary to add another step to assess absence.

In the second "rating" step, participants were given a segment and six concepts (from the pool of concepts produced in the tagging step), and were instructed to rate each concept's degree of presence or absence from 1 (absent) to 4 (present). The degree of presence is meant to give options to the participant to mark a concept as present, but to a lesser degree, if, say, it is not particularly salient, or if lack of context gives way to some doubt as to whether it really is present. Using this strategy, we can get participants to mark the absence of a concept (degree 1 of the scale) in a way that is intuitive even if one has not properly understood the instructions. For our purposes, we assume that CrowdFlower participants mark a concept as absent when they give it a rating of 1, and as present (even if minimally) if they make any other choice. After removing low-quality annotations, we get 104 segments annotated by 37 participants, for a total of 5,256.

---

[2]`http://www.crowdflower.com`

	Present	Absent
Present	32	2
Absent	24	4

Table 2.1 Contingency table of the CrowdFlower ratings against the legal experts' ratings for the rating step.

In order to ensure that annotations by CrowdFlower participants reflect a genuine understanding of the text, we also recruited legal experts to make similar taggings and judgments and to compare annotations. While the first task was the same for the experts, the second was slightly different in that there were only two options, and in that they were given oral and written instructions to only mark as absent concepts which were definitely absent. This is because the contact we had with these participants made it possible to ensure that instructions were well understood: we did not need to add options to reinforce the idea that a concept is only absent when it is completely and undoubtedly absent. In total, 5 experts tagged 82 text segments in the tagging step, producing a total of 361 tag-segment pairs, and 4 experts rated concepts on 58 segments in the rating step, producing a total of 412 tag-segment pairs.

As table table 2.1 shows, the distribution is skewed towards presence, which makes Cohen's κ a poor choice of metric (Gwet, 2008). Gwet's AC1 coefficient (Gwet, 2008) was used instead, and it revealed that CrowdFlower participants and legal experts have moderate but above-chance agreement, with a coefficient of 0.30 and p-value of less than 0.05 (indicating that there is less than 5% chance that this above-chance agreement is due to random factors).[3] As the confusion matrix of table table 2.1 shows, the error mostly comes from the fact that CrowdFlower

[3]The scenario on the tagging step does not fit any of the common inter-annotator agreement metrics. Firstly, a single item is given five values for the same property. Secondly, in our annotations, absence of annotation does not mean absence of concept; the converse would have been a common assumption in inter-annotator metrics.

participants seem much more likely to mark concepts as absent than legal experts.

2.4.2 Algorithms

Both LDA algorithms were implemented as described in the previous section. For the online LDA, we have used the implementation that is part of Gensim (Řehůřek and Sojka, 2010), and for the Gibbs sampler-LDA, we have adapted and optimized code from Mathieu Blondel (Blondel, 2010). In both cases, we used $k = 150$ topics as parameter, because observing the semantic coherence of the most probable words in each topic (as indicated by ϕ_z) suggests that greater values for k yield topics that seem less coherent and less interpretable overall. For the Gibbs sampler-LDA, we did a burn-in of 150 iterations.

The baseline chosen was the keyword heuristic: a concept is marked as present in a segment if the segment contains the word that represents it, and absent if it does not.

Each method was successively applied to our corpus, using, as queries, items from a set of concept-representing words that were both used in annotations from the rating steps and found in the corpus lexicon. In total, this set numbers 229 concepts for the legal experts' annotations and 808 concepts for the CrowdFlower annotations. Among these, 170 terms are found in both sets of annotations.

2.5 Results

Results from the application of the baseline and our methods on all concepts were compared to the gold standards obtained from the rating step using overall precision, recall, and F1-score. They are illustrated in table 2.2.

Apart from the Gibbs Sampling-LDA/Highest Rank method, all of the proposed

		CrowdFlower			Experts		
		Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
	Keyword	0.03	0.56	0.07	0.01	**1.00**	0.04
Online LDA	**Most Likely**	0.06	0.63	0.13	0.03	0.67	0.07
	Highest Rank	0.07	0.51	0.16	0.03	0.50	0.07
	Top 30 Rank	**0.18**	0.60	**0.32**	**0.15**	0.61	**0.29**
Gibbs Sampling-LDA	**Most Likely**	0.05	0.55	0.12	0.05	0.50	0.13
	Highest Rank	0.00	0.60	0.01	0.03	0.50	0.07
	Top 30 Rank	0.01	0.64	0.03	0.01	0.25	0.04
	Concrete Assign	0.08	**0.65**	0.19	0.12	0.53	0.25

Table 2.2 Performance for each method, calculated using data from the rating task.

methods improved on the baseline, except for the ones using word rankings among the Gibbs Sampling-LDA methods. This is due in particular to improvements in recall. This is to be expected, as the keyword only targets one way in which a concept can be expressed, and thus appears to be overly conservative.

Among the Gibbs Sampling-LDA methods, Concrete Assignment fares significantly better, but the best overall, both in recall and F1-score, is the Online LDA/Top 30 Rank. On this, experts and non-experts are in agreement.

2.6 Discussion

These results seem to validate this study's main hypothesis, that is, LDA methods can improve on the keyword heuristic when it comes to detection of concept expression.

This said, recall remains under 20 %, indicating that topic models are still insufficient to detect all forms of expression of a concept. As such, while it is a clear improvement on the keyword heuristic, it would seem to contradict our hypothesis that topic models can be used to detect all sorts of concept expressions.

	CrowdFlower	Legal experts
Tagging task (step 1)	0.35	0.10
Rating Task (step 2)	0.75	0.24

Table 2.3 Reuse rate in annotation tasks.

2.6.1 Quality of Annotations

While experts' and non-experts' annotations are mostly in agreement, there are important discrepancies. Experts' annotations systematically give better scores to Gibbs Sampling methods, and lower scores to Online LDA methods, than non-experts'. For instance, while the Online LDA/Top 30 Rank method beats the Gibbs Sampling/Assignment method by 0.13 in F1-scores using CrowdFlower annotations, this difference shrinks to 0.04 when using experts' annotations. These discrepancies, however, can be traced to a difference in types of heuristics employed in the tagging step: CrowdFlower participants are more likely to employ words from the excerpt as annotations (i.e. using the concept BROTHER when the word "brother" is present *verbatim* in the text segment), which favors the baseline.

In order to give evidence for this claim, we calculated the propensity of a participant to mark as present a tag that is also a word in the text segment. Specifically, we estimated the reuse rate[4] as depicted by table table 2.3.[5]). As it turns out, in the initial tagging step, CrowdFlower participants are more than three times more likely to write down a word that is present in the text. As participants in

[4]The reuse rate here is simply the number of tags which are a word in the text segment divided by the total number of tags that are words. Multi-word expressions were excluded because detecting whether they are in the text or not would be complicated.

[5]Experts' annotations were ignored because there were too few annotation instances where the queried concept's keyword was in the textual segment, and, as a result, values for the "Keyword in segment" condition were uninformative.

the rating step are only rating tags entered by people of the same group, this translates into a similar ratio in the rating task. However, as it seems that in the rating step, participants are less likely to mark as present a word which is not specifically in the text, reuse rate is inflated for both participant groups. As a result, a large majority of one-word annotations by CrowdFlower participants are already in the text, while the reverse is still true of expert annotations.

Thus, when we discriminate between tags that are present in the textual segment and those that are not, we get a much clearer picture (table table **??**). In the first case, the best heuristic is still the baseline, with Online LDA methods offering much better results than Gibbs Sampling-LDA methods. But in the second, the baseline is unusable, and while F1-scores of Online LDA methods drop by more than half, Gibbs Sampling-LDA methods stay the same or improve. Having fewer annotations where the concept's keyword is in the textual segment will penalize the Online LDA methods, but not the Gibbs Sampling-LDA ones.

As such, this discrepancy should not count as evidence against the hypothesis that CrowdFlower annotations are invalidated by their discrepancies with experts' annotations. However, it suggests that future annotations should control for the ratio of present and absent words in the rating step. Furthermore, it would be useful to test participants of a same group on the same textual segment/concept pair in order to compare in-group inter-annotator agreement with between-group inter-annotator agreement.

	Keyword in segment	**Keyword absent**
Keyword	**0.72**	0.00

Most Likely	0.21	0.13
Highest Rank	0.48	0.14
Top 30 Rank	0.67	**0.30**
Most Likely	0.00	0.12
Highest Rank	0.00	0.01
Top 30 Rank	0.00	0.03
Concrete Assignment	0.21	0.19

Table: F1-scores against CrowdFlower annotations for each method, based on presence or absence of the queried concept keyword in the textual segment. {#tbl:F1-no-cooc}

2.6.2 Improving on Topic Model Methods

In any case, while it does not solve the problem of retrieving all the textual segments where a concept is expressed, the Online LDA/Top 30 Rank method makes important headway towards a more satisfactory solution. It improves on the keyword heuristic's F1-score by 0.25 (both when experts' and non-experts' annotations are used as gold standard), and, as such, constitutes a clear improvement and a much better indicator of concept presence.

Improvements could be reached by associating different approaches to concept detection, when we know that some methods do better than others in specific contexts. For example, the keyword heuristic does slightly better than Online LDA/Top 30 Rank when the queried concept's keyword is present in the text, so it could be used in these situations, while the former method could be used in other cases. In fact, this produces a minor improvement (F1-score of 0.33 with the CrowdFlower gold standard, as compared to 0.32 for pure Online LDA/Top 30 Ranks). We can hope that including other methods for other means of expressing

a concept can contribute to further improvements.

2.7 Conclusion

In this paper, we expressed the problem of concept detection for the purpose of philosophical conceptual analysis, and sought LDA-based methods to address it. In order to evaluate them, we devised an annotation protocol and had experts and non-experts annotate a corpus.

Our results suggest that LDA-based methods and the Online LDA/Top 30 Rank method in particular, can yield important improvements over the keyword heuristic that is currently used as a concept detection heuristic in many contexts. Despite important improvement, it remains a high-precision, low-recall method. However, while more work remains to be done, this indicates that detecting concepts through topics can serve as a general-purpose method for at least some forms of concept expression that are not captured using naive keyword approaches.

As such, we suggest that further research should try to integrate other methods of detecting concept presence in textual data that focus on other means of expressing concepts in texts and discourse.

Acknowledgments.

This work is supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and from the Social Sciences and Humanities Research Council of Canada (SSHRC).

References

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Blondel, M. (2010). Latent Dirichlet Allocation in Python. *Mathieu's log.* Retrieved {http://www.mblondel.org/journal/2010/08/21/latent-dirichlet-allocation-in-python/}

Braddon-Mitchell, D. and Nola, R. (2009). Introducing the Canberra Plan. In D. Braddon-Mitchell and R. Nola (Eds.) (eds.), *Conceptual Analysis and Philosophical Naturalism* (pp. 1–20). MIT Press.

Chalmers, D. J. and Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review*, *110*(3), 315–61.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Dutilh Novaes, C. and Reck, E. (2017). Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization. *Synthese*, *194*(1), 195–215.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psy-*

chology, 61(1), 29–48.

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique.* Oxford : Oxford University Press.

Hoffman, M., Bach, F. R. and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford : Oxford University Press.

Knobe, J. and Nichols, S. (2007). An Experimental Philosophy Manifesto. In J. Knobe and S. Nichols (Eds.) (eds.), *Experimental Philosophy* (pp. 3–14). Oxford University Press.

Laurence, S. and Margolis, E. (2003). Concepts and conceptual analysis. *Philosophy and Phenomenological Research, 67*(2), 253–282.

Meunier, J. G., Biskri, I. and Forest, D. (2005). Classification and categorization in computer assisted reading and analysis of texts. In H. Cohen and C. Lefebvre (Eds.) (eds.), *Handbook of categorization in cognitive science* (pp. 955–978). Elsevier.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (pp. 46–50). Valletta, Malta : University of Malta.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (Vol. 12, pp. 44–49).

Addendum

Erreur concernant la mesure employée

Bien que l'article précédent ait été publié, il comporte une erreur importante, que je me dois ici de corriger.

Pour évaluer l'efficacité des différentes heuristiques employées pour la détection de concepts, on a opté pour le rappel, la précision et le score F1. Ces mesures sont très couramment employée en apprentissage automatique, mais sont toutes sujettes à des biais, surtout en cas de corpus de données mal équilibrées. Le rappel, qui est la proportion des individus recherchés qui sont effectivement retrouvés par l'algorithme, est toujours maximal si l'algorithme retourne toutes les instances sans discrimination. Inversement, la précision, qui est la proportion parmi les individus rappelés d'individus qui étaient recherchés, est maximale si l'algorithme ne retourne qu'une instance et qu'elle s'adonne à être effectivement recherchée. Pour éviter ces deux extrêmes, on combine ces deux mesures dans leur moyenne harmonique, qui est le score F1.

Cependant, dans certains cas, le score F1 peut également être trompeur, en particulier s'il y a un débalancement qui fait en sorte que la majorité des instances sont recherchées. Pour reprendre l'exemple de Chicco (2017), imaginons qu'on a un corpus où 95 segments de texte sur 100 mobilisent un concept C. Alors, si on prend pour algorithme une heuristique qui retourne tous les segments sans discrimination, on a 95 vrais positifs et 5 faux positifs, ce qui nous donne un score F1 de 0.975.

La méthode d'annotation du chapitre précédent visait à ce que, pour la moitié des paires concept-segments qui sont proposées aux annotateur·trices, le concept soit effectivement présent dans le segment. Ce ratio artificiellement élevé (un seg-

		Non-expert·es			Expert·es		
		Rappel	**Précision**	**F1**	**Rappel**	**Précision**	**F1**
	Mot-clé	0.09	0.56	0.07	0.01	1.00	0.04
Online LDA	**Plus probable**	0.06	0.63	0.13	0.03	0.67	0.07
	Rang supérieur	0.18	0.60	0.16	0.03	0.50	0.07
	Top 30	0.05	0.55	0.32	0.15	0.61	0.29
Échantil-lonnage Gibbs	**Plus probable**	0.05	0.55	0.12	0.05	0.50	0.13
	Rang supérieur	0.00	0.60	0.01	0.03	0.50	0.07
	Top 30	0.01	0.64	0.03	0.01	0.25	0.04
	Assignation	0.08	0.65	0.19	0.12	0.53	0.25
	Tout retourner	1.00	0.57	0.73	1.00	0.58	0.74

TABLE 2.4 Précision, rappel des principales heuristiques de l'article "Detecting Large Concept Extensions" en comparaison à l'heuristique "ToutRetourner"

		Non-expert·es MCC	**Expert·es MCC**
	Mot-clé	0.24	0.13
Online LDA	**Plus probable**	0.22	0.14
	Rang supérieur	0.27	0.19
	Top 30	0.37	0.36
Échantillonnage Gibbs	**Plus probable**	0.22	0.22
	Rang supérieur	0.10	0.18
	Top 30	0.05	0.10
	Assignation	0.18	0.30
	Tout retourner	-	-

TABLE 2.5 Corrélations de Mathew (MCC) des principales heuristiques de l'article "Detecting Large Concept Extensions"

ment mobilise un nombre très limité de concepts) visait à éviter que les annotateur·trices, voyant qu'ils marquaient beaucoup de concept comme absent, ne se mettent à s'imaginer des liens un peu trop poussés entre le segment et le concept pour pouvoir annoter le concept comme présent. Cependant, nous n'avons pas pris acte de ce ratio élevé dans l'évaluation des résultats. Aussi, suivant notre méthode d'évaluation, la meilleure heuristique aurait été de tout retourner (cf. tableau 2.4).

Une solution à ce problème est d'employer le coefficient de corrélation de Matthews, qui est reconnu pour sa résilience face aux données déséquilibrées (Boughorbel, Jarray, et El-Anbari 2017).

On voit dans le tableau 2.5 que le calcul de la MCC confirme les principales

conclusions de l'article : (1) que l'on peut, à l'aide de la LDA, obtenir de meilleurs résultats qu'avec l'heuristique du mot-clé, et ce par une marge assez importante (0.13 et 0.23 points comparé aux annotations des non-expert·es et des expert·es respectivement) ; et (2) que la meilleure heuristique est celle du Top 30 avec un modèle appris avec l'inférence variationnelle bayésienne en ligne de Hoffman, Bach, et Blei (2010) ("Online LDA-Top30").

Mesures alternatives

On peut, bien entendu, se demander si d'autres mesures auraient été appropriées. Plusieurs métriques ont été testées avant de porter notre choix sur la MCC, notamment d'autres mesures de rappel d'information comme l'exactitude (*accuracy*) et plusieurs mesures d'accord inter-juge, comme le κ de Cohen, la mesure α de Krippendorf, et la mesure AC1 de Gwet ; cependant, aucune de ces mesures ne résout de façon satisfaisante les problèmes liés aux déséquilibre dans notre corpus.

Cependant, la MCC, la mesure F1 et ces mesures ont en commun d'être des indices qui synthétisent les signaux donnés par des mesures plus simples en un seul chiffre. On peut légitimement se demander s'il ne serait pas préférable d'adopter des mesures plus simples, comme le rappel et la précision, de façon à avoir un portrait plus nuancé des performances de l'algorithme. Ce faisant, on peut adapter notre choix à des exigences particulières venant du contexte d'utilisation de l'algorithme. Par exemple, si l'on sait que l'on a besoin, disons, de rappeler environ le deux tiers des contextes qui nous intéressent, sachant qu'il y a peu de chances que le tiers restant n'apporte beaucoup d'informations pertinentes qui ne sont pas déjà détectable dans les segments rappelés, on peut fixer une limite au rappel de 0.67 et choisir l'algorithme qui obtient la meilleure précision étant donné cette limite.

Le problème, ici, en plus des difficultés qui viennent avec le choix d'une limite

arbitraire, est que nous tentons d'évaluer une tâche qui peut être mobilisée pour plusieurs contextes d'utilisation, et non pas une application particulière. Comme on l'a mentionné dans l'introduction (cf. page 8), la détection de concept ne vise pas une chaîne de traitement particulière, mais plutôt un ensemble très large d'applications ; la lier à un contexte spécifique nuirait à l'objectif général dans lequel elle s'inscrit, c'est-à-dire le développement de méthodes d'analyse conceptuelle basée sur des données textuelles. Pour un tel objectif générique, mieux vaut employer une métrique qui permet de contre-balancer la précision avec le rappel, et vice-versa, de façon à modéliser notre préférence pour une méthode généraliste qui permet d'optimiser l'une et l'autre.

Heuristique de référence

Une heuristique de référence doit correspondre à une heuristique qui est bien connue de la part des gens qui sont susceptibles d'utiliser les méthodes qui lui seront comparées. Il s'agit d'un dénominateur commun, quelque chose à quoi les gens doivent être suffisamment familier pour que les résultats soient facilement compréhensibles et évaluables.

L'heuristique du mot-clé correspond, à notre connaissance, à l'heuristique qui est la plus communément employée par les chercheur·ses en humanités numériques pour retrouver automatiquement un concept dans le texte afin d'en faire l'analyse. C'est ce que l'on trouve dans les études de ce type qui ont été faites en philosophie (McKinnon 1977, 1993 ; Meunier, Biskri, et Forest 2005 ; Chartier et al. 2008 ; Estève 2008 ; Sainte-Marie et al. 2011 ; Danis 2012 ; Sytsma et al. 2019). Un rapide examen de ce genre d'étude dans les actes des Journées d'Analyse statistique des Données Textuelles (JADT) pour 2014 et 2016 semble montrer qu'il ne s'agit pas là d'une exception (Wu 2014 ; Guaresi 2016 ; Venant et Maheux 2016 ; Bendinelli

2016) : à chaque fois, lorsque les chercheu·ses veulent trouver un concept dans le texte, illes choisissent un mot ou une expression, et cherchent ses occurrences.

Bien qu'il s'agisse plus d'une analyse d'archétypes jungiens qu'une analyse d'un concept, Beucher-Marsal et Kerneis (2016) emploient une méthode légèrement différente : ils produisent une liste de synonymes pour des mots correspondants à des archétypes qui les intéressent (*ombre* et *lumière*) et emploient les occurrences pour guider leur interprétation de la sémantique de ces archétypes dans un corpus de chansons. On peut imaginer une procédure similaire pour la détection de concept, où l'on emploierait les occurrences de plusieurs synonymes plutôt qu'un seul mot afin d'identifier la présence d'un concept[6].

Cependant, il y a des raisons de croire que cette méthode n'est peut-être pas désirable dans tous les cas. Ainsi, Sytsma et al. (2019) notent que le vocabulaire de la causalité (l'objet de leur étude) est beaucoup trop vaste pour qu'ils puissent identifier précisément tous les mots et expressions qui expriment la causalité. Par ailleurs, ils notent que chaque nouveau mot vient avec son ambiguité, et peut être employé dans des contextes où il ne sert pas à exprimer la causalité. Aussi, il est probable que les chercheur·ses préfèrent généralement s'en tenir à une seule expression (et ses différentes conjugaisons, déclinaisons et accords) par crainte d'introduire du bruit dans l'analyse.

Néanmoins, il n'est pas clair que ces risques ne soient pas compensés par les gains en rappel qu'apportent cette heuristique des synonymes. Sur notre corpus de la Cour d'appel, j'ai entraîné un ensemble de vecteurs de mots (*word embeddings*) à l'aide de l'algorithme *word2vec* (Mikolov et al. 2013). Les mots qui ont des vecteurs qui ont des distances cosines très réduites sont généralement des synonymes dans le contexte du corpus, de sorte que l'on peut trouver automatiquement des

[6]C'est ce que suggère un réviseur de cette thèse.

synonymes à n'importe quel mot du corpus en prenant les n mots dont les vecteurs sont les plus similaires. Pour chacun des mots ayant servi à l'évaluation des différentes heuristiques de détection de corpus dans "Detecting Large Concept Extensions", on peut donc faire une liste de n synonymes (on prend ici $n = 5$), et produire une liste des segments de texte où au moins un de ces 6 mots est présent. Si l'on emploie cette heuristique avec les mots employés comme annotation[7], on voit que cette heuristique obtient une corrélation de Matthews de 0.23 avec les jugements des non-expert·es (contre 0.24 pour l'heuristique du mot-clé) et de 0.22 avec les jugements des juristes (contre 0.13 pour l'heuristique du mot-clé). Bien qu'on n'observe de changement important du côté des annotations non-expert·es, il y a bien une augmentation appréciable en ce qui concerne les juristes. Néanmoins, dans tous les cas, on est encore loin des succès de l'heuristique "Online LDA-Top 30".

Il y a donc un potentiel pour des recherches futures sur la possibilité d'employer l'heuristique des synonymes pour améliorer la détection de concepts. Cependant, comme il ne s'agit pas d'une procédure communément utilisée par les gens qui font de l'analyse de concepts à partir de corpus textuels, il serait difficile de justifier de l'employer comme heuristique de référence – pour la plupart des gens, ce ne serait pas très parlant. De plus, cette heuristique mérite sa propre étude afin de bien comprendre ses effets sur la détection de concepts (par exemple, pour expliquer le contraste entre annotations expertes et non-expertes, ou pour explorer davantage de valeurs de n). En l'état des choses, il semble donc prématuré de la mettre comme point de comparaison pour les études de la présente thèse.

---

[7]On emploie les annotations du troisième article de la présente thèse, qui incluent davantage d'annotation expertes. Cf. Chartrand (s. d.)

CHAPTER III

MIXING SYNTAGMATIC AND PARADIGMATIC INFORMATION FOR CONCEPT DETECTION[1]

Mise en contexte

Dans ce dernier article est présenté un nouvel ensemble de solutions au problème de la détection du concept. Dans l'article précédent, l'absence de modélisation des concepts dans le modèle topique nous forçait à employer le conditionnement des mots par les topiques comme des indices de la relation de constitution entre concept et topique. Or, le mot est un piètre indicateur du concept. Par ailleurs, ce modèle ne rend pas compte de la nature fonctionnelle du concept – par exemple, des mots ou expressions qui sont utilisés de la même façon dans un corpus expriment vraisemblablement le même concept dans ce contexte.

Afin de palier à ce problème, cet article se penche sur le LCTM (Latent Concept Topic Model) – un modèle topique qui modélise les concepts sur l'espace des enrobages de mots, et les topiques comme constitués par ces concepts. Ce faisant, on a un modèle du concept qui, d'une part, le représente à l'aide d'indicateurs de sa fonction dans le discours, et, d'autre part, le pose explicitement comme un consti-

---

[1] A previous version of this article, which I authored alone, has been submitted to the NAACL 2019 conference.

*tuant des topiques. En plus de produire un modèle plus fidèle aux considérations théoriques exprimées dans le premier article, le présent article produit, d'une part, des heuristiques donnant de meilleures performances, et d'autre part, permettant de formuler le concept à détecter sous la forme d'un vecteur sur l'espace d'enrobages de mots (*word embeddings*). Ce faisant, cet article offre une deuxième réponse à la question de la thèse – comment détecter automatiquement un concept dans le texte – qui est supérieure à la première à la fois en termes théoriques, en termes de performance, et en termes de flexibilité.*

Abstract

In the last decades, philosophers have begun using empirical data for conceptual analysis, but corpus-based conceptual analysis has so far failed to develop, in part because of the absence of reliable methods to automatically detect concepts in textual data. Previous attempts have shown that topic models can constitute efficient concept detection heuristics, but while they leverage the syntagmatic relations in a corpus, they fail to exploit paradigmatic relations, and thus probably fail to model concepts accurately. In this article, we show that using a topic model that models concepts on a space of word embeddings (Hu and Tsujii, 2016) can lead to significant increases in concept detection performance, as well as enable the target concept to be expressed in more flexible ways using word vectors.

3.1 Introduction

Conceptual analysis has, in one form or another, long been a staple of philosophical methodology (Beaney, 2018). When considered as a method, conceptual analysis often requires the input of empirical data, be it in the form of perceptual data, like philosophical intuitions (Pust, 2000), or in other measurable forms. In the last decades, philosophers have begun performing experiments to get a grasp

of human's conceptual behaviour and reactions in order to get a more precise understanding of fundamental concepts like KNOWLEDGE, JUSTICE or RESPONSIBILITY. However, these methods are limited, as controlling for variables forces experimenters to put participants in somewhat unnatural situations and create studies with low ecological validity.

Other voices have suggested that concepts could fruitfully be studied in textual corpora (Andow, 2016; Bluhm, 2013; Chartrand, 2017; Meunier *et al.*, 2005). They argue that methods based on the distributional hypothesis, and that hail from subfields such as natural language processing, text mining and corpus linguistics, could shed light on at least some of the concepts that are objects of philosophical scrutiny.

However, these methods are not well-tuned to philosophical conceptual analysis, as they usually rely heavily on keywords as an indicator of the presence of a concept. While concepts are often associated to words—indeed, even in philosophical discussions, we usually use words as tags for concepts—words are usually associated to more than one concept, and the one that is expressed in any instance is determined by the context. Conversely, concepts can be expressed in a variety of ways. Not only can two or more words or word compounds express the same concept, but concepts don't always attach to words: they can be present in a sentence thematically, or because of an inference that one must make when decoding the sentence. For conceptual analysts who would like their idea of a concept to be representative of all its various uses, this can be an important problem.

The first problem can often be circumvented with a judicious choice of corpus, as controlling the context can often control the sense a word will espouse. As a result, recent work on concept detection has focused on the second problem in a bid to detect where a concept is expressed without the word it is most associated

with (Chartrand *et al.*, 2017, 2016; Pulizzotto *et al.*, 2016). While these efforts have yielded promising results, these methods still rely on the identification of a concept with a word in order both for modelling the constitution of higher-level discursive entities (like topics or narratives) and for representing the queried concept.

This assumption is problematic in at least two ways. On the one hand, as Chartrand (n.d.)[2] argues, these higher-level entities are not composed of words, but of concepts. There is thus the worry that representing them as composed of words makes for an imprecise model. Another concern is with the word-sense ambiguity: a single word may refer to two different concepts, depending on its pragmatic and textual context. One way textual data analysts have dealt with this problem has been to tailor corpora to fit their needs, and choose corpora where the concepts they are interested in happen to be unambiguously associated to a word or a word expression.

Our hypothesis is that these two concerns can be addressed by representing concepts not as words or word expressions, but coordinates on a word embedding space—or, to be more precise, n-dimensional vectors whose semantic properties are determined by their distances to other n-dimensional vectors that represent words ands concepts from a text corpus. In other words: representing concepts as such enable us (1) to make a better model of higher-level entities like topics, which in turn translates in better performances in concept detection and (2) to formulate queries when concepts do not perfectly match with a word or word expression in the corpus.

To test this hypothesis, we employ Hu's and Tsujii's (2016) Latent Concept Topic Model (LCTM) to construct processing chains for concept detection. The LCTM

[2]Chapter 1 of this thesis.

constructs topics as distributions over concepts, which are coordinates in a word embedding space, making for a model that is more theoretically coherent with (Chartrand, n.d.) than previous concept detection models. The processing chains can then be applied to a corpus (in this case, decisions from the Québec Court of Appeal, as in Chartrand *et al.*, 2017) and tested against human annotations.

In section 3.1.1, we review the relevant litterature. We give an overview of the state of the art in topic modelling and word embeddings, and then review hybrid models. In section section 3.2, we formulate the concept detection task and explain how the annotations forming the gold standard are gathered. In section 3.3, we describe the underlying model and functionning of LCTM, and how it links with the theory that underlies the concept detection task. In sections 3.4 and 3.5, the experiments and their application are described, and in section 3.6, we review the results, which are discussed in section section 3.7.

3.1.1 Previous work

From the beginning, topic models have tried to model concepts as an underlying dimension of the text: latent semantic indexing (Deerwester *et al.*, 1990) described documents in terms of latent "concepts". However, by the end of the 90s, Hofmann (1999) described the latent variables in his probabilistic latent semantic indexing as "class variables", and Blei *et al.* (2003) called them "topics" in his latent Dirichlet allocation (LDA) model. Gabrilovich and Markovitch (2007) resurrected the idea of a latent semantic dimension as a concept by forming representations from Wikipedia articles, but their reported success seems to hail from a mere size effect rather than Wikipedia's grouping of discourse under labels (Gottron *et al.*, 2011).

While there are numerous variations, the topic models that are well-known in

the natural language processing community treat topics as a mixture of variables that are of the same kind, that influence word occurrences in the same way, and that thus have *a priori* the same role in shaping discourse. Once learned, they capture the syntagmatic[3] relations between words. Words are syntagmatically close when they participate in the same discourse units—in other words, when they are neighbours, or when they tend to come together.

Word embeddings, on the other hand, tend to capture paradigmatic relations. Words are paradigmatically[3] close when they tend to have the same neighbours. As a result, they tend to have similar roles in discourse, and therefore be synonyms or antonyms[4]. Word embeddings evolved from language models in the early 2000s (Bengio *et al.*, 2003), but were then too computationally expensive to be applied to large corpora. Collobert and Weston (2008), followed by Mikolov and colleagues (Mikolov and Chen *et al.*, 2013; Mikolov and Sutskever *et al.*, 2013), found ways to get the computing cost down, opening the way for word embeddings to become an essential part of the natural language processing toolkit.

Given the popularity of topic models and the word embedding boom that followed word2vec (Mikolov and Chen *et al.*, 2013), it is no wonder that many attempts to combine them have been made. Several of them (Hu and Tsujii, 2016; Le and Lauw, 2017; C. Li *et al.*, 2016; Li *et al.*, 2018; Nguyen *et al.*, 2015; Peng *et al.*, 2018; Wang *et al.*, 2017; Zhang *et al.*, 2019) aim at making topic models that work well with short documents like tweets, where too few words are employed (sparsity problem). Others target the problem of homonymy/polysemy (Law *et al.*, 2017; Liu *et al.*, 2015), seek more interpretable topics (Potapenko *et al.*, 2017;

---

[3]For a more thorough account of syntagmatic and paradigmatic relations in the context of distributional semantics, cf. Sahlgren (2008). Cf. also pages ??-?? of the present thesis.

[4]Antonyms are nearly identical except on one semantic dimension, on which they are opposites. This is why they typically have very similar roles in discourse and sentences.

Zhao *et al.*, 2018), or aim at exploiting complementary representations (Bunk and Krestel, 2018; S. Li *et al.*, 2016; Moody, 2016). Often word embeddings are simply seen as a means to make a more realistic model (Batmanghelich *et al.*, 2016; Das *et al.*, 2015; Hu and Tsujii, 2016; X. Li *et al.*, 2016; Xun *et al.*, 2017).

For example, there is a lineage of models that can be seen as attempts to see how word embeddings fit in the generative story behind probabilistic topic models. Das *et al.*'s Gaussian LDA (2015) replaces the word-over-topic distribution of the LDA with coordinates on the word embedding space. A word's probability given a topic associated with such coordinates are then inferred from the corresponding word embedding's proximity using the Gaussian distribution. Batmanghelich *et al.* (2016) starts from the Gaussian LDA and replaces the Gaussian distribution with the von Mises-Fisher (vMF) distribution, which is a probability distribution over angles centered on a vector. Hu and Tsujii (2016) and X. Li *et al.* (2016) both choose not to identify topics to coordinates or vectors on the word embeddings space, but rather model topics as constituted by such objects. In the former, topics are distributed over these objects (which are called "concepts"), and word probabilities are inferred from concepts using a Gaussian distribution, while the latter identifies topics as complex von Mises-Fisher mixtures over a determined number of bases. Bunk and Krestel (2018) go for a middle-ground position, where words are both influenced by typical LDA-style topics and GLDA-style vector-topics that are situated in the word embeddings space. Perhaps more interestingly, they report no advantage in using mixture models or vMF distributions over simple GLDA-style gaussian distributions, at least in terms of topic coherence and word intrusion tasks.

Perhaps because it is specifically tailored for the needs of philosophical conceptual analysis, few attempts have been made so far at addressing the concept detection task (Chartrand *et al.*, 2017, 2016; Pulizzotto *et al.*, 2016). These papers empha-

size the inadequacy of using keywords to recall text segments where a concept is expressed, but their models still use words as stand-ins for concepts both for articulating queries and for modelling the concept-topic relationship.

This inadequacy calls for alternate models. However, given the large variety of existing topic models, we might not have to create a new one. Chartrand (n.d.) argues that higher-level discourse entities (which can arguably be modelled by LDA-style topics) are constructed from concepts rather than words, topic models that use the word embedding space to model concepts over which topics are distributed. This suggests that topics models where topics are distributed over concepts (rather than words) in the word embedding space (Hu and Tsujii, 2016; X. Li *et al.*, 2016) are more likely to accurately represent topics and their structure. One can hope that such a representation of concepts and their association with topics will yield better results on the concept detection task.

3.2 The concept detection task

As Haslanger (2012) argues, philosophical conceptual analysis can pursue different aims. In some cases, the goal is to represent the concept that we (collectively) have as we possess it. We can call this a *conceptualistic* conceptual analysis. In other cases, the idea is to represent the concept as it functions: this would be a *functionalistic* conceptual analysis. In the case of a concept that represents something, this means that our objective here is to represent the concept so as to reflect its referent rather than our common account of it. For instance, if the function of the concept is to refer to dolphins, then it would not matter if most of us thought of dolphins as fish or if we thought that they have wings: a proper functionalist analysis would still represent dolphins as wingless aquatic mammals. Finally, we make an *ameliorative* conceptual analysis when the goal of the analysis

is to produce a concept that better fulfills the role it plays in discourse, knowledge or society.

This diversity in purpose, however, branches from common grounds. Firstly, there is a sense in which conceptual analysis always is ameliorative, as the representation it aims to make is itself a new concept, meant to play (most often) new roles, if only in philosophical conversations. Secondly, no matter our purpose, it ought to start with an understanding of how this concept functions in its community's discourses and ways of life. Therefore, conceptual analysis demands a thorough picture of a concept's usage, which is where natural language processing can lend a hand.

This thorough picture demands that we be able to capture a concept in as large a variety of uses as possible. While NLP can only be of help when it comes to observing discourse, it is important to include, as much as possible, all ways by which a concept is employed in discourse. As argued by Chartrand (n.d.), much like words bind together to form sentences, concepts bind together to form higher-level entities that are reproduced in a community. These entities can be themes, narratives, arguments, etc. To understand such a higher-level entity, one needs to understand all of its components—therefore, a concept is always present when a topic or a narrative is expressed. However, a concept might not be present in the form of the word that, in proper context, we most readily associate with it (say the word "dolphin" for the concept DOLPHIN). It might present itself in the form of an anaphor ("it", "them"), an hypernym ("the animal") or a description ("these long-nosed swimmers"). It might also be implicitly present within a hidden premise, as part of a piece of information that can be inferred from the text, as part of the background knowledge that we access in order to understand what is being communicated, or even as the object about which we are implicitly talking about. Being present in different ways in discourse often means that a concept is

employed differently and, therefore, has different roles. As a result, it is important that a concept detection algorithm be able to capture the different ways a concept is present in the text.

Concept detection is thus distinct from more traditional information retrieval problems: here, it is not relevance that is sought, but presence. The challenge is not to find the most relevant passages for the expression of a concept, but to find all text segments where is it present. It is also different from such problems as word-sense extraction or ontology learning because concepts need not be associated with words.

3.3 Models

As mentioned in section 3.1.1, there are considerations which lead us to hypothesize that, for concept detection, models that represent concepts in their generative story are more likely to reflect the topic structure in such a way that it can be leveraged for concept detection. Perhaps more interestingly, explicit modelling of concepts (as opposed to simply leveraging word embedding data to direct the learning process) makes it possible to formulate queries using word combinations, which can help disambiguate the query (e.g. $bank - river$ might yield the concept of BANK as this place where we make financial transaction) or make it possible to look for new concepts.[5]

This leaves us with the LCTM (Latent Concept Topic Model, Hu and Tsujii, 2016) and MvTM (Mix von Mises-Fisher Topic Model, X. Li *et al.*, 2016) models. However, the MvTM makes counter-intuitive assumptions concerning the avail-

[5]This is also why topic models based on word embeddings are, in this context, a superior solution to algorithms that use concept databases, such as Tang *et al.* (2018), or algorithms that model concepts simply as latent variables, as El-Arini *et al.* (2012).

ability of concepts for constituting topics. There are two variants to the MvTM: the "disjoint bases" variant (MvTM_d), in which topic mixtures are made from bases that cannot be shared with other topics, and the "overlapping bases" variant (MvTM_o) where topic mixtures are partly made from bases that can be shared with other topics. If bases were meant to model concepts, then we would expect all of them to be shared by many topics. According to the authors, this serves to prevent identical topics from emerging, but language is too fluid to afford such a restriction: no theme, narrative or argument ever has had exclusive rights to a concept. As a result, LCTM seems like the better alternative.

The LCTM is an evolution of the LDA (Blei *et al.*, 2003) and GLDA (Das *et al.*, 2015) models, all three of which are probabilistic graphical models. This is to say that they rely on a generative model, which represents an abstract hypothesis of how a text is constructed and structured.

3.3.1 LDA

In the LDA model, topics are represented by two variables: a multinomial distribution over documents (θ), and multinomial distribution over words (ϕ). These distributions are designed to be sampled from the conjugate Dirichlet priors with parameters α and β respectively. In Blei's (Blei *et al.*, 2003) account, model uses this generative process, which assumes a corpus D of M documents each of length N_i:

- Draw $\theta_i \sim \text{Dirichlet}(\alpha)$, where $i \in \{1 \dots M\}$, the topic distribution for document i
- Draw $\phi_k \sim \text{Dirichlet}(\beta)$, where $k \in \{1 \dots K\}$, the word distribution for topic k
- For each of the word positions i, j, where $j \in \{1, \dots, N_i\}$, and $i \in \{1, \dots, M\}$:

- Draw a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
- Draw a word $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$.

3.3.2 GLDA

With their GLDA model, Das *et al.* (2015) replace ϕ_k with a covariance Σ_k and coordinates to a point that acts as its distribution's mean μ_k. The covariance Σ_k is sampled from an inverse Wishart distribution, and the mean μ_k is sampled from a normal distribution centered at zero (μ). Thus, GLDA's generative story goes like this:

- For each topic $k \in \{1 \dots K\}$
  - Draw a topic covariance $\Sigma_k \sim \mathcal{W}^{-1}(\mathbf{\Psi}, \nu)$
  - Draw a topic mean $\mu_k \sim \mathcal{N}(\mu, \frac{1}{\kappa}\Sigma_k)$
- For each document $i \in \{1 \dots M\}$
  - Draw a topic distribution $\theta_i \sim \text{Dirichlet}(\alpha)$
  - For each word $w \in \{1 \dots N_i\}$
    - Draw a topic $z_w \sim \text{Multinomial}(\theta_i)$
    - Draw a word vector $v_w \sim \mathcal{N}(\mu_{z_w}, \Sigma_{z_w})$ (the chosen word is the one whose word embedding is closest to v_w)

3.3.3 Word embeddings

Word embeddings have developed as a way of representing the semantic information of words in a corpus (paradigmatic relations in particular), and they are employed as such in the LCTM model.

Word embeddings tap in the power of term-term cooccurrence vectors. A term-term cooccurrence matrix is a $N \times N$ matrix M, where N is the number of word types in a corpus, and where the value of each cell $w_{i,j}$ is equal to the number of times the i^{th} and j^{th} cooccur within a window of k words. A cooccurrence vector v_i is the i^{th} row of matrix M and corresponds to the i^{th} word. Cooccurrence vectors whose cosine distance are small are typically semantically close in the sense that they are often synonyms or antonyms. In other words, they are paradigmatically related.

Because term-term cooccurrence vectors tend to be very large, especially in big corpora, there is an incentive to compress them to make them more manageable through dimensionality reduction. Thus, words are associated with a k-dimensional vector, where k is an arbitrary number, typically between 50 and 300.

Dimensionality reduction can be achieved by many means. So-called "count" methods (Baroni *et al.*, 2014; Pennington *et al.*, 2014) use methods such as singular value decomposition and matrix factorization to reduce weighted count vectors (weighting schemes include positive pointwise mutual information and local mutual information). Meanwhile "predict" methods (Bengio *et al.*, 2003; Collobert and Weston, 2008; Mikolov and Chen *et al.*, 2013; Mikolov and Sutskever *et al.*, 2013) set up neural networks that simultaneously learn to predict a word from a small context window[6] and learn vector representations for each word type.

---

[6]Classically, this meant a small window before the target word (Bengio *et al.*, 2003; Collobert and Weston, 2008) (this would be the classic "language model" paradigm), but Mikolov and Chen *et al.* (2013) have introduced the Skip-gram model, where a word is used to predict the words immediately before and after it, within a small window, and the CBOW model, where the context words are used to predict the target word. These models have since then become the norm.

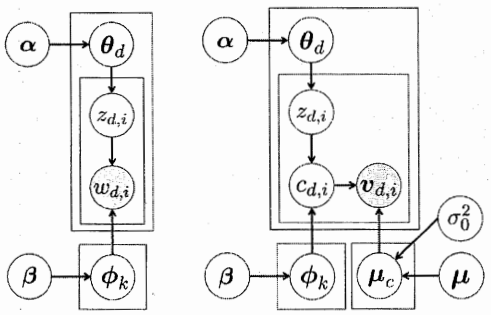
3.3.4 LCTM

With the LCTM, Hu and Tsujii (2016) act on the intuition that topics do not model the same kind of distributional similarity that are modeled with word embeddings. As they note, words that are topically close, like "neural" and "network" in a computer science corpus, will be far away on a word embedding space. This is why they see topics as distributed over other latent variables which they call *concepts*[7], and which are represented by coordinates in the word embedding space. In other words, if a concept is active at a certain point in the text, then the words whose word embeddings are close to the concept's are more likely to appear there.

The LCTM's generative model goes like this, with C being the number of concepts in the model:

- For each topic $k \in \{1 \dots K\}$
  - Draw a topic concept distribution $\phi_k \sim \text{Dirichlet}(\beta)$
- For each concept $c \in \{1 \dots C\}$
  - Draw a concept vector $\mu_c \sim \mathcal{N}(\mu, \sigma_0^2 \boldsymbol{I})$
- For each document $i \in \{1 \dots M\}$
  - Draw a topic distribution $\theta_i \sim \text{Dirichlet}(\alpha)$

[7]While here "concept" is a technical term that refers to features of the LCTM, we believe this use is justified as this concept of CONCEPT can be argued to be an *explication* (in Carnap's (1950) sense) of the concept of CONCEPT that is defended in Chartrand (n.d.). In other words, for the purpose of building an algorithm, it is a more precise, more explicit version of the latter concept, that retains some of its features and enables us to say something about the former. In particular, we assume that instantions of the technical sense of CONCEPT can tell us something about where, in the text corpus, corresponding instantiations of the non-technical sense of CONCEPT are mobilized. For example, if a technical concept, represented as coordinates in a word embedding space, has for corresponding lay concept the concept that we associate with the word "dolphin" (therefore, the lay concept DOLPHIN), then we expect this technical concept to help us determine where the lay concept DOLPHIN is mobilized.



(a) LDA. (b) LCTM.

Figure 3.1 Plate models for LDA and LCTM

- For each word $w \in 1 \ldots N_i$
  - Draw a topic $z_w \sim \text{Multinomial}(\theta_i)$
  - Draw a concept $c_w \sim \text{Multinomial}(\phi_{z_w})$
  - Draw a word vector $v_w \sim \mathcal{N}(\mu_{c_w}, \Sigma_{z_w})$ (the chosen word is the one whose word embedding is closest to v_w)

The graphical model for LDA and LCTM are shown in figure 3.1.

In relation with the problem of word-sense ambiguity mentioned in section section 3.1, it is interesting to note that the same word can be highly likely for different concepts belonging to different topics, which themselves are associated with whole documents. As such, the context of the document determines the topics, and therefore the concept to which a word can be associated. In so doing, the LCTM model can associate different concepts to the same word type depending on the context of the document it is in, and can thus disambiguate between different senses of a word.

3.4 Method

3.4.1 Inference

Given its relative simplicity and its efficiency, Gibbs sampling is by far the most popular method for learning the parameters of topic models that employ word embeddings. LCTM is no exception.

During the inference process, both concept and topic assignments for each word are sampled, using those two equations:

$$p\left(z_w = k \middle| c_w = c, \boldsymbol{z}^{-w}, \boldsymbol{c}^{-w}, \boldsymbol{v}\right) \propto \left(n_{i,k}^{-w} + \alpha_k\right) \cdot \frac{n_{k,c}^{-w} + \beta_c}{n_{k,\cdot}^{-w} + \sum_{c' \in \{1 \ldots C\}} \beta_{c'}} \tag{3.1}$$

$$p\left(c_w = c \middle| z_w = k, \boldsymbol{z}^{-w}, \boldsymbol{c}^{-w}, \boldsymbol{v}\right) \propto \left(n_{c,k}^{-w} + \beta_c\right) \cdot \mathcal{N}\left(v_w \middle| \overline{\mu}_c \sigma_c^2 \boldsymbol{I}\right) \tag{3.2}$$

3.4.2 Concept extension

Once the model has been learned, we have, for each word position, assignment to a concept and a topic, on top of information about its word type and document membership that was provided to the LCTM. Furthermore, we have vectors for each concept and we had provided a vector for each word, all in the same word embeddings space.

Concept detection formally consists in a function that yields a set of documents from a query, which can be either a word type or a vector in the word embeddings space. From the information LCTM produces, there are a number of ways we could make such a function. For instance, we could find the concepts assigned to the query word, and then find all the documents where these concepts are

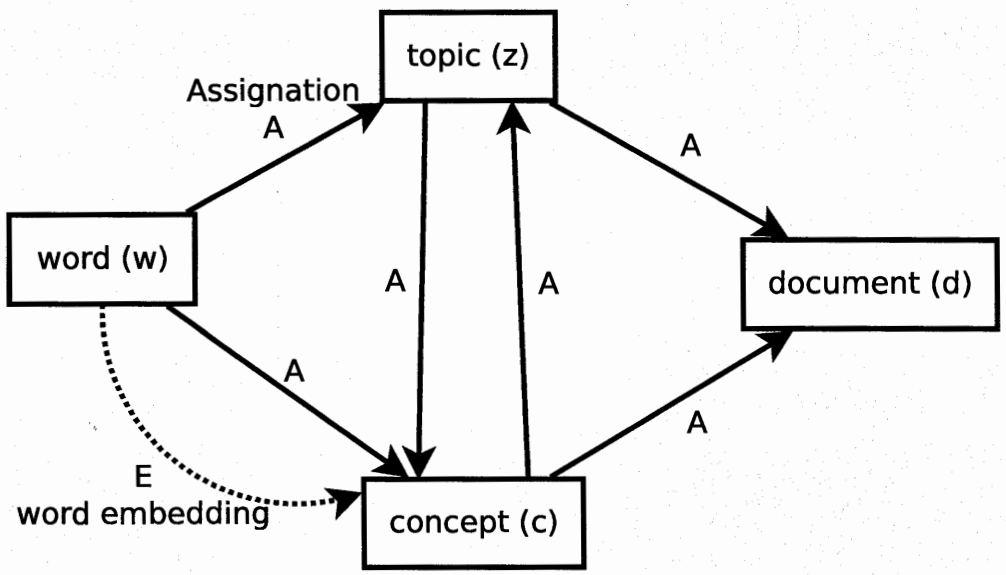


Figure 3.2 Constructing a concept extension chain from an LCTM model.

assigned. But we could also find the word vector for this word, then retrieve the closest concept (in terms of cosine similarity), find the topics where it is assigned, then find the documents where these topics are assigned.

In order to represent the variety of ways a concept extension can be obtained, we use a special notation (cf. figure 3.2). "w", "c", "z" and "d" respectively mean "word", "concept", "topic" and "document". Furthermore, "q" represents a query expressed in the form of a vector. Transitions are noted "E" or "A": "xEy" means "get the y whose vector is closest to x's vector" and "xAy" means "get all the ys which are assigned to a word where x is also assigned". Thus, the first example of the previous paragraph would be noted "wAcAd" and the second example would be "wEcAzAd". Given the nature of concept detection, the first letter is always either "w" or "q", and the last is always "d".

While the number of possible ways we can get a concept extension using data from a LCTM model is potentially infinite, it makes no sense looping over concepts and

topics. Therefore, given a word query, only 8 variations are possible: "wAcAd", "wAcAzAcAd", "wAcAzAd", "wAzAcAd", "wAzAd", "wEcAd", "wEcAzAcAd" and "wEcAzAd". Given a word vector query, only three variations are available: "qEcAd", "qEcAzAcAd" and "qEcAzAd".

3.5 Experimentation

3.5.1 Experiment 1

The first part of our research hypothesis stated that modelling concepts in a topic model would lead to a generally better modelling of the text structure. This, in turn, should lead to better concept detection performance.

To evaluate this proposition, we test all 8 methods for concept detection using LCTM as described in section 3.4. For comparison, we are also testing the "Online LDA-Top 30" and the "Gibbs sampling-Concrete Assignment" heuristics from Chartrand *et al.* (2017)[8], along with the keyword heuristic (recall all text segments where the query word is present). In this evaluation, concept extensions for 754 queries are computed and evaluated against gold standards using the Matthew correlation coefficient (MCC)[9]. All of these queries are formulated as a single word; when computing an extension from a chain that begins with "wE", we employ the word embedding corresponding to the query word.

---

[8]In this heuristic, words are associated to a topic if they are among the 30 words most likely to come up if this topic is activated. From this, we get the concept extension by recalling all the segments or documents in which any of the topics associated with the query word are activated.

[9]Here, IR standard metrics for evaluation like accuracy and F1-measure are not employed because our gold standard represents an unsual set, where annotated concept-segment pairs are much more likely to be positive than randomly chosen concept-segment pairs. Cf. section 3.5.4.

3.5.2 Experiment 2

The second part of our research hypothesis suggested that using a topic model like LCTM, that models concepts on the word embedding space, would allow us to formulate queries for concepts that are not adequately represented with a single word. To test this, we test 588 multiword expressions and represent them on the word embedding space. To do this, we exploit the compositional property of word embeddings, and represent these expressions as the sum of the vectors corresponding to the content words in the expression[10]. Because LCTM makes no assignation for multiword expressions, only the chains built for word vector queries—those beginning with "qE"—are available. They are compared with the keyword heuristic against a gold standard using the MCC.

3.5.3 Corpus & pretreatment

Our corpus is composed of 186,860 segments extracted from 5,229 French-language court decisions of the Quebec Court of Appeal. These decisions where all published between January 1, 2004 and December 31, 2014. Each segment corresponds to a numbered paragraph in this judgements, and each is parsed with a POS tagger[11]. Only verbs, nouns, adverbs and adjectives are kept. Furthermore, judges like to cite law articles, jurisprudence or doctrine in their judgements; these citations have been removed.

Prior to applying the concept detection chains, a word embedding matrix has been

---

[10]Unlike in English, where compound words are created merely by putting the words together, compound words in French often involve prepositions that further constrains how the semantic composition should be interpreted. For simplicity's sake, we ignore this information here.

[11]`http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

learned using gensim's implementation of word2vec[12]. Vector size has been set to 100. The LCTM model was learned using Hu and Tsujii's implementation,[13] with the number of topics being 150 (same as in Chartrand *et al.*, 2017) and the number of concepts being 1,000. Two LDA models were used for comparison: one is learned using Hoffman's Online LDA algorithm (2010), as implemented in gensim[14]; the other is learned using Griffiths' and Steyvers' collapsed Gibbs sampling (Griffiths and Steyvers, 2004), as implemented by Blondel (2010)[15].

3.5.4 Corpus annotations

A subset of our corpus was annotated using the same two-step method presented in Chartrand *et al.* (2017):

1. Annotators were asked to read a text segment from the corpus, and then write down five concepts that were present in it—or, in other words, that contributed to what was being said.
2. Drawing from concepts obtained in step 1, segments were paired with six concepts. Annotators were asked to rate the concept's presence from 1 to 4, 1 being completely absent and 4 being highly present. For our purpose, we consider that a concept is present if the annotator scores more than 1—the scale was employed to avoid a concept to be tagged as absent if it was weakly or very implicitly present. The draw was tweaked so that, on average, annotators would generally be compelled to say that any given concept was

[12]https://radimrehurek.com/gensim/

[13]https://github.com/weihua916/LCTM

[14]https://radimrehurek.com/gensim/

[15]Code available at `https://gist.github.com/mblondel/542786`.

present (2-4) roughly half of the time—this was to ensure that annotators would not be tempted to force a concept upon a segment to compensate for the fact that very little concept were marked as absent.

These annotations were done, on the one hand, by domain experts—jurists—and on the other hand by workers on the crowdsourcing site *Crowdflower* (which rechristened itself *Figure 8* before the end of the study). The 9 expert jurists annotated 103 segments with 1,031 annotations. As for lay workers' annotations, after screening for "spam" annotations (annotations that seemed random or did not seem to reflect an actual understanding), we were left with 3,240 annotations from 25 workers on 105 segments. While there was little overlap in the segment-concept pairs annotated by experts and non-experts, annotation patterns on the 130 pairs where there was overlap reveal that annotations of experts and non-experts correlate moderately ($r_{MCC} = 0.32$). In Experiment 1, performances were evaluated over 871 single-word concept tags used as queries (710 for non-experts and 201 for experts), and in Experiment 2, performances were evaluated over 588 multi-word expressions.

3.6 Results

3.6.1 Experiment 1

Table 3.1: Concept detection performance on single-word queries. Best scores in MCC and precision are emphasized in bold

	Non-experts MCC	**Experts MCC**
Keyword	0.14	0.14
LDA-Top30	0.37	0.34

	Non-experts MCC	Experts MCC
Gibbs LDA	0.22	0.23
wAcAd	0.31	0.30
wAcAzAcAd	0.08	0.12
wAcAzAd	0.38	0.41
wAzAcAd	0.10	0.15
wAzAd	0.45	**0.48**
wEcAd	0.14	0.18
wEcAzAcAd	0.12	0.16
wEcAzAd	**0.46**	0.45

As table 3.1 illustrates, the "wAzAd" and "wEcAzAd" heuristics were the most performant of all, outperforming the leading LDA-Top30 method from Chartrand *et al.* (2017) by scores ranging from 0,07 to 0,14 against non-expert and expert annotations.

Other trends can also be observed. Firstly, chains ending in "cAd" do not fare well: their MCCs against non-expert annotations range from 0.10 to 0.31 (0.11 to 0.30 against expert annotations) while the chains ending with "zAd" have MCCs ranging from 0.38 to 0.46 (0.41 to 0.48 against expert annotations). In fact, only and all zAd-ending chains consistently outperformed LDA and keyword methods. Secondly, shorter chains tend to have better performance than longer chains ($r = -0.48, p < 0.001$ againts non-experts, $r = -0.46, p < 0.001$ against experts).

3.6.2 Experiment 2

Table 3.2: Concept detection performance on compound word queries

	Non-experts MCC	Experts MCC
keyword	0,33	0,20
qEcAd	0,13	0,12
qEcAzAcAd	0,17	0,15
qEcAzAd	**0,46**	**0,46**

As table 3.2 shows, the performances of the "qEcAzAd" hold up with compound words, achieving similar MCCs as with single-word queries. The same can probably be said for the "qEcAd" and "qEcAzAcAd" chains: while the former does worse against expert annotations, and the latter does better against non-expert annotations, it could only be indicative of noise in the data. The scores of the keyword heuristic, on the other hand, have seen a significant uptick, especially against non-expert annotations, where performance has more than doubled.

3.7 Discussion

We made two claims about LCTM: (1) it makes for a better model of higher-level entities like topics, which in turn translates in better performances in concept detection and (2) it allows us to formulate queries when concepts do not perfectly match with a word or word expression in the corpus.

3.7.1 Modelling and concept detection

Concerning the first claim, results from Experiment 1 seem to validate it, at least on the surface, as three of the heuristics managed to provide us with better concept detection performance than what had previously been achieved. These three chains also were more correlated with experts and non-experts than experts and non-experts annotations were correlated with each other.

Moreover, these results also suggest that the relationship between concepts and the textual contexts in which they are present should not be understood as a direct relation between words and concepts, but rather, is mediated by higher-level entites like topics. Indeed, the most successful chains are those that end by connecting those contexts (in our case, the textual segments) to topics (i.e. they end with a "zAd" operation). This confirms theoretical intuitions that we have expressed elsewhere (Chartrand, n.d.; Chartrand *et al.*, 2017, 2016).

Furthermore, it seems like it is the quality of the model that drives the success of the LCTM in comparison with the LDA, because when the same chain is used, LCTM does a lot better. The "Gibbs LDA-Concrete Assignment" employs a "wAzAd" chain, but with a LDA model learned using collapsed Gibbs sampling. Similarly, the LCTM model is learned with an adapted collapsed Gibbs sampler. Therefore, the only difference between those two methods lies in the model, and yet LCTM's "wAzAd" chain does more than twice as good as LDA's.

Other factors are also likely at play—in particular, chain length may explain why some chains are better than others. For instance, some chains seem to be achieving excessive recall ("wAcAzAcAd", "wAzAcAd" and "wEcAzAcAd" in particular). This makes intuitive sense: the "xAy" operations all make it so that for every x, there can be more than one y, as there usually is more than one token of x

in the corpus, and each token of x can be associated with a different type of y. As a result, they end up overgenerating, and thus it is no wonder that they would perform poorly in terms of the MCC. The relative success of "wAcAd" compared with "wEcAd" seems to come from the opposite excess on the part of "wEcAd": given that the "xEy" operation only select one y for every x, "wEcAd" only yields the segment assignations of single LCTM concept, which makes for a very restricted concept extension. On the other hand, with "wAcAd", individual words are likely to be associated with various concepts. As a result, word queries passing through the "wAcAd" chain yield the extension of several concepts that are likely mobilized in the topics which mobilize the queried concept—as such, they approximate the extension of a chain that would use the topic extension like "wAzAd". For the same reason, the "xEy" operation at the beginning of the "wEcAzAcAd" chain might neutralize some of this long-chain effect, which would explain why it does slightly better than "wAcAzAcAd".

3.7.2 Concept detection of multiword expressions

Concerning the second claim, it derives strong evidence from the success of the "qEcAzAd" chain, which does as well on compound words as it did on single words. This sustained performance may be somewhat surprising, given that word embeddings composition is only an approximation of a multi-word expression's meaning (e.g. Salehi *et al.*, 2015). However, single words themselves are often ambiguous (especially when they are not chosen as research term, as is the case here); it is possible that composition aliviates this ambiguity as to counter-balance the imprecision it creates.

The relative success of the keyword heuristic on multiword expression compared to single-word queries might also have to do with ambiguity. In fact, most multiword

expressions encountered among the annotation, like "*arbitre amiable compositeur*" (amiable compositeur arbitrator) and "*témoignage d'expert*" (expert testimony) belong to the technical juridical vocabulary. One can often find a precise definition for it at the beginning of a law or a contract, or a detailed discussion for its interpretation in the doctrine. Because jurists need to mitigate the risk of coming to different interpretations of the same words, it is perhaps more important than elsewhere to have technical concepts that are explicitly linked to a body of text that can be leveraged for interpretation. As a result, jurists have developed an habit of crafting expressions that can be linked to a concept as unambiguously as possible, and which are usually embedded in a set of words that can rarely be seen elsewhere. Not only are these concepts unambiguous, but often, the corresponding concept, being very technical, is also hard to mobilize without using the corresponding expression. The keyword heuristic thus employs expressions that have been refined for better precision and recall—hence its success.

3.7.3 Limitations

One of the motivations for employing LCTM was that it seemed like employing words as a stand-ins for concepts was too indirect a way to identify topics linked to said concept. One might have assumed that translating that query into a vector on a word embedding space would yield better results—but as we saw, one of the leading chains ("wAzAd") doesn't even leverage these representations. This might be because annotators themselves were determining concept presence from a word rather than a more direct expression of a concept. A fair test for determining the best way to formulate a concept query would likely require that annotators be given the task to identify the presence of concepts formulated in other ways than corresponding words or expression.

Another issue is that while testing for multiword expressions might give us a hint as to the capacity of our LCTM chains to detect concepts obtained from composition, it is not the most straightforward test for the success of concept detection. We can expect a conceptual analyst to compose concept representations to disambiguate a concept (e.g. MIND - OPINION to get the concept of MIND without contexts where "mind" is used to mean "opinion", like "in my mind, ...") or to add or remove a dimension of interest to it (e.g. MIND + REASONING to study the mind as a reasoning tool). Using composition in such ways is very different to approximating a multiword expression, as is done in Experiment 2. While its success is a good omen, we need to replicate these results with tasks that are more in line with what conceptual analysts are really likely to do.

On the more technical side, the relative success of online variational Bayes compared to collapsed Gibbs sampling (which had already been established by Chartrand *et al.*, 2017) suggests that LCTM might do even better with a different learning method. As such, it would likely be worthwhile to adapt learning online variational Bayes (Hoffman *et al.*, 2010) or hybrid variational/Gibbs sampling inference (Welling *et al.*, 2012) to the LCTM model in order to learn better models.

3.8 Conclusion

This paper sought to improve on existing concept detection methods by modelling topics in a more theorically appropriate way as constituted of concepts, and by enabling queries formulated in terms of coordinates on the word embedding space. It pursued this objective by constructing processing chains using LCTM models infered from a court decision corpus using the method described by Hu and Tsujii (2016), and evaluated their performance against annotations by legal experts and lay people.

It was successful on both counts. On single-word queries, some of the chains achieved higher performance than the previous leading method, and for reasons that seem to be due to the nature of the LCTM model. Queries formulated as compositions of word embeddings were also tested as approximation of multiword expressions and achieved equally high results, demonstrating that our method can also successfully be used with queries formulated as coordinates on the word embedding space.

References

Andow, J. (2016). Qualitative tools and experimental philosophy. *Philosophical Psychology, 29*(8), 1128–1141.

Baroni, M., Dinu, G. and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 238–247).

Batmanghelich, K., Saeedi, A., Narasimhan, K. and Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. *arXiv preprint arXiv:1604.00126.*

Beaney, M. (2018). Analysis. In E. N. Zalta (Ed.) (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 ed.). Metaphysics Research Lab, Stanford University.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*(Feb), 1137–1155.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research, 3*(Jan), 993–1022.

Blondel, M. (2010). Latent Dirichlet Allocation in Python. *Mathieu's log.* Retrieved {http://www.mblondel.org/journal/2010/08/21/latent-dirichlet-allocation-in-python/}

Bluhm, R. (2013). Don't Ask, Look! Linguistic Corpora as a Tool for Conceptual Analysis. In M. Hoeltje, T. Spitzley, and W. Spohn (Eds.) (eds.), *Was dürfen wir glauben?: Was sollen wir tun? Sektionsbeiträge des achten internationalen Kongresses der Gesellschaft für Analytische Philosophie e.V.* (pp. 7–15). DuEPublico.

Bunk, S. and Krestel, R. (2018). WELDA: Enhancing Topic Models by Incorporating Local Word Context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 293–302).

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago : University of Chicago Press.

Chartrand, L. (2017). La Philosophie Entre Intuition Et Empirie: Comment les Études du Texte Peuvent Contribuer À Renouveler la Réflexion Philosophique. *Artichaud Magazine, 2017*(8 juin).

Chartrand, L. (n.d.). *Similarity in conceptual analysis and concept as proper function.* Manuscrit soumis pour publication.

Chartrand, L., Cheung, J. C. K. and Bouguessa, M. (2017). Detecting Large Concept Extensions for Conceptual Analysis. In *Machine Learning and Data Mining in Pattern Recognition* (pp. 78–90). Springer, Cham.

Chartrand, L., Meunier, J.-G., Pulizzotto, D., González, J. L., Chartier, J.-F., Le, N. T., ... Amaya, J. T. (2016). CoFiH: A heuristic for concept discovery in computer-assisted conceptual analysis. In *JADT 2016 : 13ème Journées interna-*

tionales d'Analyse statistique des Données Textuelles (Vol. 1). Nice, France.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160–167).

Das, R., Zaheer, M. and Dyer, C. (2015). Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 795–804).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

El-Arini, K., Fox, E. B. and Guestrin, C. (2012). Concept modeling with superwords. *arXiv preprint arXiv:1204.2523.*

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007* (pp. 1606–1611).

Gottron, T., Anderka, M. and Stein, B. (2011). Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1961–1964).

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique.* Oxford : Oxford University Press.

Hoffman, M., Bach, F. R. and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems* (pp. 856–864).

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296).

Hu, W. and Tsujii, J. (2016). A latent concept topic model for robust topic inference using word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, pp. 380–386).

Law, J., Zhuo, H. H., He, J. and Rong, E. (2017). LTSG: Latent Topical Skip-Gram for Mutually Learning Topic Model and Vector Representations. *CoRR, abs/1702.07117.*

Le, T. M. V. and Lauw, H. W. (2017). Semantic Visualization for Short Texts with Word Embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (pp. 2074–2080).

Li, C., Wang, H., Zhang, Z., Sun, A. and Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 165–174).

Li, S., Chua, T.-S., Zhu, J. and Miao, C. (2016). Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 666–675).

Li, X., Chi, J., Li, C., Ouyang, J. and Fu, B. (2016). Integrating topic modeling with word embeddings by mixtures of vMFs. In *Proceedings of COLING 2016, the*

26th International Conference on Computational Linguistics: Technical Papers (pp. 151–160).

Li, X., Zhang, A., Li, C., Guo, L., Wang, W. and Ouyang, J. (2018). Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings. *The Computer Journal.*

Liu, Y., Liu, Z., Chua, T.-S. and Sun, M. (2015). Topical Word Embeddings. In *AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2418–2424).

Meunier, J. G., Biskri, I. and Forest, D. (2005). Classification and categorization in computer assisted reading and analysis of texts. In H. Cohen and C. Lefebvre (Eds.) (eds.), *Handbook of categorization in cognitive science* (pp. 955–978). Elsevier.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *CoRR, abs/1605.02019.*

Nguyen, D. Q., Billingsley, R., Du, L. and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, 3,* 299–313.

Peng, M., Xie, Q., Zhang, Y., Wang, H., Zhang, X., Huang, J. and Tian, G. (2018). Neural Sparse Topical Coding. In *Proceedings of the 56th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2332–2340). Melbourne, Australia : Association for Computational Linguistics.

Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Potapenko, A., Popov, A. and Vorontsov, K. (2017). Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. In *Conference on Artificial Intelligence and Natural Language* (pp. 167–180).

Pulizzotto, D., Lopez, J. A., Jean-Guy, J.-F. C., Tan, M. L. C. F. L. and Ngoc, L. (2016). Recherche de «périsegments» dans un contexte d'analyse conceptuelle assistée par ordinateur: le concept d'«esprit» chez Peirce. In *JEP-TALN-RECITAL 2016* (Vol. 2). Paris.

Pust, J. (2000). *Intuitions as Evidence.* New York : Routledge.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33–53.

Salehi, B., Cook, P. and Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 977–983).

Tang, Y.-K., Mao, X.-L., Huang, H., Shi, X. and Wen, G. (2018). Conceptualization topic modeling. *Multimedia Tools and Applications*, *77*(3), 3455–3471.

Wang, B., Liakata, M., Zubiaga, A. and Procter, R. (2017). A Hierarchical Topic Modelling Approach for Tweet Clustering. In *International Conference on Social Informatics* (pp. 378–390).

Welling, M., Teh, Y. W. and Kappen, H. (2012). Hybrid variational/Gibbs collapsed inference in topic models. *arXiv preprint arXiv:1206.3297*.

Xun, G., Li, Y., Zhao, W. X., Gao, J. and Zhang, A. (2017). A correlated topic model using word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4207–4213).

Zhang, X., Feng, R. and Liang, W. (2019). Short Text Topic Model with Word Embeddings and Context Information. In H. Unger, S. Sodsee, and P. Meesad (Eds.) (eds.), *Recent Advances in Information and Communication Technology 2018* (pp. 55–64). Cham : Springer International Publishing.

Zhao, H., Du, L., Buntine, W. and Zhou, M. (2018). Inter and Intra Topic Structure Learning with Word Embeddings. In *International Conference on Machine Learning* (pp. 5887–5896).

CONCLUSION

Cette thèse a tenté de développer un problème de détection de concept dans les données textuelles, et de lui donner une solution.

La première partie de ce travail – développer le problème, et lui donner une interprétation opérationnalisable – est articulée dans le premier article, "Similarity in conceptual analysis and concept as proper function". Cet article prend en charge trois tâches préalables à l'articulation d'une solution au problème de la détection de concept dans le texte : (1) il donne un portrait du contexte dans lequel s'insère cette tâche, soit l'analyse conceptuelle basée sur des données textuelles ; (2) il clarifie le sens de "concept" dans ce contexte ; et (3) il illustre comment une opérationnalisation de ces notions peut être faite pour la tâche de détection de concepts dans le texte. Comme l'article doit aborder ces travaux théoriques tout en produisant une contribution qui ne réponde pas qu'aux besoins de cette thèse, mais aussi à une question qui soit susceptible d'intéresser une partie de la communauté philosophique plus généralement, ces travaux sont articulés autour d'une question théorique : comment doit-on évaluer la similarité entre deux concepts, lorsque l'enjeu en est un d'identité – i.e. dans la perspective de déterminer à quel point deux instanciations de concepts sont les instanciations du même concept ou de concepts différents. Cette question s'ancre notamment dans l'interprétation du critère de similarité chez Carnap.

Afin de donner un portrait de l'analyse conceptuelle basée sur des données textuelles, une première section fait un tour critique des principales traditions en philosophie analytique qui ont abordé l'analyse conceptuelle en tant que méthode

et qui l'ont appliquée : la méthode des cas (Bealer, 1998 ; Machery, 2017 ; Sosa, 2007), l'explication d'inspiration carnapienne (Brun, 2016 ; Carnap, 1950 ; Dutilh Novaes et Reck, 2017) et la typologie des analyses conceptuelles de Sally Haslanger (Haslanger, 2012). Une synthèse est faite, qui note la complémentarité entre ces traditions puis dresse un portrait d'une analyse conceptuelle qui prend acte des progrès de ces différentes traditions.

Dans la deuxième section, l'article s'intéresse au problème de l'interprétation du critère de similarité chez Carnap, et le situe dans l'analyse conceptuelle basée sur des données textuelles. Ce faisant, je produis un argument en faveur d'une conception téléosémantique du concept de CONCEPT, où la fonction (au sens millikanien) est le critère principal par lequel on distingue un concept d'un autre. Ayant argué que le concept millikanien de CONCEPT, notamment tel que développé dans (Millikan, 1984, 1998) et répudié dans (Millikan, 2017), ne correspond pas à un concept de CONCEPT qui serait pertinent pour le type d'analyse conceptuelle qui nous intéresse, je propose un concept de CONCEPT calqué sur le concept millikanien de MOT.

Enfin, la troisième section tente d'illustrer une application de ce concept de CONCEPT dans le cadre de la conception d'analyse conceptuelle qui a été dépeinte, et en particulier pour la détection de concept. Se basant sur la caractérisation du concept faite dans la section précédente, je décris comment celle-ci suggère des façons de détecter des concepts pour un être humain et pour un algorithme informatique. Cela permet de lier les approches qui sont employées dans les articles suivants pour produire des annotations ou pour détecter automatiquement les concepts dans le texte avec la théorie qui est développée dans les sections portant sur l'analyse conceptuelle et la similarité entre concepts.

Les deux chapitres suivants se basent sur cette description pour produire à la

fois des méthodes automatiques pour détecter les concepts dans le texte, et pour évaluer ces méthodes à l'aide d'annotations conçues à cet effet.

La méthode d'annotation est d'abord présentée dans le deuxième article, intitulé "Detecting Large Concept Extensions for Conceptual Analysis", où elle est employée pour produire un ensemble d'annotations sur un corpus de décisions de la Cour d'appel du Québec, lesquelles permettent une évaluation des méthodes automatiques de détection du concept dans le texte. Afin d'éviter les biais de sélection, la production d'un ensemble de concepts et la validation de la présence ou l'absence du concept sont séparés, et la tâche est présentée comme l'évaluation d'un degré de présence afin d'éviter la confusion avec des notions similaires, comme la pertinence.

L'article propose ensuite plusieurs chaînes de traitement qui exploitent la modélisation des topiques par l'analyse Dirichlet latente (LDA). En l'absence de modélisation appropriée du concept, il traite le conditionnement des mots par les topiques comme un indicateur de la relation entre topique et concept. Il démontre qu'avec cette supposition, on peut tout de même beaucoup mieux détecter la présence de concept qu'en ne regardant que les passages où le mot correspondant au concept est explicitement présent.

Face aux limites de cette supposition, le dernier article, intitulé "Mixing syntagmatic and paradigmatic information for concept detection" emploie plutôt un modèle topique qui représente explicitement les concepts comme variables latentes. De surcroît, les concepts sont représentés sur l'espace des enrobages de mots, et donc sur un espace qui leur permet d'être comparés selon la similitudes de leurs fonctions avec les mots du corpus et les autres concepts. D'autre part, les concepts sont présentés comme étant les constituants des topiques, qui sont représentés comme des distributions sur les concepts. À partir de ce modèle, l'article présente sept

chaînes de traitement.

De celles-ci, trois produisent des résultats supérieurs à ceux des heuristiques de l'article précédent, indiquant la supériorité des heuristiques basées sur le LCTM plutôt que la LDA pour la détection de concept. De plus, une d'entre elles permet de formuler le concept recherché sous forme vectorielle, permettant une plus grand flexibilité.

Contributions

Ce faisant, les deux derniers articles offrent une réponse à la question de la thèse en montrant comment on peut détecter automatiquement des concepts dans le texte. En effet, les deux derniers articles ont présenté des façons de détecter le concept qui étaient mieux corrélées avec les jugements humains que l'heuristique du mot-clé, qui est couramment utilisée notamment dans les travaux qui s'inscrivent dans le courant des humanités numériques. Cependant, comme le dernier article parvient à la fois à obtenir une corrélation sensiblement meilleure et à permettre – sans perte de performance – une plus grande flexibilité dans l'articulation du concept ciblé, il doit être considéré comme donnant la meilleure réponse. Cette réponse constitue la principale contribution de la thèse, et on peut espérer qu'elle ouvrira la porte à de nouveaux progrès dans l'analyse conceptuelle basée sur des données textuelles.

Cette thèse produit aussi dans le fil de son argumentation d'autres contributions qui peuvent être mentionnées. On avait noté en introduction qu'il fallait, pour pouvoir aborder le problème de la détection de concept dans le texte, se pencher sur trois questions subsidiaires : (1) Qu'est-ce que l'analyse conceptuelle philosophique basée sur un corpus ? (2) Quel concept de concept rend compte de l'objet de l'analyse conceptuelle philosophique basée sur un corpus ? et (3) Comment peut-on instruire un humain à détecter la présence d'un concept dans des données

textuelles ? Toutes les réponses à ces questions ne constituent pas nécessairement une contribution à la philosophie. La synthèse de l'analyse conceptuelle, même si elle contient des éléments novateurs – en particulier, elle recadre la contribution de Haslanger (2012) comme une contribution critique, et elle pousse plus loin la proposition de convergence entre explication et analyse haslangerienne de Dutilh Novaes (2018) en incluant les analyses descriptivistes/fonctionalistes et conceptualiste – constitue surtout une synthèse et un commentaire sur des cadres conceptuels existants, et produit donc une contribution négligeable.

En revanche, le premier article propose également une nouvelle interprétation de la similarité conceptuelle pour l'explication carnapienne, et avance en réponse à cette nouvelle interprétation un nouveau concept de CONCEPT basé sur la téléosémantique millikanienne. Ce faisant, d'une part, il ouvre la porte à un nouveau champ d'application pour la téléosémantique : celui du discours. Comme le notent Graesser *et al.* (1997), les objets linguistiques d'un ordre de grandeur plus grand que la phrase sont souvent négligés, voire dévalorisés par la linguistique nord-américaine, et Millikan ne fait pas exception. Cependant, comme l'illustre le succès des modèles topiques en traitement automatique du langage naturel, leur pouvoir structurant sur le texte peut être mis à profit pour certaines applications. D'autre part, comme on le montre en répondant à la troisième question subsidiaire, l'extension du cadre millikanien mène à une nouvelle explication de l'hypothèse distributionnelle, qui à son tour permet de comprendre les relations paradigmatiques sous un nouveau regard. La synonymie n'est alors plus une question de similarité des dimensions sémantiques, mais de similarité de la fonction propre, laquelle explique les régularités dans la distribution des termes cooccurrents.

Cette thèse fait également une contribution, d'une part, en formulant le problème de la détection de la présence du concept dans les données textuelles, et d'autre part en présentant une méthode d'annotation pour son évaluation. Comme on

l'a proposé de par le passé (Chartrand *et al.*, 2016), la pauvreté en méthodes pour identifier et décrire les concepts dans le texte constitue un frein important à l'emploi des données textuelles en philosophie et dans les sciences humaines en général. Pour profiter aux sciences humaines, le traitement automatique de la langue naturelle doit pouvoir rendre compte des unités qui structurent le discours, et qui sont couramment mobilisées dans l'interprétation et dans la compréhension de celui-ci. Le développement d'algorithmes flexibles et efficaces pour détecter la présence de concepts pourrait ouvrir la porte à une nouvelle génération d'outils et de méthodes d'assistance à l'analyse et à l'interprétation des données textuelles.

Pour réaliser ces développements, la conception de nouvelles heuristiques de détection des concepts est une contribution importante. Cependant, pour assurer le progrès d'algorithmes de détection des concepts, la communauté scientifique a besoin d'une formulation claire et riche du problème, mais surtout d'outils pour évaluer les réponses qu'on lui propose. En ce sens, la méthode d'annotation et le corpus annoté pourraient s'avérer être des contributions plus durables que les chaînes de détection de concept proposées dans le dernier article.

La thèse amène également une contribution secondaire plus empirique : contrairement à ce que l'on pourrait penser, la présence d'un mot dans un segment de texte est un assez piètre indicateur de la présence d'un concept dans son propos, avec des corrélations rapportées allant de 0.13 à 0.24 pour des concepts exprimés avec un seul mot. Une partie de ces résultats s'explique par le fait qu'un concept peut être présent de plusieurs façons, ce qui cause un taux de rappel très bas. Cependant, lorsque les concepts ne sont pas spécifiquement choisis pour leur univocité, les mots qui leur sont associés tendent à prendre plusieurs sens dans différents contextes. C'est peut-être ce qui explique que la précision de l'heuristique du mot-clé soit suffisamment basse pour qu'une heuristique assez simple basée sur la LDA puisse s'en approcher. Comme on l'a vu dans l'annexe au deuxième

article, il est périlleux de faire une interprétation à partir de la précision et du rappel, étant donné que l'échantillonnage des annotations est biaisé de façon à en assurer la qualité. Néanmoins, ces résultats remettent en question l'idée selon laquelle l'heuristique du mot-clé est une bonne façon de circonscrire l'expression du concept dans les données textuelles.

Horizons

Les contributions de cette thèse pourraient nous amener dans des directions très diverses. En sciences humaines, les méthodes de détection du concept développées ici pourraient aider considérablement l'assistance automatique à la lecture et à l'interprétation des textes. En traitement automatique des langues naturelles, on pourrait espérer que les contributions de cette thèse à la compréhension du concept et de ses rôles dans la structuration du texte puisse mener à des méthodes qui dressent un portrait plus subtil de celui-ci et des différentes façons qu'il est mobilisé dans le langage, et qu'il en sortira des outils robustes et performants pour détecter une grande variété de concepts dans une grande variété de corpus.

Pour son auteur, cette thèse s'inscrit surtout comme un important jalon pour deux projets.

Le premier est le développement d'une méthode d'analyse conceptuelle qui soit conçue particulièrement pour exploiter les données à propos d'un concept qui se trouve dans les corpus de données textuelles. Comme on l'a vu dans l'introduction, une telle méthode pourrait considérer des limites importantes de la philosophie expérimentale. Elle permettrait d'observer les concepts dans leurs environnements naturels, et donnerait aux philosophes les moyens d'explorer les écosystèmes dans lesquels se produit le concept afin de générer de meilleures hypothèses concernant celui-ci. En continuant de cultiver le rapport au corpus et la valorisation des voix

qu'il contient, elle permettrait également de répondre à certains des problèmes systémiques décrits par Pohlhaus (2015).

Les années qui viendront seront cruciales pour ce projet. L'essor de la philosophie expérimentale a ouvert la voix à l'utilisation de données empiriques pour la pratique de la philosophie, de sorte que les philosophes sont plus ouverts à remettre en question leurs méthodes pour faire de la place aux innovations technologiques. En témoigne ceci qu'on commence à voir émerger un peu partout des propositions de nouvelles méthodes d'analyse des données textuelles en philosophie (Alfano et Higgins, s. d. ; Andow, 2016 ; Bluhm, 2013 ; Mejía-Ramos *et al.*, s. d. ; Murdock *et al.*, 2017 ; Sytsma *et al.*, 2019). On peut espérer que cette nouvelle vague d'intérêt pourra contribuer à amener le développement de l'analyse conceptuelle basée sur des données textuelles à sa maturité.

Au niveau informatique, on peut espérer que de nouvelles réponses seront apportées au problème de la détection de concept. Les résultats obtenus ici sont prometteurs et constituent une amélioration importante par rapport aux méthodes utilisées par les chercheur·ses investi·es dans l'analyse conceptuelle basée sur des corpus de données textuelles, mais on ne peut dire qu'ils garantissent que tous les aspects d'un concept présents dans un texte seront visibles à l'analyse, ni exclure que les éléments visibles ne soient attribuables au bruit du corpus. Cependant, Rome ne s'est pas bâti en un jour, et les travaux effectués ici suggèrent de nombreuses pistes de solution qui méritent d'être explorées.

Le deuxième projet serait celui d'une téléosémantique renouvelée. Comme on l'a vu dans le premier article, la téléosémantique millikanienne s'arrête à la phrase, refusant, comme beaucoup de la linguistique nord-américaine, de s'intéresser aux phénomènes linguistiques d'un ordre supérieur à celui de la phrase. Or, il semblerait que ceux-ci aient une part importante dans la structuration du langage

naturel. Par ailleurs, si la notion de fonction propre telle que l'entend Millikan est puissante, elle n'a pas nécessairement suivi les développement dans la compréhension du concept de fonction dans la sélection naturelle. Par exemple, les phénomènes d'évolution convergente et le concept de niche écologique suggèrent que les entités biologiques, les organes mais aussi les espèces puissent avoir une fonction qui soit déterminée par le contexte écologique (cf. par exemple Dussault, s. d.).

On a vu dans cette thèse comment la téléosémantique peut être fructueuse pour l'analyse conceptuelle basée sur des données textuelles. Dans le contexte du problème qu'on a étudié, elle permet, par exemple, de comprendre pourquoi un concept est présent dans l'intégralité des propos où l'entité de haut niveau qu'il contribue à constituer est développée. Elle permet également un arrimage naturel avec la sémantique distributionnelle, dont elle peut expliquer certaines hypothèses. On peut donc espérer qu'une téléosémantique renouvelée donnerait lieu à des idées qui pourraient faire profiter le traitement automatique des langues naturelles et les autres disciplines du langage.

BIBLIOGRAPHIE

Alexander, J. et Weinberg, J. M. (2007). Analytic Epistemology and Experimental Philosophy. *Philosophy Compass*, *2*(1), 56-80.

Alfano, M. et Higgins, A. (s. d.). Natural Language Processing and Semantic Network Visualization for Philosophers. Dans E. Fischer et M. Curtis (dir.), *Methodological Advances in Experimental Philosophy*. Bloomsbury.

Andow, J. (2016). Qualitative tools and experimental philosophy. *Philosophical Psychology*, *29*(8), 1128-1141.

Baroni, M., Dinu, G. et Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Dans *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, p. 238-247).

Batmanghelich, K., Saeedi, A., Narasimhan, K. et Gershman, S. (2016). Nonparametric spherical topic modeling with word embeddings. *arXiv preprint arXiv:1604.00126*.

Bealer, G. (1998). Intuition and the Autonomy of Philosophy. Dans M. DePaul et W. Ramsey (dir.), *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry* (p. 201-240). Rowman & Littlefield.

Bealer, G. et Strawson, P. F. (1992). The incoherence of empiricism. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *66*, 99-143. Récupéré de JSTOR

Beaney, M. (2018). Analysis. Dans E. N. Zalta (dir.), *The Stanford Encyclopedia*

of Philosophy (Summer 2018 éd.). Metaphysics Research Lab, Stanford University.

Bendinelli, M. (2016). Propositions textométriques pour la traduction. Application au concept DELICACY issu de la linguistique systémique fonctionnelle. Dans *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 2). Nice, France.

Bengio, Y., Courville, A. et Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence, 35*(8), 1798-1828.

Bengio, Y., Ducharme, R., Vincent, P. et Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research, 3*(Feb), 1137-1155.

Beucher-Marsal, C. et Kerneis, J. (2016). De l'ombre à la lumière des textes d'Hubert-Félix Thiéfaine : éclairage textométrique de deux archétypes. Dans *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 1). Nice, France.

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of machine Learning research, 3*(Jan), 993-1022.

Blondel, M. (2010). Latent Dirichlet Allocation in Python. *Mathieu's log.* Récupéré de {http://www.mblondel.org/journal/2010/08/21/latent-dirichlet-allocation-in-python/}

Bluhm, R. (2013). Don't Ask, Look! Linguistic Corpora as a Tool for Conceptual Analysis. Dans M. Hoeltje, T. Spitzley, et W. Spohn (Dir.) (dir.), *Was dürfen wir glauben?: Was sollen wir tun? Sektionsbeiträge des achten internationalen Kongresses der Gesellschaft für Analytische Philosophie e.V.* (p. 7-15). DuEPublico.

BonJour, L. (2006). Kornblith on Knowledge and Epistemology. *Philosophical Studies*, *127*(2), 317-335.

Boughorbel, S., Jarray, F. et El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, *12*(6), 1-17. Récupéré de Public Library of Science

Braddon-Mitchell, D. (2009). Naturalistic Analysis and the a Priori. Dans D. Braddon-Mitchell et R. Nola (dir.), *Conceptual Analysis and Philosophical Naturalism.* MIT Press.

Braddon-Mitchell, D. et Nola, R. (2009). Introducing the Canberra Plan. Dans D. Braddon-Mitchell et R. Nola (dir.), *Conceptual Analysis and Philosophical Naturalism* (p. 1-20). MIT Press.

Brun, G. (2016). Explication as a Method of Conceptual Re-Engineering. *Erkenntnis*, *81*(6), 1211-1241.

Bunk, S. et Krestel, R. (2018). WELDA: Enhancing Topic Models by Incorporating Local Word Context. Dans *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (p. 293-302).

Cappelen, H. (2012). *Philosophy without intuitions.* Oxford : Oxford University Press.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering.* Oxford : Oxford University Press.

Carnap, R. (1928). *Der logische Aufbau der Welt.* Berlin-Schlachtensee : Weltkreis Verlag.

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago : University of

Chicago Press.

Carnap, R. (1963). The Philosophy of Rudolf Carnap. Dans *The Philosophy of Rudolf Carnap* (p. 859-1013). La Salle : Open Court.

Carus, A. W. (2008). *Carnap and Twentieth-Century Thought: Explication as Enlightenment.* Cambridge : Cambridge University Press.

Chalmers, D. J. (2014). Intuitions in Philosophy: A Minimal Defense. *Philosophical Studies*, *171*(3), 535-544.

Chalmers, D. J. et Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review*, *110*(3), 315-61.

Chartier, J. F., Meunier, J. G., Danis, J. et Jendoubi, M. (2008). Le travail conceptuel collectif: une analyse assistée par ordinateur du concept d'ACCOMMODEMENT RAISONNABLE dans les journaux québécois. *Actes des JADT 2008*, 297-307.

Chartrand, L. (2017). La Philosophie Entre Intuition Et Empirie: Comment les Études du Texte Peuvent Contribuer À Renouveler la Réflexion Philosophique. *Artichaud Magazine*, *2017*(8 juin).

Chartrand, L. (s. d.-a). *Mixing syntagmatic and paradigmatic information for concept detection.* Manuscrit soumis pour publication.

Chartrand, L. (s. d.-b). *Similarity in conceptual analysis and concept as proper function.* Manuscrit soumis pour publication.

Chartrand, L., Cheung, J. C. K. et Bouguessa, M. (2017). Detecting Large Concept Extensions for Conceptual Analysis. Dans *Machine Learning and Data Mining in Pattern Recognition* (p. 78-90). Springer, Cham.

Chartrand, L., Meunier, J.-G., Pulizzotto, D., González, J. L., Chartier, J.-F.,

Le, N. T., ... Amaya, J. T. (2016). CoFiH: A heuristic for concept discovery in computer-assisted conceptual analysis. Dans *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 1). Nice, France.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, *10*(1), 35. Récupéré de BioMed Central

Clark, A. (2008). *Supersizing the mind.* (s. l.) : Oxford University Press.

Clark, A. et Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7.

Collobert, R. et Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. Dans *Proceedings of the 25th international conference on Machine learning* (p. 160-167).

Danis, J. (2012). *L'analyse conceptuelle de textes assistée par ordinateur (LACTAO): une expérimentation appliquée au concept d'évolution dans l'oeuvre d'Henri Bergson.* (mémoire de master). Université du Québec à Montréal.

Das, R., Zaheer, M. et Dyer, C. (2015). Gaussian lda for topic models with word embeddings. Dans *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1, p. 795-804).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Deutsch, M. E. (2015). *The myth of the intuitive: Experimental philosophy and philosophical method.* (s. l.) : MIT Press.

Dussault, A. C. (s. d.). *Functional Biodiversity, Context-Based Ecological Func-*

tions and the Function/Mere Effect Distinction. Manuscrit soumis pour publication.

Dutilh Novaes, C. (2018). Carnapian explication and ameliorative analysis: a systematic comparison. *Synthese*, 1-24.

Dutilh Novaes, C. et Reck, E. (2017). Carnapian Explication, Formalisms as Cognitive Tools, and the Paradox of Adequate Formalization. *Synthese*, *194*(1), 195-215.

Egré, P. (s. d.). Intentional Action and the Semantics of Gradable Expressions (On the Knobe Effect). Dans B. Copley et F. Martin (Dir.) (dir.), *Causation in Grammatical Structures.* Oxford University Press.

El-Arini, K., Fox, E. B. et Guestrin, C. (2012). Concept modeling with superwords. *arXiv preprint arXiv:1204.2523.*

Estève, R. (2008). Une analyse quantitative de la morale chez Vladimir Jankélévitch. Dans *JADT 2008 : 9ème Journées internationales d'Analyse statistique des Données Textuelles.* Lyon, France.

Evans, V. (2006). Lexical Concepts, Cognitive Models and Meaning-Construction. *Cognitive Linguistics*, *17*(4).

Fayyad, U., Piatetsky-Shapiro, G. et Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong.* Oxford : Oxford University Press.

Forest, D. (2002). *Lecture et analyse de textes philosophiques assistées par ordinateur : application d'une approche classificatoire mathématique à l'analyse thé-*

matique du Discours de la méthode et des Méditations métaphysiques de Descartes. (mémoire de master). Université du Québec à Montréal.

Gabrilovich, E. et Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. Dans *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007* (p. 1606-1611).

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121-123.

Goldman, A. (2005). Kornblith's Naturalistic Epistemology. *Philosophy and Phenomenological Research*, *71*(2), 403-410.

Goldman, A. I. (1986). *Epistemology and Cognition.* Cambridge, MA : Harvard University Press.

Goldman, A. I. (2007). Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status. *Grazer Philosophische Studien*, *74*(1), 1-26.

Gopnik, A. et Schwitzgebel, E. (1998). Whose Concepts Are They, Anyway?: The Role of Philosophical Intuition in Empirical Psychology. Dans M. R. DePaul et W. Ramsey (dir.), *Rethinking Intuition* (p. 75-91). Lanham: Rowman and Littlefield.

Gottron, T., Anderka, M. et Stein, B. (2011). Insights into explicit semantic analysis. Dans *Proceedings of the 20th ACM international conference on Information and knowledge management* (p. 1961-1964).

Graesser, A. C., Millis, K. K. et Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology*, *48*(1), 163-189.

Griffiths, T. L. et Steyvers, M. (2004). Finding scientific topics. *Proceedings of*

the National academy of Sciences, *101*(suppl 1), 5228-5235.

Guaresi, M. (2016). Cooccurrences, contrastes et caractérisation textuels. Applications à un corpus de professions de foi électorales (1958 – 2007). Dans *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 2). Nice, France.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48.

Hannon, M. (2015). The Universal Core of Knowledge. *Synthese*, *192*(3), 769-786.

Harnad, S. (2009). Concepts: The Very Idea. Canadian Philosophical Association Symposium on Machery on Doing without Concepts. Récupéré de {http://eprints.soton.ac.uk/268029/}

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique.* Oxford : Oxford University Press.

Hoffman, M., Bach, F. R. et Blei, D. M. (2010). Online learning for latent dirichlet allocation. Dans *advances in neural information processing systems* (p. 856-864).

Hofmann, T. (1999). Probabilistic latent semantic analysis. Dans *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (p. 289-296).

Hu, W. et Tsujii, J. (2016). A latent concept topic model for robust topic inference using word embeddings. Dans *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Vol. 2, p. 380-386).

Hurley, S. L. (1998). Vehicles, Contents, Conceptual Structure and Externalism.

Analysis, *58*(1), 1-6. Récupéré de Oxford University Press

Ichikawa, J. J. (2009). *Intuitions and Begging the Question.* Manuscrit soumis pour publication.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* Oxford : Oxford University Press.

Jackson, M. B. (2013). Conceptual Analysis and Epistemic Progress. *Synthese*, *190*(15), 3053-3074.

King, J. C. (1998). What is a Philosophical Analysis? *Philosophical Studies*, *90*(2), 155-179.

Kneer, M. (2018). Perspective and Epistemic State Ascriptions. *Review of Philosophy and Psychology*, *9*(2), 313-341.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, *63*(3), 190-194.

Knobe, J. (2016). Experimental Philosophy is Cognitive Science. Dans J. Sytsma et W. Buckwalter (dir.), *A Companion to Experimental Philosophy.* Blackwell.

Knobe, J. et Nichols, S. (2007). An Experimental Philosophy Manifesto. Dans J. Knobe et S. Nichols (dir.), *Experimental Philosophy* (p. 3-14). Oxford University Press.

Koch, S. (2019). Carnapian Explications, Experimental Philosophy, and Fruitful Concepts. *Inquiry: An Interdisciplinary Journal of Philosophy*, *0*(0), 1-18. doi : 10.1080/0020174X.2019.1567381

Koller, D. et Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques.* [Probabilistic Graphical Models]. Cambridge, MA : MIT Press.

Kornblith, H. et others. (2002). *Knowledge and its Place in Nature.* Oxford : Oxford University Press.

Lau, J. H. et Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368.*

Laurence, S. et Margolis, E. (2003). Concepts and conceptual analysis. *Philosophy and Phenomenological Research, 67*(2), 253-282.

Law, J., Zhuo, H. H., He, J. et Rong, E. (2017). LTSG: Latent Topical Skip-Gram for Mutually Learning Topic Model and Vector Representations. *CoRR, abs/1702.07117.*

Le, N. T., Meunier, J.-G., Chartrand, L., Pulizzotto, D., Lopez, J. A., Lareau, F. et Chartier, J.-F. (2016). Nouvelle méthode d'analyse syntactico-sémantique profonde dans la lecture et l'analyse de textes assistées par ordinateur (LATAO). Dans *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 2).

Le, T. M. V. et Lauw, H. W. (2017). Semantic Visualization for Short Texts with Word Embeddings. Dans *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (p. 2074-2080).

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics, 20*(1), 1-31.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady, 10,* 707.

Lewis, D. (1970). How to Define Theoretical Terms. *Journal of Philosophy, 67*(13), 427-446.

Li, C., Wang, H., Zhang, Z., Sun, A. et Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. Dans *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (p. 165-174).

Li, S., Chua, T.-S., Zhu, J. et Miao, C. (2016). Generative topic embedding: a continuous representation of documents. Dans *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, p. 666-675).

Li, X., Chi, J., Li, C., Ouyang, J. et Fu, B. (2016). Integrating topic modeling with word embeddings by mixtures of vMFs. Dans *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (p. 151-160).

Li, X., Zhang, A., Li, C., Guo, L., Wang, W. et Ouyang, J. (2018). Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings. *The Computer Journal.*

Liu, Y., Liu, Z., Chua, T.-S. et Sun, M. (2015). Topical Word Embeddings. Dans *AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (p. 2418-2424).

Ludwig, K. (2007). The Epistemology of Thought Experiments : First Person Versus Third Person Approaches. Dans P. A. French et H. K. Wettstein (dir.), *Midwest Studies in Philosophy* (p. 128-159). Blackwell.

Machery, E. (2009). *Doing without Concepts.* (s. l.) : Oxford University Press.

Machery, E. (2017). *Philosophy Within its Proper Bounds.* (s. l.) : Oxford University Press.

Machery, E., Mallon, R., Nichols, S. et Stich, S. P. (2004). Semantics, Cross-

Cultural Style. *Cognition*, *92*(3), 1-12.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., ... Hashimoto, T. (2017). Gettier Across Cultures. *Noûs*, *51*(3), 645-664.

McCann, H. J. (2005). Intentional Action and Intending: Recent Empirical Studies. *Philosophical Psychology*, *18*(6), 737-748.

McKinnon, A. (1973). The conquest of fate in Kierkegaard. *CIRPHO*, *1*(1), 45-58.

McKinnon, A. (1977). From co-occurrences to concepts. *Computers and the Humanities*, *11*(3), 147-155. Récupéré de Springer

McKinnon, A. (1993). Kierkegaard and" The Leap of Faith". *Kierkegaardiana*, *16*.

Mejía-Ramos, J. P., Alcock, L., Lew, K., Rago, P., Sangwin, C. et Inglis, M. (s. d.). Natural Language Processing and Semantic Network Visualization for Philosophers. Dans E. Fischer et M. Curtis (Dir.) (dir.), *Methodological Advances in Experimental Philosophy*. Bloomsbury.

Menary, R. (2007). *Cognitive Integration: Mind and Cognition Unbounded*. (s. l.) : Palgrave-Macmillan.

Menger, K. (1943). What is dimension? *The American Mathematical Monthly*, *50*(1), 2-7.

Meunier, J.-G. (2017). Theories and Models: Realism and Objectivity in Cognitive Science: Objectivity and Truth in Science. [Theories and Models: Realism and Objectivity in Cognitive Science]. Dans E. Agazzi (dir.), (p. 331-352). Cham : Springer International Publishing.

Meunier, J. G., Biskri, I. et Forest, D. (2005). Classification and categorization in computer assisted reading and analysis of texts. Dans H. Cohen et C. Lefebvre (dir.), *Handbook of categorization in cognitive science* (p. 955-978). Elsevier.

Meunier, J.-G. et Forest, D. (2009). Lecture et analyse conceptuelle assistée par ordinateur : premières expériences. Dans F. L. Priol et J.-P. Desclès (dir.), *Annotations automatiques et recherche d'information* (chap.Cognition et Traitement de l'information). Florence : Hermès science publ.

Michael, J. A. et Szigeti, A. (2018). « The Group Knobe Effect »: evidence that people intuitively attribute agency and responsibility to groups. *Philosophical Explorations*, *0*(0), 1-18.

Mikolov, T., Chen, K., Corrado, G. et Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Dans *Advances in neural information processing systems* (p. 3111-3119).

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism.* (s. l.) : MIT press.

Millikan, R. G. (1998). A common structure for concepts of individuals, stuffs, and real kinds: More Mama, more milk, and more mouse. *Behavioral and Brain Sciences*, *21*(1), 55-65.

Millikan, R. G. (2017). *Beyond Concepts: Unicepts, Language, and Natural Information.* (s. l.) : Oxford University Press.

Minsky, M. (1975). A framework for representing knowledge. Dans P. Winston (dir.), *The psychology of computer vision* (p. 211-277). New-York : McGraw-Hill.

Mizumoto, M. (2018). A Simple Linguistic Approach to the Knobe Effect, or the Knobe Effect Without Any Vignette. *Philosophical Studies*, *175*(7), 1613-1630.

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *CoRR*, *abs/1605.02019*.

Murdock, J., Allen, C. et DeDeo, S. (2017). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. *Cognition*, *159*, 117-126. Récupéré de Elsevier Bv

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA : MIT press.

Murphy, T. (2014). Experimental Philosophy: 1935-1965. Dans T. Lombrozo, J. Knobe, et S. Nichols (dir.), *Oxford Studies in Experimental Philosophy* (p. 1-325). Oxford University Press.

Nado, J. (2016). The intuition deniers. *Philosophical Studies*, *173*(3), 781-800.

Naess, A. (1938). *« Truth » as Conceived by Those who are Not Professional Philosophers*. Oslo : (n. é.).

Nguyen, D. Q., Billingsley, R., Du, L. et Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, *3*, 299-313.

Nichols, S. et Ulatowski, J. (2007). Intuitions and Individual Differences: The Knobe Effect Revisited. *Mind and Language*, *22*(4), 346-365.

Nolan, D. (2009). Platitudes and Metaphysics. Dans D. Braddon-Mitchell et R. Nola (dir.), *Conceptual Analysis and Philosophical Naturalism*. MIT Press.

Nolfi, K. (2016). *Epistemically Flawless False Beliefs*. Montréal, Québec : Fillosophie. Manuscrit soumis pour publication.

Peng, M., Xie, Q., Zhang, Y., Wang, H., Zhang, X., Huang, J. et Tian, G. (2018). Neural Sparse Topical Coding. Dans *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (p. 2332-2340). Melbourne, Australia : Association for Computational Linguistics.

Pennington, J., Socher, R. et Manning, C. (2014). Glove: Global vectors for word representation. Dans *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (p. 1532-1543).

Pettit, D. et Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language*, *24*(5), 586-604.

Pinder, M. (2017). Does Experimental Philosophy Have a Role to Play in Carnapian Explication? *Ratio*, *30*(4), 443-461.

Plantinga, A. (1993). *Warrant and Proper Function.* Oxford : Oxford University Press.

Pohlhaus, G. (2015). Different Voices, Perfect Storms, and Asking Grandma What She Thinks: Situating Experimental Philosophy in Relation to Feminist Philosophy. *Feminist Philosophy Quarterly*, *1*(1).

Potapenko, A., Popov, A. et Vorontsov, K. (2017). Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. Dans *Conference on Artificial Intelligence and Natural Language* (p. 167-180).

Pulizzotto, D., Lopez, J. A., Jean-Guy, J.-F. C., Tan, M. L. C. F. L. et Ngoc, L. (2016). Recherche de «périsegments» dans un contexte d'analyse conceptuelle assistée par ordinateur: le concept d'«esprit» chez Peirce. Dans *JEP-TALN-RECITAL 2016* (Vol. 2). Paris.

Pust, J. (2000). *Intuitions as Evidence.* New York : Routledge.

Quine, W. V. (1971). Epistemology naturalized. *Akten des XIV. Internationalen Kongresses für Philosophie, 6*, 87-103.

Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. Dans G. Williams (dir.), *La linguistique de corpus* (p. 31-45). Presses universitaires de Rennes.

Rey, G. (2018). The Analytic/Synthetic Distinction. Dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 éd.). Metaphysics Research Lab, Stanford University.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology, 4*(3), 328-350.

Rupert, R. D. (2009). *Cognitive systems and the extended mind.* (s. l.) : Oxford University Press, USA.

Rysiew, P. (2017). Naturalism in Epistemology. Dans E. N. Zalta (Dir.) (dir.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 éd.). Metaphysics Research Lab, Stanford University.

Řehůřek, R. et Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. Dans *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (p. 46-50). Valletta, Malta : University of Malta.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies, 20*, 33-53.

Sainte-Marie, M. B., Meunier, J.-G., Payette, N. et Chartier, J.-F. (2011). The concept of evolution in the Origin of Species: a computer-assisted analysis. *Literary and linguistic computing, 26*(3), 329-334.

Salehi, B., Cook, P. et Baldwin, T. (2015). A word embedding approach to pre-

dicting the compositionality of multiword expressions. Dans *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 977-983).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. Dans *Proceedings of the international conference on new methods in language processing* (Vol. 12, p. 44-49).

Schwartz, P. H. (1999). Proper Function and Recent Selection. *Philosophy of Science*, *66*(3), 210-222.

Searle, J. R. (1958). Proper names. *Mind*, *67*(266), 166-173.

Shepherd, J. et Justus, J. (2015). X-Phi and Carnapian Explication. *Erkenntnis*, *80*(2), 381-402. Récupéré de Springer Netherlands

Sosa, E. (2007). Experimental Philosophy and Philosophical Intuition. *Philosophical Studies*, *132*(1), 99-107. Récupéré de Springer

Stalnaker, R. (1976). Propositions. *Issues in the Philosophy of Language*, 79-91.

Swain, S., Alexander, J. et Weinberg, J. (2008). The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp. *Philosophy and Phenomenological Research*, *76*(1), 138-155.

Sytsma, J., Bluhm, R., Willemsen, P. et Reuter, K. (2019). Causal Attributions and Corpus Analysis. Dans E. Fischer (dir.), *Methodological Advances in Experimental Philosophy*. London : Bloomsbury Press.

Tang, Y.-K., Mao, X.-L., Huang, H., Shi, X. et Wen, G. (2018). Conceptualization topic modeling. *Multimedia Tools and Applications*, *77*(3), 3455-3471.

Trask, A., Michalak, P. et Liu, J. (2015). sense2vec-A fast and accurate method for

word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388.*

Van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & society, 4*(2), 249-283.

Venant, F. et Maheux, J.-F. (2016). Réforme de l'enseignement au Québec : une visite guidée par la textométrie. Dans *JADT 2014 : 12ème Journées internationales d'Analyse statistique des Données Textuelles.* Paris, France.

Von Eckardt, B. (1995). *What is cognitive science?* Cambridge, MA : MIT press.

Wang, B., Liakata, M., Zubiaga, A. et Procter, R. (2017). A Hierarchical Topic Modelling Approach for Tweet Clustering. Dans *International Conference on Social Informatics* (p. 378-390).

Weinberg, J. M., Nichols, S. et Stich, S. (2001). Normativity and Epistemic Intuitions. *Philosophical Topics, 29*(1-2), 429-460.

Welling, M., Teh, Y. W. et Kappen, H. (2012). Hybrid variational/Gibbs collapsed inference in topic models. *arXiv preprint arXiv:1206.3297.*

Wellisch, H. H. (1986). The oldest printed indexes. *The Indexer, 15*(2), 73-82.

Williamson, T. (2008). *The Philosophy of Philosophy.* Malden, MA : Wiley-Blackwell.

Williamson, T. (2013). How Deep is the Distinction Between A Priori and A Posteriori Knowledge? Dans A. Casullo et J. C. Thurow (dir.), *The A Priori in Philosophy* (p. 291-312). Oxford University Press.

Wisdom, J. (2017). Proper Function Moral Realism. *European Journal of Philosophy, 25*(4), 1660-1674.

Wittgenstein, L., Anscombe, G. E. M. et Rhees, R. (2001). *Philosophical Investigations* (3e éd.). Malden, MA : Blackwell.

Wright, J. C. et Bengson, J. (2009). Asymmetries in Judgments of Responsibility and Intentional Action. *Mind and Language, 24*(1), 24-50. Récupéré de Wiley-Blackwell

Wu, L.-C. (2014). L'opinion face à la crise Google dans les pays sinophones. Dans *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles.* Paris, France.

Xun, G., Li, Y., Zhao, W. X., Gao, J. et Zhang, A. (2017). A correlated topic model using word embeddings. Dans *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (p. 4207-4213).

Young, L., Cushman, F., Adolphs, R., Tranel, D. et Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status?: Neuropsychological evidence. *Journal of cognition and culture, 6*(1-2), 291-304.

Zhang, X., Feng, R. et Liang, W. (2019). Short Text Topic Model with Word Embeddings and Context Information. Dans H. Unger, S. Sodsee, et P. Meesad (dir.), *Recent Advances in Information and Communication Technology 2018* (p. 55-64). Cham : Springer International Publishing.

Zhao, H., Du, L., Buntine, W. et Zhou, M. (2018). Inter and Intra Topic Structure Learning with Word Embeddings. Dans *International Conference on Machine Learning* (p. 5887-5896).