Utilisation de la méthode distributionnelle pour la constitution de classes sémantiques d'une liste de formes du lexique scientifique transdisciplinaire

Emmanuelle Dusserre

Utilisation de la méthode distributionnelle pour la constitution de classes sémantiques d'une liste de formes du lexique scientifique transdisciplinaire

Dusserre
Emmanuelle

Sous la direction de Patrick Drouin et
Agnès Tutin

Laboratoire : OLST et LIDILEM
UFR LLASIC

Mémoire de master 2 recherche - 30 crédits – Mention Sciences du langage

Spécialité ou Parcours : Industries de la langue

Année universitaire 2015-2016

Utilisation de la méthode distributionnelle pour la constitution de classes sémantiques d'une liste de formes du lexique scientifique transdisciplinaire

Dusserre
Emmanuelle

Sous la direction de Patrick Drouin et
Agnès Tutin

Laboratoire : OLST et LIDILEM
UFR LLASIC

Mémoire de master 2 recherche - 30 crédits – Mention Sciences du langage

Spécialité ou Parcours : Industries de la langue

Année universitaire 2015-2016

Table des matières

Remerciement

La réalisation de tout projet est une aventure unique, qui nous permet d'accroître nos connaissances, notamment grâce à l'expérience des autres. Je souhaite donc remercier :

Patrick Drouin : Mon directeur de mémoire, pour m'avoir accueilli chaleureusement à l'Observatoire de Linguistique Sens-texte de Montréal. Pour son écoute, sa patience et le savoir qu'il m'a enseigné. Également pour m'avoir permis de réaliser ce projet et pour la confiance qu'il m'a accordée.

Agnès Tutin : Ma co-directrice de mémoire, pour tous les conseils qu'elle m'a apportés, ses précieuses relectures et son soutien tout au long du projet.

Les membres de l'OLST : Pour toutes les réponses à mes questions qu'ils m'ont apportées, leur disponibilité et leur soutien. Merci à Gabriel Bernier-Colborne pour son aide pour la programmation.

L'équipe pédagogique du master Industries de la langue : Car ils ont fait preuve d'une grande disponibilité pour leurs étudiants tout au long de la formation, ce qui nous a permis d'apprendre dans de bonnes conditions.

Mes collègues du master Industries de la langue : Pour les discussions que nous avons pu échanger, leurs conseils et leur soutien.

Mes proches : Pour leurs relectures, leur soutien permanent notamment pour ce séjour à Montréal.

Je tiens particulièrement à remercier les enseignants-chercheurs que j'ai pu rencontrer tout au long de ma formation et également les membres de l'OLST qui ont su me faire partager leur passion du métier.

Déclaration anti-plagiat

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : DUSSERRE PRENOM : Emmanuelle

DATE : 17/06/2016 SIGNATURE : [redacted]

Mise à jour avril 2016

Glossaire

Hypothèse distributionnelle : Hypothèse selon laquelle les mots qui apparaissent dans les mêmes contextes linguistiques partagent des significations similaires. (Harris 1954)

Word2vec : Ensemble de modèles permettant de créer des regroupements de mots. Ces modèles utilisent des couches de réseaux de neurones artificiels dits peu profonds. Ces derniers sont entrainés afin de reconstruire les contextes linguistiques des mots. Word2vec a été mis au point par Mikolov et *al.* (2013 a, b, c) et son équipe à Google.

Lexique Scientifique Transdisciplinaire : Lexique commun à diverses disciplines scientifiques. Il contient des mots simples tels que : *analyse*, *hypothèse*, *rechercher*, *montrer*, ainsi que des collocations du type : *participer à une activité*.

Classification automatique : Ou *clustering* en anglais, consiste à attribuer à chaque objet une classe. Cela se fait en se basant sur des données statistiques, notamment grâce à l'apprentissage automatique.

Introduction

L'hypothèse distributionnelle (Harris 1954), démontre que les mots sémantiquement proches ont tendance à partager des contextes similaires. Celle-ci a été, au cours des dernières années, adaptée sous forme d'algorithme. Cela a ainsi permis de regrouper ensemble les mots partageant des caractéristiques sémantiques. L'analyse distributionnelle consiste donc à observer les contextes d'un mot. Par exemple, si l'on prend les mots *livre* et *bouquin*, nous nous apercevons assez rapidement qu'ils partagent plusieurs contextes tels que « feuilleter un livre » et « feuilleter un bouquin ». Ils apparaitront en effet, régulièrement dans le même type de phrases et seront donc considérés, conformément à l'hypothèse, comme sémantiquement proches.

La méthode distributionnelle a su faire ses preuves au cours des dernières années pour le traitement automatique des langues. Elle est notamment performante dans la recherche de synonymie et certains chercheurs (Landauer & Dumais, 1997) ont d'ailleurs évalué le modèle en lui faisant passer le TOEFL[1]. Cela consistait à proposer un mot au modèle afin que celui-ci réponde comme un humain, en proposant les synonymes du mot. Les résultats étaient impressionnants, atteignant les 92% de réussite. D'autres chercheurs (Landauer, Foltz & Laham 1997) ont également testé l'analyse distributionnelle à l'aide du « Priming effect » (Effet d'amorçage). Ces travaux ont permis d'observer que cette méthode pouvait se comporter comme un cerveau humain. Le « Priming effect » consiste à faciliter le traitement d'un stimulus par le traitement d'un autre stimulus. Par exemple, on peut créer un effet de facilitation en activant dans un premier temps une zone cérébrale, puis dans un second temps, une autre zone proche afin que la deuxième information soit plus rapidement traitée. Dans le cas de la méthode distributionnelle, il s'agit de donner un mot à la machine tel que *pomme*, puis un second. Si le second est sémantiquement proche du premier alors le traitement sera plus rapide. Si par exemple, le deuxième mot est *poire*, il sera plus rapidement traité que si le mot avait été *table*. Les résultats pour le « Priming effect » ont été satisfaisants, ce qui laisse entrevoir toutes les possibilités qu'engendre l'analyse distributionnelle.

L'analyse distributionnelle est de nos jours beaucoup employée pour la recherche de documents avec expansion sémantique de requêtes (Picton, Fabre & Bourigault 2008). Cela

[1] TOEFL : Test of English as a Foreign Language, est un test visant à évaluer l'aptitude à comprendre et utiliser la langue anglaise (Source : Wikipédia).

consiste à rechercher un mot et obtenir des réponses plus larges que le simple mot cible. Par exemple, si l'on saisit la requête *labrador*, le moteur de recherche pourra également proposer *chien*. La méthode distributionnelle est également employée pour la classification automatique. Cela permet de regrouper ensemble les mots, les phrases ou les documents comportant des caractéristiques communes. Cette technique est actuellement très demandée. Par exemple, elle peut être utilisée pour le marketing lorsque l'on veut rechercher les différents profils constituant une clientèle ou créer un panel de représentation. Dans le domaine du TAL, Grouin & Forest (2012) ont également exploité l'analyse distributionnelle afin de regrouper ensemble les textes ayant le même genre ou le même domaine.

L'analyse distributionnelle et la classification automatique ont récemment beaucoup évolué grâce à l'accroissements des données disponibles. Un modèle, basé sur les réseaux de neurones artificiels permettant de réaliser des calculs performants sur des énormes quantités de données a été élaboré. Ce dernier est un outil appelé word2vec qui a été développé par Mikolov (2013) et son équipe à Google. Il permet d'exploiter la méthode distributionnelle basée sur les réseaux de neurones artificiels. Les résultats obtenus grâce à cette technique sont prometteurs et ont été repris dans de nombreux travaux de chercheurs pour extraire des mots entretenant des relations sémantiques.

Dans notre étude, nous faisons l'hypothèse que la méthode distributionnelle peut nous permettre d'organiser une liste de mots brute. En effet, nous souhaitons utiliser cette technique sur une liste de noms afin de générer des regroupements entre les noms sémantiquement proches. Pour ce faire, nous utiliserons trois techniques de méthode distributionnelle : deux techniques « classiques » appelées Pointwise Mutual Information (PMI) et Singular Value Decomposition (SVD) et une troisième qui est celle de word2vec, basée sur les réseaux de neurones artificiels.

Nous souhaitons ainsi répondre à notre problématique intitulée :

« Utilisation de la méthode distributionnelle pour la constitution de classes sémantiques d'une liste de formes du lexique scientifique transdisciplinaire ».

Dans un premier temps, nous appliquerons ces trois techniques sur un corpus contenant des thèses et des articles recouvrant neuf disciplines scientifiques. Puis, nous utiliserons un algorithme de classification automatique sur une liste de noms du lexique scientifique transdisciplinaire (LST), constituée par Drouin (2010) et comportant 461 formes. Il convient avant tout, d'expliquer brièvement le LST afin de mieux cerner notre projet. Ce lexique est composé des formes appartenant à un genre commun, c'est-à-dire transdisciplinaires aux

domaines des écrits scientifiques. De plus, ce type de discours n'est ni terminologique, ni thématique et non intra-disciplinaire. L'étude de ces formes permet de cerner avec précision les écrits de méthodologies scientifiques et ainsi faciliter leurs enseignements aux locuteurs natifs et aux apprenants étrangers.

Nous pensons qu'organiser de manière sémantique le lexique scientifique transdisciplinaire serait un avantage indéniable pour son utilisation. Sur le plan didactique, cela faciliterait notamment son emploi pour l'enseignement et l'apprentissage. Nous sommes d'avis que la méthode distributionnelle sera un véritable atout pour y parvenir.

Nous nous appuierons essentiellement sur la méthode distributionnelle basée sur les réseaux de neurones, word2vec, dans la réalisation de nos travaux. En effet, comme l'ont montré Levy et Golderg (2015), il semblerait qu'elle soit la technique la plus robuste et qu'elle ne nécessite pas de connaissances expertes pour l'ajustement des différents paramètres.

Nous sommes convaincue qu'il est nécessaire d'apporter une description sémantique manuelle aux différentes classes que nous obtiendrons avec la classification automatique. Dans un premier temps nous souhaitons observer quels types de regroupements nous obtiendrons, c'est-à-dire quel type de relations sémantiques sont présentes entre les mots, afin d'apporter des descriptions pour chacune des classes. De plus, nous voulons les étiqueter à l'aide du dictionnaire hiérarchique de Polguère (2007). Nous pensons que la méthode distributionnelle permet essentiellement d'extraire des relations de quasi-synonymie, bien que certaines études montrent qu'avec word2vec nous pouvons parfois relever d'autres types de relations telles que la méronymie. Ensuite, une approche manuelle nous permettra d'observer quel type d'erreur nous obtiendrons et ainsi apporter un regard critique sur nos résultats.

Tout au long de nos travaux, nous nous servirons de classes sémantiques construites de manière semi-automatique par Hatier et *al.* (2014) et Hatier et *al.* (2016). Cela nous permettra d'évaluer la qualité de nos sorties et sera un appui afin d'apporter les descriptions sémantiques nécessaires à nos regroupements de noms.

Dans un premier chapitre, nous feront l'état de l'art de ce projet, en nous basant sur les différents travaux menés d'une part sur le LST, et d'autre part, sur les différentes techniques de méthodes distributionnelles. Cela nous permettra de situer avec précision comment s'insère notre étude parmi les précédents projets et ce qu'elle apportera. Dans le second chapitre, nous décrirons la méthodologie, c'est-à-dire les différentes étapes qui nous ont permis de constituer nos classes sémantiques, à partir de la liste de noms brute. Ce chapitre montrera également de quelle manière nous avons procédé afin de leur apporter une description sémantique. Ensuite,

le dernier chapitre présentera les résultats définitifs de notre étude. Il contiendra de ce fait les différentes classes sémantiques obtenues ainsi que leur description sémantique. Dans une ultime partie nous procèderons à une analyse des résultats au cours de la discussion. Cela permettra d'apporter un regard critique sur nos travaux et ainsi conclure sur l'intérêt de cette étude, puis entrevoir quelques perspectives.

Description des laboratoires

Le projet s'est déroulé à l'Observatoire de linguistique Sens-texte (OLST)[2] de l'Université de Montréal. Ce laboratoire a été créé en 1997. Il est rattaché au Département de linguistique et de traduction et l'actuel directeur est Patrick Drouin. Les recherches menées par les différents membres s'articulent autour de la terminologie, la linguistique et la didactique. Bien que les projets menés ne s'en tiennent pas uniquement à la théorie Sens-Texte (TST), l'OLST en tire tout de même son nom et un grand nombre de ses travaux.

L'OLST possède différents axes de recherche, tels que la lexicologie et la lexicographie, la terminologie, le traitement de corpus, la didactique des langues ou encore les applications documentaires. Ce laboratoire possède donc un très grand éventail de thématiques de recherche, ce qui nous a notamment permis de réaliser ce projet orienté sur la sémantique tout en utilisant des outils informatiques.

Nos travaux ont de plus, bénéficié de l'aide du Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (LIDILEM[3]) de l'Université Grenoble Alpes. Il compte une soixantaine de membres ainsi qu'environ soixante-dix doctorants. La directrice est Marinette Matthey.

Les études menées par le laboratoire portent sur les sciences du langage et regroupent différentes disciplines telles que le traitement automatique des langues, la didactique des langues ou encore de la constitution de corpus. Il comporte trois axes principaux de recherche qui sont :

- Les descriptions linguistiques, le traitement automatique des langues, corpus
- La sociolinguistique, Acquisition, Multimodalité
- La didactique des langues et ingénierie

Le soutien de la professeure Agnès Tutin ainsi que les différents travaux du LIDILEM portant sur la description du Lexiques Scientifique Transdisciplinaire nous ont permis de mieux appréhender la description sémantique des classes de noms.

[2] OLST : http://olst.ling.umontreal.ca/?cat=1 (consulté : mai 2016)
[3] LIDILEM : http://lidilem.u-grenoble3.fr/presentation/ (consulté : mai 2016)

Chapitre 1 :

État de l'art

À travers cet état de l'art, nous survolerons les différents travaux autour de notre problématique qui s'intitule :

« Utilisation de la méthode distributionnelle pour la constitution de classes sémantiques d'une liste de formes du lexique scientifique transdisciplinaire ».

Afin de mieux circonscrire ce projet, dans une première partie, nous expliquerons de manière approfondie le Lexique Transdisciplinaire et présenterons les travaux principaux qui se sont articulés autour de ce thème. Dans une seconde partie, nous aborderons l'hypothèse distributionnelle et comment elle a été reprise ces dernières années, au cours de différents projets. Puis, nous décrirons les différents modèles que nous utiliserons pour nos travaux, à savoir une méthode plutôt classique et une autre basée sur les réseaux de neurones artificiels.
Enfin, nous expliquerons comment notre projet s'insère parmi ces différents travaux et clarifierons la démarche que nous suivrons afin d'obtenir nos résultats et de les analyser.

1. Le Lexique Scientifique Transdisciplinaire

1.1 L'origine du Lexique Scientifique Transdisciplinaire

Dans cette première partie, nous nous intéresserons au Lexique Scientifique Transdisciplinaire (LST). Ce dernier regroupe des mots appartenant à divers domaines scientifiques. Le lexique est plutôt constitué de mots tels que *analyse* ou *hypothèse* et non de mots tels que *humanité* faisant référence à l'anthropologie par exemple. Ces formes sont spécifiques aux écrits scientifiques et sont collectées en fonction de leur fréquence. Cela signifie qu'il faut qu'elles apparaissent fréquemment dans le discours scientifique afin d'être retenues. Ce vocabulaire contient aussi des expressions du type : *point de vue* ou *soutenir une hypothèse*. Nous allons ainsi, caractériser finement ce dernier afin de pouvoir mener à bien notre projet.

Le LST tire son origine de travaux débutant dans les années 1970, tels que ceux de Phal (1971) pour la langue française ou Coxhead (1998, 2000), plus récemment, pour la langue anglaise.
En 1971, Phal publie le Vocabulaire Général d'Orientation Scientifique (V.G.O.S), réalisé pour la langue française, constitué de manière semi-automatique à partir d'un corpus de 1 794 500

mots ne comptant que des textes issus de manuels scolaires relevant de domaines scientifiques : la physique, la chimie et les sciences naturelles. Le chercheur en extrait une liste de vocabulaire qui exclut les noms propres, les formes du français fondamental de Gougenheim *et al*. (1956), ou encore les unités de mesure. Il apporte ainsi, de par ses travaux, les premiers critères de sélection pour constituer une liste de vocabulaire appartenant au genre scientifique. Il nous donne la description suivante afin de caractériser sa liste :

« Le vocabulaire scientifique général est (...) commun à toutes les spécialités. Il sert à exprimer les notions élémentaires dont elles ont toutes également besoin (mesure, poids, rapport, vitesse, etc.) et les opérations intellectuelles que suppose toute démarche méthodique de la pensée (hypothèse, mise en relation, déduction et induction, etc.) ». (Phal 1971 : 9)

L'utilisation du V.G.O.S était essentiellement destinée à des fins didactiques. Il également a été repris pour des travaux de constitution de thesaurus[4], comme l'a par exemple fait Pecman (2004), lorsqu'elle a élaboré le lexique scientifique général.

Coxhead (1998, 2000), a publié « L'Academic Word List » consacrée à la langue anglaise.
Cette dernière a été élaborée afin de construire des méthodes d'enseignement fiables et de pouvoir orienter les étudiants vers les mots réellement utiles pour leur formation. L'utilisation de cette liste est notamment facilitée par le classement par fréquence des mots. En effet, les étudiants ou les enseignants peuvent se focaliser uniquement sur les mots les plus fréquents de l'anglais et ainsi cibler davantage les apprentissages. Enfin, l'Academic Word List peut être employée pour l'enseignement des suffixes et des préfixes étant composée d'une grande quantité de mots d'origines latines et grecques.
Elle a été constituée à partir d'un corpus de 3,5 millions de mots provenant d'écrits scientifiques. Pour ce faire, le corpus a été divisé en plusieurs tranches principales qui sont : le droit, les sciences humaines, le commerce et les sciences, puis il a été divisé en vingt-huit « sous-sujets ».
Les mots de la liste ont été extraits du corpus selon trois critères qui sont :

- Un nombre d'apparition spécifique, c'est-à-dire n'appartenant pas aux des deux milles mots les plus fréquents de l'anglais, recensés dans l'ouvrage de West (1853)

[4] Thésaurus : Répertoire alphabétique des mots d'une langue, d'un domaine scientifique, technique, etc.

- La répartition, qui signifie qu'une forme doit apparaître au moins dix fois dans les quatre principales sections du corpus et au moins quinze dans les vingt-huit « sous-sujets »
- Enfin, la fréquence : un mot doit apparaître au moins cent fois dans tout le corpus académique de référence

Le Lexique Scientifique Transdisciplinaire tire donc essentiellement son origine des premiers travaux de Phal (1971) et de Coxhead (1998-200). Néanmoins, ce n'est qu'à partir des travaux de Tutin (2007) et de Drouin (2007) que le LST a réellement émergé en tant que tel. De plus, les recherches de Pecman (2004) ont apporté une grande avancée pour le LST concernant les aspects théoriques. En effet, cette dernière a réalisé une étude sur la langue scientifique générale afin d'observer les invariants du discours scientifique. Elle a notamment démontré que le discours commun aux écrits scientifiques se réalise à travers une phraséologie codifiée renvoyant aux raisonnements scientifiques tels que le positionnement de l'auteur ou la présentation des résultats.
Nous allons, au cours de la section suivante, présenter les caractéristiques du LST actuel. Pour y parvenir, nous expliquerons comment les chercheurs le caractérisent, puis les techniques utilisées afin de sélectionner ses différentes formes. Nous montrerons aussi les fins auxquelles il peut être destiné. Nous illustrerons cela par les travaux menés sur le LST.

1.2 Circonscrire le Lexique Scientifique Transdisciplinaire

Les articles scientifiques portant sur le LST présentent de nombreux ensembles lexicaux. Voici ceux retenus dans Tutin (2007) :

- Le lexique de la langue générale, celui-ci n'est pas lié à la discipline ou au genre d'écrits, il contient par exemple : *à, en, parents, demain* etc.
- Le lexique terminologique renvoyant aux objets de l'étude, comme par exemple, dans un texte d'économie avec : *emploi, travail, bien, services* etc.
- Le lexique abstrait ou non spécialisé, ce qui signifie qu'on peut le retrouver dans des écrits différents des écrits scientifiques, tels que : *soulever un problème, homogénéité, dimension* etc.
- Le lexique méthodologique disciplinaire est propre à un domaine particulier, comme par exemple pour les sciences sociales : *panel, échantillon, analyses longitudinales* etc.

- Le lexique propre aux écrits scientifiques, il est celui qui nous intéresse pour notre étude, contenant des mots tels que : *général*, *montrer*, *appartenir*, *analyser*, *problème* etc.

Le lexique propre aux écrits scientifiques est difficile à circonscrire. Pourtant il existe bien cette notion de lexique partagé par les écrits scientifiques que l'on peut qualifier de transdisciplinaire. En effet, les différentes disciplines scientifiques contiennent d'une certaine manière les mêmes codes, c'est-à-dire un vocabulaire partagé permettant d'atteindre les mêmes buts. Plus précisément, cela signifie qu'il est composé d'éléments qui renvoient à des pratiques intellectuelles ou des raisonnements, référant à des procédures communes telles que : l'argumentation, l'analyse ou le développement. Cette notion de transdisciplinarité est essentielle et elle est davantage corrélée au sous-genre comme les articles scientifiques ou les manuels, qu'au domaine scientifique.
Le LST est défini comme étant méta-scientifique, cela signifie qu'il permet de référer aux notions scientifiques fondamentales, aux objets et à la pensée scientifique. Il regroupe également les différents éléments du métadiscours, comme les marques d'atténuations et les connecteurs. Voici quelques exemples de mots appartenant à la liste : *mener, intervenir, annexe, genre, forte augmentation, etc.*

Le LST est essentiellement utilisé à des fins didactiques. En effet, il constitue une aide pour la rédaction de parties méthodologiques, par exemple. Cela est valable aussi bien pour une langue maternelle que pour l'apprentissage d'une langue étrangère.
Pecman (2004) présente la construction d'un dictionnaire bilingue (français-anglais) phraséologiques multifonction, basé sur la langue scientifique générale biologie, physique). La chercheuse a souhaité mener ces travaux, car elle est partie du constat qu'il n'est absolument pas naturel de maitriser ce type de phraséologie pour un étranger ou même un locuteur natif.
Nous pouvons aussi retrouver cet aspect didactique des langues étrangères avec les travaux de Drouin (2010), qui a constitué de manière automatique un Lexique Scientifique Transdisciplinaire bilingue permettant d'obtenir des paires de collocations anglaises-françaises. Les travaux de l'équipe du LIDILEM (Hatier & Yan (2015), Tran 2014) ont aussi mis en évidence l'utilité du LST pour aider les locuteurs natifs et étrangers du français dans l'apprentissage de la phraséologie des méthodologies scientifiques.

Dans le domaine du traitement automatique des langues (TAL), le LST constitue un socle important afin d'analyser l'information scientifique. En effet, pouvoir accéder à ce genre de phraséologie et le traiter est un véritable enjeu pour le TAL. Nous pouvons notamment observer cet intérêt dans le projet TermITH de Jacquey et *al*. (2013) ainsi que celui de Hatier et *al.* (2016) où il s'agit d'améliorer l'extraction automatique de termes en exploitant le LST.

1.3 Méthodes de constitution du Lexique Scientifique Transdisciplinaire

La sélection des différents éléments du LST est compliquée. C'est pourquoi il a été nécessaire de développer des techniques de traitement automatique des langues, afin de rendre le travail moins fastidieux. Nous allons présenter, dans un premier temps, comment Drouin (2007) a réalisé son LST. Puis dans un second temps, nous montrerons quelle approche, Hatier *et al.* (2014) et Hatier et *al.* (2016) ont utilisée quant à eux.

1.4 La constitution du LST de Drouin (2007)

Drouin (2007) a constitué un corpus transdisciplinaire afin de pouvoir en extraire les différents membres du LST. Pour cela, il a constitué un corpus de plus de 2 300 000 mots comportant neuf disciplines différentes qui sont : l'anthropologie, la chimie, l'informatique, l'ingénierie, la géographie, l'histoire, le droit, la physique et la psychologie.

Afin de pouvoir filtrer les différents candidats appartenant au LST, il a utilisé un lexique de référence qui est le V.G.O.S de Phal (1971), le but étant de voir comment il pouvait enrichir ce dernier avec un minimum d'intervention.

Drouin (2007) a utilisé une méthode statistique appelée calcul des spécificités (Lafon, 1980). Ce calcul permet d'évaluer la répartition d'une forme au sein d'un ensemble textuel. Cela signifie que si la spécificité d'un mot appartenant au corpus d'analyse est haute, alors la fréquence du mot se démarque significativement par rapport au corpus de référence. Ainsi, dans ses travaux, le seuil de fréquence était fixé à 3,09. Cela limitait à une chance sur 1000 d'obtenir un résultat dû au hasard.

Drouin (2007) a ajouté une autre contrainte afin de sélectionner les éléments du LST : la répartition, c'est-à-dire que le corpus d'analyse est divisé en 100 sous-corpus, afin que les mots

relevés n'appartiennent pas à un domaine en particulier mais à un ensemble textuel. Pour qu'un mot soit sélectionné il doit apparaître dans un minimum de 50 sous-corpus.

Concernant les mots simples, Drouin (2007), a recensé les noms, les verbes, les adjectifs et les adverbes. Voici des exemples des résultats obtenus :

- 461 noms : *argument, conclusion, enjeu, résultat, analyse, constat, objectif* et *difficulté.*
- 356 verbes *: varier, reposer, éclairer, ressortir, transmettre, fonder, tester* et *citer.*
- 178 adjectifs : *fin, minimal, fiable, technique, supérieur, absent, progressif* et *homogène*.

Drouin (2007) avait également extrait des collocations appartenant au LST. Pour ce faire, il s'est limité à extraire les collocations autour des 5 premières entrées de sa liste du LST, étant donné qu'il ne s'agissait pas de l'objectif principal de ses travaux. Les contraintes afin de sélectionner les collocations étaient : la spécificité de la collocation dans le corpus d'analyse par rapport au corpus de référence. Il n'y avait cependant pas de critère de répartition et les résultats obtenus sont par exemple : *supporter un travail, mettre en jeu, engager un jeu, exprimer une fonction* etc.

1.5 La constitution du LST de Hatier et *al.* (2014 et 2016)

Lors de leurs travaux d'extraction des noms simples, Hatier, Tutin, Jacques, Jacquey & Kister (2014) ont utilisé un corpus en sciences humaines et sociales, issu du projet Scientext[5]. Il comporte cinq cents articles dans dix disciplines différentes des sciences humaines, qui sont : l'anthropologie, l'économie, l'histoire, la linguistique, la psychologie, les sciences de l'information, les sciences politiques, la géographie, la sociologie et les sciences de l'éducation. Ces chercheurs ont choisi un corpus en sciences humaines car cela n'avait jamais été fait auparavant. En effet, la plupart des travaux ont été réalisés sur des corpus de sciences « dures ». Pourtant, la demande en sciences humaines est forte, notamment car elles constituent un intérêt didactique important pour les étudiants.

[5] Scientext : http://scientext.msh-alpes.fr/scientext-site/spip.php?article1 (consulté : mai 2016)

Pour ce faire, ils ont également utilisé les critères de répartitions, de fréquence et de spécificité de (Drouin, 2007/b). En conséquence, un mot est transversal aux différentes disciplines, s'il apparait au moins dans cinquante des cent tranches du corpus d'analyse et dans cinq des dix disciplines présentes.

Néanmoins, dans leurs travaux ils se sont assez rapidement aperçus qu'il est difficile de sélectionner les formes du LST sans l'approbation d'une approche manuelle. En effet, comme Hatier (2013) le souligne, lorsqu'il a procédé à l'extraction des mots simples, les résultats comportaient beaucoup de bruit. D'après ses travaux, aucun calcul n'est suffisant pour valider ou non l'appartenance d'un mot au LST. Pour pallier cette difficulté, le chercheur a notamment eu recours à des juges experts devant choisir si un mot appartenait au LST, au lexique terminologique ou à la langue générale.

A la suite de cela, il a obtenu les résultats suivants :

- 493 noms : *corpus, cible, appartenance, divergence, mutation, pôle, façon et information.*
- 343 verbes : *analyser, observer, montrer, expliquer, noter, mener, marquer* et *lier.*
- 273 adjectifs *: empirique, objectif, théorique, fondamental, majeur, intense, grand* et *futur.*
- 213 adverbes : *explicitement, nécessairement, essentiellement, positivement, ailleurs, environ, initialement et strictement.*

Ensuite, les chercheurs Tutin, Tran, Kraif & Hatier (s.d.) ont également travaillé sur les collocations qui sont considérées comme centrales pour le LST. Ils ont ainsi réalisé des travaux d'extraction automatique de ces collocations. Il s'agissait d'extraire grâce au Lexicoscope[6], dans un texte parsé avec l'outil Xip[7] en utilisant une fréquence de sept occurrences et une répartition au sein de trois disciplines. Le but était de repérer différentes structures, telles que : « Nom Préposition » « Nom Adjectif », « Nom Verbe », etc. Il y a par exemple *absence d'effet* pour « Nom Préposition Nom » ou *participer à une activité* pour un « Verbe Préposition

[6] Lexicoscope : Le lexiscope est un outil d'exploration de la combinatoire du lexique, initialement développé dans le cadre du projet Emolex. (Source : http://phraseotext.u-grenoble3.fr/lexicoscope/ consulté mai 2016)

[7] Xip : Analyseur syntaxique : https://open.xerox.com/Services/XIPParser/Pages/Using%20XIP (consulté : mai 2016)

Nom ».

Lors de leurs récents travaux, Hatier et *al.* (2016), ont procédé à la description sémantique des éléments de chaque catégorie syntaxique du LST. Ils ont créé des classes et sous-classes sémantiques en se basant sur plusieurs critères, ainsi que sur le *Dictionnaire Electronique des Mots* (Dubois & Dubois-Charlier 2010) et le *Lexique des Verbes du Français* (Dubois & Dubois-Charlier 1997). Ils ont également utilisé le Lexicoscope permettant d'observer notamment les cooccurrents. Enfin, ils ont attribué une définition aux différents mots, ce qui permet de traiter les cas de polysémie.

Voici ci-dessous un échantillon de leurs travaux, où l'on voit comment chacun des mots est réparti au sein des classes et sous-classes, ainsi que la définition à laquelle il correspond :

Meaning Identifier	Lemma	Part of speech	Class	Subclass	Definition / gloss
développement-1	*développement* (‘development’)	noun	{*communication*} (‘communication’)	{*document*} (‘document’)	*Exposé* (‘report’)
développement-2	*développement* (‘development’)	noun	{*processus_évolutif*} (‘progressive process’)	{*augmentation*} (‘increase’)	*Croissance* (‘growth’)
développer-2	*développer* (‘to develop’)	verb	{*processus_évolutif*} (‘progressive process’)	{*augmentation*} (‘increase’)	*Donner de l’extension* ’to expand’
strict	*strict* (‘strict’)	adjective	{*modalité*} (‘modality’)	{*restriction*} (‘restriction’)	*Limité* (‘limited’)
strictement	*strictement* (‘strictly’)	adverb	{*modalité*} (‘modality’)	{*restriction*} (‘restriction’)	*Rigoureusement* (‘strictly’)

Tableau 1 : Extrait des classes et sous-classes sémantiques des éléments du LST de Hatier et *al.* (2016)

2. L'analyse distributionnelle

Nous avons pu, dans la partie précédente, définir de façon plus précise ce qu'est le Lexique Scientifique Transdisciplinaire. Dans le cadre de notre projet, nous souhaitions utiliser la méthode distributionnelle[8] sur ce lexique, afin d'explorer les relations sémantiques entre les mots. Nous allons, dans cette seconde partie, décrire l'hypothèse distributionnelle, puis, nous expliquerons les méthodes que nous avons utilisées pour nos expérimentations ainsi que les travaux orientés autour de celles-ci. Nous avons, tout d'abord, exploité une méthode plus classique basée sur l'utilisation de matrices : Pointwise Mutual Information (PMI) et Singular Value Decomposition (SVD). Ensuite, nous décrirons davantage la seconde méthode appelée word2vec, centrale dans notre projet.

Concernant l'hypothèse distributionnelle, nous nous appuierons sur les travaux des chercheurs Fabre (2012, 2015), Perinet & Hamon (2015), puis ceux de Levy & Goldgerb (2014) pour les modèles PMI et SVD. Enfin, nous expliquerons l'outil word2vec en citant Mikolov, Corrado, Chen & Dean (2013a), Mikolov, Sustkever, Chen, Corrado, & Dean (2013b), (Mikolov, Yih & Zweg (2013c), et Bernier-Colborne (2014).

1.6 L'histoire de l'hypothèse distributionnelle

L'hypothèse distributionnelle remonte à Harris (1954). Ce dernier avait mis en évidence que si des mots peuvent se substituer dans le même contexte, c'est qu'ils sont proches sémantiquement. Ainsi, l'hypothèse démontre que la similarité des sens est corrélée avec celle de la distribution. Si nous prenons l'exemple : « Les topinambours sont dans le panier à légumes » et « J'ai mis des topinambours dans la soupe », même si nous ne connaissons pas le mot « topinambour », nous arrivons tout de même à deviner qu'il s'agit d'un légume. L'idée de la sémantique distributionnelle est donc en quelque sorte de projeter ces processus cognitifs par ordinateur. Cela représente un véritable enjeu pour la linguistique computationnelle, car cette technique

[8] Méthode distributionnelle : Nous entendons par méthode distributionnelle, une mise en application de l'analyse distributionnelle.

peut permettre la création de ressources sémantiques de façon automatique, en arrivant à extraire des unités analogues dans un corpus. Depuis ces vingt dernières années, ce modèle a largement été popularisé, car aujourd'hui nous pouvons travailler sur de grosses quantités de données grâce à l'extension du web. La méthode distributionnelle a ainsi permis par exemple la création de thesaurus automatiques pour la traduction automatique (Grefenstette 1994).

La mise en œuvre du modèle s'opère généralement en quatre étapes, que nous allons expliquer.

a) Chaque cooccurrent va être collecté, puis compté.
b) On va ensuite regarder avec quoi il cooccurre le plus au sein du corpus. Nous pouvons prendre l'exemple de *chat*, *chien* et *avion*. Ainsi, *chat* va avoir beaucoup plus de contextes en commun avec *chien* au sein d'un corpus.
c) Nous utilisons ensuite une matrice afin de représenter cela. Elle peut par exemple se composer d'un ensemble de vecteurs définissant le nombre de fois où les mots apparaissent dans le même contexte. Ainsi, *chat* peut être représenté par un vecteur numérique qui va indiquer qu'il apparait dans le corpus par exemple, cinq fois dans les mêmes contextes que *chien*, contre zéro avec *avion*.
d) Une fois cette matrice réalisée, il faut calculer la distance entre les cooccurrents. Pour ce faire, nous allons les placer sur un graphe et calculer la distance entre les vecteurs avec par exemple le cosinus des angles de ces derniers. Enfin, cela permet, d'obtenir des groupes de mots sémantiquement proches.

1.7 Le choix des paramètres :

De nombreux travaux ont repris la méthode distributionnelle, tels que ce de Fabre (2012) et Perinet & Hamon (2015). Ces chercheurs ont montré qu'il est possible de faire varier différents paramètres de la méthode afin de modifier les résultats obtenus. Le choix de ces paramètres est extrêmement important et affecte tout le fonctionnement de la méthode et dépend entièrement des objectifs.

Tout d'abord le choix des corpus de travail est essentiel. Par exemple, Fabre (2015) travaille sur des corpus spécialisés car cela permet de mieux gérer la polysémie des mots. Elle obtient ainsi des résultats plus pertinents.
Ensuite, le choix du contexte est très important, à savoir si l'on souhaite par exemple travailler sur peu de documents, ou une grande quantité de données. Effectivement, si nous souhaitons par exemple, travailler sur les animaux et nous prenons un corpus de vingt-quatre encyclopédies, il est fort probable qu'il y ait peu de différence entre les cooccurrents des animaux, car le contexte sera trop large.
La méthode distributionnelle est composée de trois familles de modèles permettant entre autres de faire varier le contexte des mots pris en compte. Dans un premier temps, il y a le Latent Semantic Analysis (Landauer and Dumais, 1997) qui va comparer les mots au sein d'un paragraphe ou de documents. Ensuite, le Word Based Model va quant à lui considérer les mots dans une fenêtre, c'est-à-dire en prenant par exemple deux mots avant et après le mot cible. Enfin, nous avons la dernière famille basée sur les modèles syntaxiques, qui vont quant à eux, observer les relations de dépendance entre les mots.
De plus, il est possible d'ajuster les paramètres concernant les mesures statistiques entre les cooccurrents. Nous pouvons, par exemple, utiliser un calcul assez connu appelé Jaccard, souvent repris dans les travaux de Périnet & Hamon (2015). On recense aussi le loglikelihood ou encore le PointWise Mutual Information qui est actuellement un des plus utilisés et donne de très bons résultats. Ces mesures servent à montrer qu'il ne s'agit pas du simple fait du hasard lorsque deux formes ont tendance à apparaître ensemble.

1.8 Le Latent Semantic Analysis

Nous allons davantage expliquer un modèle reconnu appelé le Latent Semantic Analysis (Landauer and Dumais, 1997)**.** Ce dernier permet d'obtenir des relations paradigmatiques entre les mots, telles que la synonymie, l'antonymie ou la méronymie et nous intéressent davantage dans le cadre de nos recherches.
Cet algorithme, est conçu pour créer une grande matrice à plusieurs dimensions, où les mots cooccurrent entre eux, et sur une troisième dimension de quelle manière ils cooccurrent. La matrice générée est immense et le but du Latent Semantic Analysis est donc dans un premier temps de la réduire. Pour ce faire, il va faire appel à d'autres algorithmes tels que le Non-Negative Matrix Factorization. Cela donne la possibilité de travailler sur une matrice dite moins

« éparpillée ». L'enjeu de ce modèle est de pouvoir capturer en quelque sorte la transitivité entre les mots. Cela signifie qu'il est possible, par exemple, au sein d'un texte d'attribuer les mêmes propriétés au mot *chien* que *Berger allemand*. Les résultats de ce modèle sont assez impressionnants et peuvent nous faire penser à de l'intelligence artificielle.

L'évaluation de tels systèmes est assez complexe. Les experts, tels que (Landauer and Dumais, 1997), ont par exemple eu recours au TOEFL (Test of English as a Foreign Language) pour observer comment le modèle répondait à la place d'un humain. Les résultats obtenus ont été très satisfaisants, arrivant au même niveau que les personnes passant l'examen à l'université, à savoir environ 90% de réussite. D'autres types d'évaluation ont pu montrer que le Semantic Latent Analysis pouvait « se comporter » comme le cerveau humain, tel que le test du Semantic Priming, utilisé par (Balota, Yap, Cortese, and Watson, 2007). Ce test que le temps d'activation d'une zone cérébrale sera plus court pour un mot, si juste avant nous avons entendu un mot proche sémantiquement. L'algorithme est dans une certaine mesure capable de se comporter de la même manière et de réaliser une sorte de « Priming » en associant les formes entre elles.

1.9 Travaux autour de l'analyse distributionnelle

L'analyse distributionnelle a souvent été reprise et de nombreux travaux ont été réalisés, afin de savoir quels étaient les paramètres les plus pertinents. Nous avons par exemple, Périnet & Hamon (2014) qui montrent à la suite de leurs expériences que pour les corpus de spécialité les meilleurs paramètres sont :

- Utiliser la mesure appelée test de Jaccard qui permet d'évaluer la similarité entre deux ensembles, ici A et B.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Le contexte le plus performant est celui de deux mots de chaque côté de celui ciblé. Par exemple, dans la phrase *Eugénie étudiait la sociologie depuis quatre ans*, pour le mot *sociologie*, on prendra en considération les mots *étudiait, la, depuis* et *quatre*.

Ensuite, les travaux de Morlane-Hondère & Fabre (2012) basés sur le test de substituabilité[9], reprennent également la méthode distributionnelle. Ces derniers se sont intéressés à la façon dont l'analyse distributionnelle permet d'extraire les relations paradigmatiques entre les mots. Pour cela, ils ont comparé les données qu'ils ont obtenues sur corpus, avec des lexiques externes tels que « JeuxdeMots » (Lafourcade 2008). Le fait de comparer les données obtenues par la méthode distributionnelle et des lexiques déjà constitués leur ont permis d'observer pour quelles raisons des paires présentes dans les lexiques ne sont pas relevées par une analyse manuelle. Ils concluent dans leurs travaux que la proximité sémantique relevée par la méthode distributionnelle dépasse celle contenue dans les lexiques externes qu'ils ont utilisés et peut être une aide véritable pour la constitution de lexique.

Périnet (2015), au cours de ses travaux, a essayé d'améliorer la pertinence des résultats de la méthode distributionnelle. Pour cela, elle a prouvé que le fait de réduire la dispersion des données donnait des regroupements sémantiques plus homogènes. Pour obtenir ces résultats, elle a utilisé « l'abstraction des contextes ». Cela consiste à supprimer, sur la matrice avec toutes les occurrences des mots cibles, celles qui occurrent à zéro. En effet, il existe beaucoup de mots cibles du texte n'ayant pas d'association les uns avec les autres. Cette technique permet donc d'obtenir des matrices réduites et de ne conserver que ce qu'il y a de plus pertinent.

Ces différents travaux peuvent ainsi constituer une base pour nos futures recherches, permettant de voir ce qui s'adapte le mieux à nos projets, et faciliter la définition de nos paramètres. En outre, ceux qui se rapprocheront sans doute le plus des nôtres sont ceux de Bernier-Colborne (2014) à l'Université de Montréal, le but étant de repérer automatiquement les relations sémantiques à partir de corpus spécialisés, constitués d'articles extraits de TALN (Traitement Automatique de la Langue Naturelle). Le chercheur a utilisé un modèle appelé Hyperspace Analogue to Language (HAL), qui ne tient pas compte des relations syntaxiques. Cela consiste à chercher, en se basant sur la cooccurrence, à quelle fréquence un mot cible cooccurre avec un autre, pour, par la suite, définir des groupes de relations paradigmatiques tels que des quasi-

[9] Le test de substituabilité : « le critère clé auquel les lexicologues ont recours pour identifier la plupart des relations de nature paradigmatique entre mots (Cruse 1986 ; Murphy 2003). » (Morlane-Hondère, Fabre 2012, p 1)

synonymes, des antonymes et des hyperonymes.

1.10 Les modèles SVD et PMI

Dans cette section, nous décrirons, sans trop entrer dans les détails, les deux méthodes distributionnelles dites plutôt classiques, que sont le Singular Value Decomposition et le Pointwise Mutual Information. Ces deux modèles sont exploités dans nos travaux et s'appliquent bien pour l'extraction de relations paradigmatiques. Nous nous appuierons sur les travaux de Levy et Goldberg (2014 et 2015).

Ces deux méthodes sont basées sur la méthode distributionnelle où nous avons une matrice *M*. Sur cette matrice chaque ligne *i* est un mot et chaque colonne *j* est un contexte où le mot apparaît, c'est-à-dire un cooccurrent du mot. Par exemple, nous pouvons avoir un mot *i*, comme *chien* sur la ligne, et sur chaque colonne *j* nous aurions ses cooccurrents tels que : *chat, croquette, niche* etc. L'intérêt est de calculer la mesure d'association *Mij* entre le mot et son contexte. Pour ce faire, nous pouvons utiliser le Pointwise Mutual Information, qui va mesurer l'association entre un mot *w* et son contexte *c*. Cette mesure calcule en fait le log du ratio entre la probabilité conjointe du mot et du contexte, c'est-à-dire à quelle fréquence ils occurrent ensemble :

$$PMI(w,c) = \log \frac{\#(w,c) \cdot |D|}{\#(w) \cdot \#(c)}$$

Ici, *D* représente l'ensemble des paires de mots et contextes. À la suite de ce calcul, on obtient des vecteurs qui sont dits « sparse », ce qui signifie que sur la matrice beaucoup de « cases » sont à 0 ou vides.

Levy et Goldberg (2015), notent que cette méthode permet d'obtenir de bons résultats. Néanmoins, il a également été prouvé que de travailler sur des dimensions réduites de vecteurs pouvaient aussi améliorer les performances. Le modèle Singular Value Decomposition apporte ainsi cette possibilité. En effet, grâce à un algorithme il va générer trois matrices à partir de la matrice principale *M*. Cela permet de travailler sur des vecteurs qui sont dits « denses ». Dans le domaine, la matrice PMI est souvent utilisée par le modèle SVD afin de « compresser » les

vecteurs et de créer les trois autres matrices.

3. Word2vec

Dans cette section, nous allons présenter une seconde méthode distributionnelle permettant d'extraire des relations entre les mots issues des travaux de Mikolov et *al.* (2013a) Mikolov et *al.* (2013b) et Mikolov et *al.* (2013c). Nous expliquerons comment cette technique fonctionne et décrirons l'outil word2vec qui lui est associé. Pour ce faire, nous nous appuierons sur les travaux de ces derniers ainsi que d'autres chercheurs qui ont repris cette méthode tels que Janod, Morchid, Dufour & Linares (2015), Périnet (2015) ou encore Bernier-Colborne (2015).

1.11 L'émergence d'un nouveau modèle

Mikolov et *al*. (2013), pour développer leur modèle, sont partis d'un premier constat, à savoir qu'il est impossible d'entrainer les modèles « n-gramme[10] classiques » existants sur des milliers de données. Mais depuis le développement du web, le volume des données s'est fortement accrue et il est donc possible de disposer d'immenses corpus. Ainsi, depuis quelques temps, les modèles de réseaux de neurones artificiels refont surface. En effet, ces modèles permettent l'apprentissage de données de manière très performante, mais à la condition qu'il y ait une quantité de données suffisante. Ces conditions réunies ont permis le développement de l'outil word2cec dont le but est d'utiliser des techniques pour l'apprentissage de vecteurs de mots provenant de grosses quantités de données afin de prédire un contexte ou simplement un mot. Cette méthode s'inscrit, en quelque sorte, dans la continuité de l'hypothèse distributionnelle, avec l'ajout de cette notion de prédiction. Nous allons expliquer son fonctionnement, sans entrer profondément dans la complexité du modèle.

Word2vec dispose d'une version en libre accès sur Internet, le code et des démonstrations peuvent être téléchargés sur Google[11]. Cet outil, fait ce que l'on appelle de l'apprentissage de représentation, plus précisément de distribution de mots. Ce n'est pas ce que l'on peut appeler de l'apprentissage profond, il s'agit de réseaux de neurones peu profonds, car il ne possède

[10] N-gramme : est une sous-séquence de *n* éléments construite à partir d'une séquence donnée.

[11] Word2vec : https://code.google.com/archive/p/word2vec/ (consulté novembre 2015)

qu'une seule couche cachée. Le logiciel va apprendre des représentations de mots, c'est-à-dire des vecteurs, afin de pouvoir prédire leurs contextes, ou inversement en fonction du contexte, pouvoir prédire un mot. Il permet ainsi de capturer aussi bien des régularités sémantiques que syntaxiques. Cet outil propose deux architectures possibles : nous avons d'une part, celle dite en « sac-de-mots » continus (continuous bag of words, CBOW) et d'autre part l'architecture en Skip-gram. Les deux disposent de trois couches : une entrée, une couche cachée et une sortie :

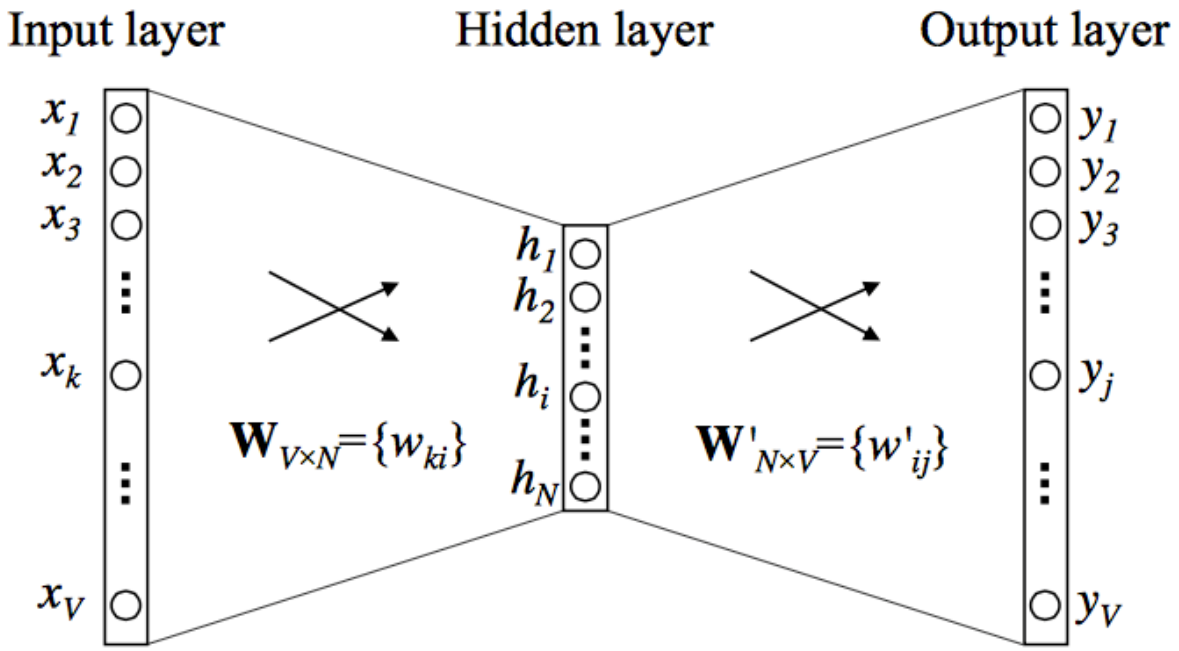

Figure 1 : Schéma de architecture word2vec.[12]

La figure 1 illustre l'architecture du modèle avec les trois couches. La couche d'entrée est soit composée d'un seul mot dans le cadre du modèle Skip-gram, soit d'un « sac de mots » pour l'architecture CBOW. On y voit aussi la couche cachée où sont projetés les mots de la couche d'entrée dans la matrice de poids. Cette matrice de poids est partagée par tous les mots. Enfin, la couche de sortie contient en quelque sorte la prédiction de ce que l'on recherche, composée également de neurones artificiels. Dans chacune des architectures, nous avons donc une

[12] Schéma de word2vec : Issu du site, http://www.folgertkarsdorp.nl/word2vec-an-introduction/, (consulté novembre 2015).

première phase, où le modèle va construire son vocabulaire en s'entrainant sur d'énormes données de textes afin de pouvoir ainsi apprendre les vecteurs.

Pour tenter d'éclaircir le fonctionnement, nous allons maintenant décrire les deux architectures existantes.

1.12 L'architecture « Skip-Gram »

Cette architecture est celle qui retiendra le plus notre attention, car elle fournit les meilleurs résultats et nous souhaitons utiliser celle-ci pour nos travaux. Effectivement, elle est plus adaptée pour les représentations sémantiques et les mots peu fréquents, que l'autre architecture, comme le soulignent, dans leurs travaux Janod et *al.* (2015). Le modèle Skip-Gram va en fait prédire le contexte d'où est issu un mot. Pour ce faire, dans la couche d'entrée, nous avons un vecteur contenant un seul mot.

Dans un premier temps, il va être projeté dans la couche cachée au moyen d'une matrice de poids. Cette matrice va, en fait, attribuer des poids à chacun des vecteurs afin d'être projetés dans les neurones de la couche cachée. Ensuite, avec une seconde matrice, ils vont être projetés dans la couche de sortie. Le vecteur d'entrée va être comparé à chacun des mots du contexte afin de savoir si les prédictions étaient correctes ou non. Pour cela, il va « s'autocorriger » par ce que l'on appelle la rétro-propagation du gradient. La rétro-propagation du gradient est une méthode qu'utilise word2vec automatiquement, pour rééquilibrer les poids de la première matrice, afin d'obtenir les meilleures prédictions possibles. En effet, le but étant de se rapprocher le plus possible du bon contexte pour un mot, pour cela elle va soit augmenter, soit diminuer les poids de chacun des vecteurs. Ainsi, la méthode va se baser sur ce qui s'appelle une fonction de coût, c'est-à-dire que plus elle va être proche de zéro, plus le modèle sera performant et aura donc les bonnes prédictions. De ce fait, concernant l'architecture Skip-gram, le mot d'entrée va se rapprocher le plus possible des mots présents dans le contexte.

1.13 L'architecture en « sac de mots » continus

Cette architecture est la même que la précédente, mais cette fois-ci c'est un réseau de neurones artificiels qui prédit un mot à partir de son contexte. Ainsi, la couche d'entrée représente la présence ou l'absence des mots dans le contexte de manière binaire. Si le mot est

absent du contexte alors on lui attribue un 1 et 0 s'il est absent. Ensuite, comme précédemment, le mot est projeté dans la matrice des poids du modèle et donc dans la couche cachée. Puis, dans un troisième temps, la somme ou moyenne de ces représentations, est passée dans la couche de sortie. Il va également procéder à la rétro-propagation du gradient afin de corriger sa représentation. Il va ainsi tenter de faire baisser la fonction de coût afin d'obtenir le « bon » mot à partir du contexte donné. Ce modèle est plus utilisé pour les relations syntaxiques.

1.14 Description des paramètres

L'outil word2vec permet en plus de choisir parmi ces deux structures, de pouvoir faire varier différents paramètres. Tout comme pour l'analyse distributionnelle, le choix des paramètres est très important. Dans un premier temps, il est possible de l'améliorer en utilisant ce que l'on appelle les calculs parallèles, fréquemment utilisés pour les modèles en réseaux de neurones aujourd'hui, avec des « multiple-CPU-machine[13] », permettant d'accroitre les performances de calculs. Il existe également plusieurs types d'algorithmes d'entrainement sur les corpus. Nous avons, d'une part, le « Hierarchical softmax » qui est meilleur pour les mots rares, et le « Negative sampling » qui lui est plus performant sur les mots fréquents avec des petites dimensions de vecteurs, selon Mikolov et *al.* (2013b). Il est donc également possible de faire varier la dimension des vecteurs et aussi le contexte des mots, c'est-à-dire la taille de la fenêtre que l'on veut choisir. Word2vec contient également une chaine de prétraitements assez complexe, améliorant les performances.

1.15 Résultats et travaux menés autour de Word2vec

Mikolov et *al.* (2013a) se sont aperçu dans leurs travaux que leur modèle pouvait rapidement proposer des résultats assez prometteurs. Dans un premier temps, ils ont obtenu les mêmes matrices de représentation de vecteurs que l'on peut avoir avec les méthodes distributionnelles où il est très facile par exemple, d'associer *France* avec *Italie*. Ils se sont également rendus compte qu'il existait d'autres relations sémantiques entre les mots à exploiter. Par exemple, en anglais, pour le mot *big*, *bigger* lui est similaire comme *smaller* l'est à *small*.

[13] Multiple-CPU-machine : réfère à la capacité d'un système à soutenir un processeur et/ou la capacité à répartir les tâches entre les processeurs. (Source : Wikipédia)

Ils ont donc amélioré leur système de manière à ce qu'il puisse répondre à des questions du type, « Quel est le mot qui est similaire à *small* de la même manière que *bigger* l'est à *big* ? » et ainsi créer des analogies. De là, ils ont pu augmenter leur jeu de questions, de façon à obtenir aussi bien des relations sémantiques que syntaxiques. Lors de certains de leurs travaux, en entrainant leurs vecteurs de mots sur un corpus de Google news, ils ont obtenu des résultats assez impressionnants, notamment certains permettant de réaliser des analogies entre une capitale et son pays.

Les modèles neuronaux ont été régulièrement repris, souvent à titre de comparaison avec d'autres modèles distributionnels plus « classiques », notamment avec les travaux de Bernier-Colborne (2015) qui vont nous intéresser ici. Le chercheur a utilisé word2vec afin d'identifier les relations paradigmatiques, telles que la synonymie, en construisant ce qu'il appelle des thésaurus distributionnels. Il s'agit en fait d'un graphe contenant les « k » plus proches voisins pour un mot (c'est-à-dire les mots les plus proches sémantiquement du mot cible où k est un nombre prédéfini de voisins que l'on souhaite obtenir). Cela permet de construire en quelque sorte un réseau pour un mot. Pour le réaliser, il a donc comparé deux modèles : « HAL », un modèle distributionnel à fenêtre graphique plus classique, et word2vec, afin de voir s'il obtenait des relations sémantiques différentes entre les mots. Les modèles présentaient quelques divergences et le meilleur rappel a ici été obtenu par word2vec.

Nous avons donc ici observé deux méthodes d'extraction de relations sémantiques. Nous pouvons voir que depuis l'apparition de l'outil word2vec en 2013, d'autres travaux ont été réalisés et ont montré que si les paramètres de la méthode distributionnelle plutôt classique étaient bien ajustés, on pouvait obtenir de très bons résultats, similaires à ceux de word2vec. Dans sa thèse concernant l'analyse distributionnelle appliquée aux textes de spécialité, Périnet (2015) compare également ces deux méthodes et obtient de meilleurs résultats avec les méthodes distributionnelles classiques qu'avec word2vec. Cependant, word2vec a prouvé qu'il était plus robuste et si les paramétrages sont mal ajustés, les résultats restent assez satisfaisants contrairement aux méthodes distributionnelles classiques telles que les PMI et SVD.

4. Paradigme de l'étude

Dans cet état de l'art, nous avons pu noter que la méthode distributionnelle a fait l'objet de nombreux travaux en TAL, au cours de ces dernières décennies. Elle a su faire ses preuves concernant l'extraction des relations sémantiques entre les mots.

Dans notre projet, nous souhaitons aborder une application nouvelle pour la méthode distributionnelle. En effet, nous l'utilisons afin d'organiser sémantiquement la liste de noms brute appartenant au LST constituée par Drouin (2010). Nous pensons donc qu'il serait intéressant d'utiliser cette approche afin de créer des classes sémantiques, de la documenter davantage et de ce fait la rendre de ce fait plus facile d'utilisation.

En effet, le LST, comme nous l'avons dit précédemment a un réel intérêt. D'une part, il permet d'apporter un soutien didactique dans la rédaction de méthodologie pour les apprenants d'une langue étrangère et également pour les locuteurs natifs, d'autre part, maitriser la manipulation du LST pour le TAL est crucial si l'on veut extraire et traiter des données appartenant aux sciences de l'information par exemple. Ainsi, constituer des classes sémantiques facilitera considérablement son usage, aussi bien pour les étudiants que pour les enseignants ou les chercheurs.

Pour parvenir à notre objectif, nous utilisons les trois méthodes distributionnelles que nous avons décrites précédemment, à savoir les PMI, les SVD et l'outil word2vec. Nous appliquons ces méthodes sur un corpus de référence, créé également par Drouin (2010). Nous évaluons nos résultats selon les trois méthodes et conservons word2vec qui fournit les meilleurs résultats.

L'objectif est d'observer en quoi la méthode distributionnelle peut nous permettre d'organiser sémantiquement une liste de mots et ainsi, à terme pouvoir apporter une réelle description sémantique aux noms de ce lexique. Nous construisons donc des classes sémantiques de noms, à partir des données résultantes de l'outil word2vec, auxquelles nous appliquons un algorithme de classification automatique. Cela nous permet d'obtenir trente classes regroupant les noms proches sémantiquement. Nous nous basons également sur des classes sémantiques de noms constituées manuellement par Hatier et *al.* (2014), afin d'évaluer nos sorties obtenues.

Enfin, pour chaque classe obtenue de façon automatique nous apportons une description sémantique manuelle en nous appuyant sur le dictionnaire hiérarchique de Polguère (2007).

Chapitre 2 :

Méthodologie

Dans de ce chapitre, nous montrerons les différentes étapes qui nous ont permis de mener à bien notre projet. L'objectif final de nos expérimentations est de pouvoir, en utilisant la méthode distributionnelle, organiser de manière sémantique une liste brute de noms.
Dans un premier temps, nous présenterons les deux corpus de travail qui nous ont servi à réaliser nos expérimentations, c'est-à-dire d'une part le corpus de travail et d'autre part celui d'évaluation. Puis, nous décrirons l'outil Hyperwords, permettant d'utiliser trois techniques de méthode distributionnelle, à savoir les PMI, SVD et word2vec que nous avons présentées précédemment. Nous expliquerons également de quelle façon nous avons exploité ces trois méthodes sur nos corpus. Ensuite, dans une troisième partie, nous montrerons de quelle manière nous avons utilisé la classification automatique sur notre liste de noms, ce qui nous a permis de générer trente clusters[14] de noms. Enfin, dans une dernière partie, nous expliquerons les techniques que nous avons utilisées afin d'analyser ces clusters et de quelle manière nous leur avons apporté une description sémantique. Pour cela, nous présenterons notamment les classes sémantiques du LST de Hatier et *al*. (2014) et Hatier et *al.* (2016) et le dictionnaire hiérarchique de Polguère (2007).

Partie 1 : Élaboration des corpus de travail

Dans cette première partie de méthodologie, nous présenterons les deux corpus que nous avons constitués pour nos différents travaux. Dans un premier temps, il s'agit du corpus d'étude pour l'ensemble de notre projet, issu des travaux de Drouin (2007). Puis, dans un deuxième temps, d'un second corpus provenant du projet Scientext (2010), utilisé pour l'évaluation des clusters que nous avons générés. Le premier corpus est utilisé afin d'organiser la liste des noms et créer les clusters. Le second est exploité uniquement dans le but de savoir quelle technique de méthode distributionnelle apporte les meilleurs résultats.

1.1 Le corpus d'étude

Dans ses travaux concernant l'extraction du Lexique Scientifique Transdisciplinaire (LST), Drouin (2007) a construit un corpus scientifique transdisciplinaire ainsi qu'une liste de mots

[14] Cluster : « paquet » contenant un ensemble de données homogènes, obtenues par des méthodes statistiques. (Source : Wikipédia)

appartenant au LST. Nous allons dans la partie suivante décrire comment il a procédé à l'élaboration de ce corpus ainsi qu'à l'extraction des noms du LST, constituant la liste que nous souhaitons organiser dans le cadre de ce projet.

1.1.1 Description du corpus

Le corpus est issu d'un projet s'inscrivant dans la suite des travaux de Drouin (2007), qui ont permis de constituer le LST. Celui-ci avait pour but de montrer qu'il est possible, à partir d'un corpus bilingue (anglais et français), de générer une liste de collocations transdisciplinaires, autour des mots appartenant au LST.

Pour y parvenir, Drouin (2010) a construit deux corpus : d'une part, un corpus transdisciplinaire et, d'autre part, un corpus de référence. Nous nous intéresserons dans notre cas au premier uniquement. Celui-ci est constitué de thèses et d'articles scientifiques et recouvre neuf disciplines scientifiques qui sont : l'anthropologie, la chimie, l'informatique, l'ingénierie, la géographie, l'histoire, le droit, la physique et la psychologie. Chaque discipline compte environ 200 000 mots dans les thèses et autant dans les articles, ce qui constitue exactement 5 736133 mots. À l'origine, pour les fins du projet, le corpus comptait autant de mots en anglais. Dans notre cas nous avons seulement utilisé la partie en langue française.

Disciplines	**Articles**	**Thèses**
Anthropologie	233 699	254 956
Chimie	213 239	191 034
Informatique	207 445	247 612
Ingénierie	238 868	145 252
Géographie	227 715	220 653
Histoire	245 014	320 267
Droit	234 784	374 830
Physique	214 546	197 867
Psychologie	245 292	360 473

Tableau 2 : Nombre de mots par domaine.

Dans le cadre de nos travaux, pour améliorer les résultats, nous avons utilisé une version lemmatisée du corpus, qui a été réalisée par Drouin (2010) avec l'outil TreeTagger (Schmid 2004). Cet outil permet pour chaque mot du corpus d'obtenir son lemme et de l'étiqueter grammaticalement. Le corpus que nous utilisons dans nos expérimentations contient uniquement les lemmes.

Voici ci-dessous un extrait du corpus lemmatisé :

ce étude préliminaire conduire en principe à le fouille de sauvetage du site qui avoir priori ne être pas sauvegarder (sauf cas exceptionnel) , afin de documenter scientifiquement le donnée relatif au vestige d'occupations [15] humain il exister un troisième type d'intervention qui se dérouler après un pillage , un tentative de pillage , un début de destruction intentionnel ou non de site (vandalisme , accident , catastrophe naturel , etc .

1.1.2. La création de la liste de noms du LST

Au cours de ce même projet, Drouin (2007) a extrait une liste de mots du LST afin de pouvoir par la suite en générer les collocations.
Pour ce faire, il a extrait les mots de façon automatique en les sélectionnant selon trois critères :
Dans un premier temps, le chercheur a sélectionné les mots selon une mesure de spécificité, celle de Lafon (1980), qui permet de calculer la spécificité d'une forme sur la base de sa fréquence dans le corpus, par rapport à un corpus de référence. Ainsi, les mots retenus devaient avoir une spécificité positive, le choix était parmi : positif, négatif et neutre. Cette mesure signifie que la fréquence du mot est plus haute que celle qui aurait été attendue dans le corpus de référence. Nous parlons alors de surreprésentation.
Dans un deuxième temps, les mots sont sélectionnés en fonction de leur répartition à travers le corpus. Pour cela, Drouin (2010) a découpé le corpus en plusieurs sous-corpus de différentes tailles et s'est intéressé à la fréquence relative de chaque mot et non à la fréquence absolue. Le mot est ajouté à la liste du LST s'il apparaît dans tous les différents sous-corpus.

[15] Occupations : Dans ce cas présent, le mot n'est pas étiqueté car Tree-Tagger ne le connaît pas.

La liste qui nous intéresse dans notre projet est uniquement celle de la catégorie des noms, qui compte 461 formes et a été légèrement revue depuis 2010. Elle est composée de noms tels que : *modèle, mouvement, mécanisme, schéma, augmentation, mémoire, règle, schéma, liste, minimum, valeur, formule,* etc.
La version complète est disponible sur le site internet de l'Observatoire de Linguistique Sens-Texte.[16]

1.2 Constitution du corpus d'évaluation

Afin de pouvoir réaliser une évaluation des clusters que nous avons générés dans la suite du projet, nous avons procédé à la constitution d'un second corpus scientifique transdisciplinaire : le corpus d'évaluation. En effet, l'outil Hyperwords génère des matrices qui sont utilisées par l'algorithme de classification automatique. Nous voulions évaluer les clusters que nous obtenions selon les trois méthodes proposées par l'outil, dans le but de ne conserver que la plus performante. De ce fait, nous disposions d'un programme permettant de calculer la correspondance entre des clusters et des classes sémantiques. Nous avions également les classes sémantiques des noms du LST réalisées par Hatier et *al*. (2014) et Hatier et *al*. (2016). L'objectif était donc d'utiliser ces calculs afin de voir quelle méthode distributionnelle permettait d'obtenir la meilleure correspondance entre les clusters générés et les classes sémantiques. Ainsi, nous préférions réaliser ces clusters avec le corpus ayant permis d'extraire les noms du LST de Hatier et *al.* (2014), à savoir le corpus Scientext. Cela explique pourquoi nous avions besoin de ce second corpus.
Nous allons, dans cette partie, présenter le corpus d'évaluation et expliquer les différentes étapes qui nous ont permis de le constituer.

1.2.1. Description du corpus Scientext

À l'origine, le corpus que nous avons utilisé et modifié est issu du projet appelé Scientext (2010). Ce dernier est un projet qui a été réalisé par différents laboratoires tels que le LIDILEM

[16] Lexique scientifique transdisciplinaire de l'observatoire de Linguistique Sens-Texte : http://olst.ling.umontreal.ca/lexitrans/ (consulté avril 2016)

(Laboratoire de Linguistique et de Didactique des Langues Etrangères et Maternelles), à Grenoble et le LiCoRN (Linguistique de corpus) de Lorient. Scientext a été lancé dans le but de pouvoir mettre à disposition des chercheurs et des étudiants des corpus regroupant des écrits scientifiques sur une base Internet. Cette base dispose donc d'un corpus qui est constitué de textes en anglais et français et regroupe trois familles de disciplines qui sont : les sciences humaines, les sciences expérimentales et les sciences appliquées. Le corpus contient également différents genres textuels comme par exemple des thèses et des articles.

Nous avons quant à nous choisi de sélectionner une partie de ce corpus afin de pouvoir en constituer un plus adapté. En effet, nous souhaitions obtenir un corpus scientifique transdisciplinaire ne contenant pas les sciences expérimentales et les sciences appliquées mais les sciences humaines et sociales, tel que l'avait fait Hatier (2013) pour constituer le LST. De ce fait, nous avons conservé dix disciplines appartenant aux sciences humaines qui sont : l'anthropologie, l'économie, la géographie, l'histoire, la linguistique, la psychologie, les sciences de l'éducation, les sciences politiques, les sciences de l'information et la sociologie. Ces différents textes sont rédigés en français et sont constitués uniquement d'articles.

Les textes que nous avons récupérés à partir de la base Scientext, sont en XML TEI[17]. Les métadonnées sont séparées et il y avait différents niveaux de structures marqués par des balises dans les écrits. Nous souhaitions travailler sur un corpus sans annotation et ne pas conserver les métadonnées. Nous avons donc dû procéder à la conversion des textes au format XML TEI de Scientext en textes au format texte.

1.2.2. Conversion du corpus XML au format txt

a. Utilisation du langage Python

Le langage Python, crée dans les années 1980 aux Pays-Bas par Guido Van Rossum, dispose de bibliothèques spécialisées permettant de faciliter ce genre de travaux, à savoir convertir des textes XML en texte. Il est en effet réputé pour être utilisé comme langage de script permettant

[17] XML TEI (Extensible Markup Language Text Encoding Initiative) : XML : langage balisé extensible, métalangage de description de textes. TEI, vise à définir des recommandations d'encodages textuelles. (Source : Wikipédia)

d'accomplir les tâches simples mais fastidieuses. Nous pouvons noter que la plupart des travaux effectués dans le cadre de ce mémoire ont été réalisés avec des programmes en langage Python. De plus, Python dispose d'une bibliothèque appelée Beautiful Soup permettant de « parser [18]» les données HTML ou XML dans les différents fichiers. Elle a été créée par Leonard Richardson. Nous avons utilisé une version de 2013.
Dans le cas présent, le script créé prend en entrée tous les fichiers au format XML TEI contenus dans le répertoire courant. Il permet d'aller chercher uniquement le texte qui nous intéressait, c'est-à-dire qui est contenu dans les balises afin de les mettre dans un fichier de sortie texte au format texte (txt) en utilisant un encodage UTF-8[19].

Voici un exemple, issu du corpus d'évaluation :

* Ce constat empirique renvoie à la double question que pose cet article : comment le risque participe-t-il à la production des rapports sociaux de sexe, et comment le genre<ref target="#_ftn9">[9] </ref> façonne-t-il les rapports subjectifs au risque ? *

Cependant, les formats d'encodage des corpus originaux XML TEI comportaient quelques « variations ». En effet, les apostrophes n'étaient pas toujours encodées de la même manière, il a donc fallu également procéder à un travail d'homogénéisation.

b. Utilisation de la bibliothèque Beautiful Soup.

Beautiful Soup permet ainsi de récupérer toutes les données qui nous intéressent, de manière assez simple en évitant des expressions régulières qui peuvent rapidement devenir fastidieuses. Le parseur utilisé, « html.parser », pour notre constitution de corpus est celui pour le HTML. Beautiful Soup, permet ainsi de parcourir toutes ces balises grâce à une fonction : soup.find_all et pour chaque balise trouvée, de ne récupérer que son contenu. Enfin le script écrit les données récupérées dans un fichier texte de sortie afin de conserver les deux versions XML TEI et texte.

[18] Parser : issu de l'anglais, synonyme de « analyser ».

[19] UTF-8 (Universal Character Set) : codage de caractères informatiques.

Nous avons réalisé notre corpus scientifique transdisciplinaire d'évaluation, contenant dix disciplines et 3 679 136 formes en français. Ce dernier nous servira donc, au cours de notre projet, pour l'évaluation de nos clusters.

Partie 2 : Hyperwords

Dans cette seconde partie, nous présenterons comment nous avons exploité l'outil Hyperwords, permettant d'utiliser la méthode distributionnelle, sur notre corpus scientifique transdisciplinaire. Pour cela, nous expliquerons tout d'abord, le fonctionnement de l'outil, puis, comment nous l'avons adapté à notre corpus et enfin, quels paramètres nous avons sélectionnés pour les fins notre projet.

Néanmoins, il convient de noter que notre choix s'est porté sur cet outil car il permet d'exploiter trois techniques de méthodes distributionnelles : Pointwise Mutal Information (PMI), Singular Value Decomposition (SVD) et word2vec, c'est-à-dire deux méthodes plutôt standard et une seconde basée sur les réseaux de neurones artificiels. Hyperwords ne contient qu'un seul script permettant d'utiliser les trois méthodes avec des paramètres relativement faciles à moduler, ainsi que des scripts d'évaluation que nous pouvions adapter à notre étude afin d'analyser nos résultats.

Cela nous permettait donc de faciliter la comparaison entre les trois méthodes et de ne conserver que celle générant les meilleurs résultats.

2.1 Description de l'outil Hyperwords

Nous commencerons par décrire l'outil Hyperwords, que nous avons utilisé, afin de pouvoir exploiter la méthode distributionnelle sur notre corpus.

Hyperwords a été mis au point par Omer Levy en 2015 et il est possible de récupérer le code source libre de l'outil sur le site BitBucket (site de gestion de versions collaboratives [20]). Nous pouvons noter que cet outil est très récent et de ce fait peu documenté.

[20] BitBucket : https://bitbucket.org/omerlevy/hyperwords (Consulté février 2016)

2.1.1. Un outil qui exploite l'hypothèse distributionnelle

Hyperwords propose, par le biais de scripts Shell [21] et Python, d'utiliser la méthode distributionnelle en exploitant trois techniques, ce qui s'adaptait parfaitement à nos recherches. Pour ce faire, il faut télécharger l'outil qui comprend les différents scripts de traitement du corpus ainsi que le script de la chaine de traitement. Un corpus de test est également disponible pour l'anglais, ainsi que deux scripts d'évaluation des résultats aussi réservés pour la langue anglaise.

Avant d'expliquer le fonctionnement d'Hyperwords, nous voulions revenir brièvement sur ce que sont les modèles PMI, SVD et l'architecture de word2vec exploités avec cet outil. Pour cela, nous nous appuierons sur les travaux de Levy & Goldgerb (2014).

Les méthodes PMI, Pointwise Mutal Information et SVD, Singular Value Decomposition ont pour objectif de calculer la proximité sémantique entre les mots. Ces dernières reprennent en fait, les principes de la méthode distributionnelle, c'est-à-dire une matrice *M* où chaque ligne *i* est un mot et chaque colonne *j* est un contexte où ce mot apparaît. Les mots sont en fait représentés par des vecteurs.

En ce qui concerne les PMI, il s'agit d'un calcul mesurant l'association entre un mot et son contexte. Pour y parvenir, le modèle va calculer le « log » du ratio entre leur probabilité conjointe, ce qui signifie la fréquence à laquelle ils apparaissent ensemble. Concernant les PMI, les vecteurs sont dits « sparse[22] », ce qui veut dire qu'il y a beaucoup de « cases » vides dans la matrice.

Le modèle SVD, va quant à lui permettre de travailler avec des vecteurs dits « denses », ayant des dimensions plus petites. Il va pour cela diviser la matrice (*M*) en produit de trois matrices.

SGNS[23], c'est-à-dire Skip-Gram with Negative Sampling, est une méthode inspirée des réseaux de neurones artificiels, permettant d'exploiter la méthode distributionnelle d'une autre manière. L'algorithme utilisé s'appelle le « negative sampling » et a pour but, tout comme la méthode distributionnelle, de montrer que des mots qui ont des contextes similaires ont plus de probabilités d'être dans les mêmes ensembles. Pour cela, il va, pour une paire mots-contextes

[21] Shell : langage de script

[22] Sparse : peu dense (Source : Termium)

[23] SGNS : Modèle tiré de l'outil Word2vec.

ciblée, maximiser la probabilité que cette paire appartienne à un ensemble de paires préalablement construit, pendant qu'il va maximiser la probabilité que des exemples « négatifs » de paires n'y appartiennent pas. Ces exemples négatifs sont prélevés de manières aléatoires et constituent des paires qui ne sont en fait pas observées.

2.1.2 Chaine de traitement

L'utilisation d'Hyperwords nécessite l'installation de modules supplémentaires qui sont NumPy, SciPy, Sparsesvd et Docopt. Ces derniers ont pour but de prendre en charge les manipulations des différentes matrices. De plus, il faut également télécharger le dossier contenant les scripts de word2vec, permettant l'installation et l'utilisation de l'outil.
Il est ensuite possible de lancer Hyperwords par le biais d'une chaine de traitement. Cette dernière se décompose en sept étapes différentes que nous allons décrire ici. La chaîne de traitement permet d'appeler les scripts pour parvenir à terme à générer les différentes matrices pour les trois modèles différents. Voici brièvement les sept étapes ci-dessous :

1 Nettoyer le corpus afin de supprimer tous les caractères non alphanumériques du texte d'entrée.
2 Extraire une collection des contextes de mots des différentes paires du corpus.
3 Appel d'un script afin de compter les occurrences des différentes paires.
4 Créer le vocabulaire avec la distribution unigramme [24] des mots et leur contexte.
5 La chaine de traitement va dans une cinquième étape créer une matrice PMI à l'aide des paires comptées précédemment et le vocabulaire créé. Cette étape n'est nécessaire que pour les PMI et SVD, tout comme pour la sixième.
6 Factorisation de la matrice PMI afin d'obtenir en sortie trois matrices denses dites « NumPy », selon le modèle SVD.
7 La chaîne de traitement génère les vecteurs selon la méthode SGNS. Pour cela, il utilise un script extérieur appartenant à l'outil word2vec. Celui-ci prend en entrée les différentes paires et le vocabulaire, puis, il va à partir de cela créer la matrice SGNS.

[24] Unigramme : séquence de mots sans historique, selon le modèle de Markov caché.

2.1.3 Vers une version française d'Hyperwords

Hyperwords est un outil qui a été développé afin de traiter des corpus anglais. Cela nous a confrontée à plusieurs problèmes, avant de pouvoir le prendre en main et l'adapter à nos corpus français.

Dans un premier temps, nous avons dû modifier le script de nettoyage du corpus. Celui-ci permettait de transformer les caractères en ASCII, puis de les passer en minuscules et par la suite éliminer tous les caractères non alphanumériques. Nous avons donc revu ces traitements, changé les différentes commandes afin de garder les caractères accentués ainsi que les « ç ». De plus, nous ne transformons plus le texte en caractères ASCII mais conservons un format UTF-8 dans les différents fichiers et nous supprimons également les caractères non alphanumériques.

Un autre problème était le découpage en tokens [25] des mots. Effectivement, cela se fait au moment de l'appel du script Python qui permet d'extraire les différentes paires à partir du corpus nettoyé. Dans la version originale, les mots étaient uniquement découpés en fonction des espaces. Nous voulions changer cela afin de pouvoir également découper les mots selon les apostrophes. Pour ce faire, nous avons ajouté une étape dans le nettoyage du corpus, consistant à remplacer toutes les apostrophes par des espaces. Le découpage dans le script constituant les différentes paires, découpe de ce fait ces mots, comme par exemple pour *l'anthropologie* qui devient deux mots *l* et *anthropologie*. Enfin, nous avons décidé de supprimer tous les chiffres car ils ne nous étaient pas utiles. Néanmoins, nous avons fait le choix de conserver les tirets, «-», dans notre corpus de travail et donc garder les formes entières comme par exemple *arc-en-ciel*.

2.2 Utilisation d'Hyperwords sur notre corpus.

Après ces différentes étapes, nous avons donc pu lancer la chaîne de traitement sur le corpus de travail scientifique transdisciplinaire.

[25] Découpage en tokens : La « Tokenization » en anglais, est le processus de découpage d'un flux textuel en éléments significatifs tel que des mots, phrases, symboles appelés tokens.

Afin de pouvoir interroger les différents résultats, nous avons construit trois scripts en Python qui lisent les différentes matrices et permettent d'obtenir pour un mot ses neuf plus proches voisins, selon les trois méthodes PMI, SVD et SGNS (le chiffre neuf est un paramètre par défaut). Pour ce faire, nous avons utilisé la fonction « closest » mise à notre disposition par les modules d'Hyperwords.

Ces scripts nous permettent de comparer nos sorties quand nous faisons varier les différents paramètres disponibles. Nous allons détailler comment nous avons sélectionné les paramètres, qui nous ont permis d'obtenir les résultats les plus performants pour chacun des modèles.

2.2.1 Le choix des paramètres

L'outil Hyperwords contient un grand nombre de paramètres qu'il peut être intéressant de faire varier, dans le but d'améliorer les sorties obtenues. De ce fait, nous avons modulé ceux qui ont prouvé qu'ils pouvaient apporter des différences significatives dans des travaux précédents de Levy et Goldberg (2014.

Tout d'abord, concernant la méthode SGNS, les paramètres que nous avons changés sont :

1. Le nombre minimal d'occurrences d'un mot
2. La taille de la fenêtre
3. Le nombre d'itération

Le nombre minimal d'occurrence d'un mot dans le corpus est celui déterminant si un mot peut appartenir à la liste du vocabulaire. Le paramètre est appelé « min count » dans l'outil Hyperwords.

De plus, nous avons fait varier la taille de la fenêtre de mots pris en compte pour créer les paires, c'est-à-dire combien de mots avant et après le mot cible sont observés par l'algorithme. Il s'agit du paramètre « win ».

Ensuite, nous avons modulé le nombre d'itération, ce qui signifie, combien de fois le modèle parcourt le corpus pour réaliser son apprentissage. En effet, celui-ci fonctionne sur un modèle de réseau de neurones ce qui nécessite une phase préalable d'apprentissage sur le corpus. Le paramètre est appelé « iters » dans Hyperwords. Nous avons donc généré dix sorties différentes pour le modèle SGNS.

Ensuite, nous avons utilisé une approche similaire pour les matrices SVD et PMI afin d'obtenir les résultats les plus pertinents possibles. Nous avons aussi fait varier la taille de la fenêtre dans laquelle est contenu le mot cible.
Néanmoins, nous avons souhaité laisser le paramètre « min count », du nombre minimal d'occurrences à 1 pour la création du vocabulaire puisqu'il a déjà été montré dans des travaux de Levy et Goldberg (2014) qu'il s'agit de la valeur qui permet d'obtenir les résultats les plus performants pour ces deux modèles.

Il a été également prouvé dans des travaux de Levy, Goldberg et Dagan (2015) qu'il pouvait être pertinent de faire varier un paramètre appelé « eig » (Eigenvalue Weighting) se rapportant aux poids des mots, pour le modèle SVD. En effet, ce dernier était réglé par défaut à 0,5 et les résultats générés étaient peu pertinents. La précision calculée sur la sortie était aux alentours de 20%, ce qui nous a contraint à passer le paramètre à 1.

Nous avons donc obtenu trois sorties différentes pour les modèles PMI et six pour les SVD.

2.2.2 Paramètres d'évaluation

Afin de pouvoir déterminer quels étaient les paramètres, qui permettaient d'obtenir les résultats les plus performants pour les trois modèles, nous avons procédé à une petite évaluation de chacune des sorties obtenues en variant les paramètres. Nous avons calculé la précision pour chacun des résultats générés, c'est-à-dire le nombre de réponses justes divisé par le nombre de réponses produites par le système.
Le calcul a été réalisé avec six mots différents, pour lesquels nous obtenions à chaque fois ses neuf plus proches voisins, ce qui signifie donc cinquante-quatre réponses par modèle testé.
Pour déterminer si un mot en sortie est considéré comme « juste », nous avons adopté une approche manuelle. En effet, il semblait difficile d'observer les résultats obtenus de manière assez fine avec uniquement des analyses statistiques. De ce fait, nous avons mis en place quelques critères manuels pour observer nos résultats, qui sont les suivants :

- Un mot est considéré comme du bruit s'il n'est pas en français. Effectivement, le corpus contient des passages en anglais, espagnol ou allemand par exemple.
- S'il s'agit d'un nombre tel que « 2005)[20 ». Ou d'erreur « graphique » comme « solides[9 » .

- Un mot est considéré comme « juste » s'il a un lien paradigmatique avec le mot cible. Dans les sorties observées il s'agissait essentiellement de quasi-synonymes. Par exemple : pour *analyse* nous obtenions des noms tels que : *étude*, *approche*, *description* etc.
- Enfin dans quelques cas plus rares, il y avait des liens syntagmatiques (de cooccurrence) entre le mot cible et les différentes sorties. Dans ce cas présent le mot était également considéré comme juste. Par exemple avec : *formuler* et *hypothèses*.

Voici une sortie présentant les neuf mots obtenus à l'aide des matrices SGNS pour le mot *science*. Les formes qui sont considérées comme du bruit sont en italiques. En effet, *finalités* n'entretient aucune relation avec le mot science et *sic* est considéré comme une erreur puisqu'il n'a pas de signification. Enfin, *psychologie*, *biologie*, *anthropologie*, *ethnographie*, *philosophie*, *neuroscience* et *mathématique* sont considérés comme justes puisqu'ils sont des hyponymes de sciences.

psychologie
finalités
biologie
anthropologie
ethnographie
sic
philosophie
neuroscience
mathématique

Voici ci-dessous les tableaux récapitulatifs présentant la mesure de la précision selon chaque modèle et chaque paramètre que nous avons fait varier.

PMI	Win 1	Win 2	Win 5
Min-count 1	79,63%	**81,48%**	62,96%

Tableau 3 : Précision obtenue selon la variation des paramètres pour le modèle PMI.

SVD	Win 1	Win 2	Win 5
Min-count 1 eig = 1	25,93%	**51,85%**	30,71%

Tableau 4 : Précision obtenue selon la variation des paramètres pour le modèle SVD.

NS	Win 2	Win 5
Min-count 1 Itération 5	51,85%	42,59%
Min-count 5 Itération 5	61,11%	42,59%
Min-count 1 Itération 8	**68,51%**	51,85%
Min-count 10 Itération 8	64,81%	35,18%
Min-count 5 Itération 8	64,81%	42,59%

Tableau 5 : Précision obtenue selon la variation des paramètres pour le modèle SGNS.

À la suite de cette série de traitements d'Hyperwords, les paramètres qui ont été retenus pour les trois modèles sont les suivants :

Pour le modèle PMI, le meilleur résultat a été obtenu avec la fenêtre de contexte réglée à deux mots, où la précision est de 81,48%.

Concernant le modèle SVD, les résultats ont également été améliorés lorsque la fenêtre de contexte était paramétrée à deux mots et le paramètre « eig » égal à 1 comme mentionné précédemment. La précision la plus haute est ici de 51,85%.

Enfin, les paramètres retenus pour le modèle SGNS sont : 8 itérations pour l'apprentissage sur le corpus, un minimum d'une occurrence pour un mot afin qu'il appartienne au vocabulaire,

ainsi qu'une fenêtre de contexte de deux mots. Cela a ainsi permis d'atteindre une précision de 68,51% pour ce modèle.
Les meilleurs résultats ont donc été atteints avec le modèle PMI, qui obtient la plus haute précision pour cette tâche d'évaluation.

PARTIE 3 : Constitution des clusters sémantiques

Un des objectifs de notre projet, était de pouvoir parvenir à constituer des clusters à partir de la méthode distributionnelle. De ce fait, après avoir fait une brève description du « clustering », nous allons présenter dans cette troisième partie les différents outils que nous avons exploités pour constituer les clusters. Ensuite, nous expliquerons la méthodologie employée pour choisir la méthode (PMI, SVD, SGNS) la plus efficace pour la construction des clusters.

3.1 Utilisation de deux outils : Hclust et K-Means.

Nous allons dans un premier temps, procéder à une brève explication du clustering puis nous présenterons les deux outils que nous avons utilisés, appelés Hclust et K-means.

3.1.1 Le clustering

Le clustering est aussi appelé « partitionnement de données » en français. Il s'agit de méthodes statistiques qui permettent de générer des ensembles homogènes et disjoints à partir d'analyses de données. Ces ensembles sont composés par des éléments qui partagent des caractéristiques communes et sont définis par des mesures de distances entre les différents objets.
Ce système de clustering est beaucoup employé de nos jours, notamment par les entreprises pour ce qui s'appelle la classification automatique. En effet, elle est par exemple utilisée pour le marketing lorsque l'on veut rechercher les différents profils constituant une clientèle ou encore créer un panel de représentation. Cela peut également s'appliquer à de nombreux domaines tels que le monde médical afin de pouvoir regrouper ensemble, tous les patients qui ont les mêmes pathologies et leur appliquer les mêmes traitements etc.

Dans le cadre de notre projet, nous avons décidé d'utiliser un clustering utilisant des algorithmes basés sur la méthode distributionnelle appelés Hclust et K-means. Ils utilisent les matrices générées par Hyperwords afin de générer des clusters. Nous souhaitons de ce fait, obtenir des regroupements contenant des noms et leurs « plus proches voisins sémantiques » afin de pouvoir organiser notre liste de noms brute.

3.1.2. Hclust

Hclust est une bibliothèque disponible sous Python 2.7 et qui a été développée sous Windows 7. Cette dernière nécessite l'installation préalable des modules qui sont Matplotlib, NumPy et SciPy afin de pouvoir manipuler les différentes matrices.

Hclust contient des fonctions Python qui permettent de faire de la classification hiérarchique, c'est-à-dire répartir des objets au sein de différentes classes. Cette bibliothèque permet notamment de générer des « grappes » hiérarchiques à partir de matrices de distances, fournir des statistiques de calcul en utilisant des clusters ou encore de visualiser des clusters avec dendrogrammes[26].

En ce qui concerne notre projet, nous avons utilisé une version adaptée de Hclust par G. Bernier-Colborne, sous Python. Nous avons créé un script Python intégrant la fonction Hclust, qui prend en entrée un fichier contenant une liste de mots et retourne selon les trois modèles, PMI, SVD et SGNS les différents clusters. En effet, le but étant de créer les clusters en fonction de la proximité des mots calculée préalablement par Hyperwords. La fonction de Hclust qui est appelée permet également de définir combien de clusters nous souhaitons obtenir en sortie.

3.1.3. K-means

K-means est à l'origine un algorithme de MacQueen (1967), permettant de faire de manière simple de l'apprentissage non supervisé afin de générer des clusters, où *K* est en fait le nombre de clusters. Cet algorithme permet de rapprocher les mots qui ont des similitudes en exploitant des vecteurs. Pour ce faire il est seulement nécessaire de lui préciser K.

[26] Dendrogramme : Représentation graphique, sous forme d'un arbre, des regroupements successifs opérés par agrégation binaire d'éléments dans la classification ascendante hiérarchique. (Source : Termium)

Dans le cadre de notre projet, nous avons utilisé une version disponible sur la plateforme web Kaggle[27], qui propose un code source libre en langage Python. Ce script fait appel à la bibliothèque Scikit-learn qui dispose d'une fonction « KMeans ». La fonction prend comme paramètre les vecteurs de mots qui sont contenus dans notre cas dans les différentes matrices préalablement construites par Hyperwords. Il convient également de lui préciser K le nombre de clusters.
De la même manière qu'avec Hclust, nous avons généré trois sorties différentes de clusters selon nos modèles PMI, SVD et SGNS.

3.2 Méthode d'évaluation des résultats de Hclust et Kmeans

Nous allons montrer, parmi les modèles PMI, SVD et SGNS lequel permet d'obtenir les meilleurs résultats pour les clusters. Pour cela, nous expliquerons la démarche évaluative que nous avons adoptée. En effet, déterminer la pertinence des résultats de manière manuelle ne peut pas être neutre. De ce fait nous avons essayé d'automatiser la méthode en faisant appel à une fonction d'évaluation de clusters.

3.2.1. Classes sémantiques

Afin de pouvoir déterminer quelle était la méthode distributionnelle que nous souhaitions conserver pour la suite de nos travaux, nous avons décidé de comparer les clusters que nous obtenions en sortie, avec des classes sémantiques construites de manière manuelle.
Pour cela, nous avons utilisé le corpus d'évaluation que nous avons constitué, comme nous l'avons expliqué en première partie. Dans un premier temps, nous avons lancé la chaine de traitement d'Hyperwords avec ce corpus, afin que les trois modèles, PMI, SVD et SGNS puissent générer leurs matrices à partir des mots de ce corpus.

Dans un second temps, nous avons utilisé la liste de noms du LST constitués par Hatier, Tutin, Jacques, Jacquery et Kister (2014) pour élaborer nos différents clusters. Cette liste a été réalisée

[27] Kaggle : plateforme web organisant des compétitions en sciences de données (Source : Wikipédia).

par les chercheurs, membres du LIDILEM, dans le cadre du projet « Extraction et traitement sémantique des noms simples du lexique scientifique transdisciplinaire ». Les mots de la liste ont été extraits par Hatier (2013) et répondaient aux mêmes critères de spécificité et répartition que Drouin (2007), à savoir une spécificité positive, un seuil d'occurrence supérieur à 50 parmi les 100 tranches qui le composent et un seuil de disciplines supérieur à 5 parmi les 10 disciplines scientifiques du corpus. Cette liste a notamment été révisée manuellement par des juges experts, afin de filtrer le bruit. La liste des noms du LST qu'ils ont réalisée contient 493 noms.

Ce qui nous a incité à utiliser cette liste de noms du LST est, que lors de ce projet, Hatier et *al*. (2014) ont apporté une description sémantique fine des noms de la liste. En effet, ils ont constitué 18 classes sémantiques également affinées en 68 sous-classes sémantiques, des noms, de façon manuelle. Néanmoins, il convient de nuancer cela, en effet nous ne considérons pas cette liste comme un « standard » mais plutôt comme une base fiable à laquelle nous pouvons comparer nos résultats. De plus, avant d'apporter une description manuelle à leurs classes sémantiques, Hatier al. (2014) ont également fait des expérimentations semi-automatiques. Ces classes ne sont donc pas totalement manuelles, bien que nous utilisons cette appellation tout au long de nos travaux pour plus de clarté.
Afin de constituer ces classes sémantiques des noms, ils ont déterminé quatre classes : support, processus, artefact et relation. Chacune contenait trois noms dits prototypes de la classe. Les noms ont été automatiquement affectés à une classe en fonction de quatre critères définis : la fréquence du nom par partie textuelle (résumé, introduction…), les relations de déterminations (possessif, démonstratif...), les relations avec les prépositions (*en*, *dans*…) et les relations lexico-syntaxiques les plus fréquentes (gouverneur, dépendant). Par la suite, ces classes ont été observées manuellement, analysées et affinées afin d'obtenir les 18 classes définitives.

Ensuite, avons choisi de procéder au clustering en lançant les algorithmes Hclust et K-means sur la liste du LST. À partir de cela, nous pouvions comparer les clusters générés avec leurs classes sémantiques. Nous avons donc préalablement fixé le nombre de clusters à 18 dans les fonctions Hclust et K-means pour obtenir autant de classes sémantiques que de clusters. Nous avons souhaité travaillé sur les classes et non les sous-classes car ces dernières étaient trop fines et nous aurions eu plus de difficulté à obtenir des similitudes entre nos clusters et les classes manuelles.

3.2.2. Résultats et conclusion de l'évaluation

Tout d'abord, en observant les premières sorties de K-means, nous avons préféré abandonner cette fonction et nous concentrer uniquement sur la fonction Hclust. En effet, il n'a pas été nécessaire d'essayer de l'évaluer pour s'apercevoir que les résultats n'étaient pas pertinents. K-means, selon les trois méthodes, fournissait des clusters contenant des mots qui partageaient des liens sémantiques relativement flous. Nous pouvons expliquer cela car il a été montré, comme l'explique la documentation sur le site de Kaggle, que de bons résultats sont obtenus avec cet outil essentiellement s'il y a peu de mots par cluster. Effectivement, il est signifié qu'il est préférable d'avoir environ cinq mots par clusters. Cependant, nous avions besoin d'avoir 18 clusters pour notre évaluation et la liste de noms du LST contient 493 mots, ce qui fait une moyenne de plus de 27 mots par clusters, cela peut expliquer nos résultats peu concluants.
Voici ci-dessous un exemple des sorties que nous avons obtenues avec K-means :

Cluster 0 :

0 commentaire discussion observation affirmation définition expression forme formule terme corpus essai étude note publication thèse volume colloque présentation synthèse figure illustration image motif schéma symbole bibliographie conclusion développement introduction section genre type diversité ensemble groupe partie totalité cycle série cadre environnement milieu pôle portée terrain plan plan axe sens voie caractère concentration expérience défaut avantage maîtrise pertinence absence état identité qualité tendance problème composition distribution structuration unité capacité faculté fonction nécessité classe groupe échantillon contenu marque matériau objet point population indice loi modèle dynamique sens mesure outil technique approche formule logique programme but projet indice distance divergence analyse réponse condition point recherche faculté mouvement organisation université explication généralisation division avancée multiplication communication évolution interaction implication pratique division fréquence différence distance minorité résultat taux chiffre siècle

Cluster 1 :

1 compétence décision sélection catégorisation structuration identification perception appréciation appréhension compréhension évaluation reconnaissance attribution expertise résolution intégration construction élaboration production réalisation développement fonctionnement traitement application manipulation utilisation discussion article terrain corpus observation estimation analyse étude questionnement contribution enquête entretien expérience expérimentation investigation questionnaire recherche test échantillon statistique

Concernant le script de Hclust, il contient également une fonction d'évaluation qui permet d'évaluer les données en les comparant avec une liste. Nous pouvions de ce fait l'utiliser pour comparer les clusters obtenus aux classes sémantiques constituées manuellement.
Nous avons ainsi intégré cette évaluation à notre script de génération des clusters. Celle-ci permet de calculer l'entropie, la pureté, l'information mutuelle spécifique par le biais de l'information mutuelle ajustée (IMA) et l'information mutuelle normalisée (IMN), et enfin le « rand index ajusté » (RI).
Avant de présenter les résultats, nous allons succinctement rappeler à quoi réfèrent ces différentes mesures. Tout d'abord, l'entropie permet de mesurer le degré d'aléatoire : plus elle se rapproche de zéro, plus le modèle probabiliste est jugé performant. La pureté calcule la moyenne pour tous les clusters, du degré de la correspondance entre les mots contenus dans les classes de comparaison et les clusters. Cette mesure, est de notre point de vue la plus évidente à interpréter car elle fournit une comparaison directe entre les clusters et les classes sémantiques.
L'information mutuelle spécifique permet d'obtenir un rapport entre deux éléments afin de calculer le degré de corrélation entre les deux. Ce calcul, est ici utilisé par deux versions de la mesure qui sont : l'information mutuelle ajustée et l'information mutuelle normalisée.
Le « rand index ajusté » permet de mesurer d'une autre manière la similarité entre deux éléments observés.

Voici ci-dessous les résultats obtenus selon chacun des trois modèles :

PMI	Entropie	Pureté	IMA	IMN	RI
	0.6232270786783 0707	0.3638253638253 6385	0.1664052199598 0798	0.2719006401661 9969	0.0697819591091 0918

Tableau 6 : Résultats de l'évaluation des clusters pour les modèles PMI

SVD	Entropie	Pureté	IMA	IMN	RI
	0.6370634209919 9864	0.3367983367983 368	0.1537912776275 7633	0.2585942978724 8434	0.0649863593848 3359

Tableau 7 : Résultats de l'évaluation des clusters pour les modèles SVD

SGNS	Entropie	Pureté	IMA	IMN	RI
	0.6190632621227411	**0.37422037422037424**	**0.18245944768319561**	**0.28179397248956628**	**0.08729539118748676**

Tableau 8 : Résultats de l'évaluation des clusters pour les modèles SVD

Les trois tableaux présentent les résultats obtenus lorsque l'on procède à la comparaison des clusters obtenus avec Hclust et des classes sémantiques constituées manuellement par Hatier et *al.* (2013). Nous avons ainsi, pour chacun des modèles, PMI, SVD et SGNS les différentes mesures permettant d'évaluer la proximité entre les classes sémantiques manuelles et les clusters que nous avons générés.

Nous pouvons tout d'abord noter que les résultats sont assez faibles pour les trois modèles, PMI, SVD et SGNS, ce qui signifie que les clusters et les classes sémantiques se ressemblent peu. Néanmoins, les meilleurs résultats sont ceux du modèle SGNS de word2vec, qui possède la plus faible entropie (0,619). La pureté (0,374), l'information mutuelle spécifique (0,182), l'information mutuelle normalisée (0,282) et le « rand index ajusté » (0,087) sont supérieurs aux modèles PMI et SVD.

Nous avons tout de même souhaité examiner manuellement les différents clusters obtenus par les trois modèles, afin de pouvoir confirmer ou infirmer les différents résultats. Nous avons pu ainsi, estimer lequel des modèles proposait des ensembles de mots avec les liens sémantiques les plus visibles. Bien que les clusters soient très chargés, nous avons été confortés dans l'idée que le modèle SGNS, générait les clusters où les liens sémantiques entre les mots étaient les plus évidents à cerner. Nous avions en effet, beaucoup de relations de quasi-synonymie entre les mots et moins de bruit qu'avec les PMI et SVD. Par bruit, nous entendons qu'il y avait moins de cas où des noms se retrouvaient dans un cluster sans partager de liens sémantiques avec les autres membres.

Nous avons donc décidé de conserver le modèle SGNS afin de pouvoir poursuivre notre projet sur la liste de noms du LST de Drouin (2007).

Voici des exemples de clusters que nous obtenions en sortie pour la méthode SGNS

cluster 1

approche démarche perspective réflexion théorie vision description interprétation conception représentation définition lecture

classe 15

présentation sélection catégorisation appréciation appréhension compréhension évaluation attribution résolution élaboration production réalisation fonctionnement traitement manipulation utilisation

Partie 4 : Méthode de description sémantique des résultats

Après avoir utilisé l'algorithme de classification automatique Hclust et la méthode distributionnelle SGNS sur le corpus d'étude, nous avons obtenu trente clusters. L'enjeu consistait donc à apporter une description de chacun d'entre eux. Au travers de cette description, nous souhaitions analyser plusieurs choses :

Dans un premier temps, nous voulions analyser quel type de relation nous avions entre les noms, c'est-à-dire des relations de quasi-synonymies, des relations d'antonymie ou encore des collocations.

Dans un second temps, nous voulions mettre en évidence les noms pouvant être éliminés du LST. Enfin, nous souhaitions apporter une description sémantique précise et, quand cela était possible, attribuer une étiquette sémantique à chaque cluster.

Nous présenterons dans cette partie, les différentes approches que nous avons utilisées afin de décrire ces clusters. Nous décrirons également le dictionnaire hiérarchique de Polguère qui nous a permis de leur attribuer une étiquette sémantique.

4.1 Description sémantique des noms

Au premier abord, il est difficile de trouver les liens sémantiques entre les mots manuellement. Ainsi, nous avons dû trouver une solution pour y parvenir. Pour ce faire, nous avons utilisé les classes et sous classes sémantiques des noms constituées par Hatier et *al.* (2014). Ces dernières nous ont aidé à cerner des liens entre les noms des clusters.

Voici ci-dessous un exemple de classe divisée en sous-classes constituées par Hatier et *al.* (2014) :

communication_contenu : , - contenu dans un processus de communication, - Par/dans ce N, l'auteur précise,	**Formulation**: formulation, 'élément de contenu' - Dans l'article, l'auteur utilise un N	affirmation, définition 02, expression 01, forme 04, formulation, formule 01, mot 01, terme 02, proposition 01
	Discussion: 'discussion, réponse à un acte de communication précédent' - Ce N répond à ...,	commentaire 01, commentaire 02, discussion 01, observation 03, remarque, débat, discussion 02

Tableau 9 : Exemple classes sémantiques Hatier et *al.* (2014)

Nous avons, dans cet exemple-ci, la classe *Communication contenu* divisée en deux sous classes *Formulation* et *Discussion*. Elle contient des noms tels que : *affirmation, définition, expression* et *débat*.

Cela nous a permis de comparer manuellement les noms de nos clusters avec ceux des classes sémantiques de Hatier et *al.* (2014). Nous avons pu réaliser des regroupements sémantiques entre les noms proches sémantiquement, au sein des clusters. En voici un exemple ci-dessous :

Cluster 3

Regroupement 1

argument 3	8	8.3
but 3	8	8.6
conclusion 3	8	8.3
enjeu 3	8	8.6
hypothèse 3	8	8.3
objectif 3	8	8.6

problème 3	8	8.10
problématique 3	8	8.10
question 3	8	8.10
résultat 3	8	8.3

Regroupement 2

aspect 3	6	6.3
avantage 3	6	6.2
problème 3	6	6.7
difficulté 3	6	6.7

Inclassable

remarque 3	1	1.2
constat 3	10	10.3

Pour chacun des clusters, nous regroupions les noms qui appartenaient à la même classe sémantique de Hatier et *al. (*2014) au sein d'un même regroupement. Ci-dessus, nous avons un cluster divisé en deux regroupements :

- Dans le premier regroupement, les noms appartiennent à la classe sémantique 8 de Hatier et *al.* (2014) qui s'intitule : *Renvoi aux observables et objets construits par l'activité scientifique*. Il contient les noms tels que : *conclusion, hypothèse* et *résultat.*
- Le deuxième regroupement correspond à la classe sémantique 6 nommée : *État ou qualité d'une entité abstraite ou concrète, - avoir Dét N*. Nous avons, par exemple, les noms : *aspect, avantage* et *difficulté.*

La première colonne correspond aux noms du LST appartenant au cluster. La seconde colonne représente la classe sémantique manuelle à laquelle appartient le nom et la troisième correspond à la sous-classe.

Nous observons, dans cet exemple, deux noms, *remarque* et *constat,* que nous avons isolés. En effet, ils n'appartiennent pas, selon les classes sémantiques, à un des deux regroupements.

Cette approche nous a permis de mettre en relief des regroupements sémantiques entre les noms et d'avoir un premier regard sur ce que l'on obtenait à l'intérieur des différents clusters.
De plus, nous avons pu observer des différences entre les classes sémantiques de Hatier et *al.* (2014) et nos clusters. Effectivement, en utilisant ces classes sémantiques, nous avons remarqué qu'il y avait plusieurs noms dans notre liste qui n'étaient pas présents dans celle d'Hatier et *al.* (2014). De ce fait, nous avons passé en revue ces différents noms, mis en relief par la comparaison, afin de juger s'il fallait que nous en supprimions ou non. Cela nous permis d'éliminer les noms qui ne correspondaient pas à des éléments du lexique scientifique transdisciplinaire, tels que : *acteur*, *action*, *passage*, *peine*, *grâce* ou *idéal*.

4.2 Le dictionnaire hiérarchique

Dans un second temps, nous avons voulu apporter une étiquette sémantique aux différents clusters. Pour cela, nous avons réfléchi à différentes ressources lexicales pouvant se prêter à la tâche. Celle qui nous a paru la plus adaptée est le jeu d'étiquettes sémantiques du dictionnaire hiérarchique de Polguère (2007).
Ce dictionnaire a été réalisé en 2007 au cours du projet DicoPop [28] et propose une interface web disponible sur le site de L'OLST. Cette dernière permet d'accéder à une hiérarchie d'étiquettes sémantiques, disponibles sur l'interface DicoPop, proposant des distinctions sémantiques assez fines entre les différentes unités lexicales. En effet, elle contient 790 étiquettes sémantiques, dont 706 nominales, 38 adjectivales et 40 verbales.
Cela nous a permis, dans la majorité des cas, de distinguer les différents regroupements sémantiques que nous avions. En effet, la présentation hiérarchique facilite l'utilisation du dictionnaire. Nous pouvions partir d'un domaine assez général tel que *Entité* par exemple et ainsi arriver jusqu'à *Document écrit*. Cela permet donc d'étiqueter avec précision nos regroupements.

Voici par exemple ci-dessous, un exemple de hiérarchie d'étiquettes :

entité

accumulation

[28] DicoPop : http://olst.ling.umontreal.ca/dicopop/nomenclature.php (consulté juin 2016)

accumulation de matière ayant une certaine forme
dépôt de matière

Dans cet exemple, nous avons l'entrée qui est « *entité* », qui contient « *accumulation* » qui elle-même contient deux entrées « *accumulation de matière ayant une certaine forme* » et « *dépôt de matière* ».
Pour chacun des clusters, ou des regroupements sémantiques dans les clusters, nous avons attribué l'étiquette sémantique qui nous semblait la plus appropriée. Dans certains cas il nous est arrivé de ne pas trouver d'étiquette adéquate.

Tableau 10 : Exemple 1 de description d'un cluster.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 1	ENTITE ENTITE INFORMATIONNELLE IDEE **ENSEMBLE D'IDEES**	*approche* *démarche* *logique* *perspective* *théorie*	Aucun	Aucun

Ci-dessus, nous avons un exemple où nous avons attribué l'étiquette « ENSEMBLE D'IDEES » à la totalité du cluster 1.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 4 regroupement 1	FAIT **RELATION FACTUELLE**	*effet* *condition* *critère* *conséquence* *cause*	*raison*	*réalité* *tendance*
Cluster 4 regroupement 2	FAIT CARACTERISTIQUE FAÇON DE FAIRE	*façon* *manière* *modalité*		

	FAIRE QQCH. D'UN CERTAINE FAÇON	*mode*		

Tableau 11 : Exemple 2 de description d'un cluster.

Ci-dessus nous avons un autre exemple où nous avons divisé le cluster en deux regroupements sémantiques. Le premier correspond à l'étiquette RELATION FACTUELLE et le second à l'étiquette FAIRE QUELQUE CHOSE D'UNE CERTAINE FAÇON.

Enfin, pour affiner et préciser davantage nos descriptions sémantiques, nous avons justifié les regroupements sémantiques et décrit les liens que nous avions entre les noms pour chaque cluster. Pour cela, nous avons notamment utilisé le dictionnaire électronique des mots[29] de Jean Dubois et Françoise Charlier-Dubois (2014), disponible sur le site du RALI de l'université de Montréal. Il avait notamment été utilisé par l'équipe de Jacques, Tutin, Tran, Hatier et Cavalla, dans le cadre du projet TermITH sur le LST. Ce dictionnaire propose une utilisation qui permet de cerner avec précision les différents sens d'un nom. Il arrivait parfois que nous ne trouvions pas, au premier regard, pour quelle raison un nom se retrouvait dans un cluster. Ce n'est qu'en consultant ses différents sens que nous parvenions à établir le lien sémantique.

[29] Le dictionnaire électronique des mots : http://rali.iro.umontreal.ca/rali/?q=fr/dem (consulté juin 2016)

Chapitre 3 :

Résultats

5. Brève présentation des résultats

La méthode distributionnelle que nous avons utilisée nous a permis d'obtenir des regroupements sémantiques de noms. Pour cela nous avons donc exploité le modèle SGNS qui utilise l'outil word2vec.

Dans un deuxième temps, nous avons également employé un algorithme de classification automatique avec lequel nous avons généré trente clusters. Nous allons donc au travers de cette partie présenter les résultats que nous avons obtenus, puis nous les décrirons et les commenterons.

Nous avons classé chacun des clusters dans un tableau de la manière suivante :

- Le numéro du cluster :

Il arrive parfois que le cluster soit découpé en différents regroupements sémantiques. En effet, nous avons regroupé chacun des mots proches sémantiquement ensemble, en nous servant des classes constituées manuellement par Hatier, Tutin, Jacques, Jacquey et Kister (2014).

- L'étiquette sémantique :

Nous avons attribué à chaque cluster, ou regroupement au sein du cluster, l'étiquette sémantique qui correspond le mieux dans le dictionnaire hiérarchique de Polguère.

Les étiquettes sont présentées de façon hiérarchique, ce qui signifie que l'étiquette la plus « basse » est celle qui s'attribue au regroupement. Cette étiquette appartient donc à des ensembles plus généraux situés au-dessus. Nous avons choisi une police en gras pour mieux la distinguer.

- La liste des noms appartenant au cluster ou au regroupement.
- Les erreurs du Lexique Scientifique Transdisciplinaire (LST), c'est-à-dire les noms que nous avons décidé d'éliminer de la liste des noms du LST.
- Les inclassables :

Dans cette colonne nous avons tous les noms que nous ne pouvons pas classer avec les autres membres du cluster car ils sont trop éloignés sémantiquement.

Enfin, pour chacun des clusters, nous avons ajouté une explication de ce que nous observons, ainsi qu'une description sémantique plus précise et complète. Afin d'approfondir ces descriptions, nous nous sommes également servie du Dictionnaire Électronique des Mots de Jean Dubois et Françoise Charlier-Dubois (2010). Ce dictionnaire nous a notamment permis d'établir des liens avec des sens polysémiques entre les mots que nous ne cernions pas au premier abord.

6. La liste des clusters

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 1	ENTITE ENTITE INFORMATIONNELLE IDEE **ENSEMBLE D'IDEES**	*approche* *démarche* *logique* *perspective* *théorie*		

Explications du cluster 1 :

Le cluster 1 est assez homogène, l'étiquette qui correspond le mieux est ENSEMBLE D'IDEES. Ces noms sont des concepts résultant de la méthode scientifique. Nous pouvons néanmoins noter que le nom *approche* semble se distinguer du reste du groupe. Alors que les autres noms correspondent plutôt à un ensemble d'idées et de raisonnements rationnels et cohérents, *approche,* lui est un peu différent. En effet il correspond plutôt à la façon d'aborder un sujet. Les quatre autres noms partagent des caractéristiques communes assez fines qui peuvent évoquer de la quasi-synonymie.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 2	INSTITUTION INSTRUMENT **OUTIL**	*instrument* *méthode* *outil* *programme* *technique* *application* *solution*		*projet* *traitement*

Explications du cluster 2 :

Nous avons choisi l'étiquette OUTIL pour le cluster 2. Celle-ci s'applique assez bien puisque nous avons des noms qui renvoient à ce qui permet de réaliser une activité scientifique ou de répondre à un problème. Au-delà d'outils, il s'agit de procédés. C'est pourquoi nous avons choisi de garder le nom *solution* qui est sémantiquement plus abstrait et ne correspond pas directement à un outil comme les autres noms. Ce nom répond tout de même à la définition « répondre à un problème ».

Nous avons écarté les deux noms *projet* et *traitement* qui sont plus éloignés du regroupement. Nous pourrions éventuellement faire le lien avec *traitement*, qui peut se rapprocher de « répondre à un problème », mais ce lien semble néanmoins difficile à établir.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 3		*argument* *but* *conclusion* *enjeu* *hypothèse* *objectif* *problème* *problématique* *question* *résultat* *difficulté* *remarque* *constat*	*aspect* *avantage*	

Explications du cluster 3 :

Le cluster 3 est très chargé et de ce fait lui attribuer une seule étiquette serait trop réducteur, bien que l'étiquette ENSEMBLE D'IDEES du dictionnaire hiérarchique de Polguère puisse s'y prêter.

Nous avons ici des noms qui renvoient à une proposition à vérifier ou un problème à résoudre dans le cheminement de l'activité scientifique. Dans le Dictionnaire électronique des mots (DEM), ces mots correspondent pour beaucoup à la notion de preuve. Ils mettent en relief les objectifs à atteindre et les questions qui se posent tout au long d'un projet scientifique et qui vont permettre de le faire avancer.

Les liens avec les noms *aspect* et *avantage* sont un peu moins étroits. Nous avons de plus, décidé de les écarter du LST car ils ne correspondaient pas aux critères scientifiques et transdisciplinaires du lexique.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 4 regroupement 1	FAIT **RELATION FACTUELLE**	*effet* *condition* *critère* *conséquence* *cause*	*raison*	*réalité* *tendance*
Cluster 4 regroupement 2	FAIT CARACTERISTIQUE FAÇON DE FAIRE **FAIRE QQCH. D'UN CERTAINE FAÇON**	*façon* *manière* *modalité* *mode*		

Explications du cluster 4 :

Nous avons divisé le cluster 4 en deux regroupements car les liens entre tous les noms n'étaient pas évidents. Nous avons donc d'un côté un premier regroupement qui correspond aux étiquettes CAUSATION, CONSEQUENCE et RAISON. Et d'un autre côté nous avons attribué au regroupement l'étiquette FAIRE QQCH. D'UN CERTAINE FAÇON.

Dans le premier regroupement, nous avons les noms renvoyant à ce qui peut faire varier l'activité scientifique. Ces derniers relèvent plutôt de la cause avec : *Effet, critère, cause*, et *condition*. Ensuite, il y a des noms entretenant une opposition avec une relation d'antonymie avec les noms *conséquence* et également *effet* qui est polysémique puisqu'il correspond à la fois à la cause et la conséquence.

Le second regroupement contient des noms quasi-synonymes permettant de qualifier la façon de procéder, la méthode employée pour parvenir à réaliser une activité scientifique.

Nous avons supprimé *raison* du LST et isolé les noms *tendance* et *réalité*, avec lesquels les liens sémantiques étaient vraiment difficiles à établir.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 5 Regroupement 1	ENTITE ENSEMBLE CONNAISSANCE **ENSEMBLE DE CONNAISSANCES**	*loi* *norme* *notion* *principe* *procédure* *règle*		*idée*
Cluster 5 Regroupement 2		*exigence* *nécessité* *possibilité*		
Cluster 5 Regroupement 3		*travail* *tâche* *opération* *stratégie*		

		expérience *compétence*		

Explications du cluster 5 :

Le cluster 5 contient beaucoup de noms et leurs liens sémantiques sont assez hétérogènes. Néanmoins on peut en dégager trois regroupements.

Les noms du premier regroupement correspondent en quelque sorte aux références sur lesquelles il faut s'appuyer dans un projet scientifique. En effet, les mots tel que *notion, norme, loi*, constituent la base, le socle auquel se référer. Nous avons attribué l'étiquette ENSEMBLE DE CONNAISSANCES, bien que celle-ci soit un peu large et n'explicite pas l'idée de « concept auquel se référer ».

Le deuxième regroupement est intéressant mais il semble difficile de lui attribuer une étiquette. Nous avons des noms soulevant le besoin de quelque chose, de répondre à une problématique scientifique avec : *exigence*, *possibilité* et *nécessité*.

Dans le troisième regroupement, nous avons les noms qui permettent d'exprimer la réponse à un besoin (exprimée dans le deuxième regroupement), c'est-à-dire quels moyens peuvent être envisagés pour y répondre avec : *travail*, *tâche*, *opération*, *stratégie*, *compétence* et *expérience*. Concernant cet ensemble, nous n'avons pas trouvé d'étiquette sémantique correspondant. Effectivement, bien que nous puissions établir des liens sémantiques entre les différents mots, il était par exemple difficile de regrouper *expérience* et *compétence* avec *travail*.

Le lien sémantique entre les trois regroupements est quant à lui difficile à visualiser.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**

Cluster 6	ENTITE ENTITE INFORMATIONNELLE CONTENU INFORMATIONNEL QU'ON COMMUNIQUE CONTENU INFORMATIONNEL QU'ON COMMUNIQUE OU SUPPORT PHYSIQUE DE CE CONTENU DOCUMENT DOCUMENT ECRIT **TEXTE**	*article* *synthèse* *texte* *thèse* *chapitre* *document* *introduction* *ouvrage* *recherche* *thèse* *synthèse* *littérature*	*œuvre* *mois*	*réflexion* *proposition* *auteur*

Explications du cluster 6 :

L'étiquette qui correspond le mieux au cluster 6 est TEXTE, on a en effet ici tout ce qui correspond au genre textuel, ce que l'on peut retrouver dans la structure d'un document écrit ou une œuvre littéraire.

Nous avons placé le nom *auteur* dans les inclassables car il ne correspond pas à l'étiquette TEXTE. Cependant, il reste cohérent avec le reste du groupe, puisque le nom *auteur* est bien celui qui réalise le document écrit.

Les noms *réflexion* et *proposition* restent également en accord avec le cluster mais ne sont pas seulement des textes mais aussi des ensembles d'idées. C'est pourquoi nous avons préféré les isoler.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 7	FAIT ACTION **ACTE COGNITIF**	*analyse* *présentation* *étude*	*détail* *propos* *histoire*	*partie* *sens*

		description *observation* *discussion* *présentation* *comparaison*		

Explications du cluster 7 :

L'étiquette qui correspond le mieux au cluster 7 est ACTE COGNITIF, bien que celle-ci reste très vague. Plus précisément, les noms que nous avons ici permettent de décrire avec précision les entités appartenant à une activité scientifique. De ce fait ils permettent de montrer et mettre en relief les éléments pertinents.

Le nom *sens* est un peu en marge du reste du groupe, il permet plutôt de mettre l'accent sur la signification de quelque chose. C'est pourquoi nous avons préféré l'isoler. Ce cluster reste cependant homogène, bien que nous ayons également écarté *partie* du regroupement.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 8		*activité* *décision*	*accord* *acte* *acteur* *action* *conflit* *demande* *emploi* *entreprise* *institution* *service* *société* *établissement* *pratique* *situation*	

Explications du cluster 8:

Le cluster 8 est assez intéressant. En effet il regroupe tout ce qui relève du domaine du travail avec des noms tels que : *société, emploi* ou encore *demande*. Néanmoins, cela permet surtout de mettre en relief un certain nombre de noms qui ne sont pas transdisciplinaires car ils apparaissent majoritairement dans les corpus de textes d'économie. De ce fait, nous les avons éliminés de la liste des noms du LST. Nous avons cependant décidé de conserver uniquement *activité* et *décision*, qui peuvent être déplacés dans un autre regroupement.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 9 regroupement 1	ENTITE ETRE VIVANT ETRE ANIME ETRE HUMAIN **INDIVIDU ET ENSEMBLE D'INDIVIDUS**	*personne* *homme* *population* *communauté* *individu* *sujet* *objet*	*dispositif* *vie*	*système*
Cluster 9 regroupement 2	ENTITE LIEU **ESPACE**	*terrain* *territoire* *espace* *environnement* *monde* *pays*		

Explications du cluster 9 :

En utilisant le dictionnaire hiérarchique le cluster 9 est divisé en deux regroupements. Nous avons, d'une part, LES INDIVIDUS ET ENSEMBLE DES INDIVIDUS et d'autre part, les ESPACES. Cependant, nous pourrions les regrouper en un seul groupe sémantique qui contiendrait à la fois

les espaces géographiques et les ensembles humains, car les liens entre les noms peuvent être visualisés.

Nous avons placé le nom *système* dans les inclassables. Néanmoins, il peut être employé comme un ensemble organisé socialement tel qu'un gouvernement par exemple, ce qui explique sa proximité avec le deuxième regroupement, mais il ne peut pas être étiqueté comme un espace.

Bien que *vie* puisse avoir un lien sémantique avec le premier regroupement, il est effectivement assez proche *d'être animé* ou *être vivant* mais ne correspond pas à un individu. Nous avons souhaité l'écarter du LST. Il en va de même pour *dispositif*, qui est proche du second regroupement. Nous pouvons ajouter que tous les noms de ce cluster sont discutables et pourraient être considérer comme des objets des sciences humaines et donc ne pas appartenir au LST.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 10	ENTITE ENSEMBLE **ENSEMBLE D'INDIVIDUS ET ENSEMBLE DE CHOSES REMARQUABLES**	*catégorie* *classe* *groupe* *famille* *liste*		

Explications du cluster 10 :

Concernant le cluster 10, l'étiquette qui correspond le mieux est ENSEMBLE D'INDIVIDUS ET ENSEMBLE DE CHOSES REMARQUABLES. Les noms sont regroupés de façon homogène. Ils servent essentiellement à nommer des ensembles sociaux. Ils pourraient ainsi être regroupés avec le cluster 9. En effet, ce dernier contient des ensembles humains. Ainsi, nous pourrions voir les liens avec les noms tels que *famille, groupe* ou encore *classe*. Néanmoins, il semble davantage intéressant de garder ces deux clusters séparés. En effet, tous les noms du clusters 10 sont polysémiques et peuvent également s'attribuer à des entités autres que humaines.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 11	ENTITE ENTITE INFORMATIONNELLE CONTENU INFORMATIONNEL QU'ON COMMUNIQUE CONTENU INFORMATIONNEL QU'ON COMMUNIQUE OU SUPPORT PHYSIQUE DE CE CONTENU DOCUMENT **PARTIE DE DOCUMENT ECRIT**	*terme* *contenu* *signe* *expression* *mot* *terme* *titre* *nom* *moment* *date*	*chose* *courant* *issue* *doute* *lieu* *fait*	

Explications du cluster 11:

Le cluster 11 contient des entités référant à des parties de textes scientifiques. Nous pourrions ainsi regrouper ce cluster avec le 6. Néanmoins, au-delà de cela, ces noms peuvent également appartenir aux phrases et non seulement aux textes, ce qui les distingue légèrement du cluster 6. C'est pour cela que nous avons attribué l'étiquette PARTIE DE DOCUMENT ECRIT et non l'étiquette TEXTE.

Il y a deux noms, *date* et *moment*, qui sont légèrement en marge du cluster, ils renvoient à la notion de temps contrairement aux autres membres du regroupement. Cependant, nous les entendons, dans le cas présent, comme des parties de texte, bien que cela soit discutable.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 12	ENTITE OCCUPATION SOCIALE FONCTION SOCIALE **FONCTION SOCIALE**	*rôle* *statut* *intérêt* *contrôle*	*considération* *réponse*	

	OU INDIVIDU EXERÇANT CETTE FONCTION	*fonctionnement* *position*		

Explications du cluster 12 :

L'étiquette correspondant le mieux au cluster 12 est FONCTION SOCIALE OU INDIVIDU EXERÇANT CETTE FONCTION. Dans ce regroupement les noms servent effectivement à dénommer la fonction, à décrire l'importance d'une entité ou d'une personne. Cette étiquette est néanmoins discutable. Nous n'avons en effet pas observé les mots en contexte. De ce fait, pour le mot rôle par exemple, il correspond plutôt à une fonction en générale, telle que la fonction d'un objet dans une expérimentation.

Dans ce cluster, nous avons d'un côté les noms *rôle, statut*, et *position* qui sont plus proches sémantiquement. De l'autre côté, les noms *fonctionnement* et *contrôle* semblent partager davantage de caractéristiques sémantiques entre eux qu'avec le reste du groupe.

Concernant le nom *intérêt*, il se démarque du reste des noms, mais nous pouvons établir des liens sémantiques avec *rôle, statut* et *position*. Nous pouvons l'entendre au sens d'« intérêt d'une personne » en fonction de son positionnement social, bien sûr cela est discutable.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 13		*compréhension* *conception* *définition* *interprétation* *lecture* *représentation* *vision* *discours* *image*	*reconnaissance*	

Explications du cluster 13 :

Le cluster 13 contient des noms dont les liens sémantiques sont très intéressants mais il n'y a pas d'étiquette dans le dictionnaire hiérarchique qui s'y prête pour les regrouper. Nous

pourrions en proposer une appelée « Procédé permettant la description d'un phénomène scientifique ».

Nous avons ici les noms permettant d'expliquer un phénomène scientifique, soit en exposant des faits comme avec les noms : *image* et *discours*, soit en donnant un point de vue avec les noms : *vision, conception* ou encore en expliquant avec précision quelque chose avec les noms : *interprétation, définition*.

Ces noms expriment le désir de comprendre, d'aller vers un but précis en passant par l'explication. Nous pouvons, de plus, noter qu'il y a une interférence entre *vision* et *image,* dû à la polysémie de ces deux noms.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 14	FAIT ACTION **ACTION DE MODIFIER UN PARAMETRE** **i. ACTION D'AUGMENTER UN PARAMETRE** **ou** **ii. ACTION DE DIMINUER UN PARAMETRE**	*amélioration* *changement* *développement* *modification* *création* *organisation* *recours* *gestion* *intervention* *maintien*		

Explications du cluster 14 :

Le cluster 14 est homogène et correspond aux étiquettes du dictionnaire hiérarchique ACTION D'AUGMENTER UN PARAMETRE ou ACTION DE DIMINUER UN PARAMETRE. Les noms permettent de décrire les différentes actions qui vont faire varier un paramètre scientifique.

Le nom *maintien* n'entre pas exactement dans l'action de diminuer ou augmenter un paramètre mais il possède tout de même des liens sémantiques forts avec le reste du groupe. Il agit en effet sur les paramètres scientifiques d'une expérimentation. C'est pourquoi nous avons souhaité le conserver avec les autres éléments du groupe.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 15 regroupement 1	ENTITE OBJET PHYSIQUE **QQCH. QUI EST DANS UNE CERTAINE RELATION AVEC QQCH. D'AUTRE**	*appartenance* *association* *identité* *opposition* *constitution* *distinction* *représentant* *membre*		

Explications du cluster 15 :

L'étiquette correspondant à ce cluster selon le dictionnaire hiérarchique est QQCH. QUI EST DANS UNE CERTAINE RELATION AVEC QQCH. D'AUTRE. Nous avons ici des noms qui entretiennent des relations relavant plutôt de la quasi-synonymie, hormis *constitution* et *distinction* qui sont antonymes.

Les deux noms *représentant* et *membre* sont plus difficiles à classer sous cette étiquette, bien que l'on puisse voir un lien sémantique. En effet, ils renvoient également à une idée d'appartenance à quelque chose, donc tout de même à une notion de relation.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 16		*esprit* *liberté* *protection*		

Explications du cluster 16 :

Le cluster 16 ne contient que trois mots, qui n'ont pas de relation sémantique très évidente. Nous pouvons peut-être plutôt noter des relations de cooccurrence telles que *la protection de la liberté*, ou encore un lien entre *esprit* et *liberté* renvoyant tous les deux à quelque chose que l'on ne peut pas contrôler. Nous pensons cependant que ce cluster n'apporte pas d'intérêt dans le cadre de notre étude. En effet, il comporte trop peu de mot et les liens entre eux restent très flous.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 17	FAIT CARACTERISTIQUE CARACTERE [DE QQCH.] **CARACTERE NEGATIF OU CARACTERE POSITIF**	*absence* *biais* *complexité* *importance* *manque* *qualité* *stabilité* *efficacité* *apport*		

Explications du cluster 17 :

Les noms du cluster 17 servent à décrire des données scientifiques, l'étiquette CARACTERE NEGATIF OU POSITIF du dictionnaire hiérarchique y correspond bien. Ces noms permettent en effet de décrire positivement ou négativement des éléments scientifiques, davantage encore tout ce qui peut faire varier des résultats. Nous pouvons cependant noter que l'étiquette s'applique moins bien pour les mots *efficacité* et *apport*.

Ce cluster est homogène, nous n'avons pas d'inclassable.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 18	ENTITE LIEU **ESPACE**	*base* *côté* *extérieur*	*aide* *dehors*	

		face *fond* *intérieur* *support*		

Explications du cluster 18 :

Le cluster 18 est homogène. Il contient des noms permettant de situer des objets dans l'espace. Nous avons donc attribué l'étiquette ESPACE, qui semble la mieux correspondre dans le dictionnaire hiérarchique. Cependant, ces noms sont très abstraits et permettent davantage de situer un objet dans l'espace plutôt qu'un lieu.

Il est assez intéressant de noter la polysémie du nom *support*. En effet, nous avons supprimé *aide* du LST, bien qu'il entretienne une relation de quasi-synonymie avec un des sens du nom *support*.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 19		*existence* *présence* *totalité* *ensemble* *exception* *apparition* *disparition* *départ* *origine* *ouverture* *majorité*		*est* *proximité*

		moitié		

Explications du cluster 19 :

Le cluster 19 est très hétérogène et il est compliqué de lui attribuer une description sémantique. Nous avons ici une série de quasi-synonymes et antonymes, ils semblent correspondre, pour la majorité, à des positions sur un axe :

- Quasi-synonymes :

existence
présence

départ
origine
ouverture

totalité
ensemble

- Antonymes :

Totalité et *ensemble* antonymes d'*exception* .

apparition
disparition

majorité
moitié

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 20		*dépendance* *impact* *influence* *rupture* *répartition* *distribution* *augmentation* *perte* *variation* *évolution* *différence* *taux* *écart* *séparation*		

Explications du cluster 20 :

Le cluster 20 est homogène et très intéressant. Il contient tous les noms permettant de décrire les données d'expérimentation et comment celles-ci se comportent, c'est-à-dire leur évolution, leur impact par exemple. Il se recoupe d'une certaine manière avec le cluster 17 permettant également de décrire des données scientifiques.
Cependant, il n'y a aucune étiquette dans le dictionnaire hiérarchique qui corresponde.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 21 Regroupement 1	FAIT ACTION ACTION PHYSIQUE **ACTION SUR UN OBJET**	*construction* *diffusion* *choix* *intégration* *production* *réalisation* *usage*	*accès* *référence* *Identification* *détermination*	

		utilisation *transformation* *transfert* *communication* *extension* *échange*		

Explications du cluster 21 :

Le cluster 21 contient beaucoup de noms mais il est assez homogène et les liens sémantiques sont intéressants. En effet, il contient tous les noms faisant référence à la réalisation d'un projet scientifique, il regroupe tous les processus du début à la fin. L'étiquette qui se prête le mieux pour le décrire est ACTION SUR UN OBJET. Néanmoins, celle-ci est très vague et ne décrit pas précisément les différents éléments du regroupement.
Nous avons les premières étapes du projet avec les noms tels que : *construction, réalisation, production, choix*. Les noms faisant référence à la finalité du projet : *usage, utilisation, intégration*. Enfin les noms renvoyant à la communication et au maintien du projet scientifique : *transformation, transfert, communication, extension, échange.*

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 22 Regroupement 1	ENTITE LIEU **ESPACE**	*champ* *domaine* *secteur* *zone* *région*	*couche* *ligne*	*temps* *moyen* *niveau* *état*
Cluster 22 regroupement 2	CARACTERISTIQUE **MOMENT**	*période* *siècle* *année* *dernier* *début*		

		fin *heure* *jour* *premier* *époque* *milieu*		

Explications du cluster 22 :

Le cluster 22 est scindé en deux regroupements. Nous avons d'une part le premier que nous avons placé sous l'étiquette ESPACE. Cependant, certains noms comme *domaine* et *champ* correspondent davantage à des *groupes* qu'à des espaces. Ils sont en quelque sorte polysémiques.

Ensuite, le second regroupement correspond à l'étiquette MOMENT du dictionnaire hiérarchique. Nous n'avons pas regroupé quatre noms *temps, moyen, niveau* et *état* qui ne s'apparentent pas aux deux regroupements. Bien que, *état* puisse éventuellement être placé dans le premier regroupement car il entretient une relation sémantique avec *zone* ou *région,* dans le sens de *pays.*

Ce qui est intéressant dans ce cluster, c'est que les deux regroupements représentent des « groupes », des « zones », soit un regroupement plutôt dans l'espace et un second plutôt dans le temps.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 23 Regroupement 1	ENTITE FIGURE **FIGURE OU FORME**	*forme* *tableau* *indice* *angle* *figure*	*source*	*réaction*

Cluster 23 Regroupement 2		*distance* *direction* *étape* *événement* *phase* *section* *structure* *transition* *déplacement* *surface* *section*		
Cluster 23 Regroupement 3	FAIT CARACTERISTIQUE PROPRIETE **AVOIR UNE CERTAINE PROPRIETE**	*type* *profil* *composition* *espèce* *matière* *produit*		

Explications du cluster 23 :

Le cluster 23 est très chargé, et de ce fait, difficile à analyser sémantiquement. Nous avons distingué trois regroupements différents.

Le premier regroupement correspond à des entités concrètes faisant référence à des représentations scientifiques avec : *figure*, *angle*, *forme*, *tableau*. Nous l'avons étiqueté par FIGURE OU FORME. Ces noms peuvent également nous évoquer les éléments d'un texte ce qui pourrait se regrouper avec le cluster 6 ou le cluster 10.

Le second regroupement contient des noms correspondant à la fois à des moments comme avec *transition, événement, phase* et à des emplacements avec par exemple les noms : *direction*, *surface*, *déplacement*. Néanmoins, ce deuxième regroupement est intéressant puisque tous les

noms semblent polysémiques et peuvent de ce fait faire aussi bien référence à des moments qu'à des espaces. Cela peut nous faire penser au cluster 22, qui présentait également des noms aux caractéristiques similaires. Cependant, nous n'avons pas attribué d'étiquette car cela serait trop réducteur.

Le troisième regroupement contient des noms servant à décrire des entités scientifiques avec des noms tels que : *profil, type* et *espèce*. Nous lui avons attribué l'étiquette AVOIR UNE CERTAINE PROPRIETE qui semblait la mieux correspondre. Néanmoins celle-ci ne s'applique pas de manière très nette à l'ensemble des noms du regroupement.

Enfin, nous pouvons noter que le mot *forme* est polysémique et pourrait également appartenir au troisième regroupement.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 24	FAIT CARACTERISTIQUE **MOMENT**	*période* *siècle* *année* *dernier* *début* *fin* *heure* *jour* *époque* *milieu*	*cœur* *pouvoir* *essentiel* *premier*	*cas* *exemple* *suite* *série*

Explications du cluster 24 :

Le cluster 24 regroupe tous les noms faisant référence à des moments, l'étiquette du dictionnaire hiérarchique MOMENT s'y prête bien pour le décrire.

Le nom *milieu* reste cependant un peu en marge du regroupement. Nous l'entendons aussi comme un moment tel que le « milieu d'une époque ».

Nous avons placé *suite* et *série* dans les noms inclassables, bien qu'ils ne soient pas pour autant incohérents avec le reste du cluster. En effet, nous pouvons y voir une relation sémantique du

type « une série de moments » ou encore, « une suite de moments ». Cependant, les liens sémantiques avec *cas* et *exemple* sont plus difficiles à cerner.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 25 Regroupement 1	PROCESSUS PROCESSUS PHYSIQUE **MOUVEMENT DE QQCH.**	*phénomène* *mécanisme* *dynamique* *mouvement* *processus*		
Cluster 25 Regroupement 2	FAIT CARACTERISTIQUE **STRUCTURE**	*modèle* *réseau* *schéma*		

Explications du cluster 25 :

Nous avons deux regroupements qui se distinguent dans le cluster 25. D'une part, le premier groupe porte l'étiquette MOUVEMENT DE QQCH du dictionnaire hiérarchique et d'autre part, nous avons étiqueté le second groupe par STRUCTURE.

Néanmoins, nous pourrions faire un lien sémantique entre les deux regroupements, notamment car le nom *mécanisme* du regroupement 1 peut être relié à *modèle*, *réseau* et *schéma* du regroupement 2.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 26 Regroupement 1	ENTITE LIEU **ESPACE**	*cadre* *contexte* *orientation* *ordre*	*clé* *culture* *formation* *échelle*	
Cluster 26 Regroupement 2	ENTITE PROPRIETE **AVOIR UNE**	*caractère* *caractéristique* *nature*		

	CERTAINE PROPRIETE	*particularité* *propriété* *unité*		

Explications du cluster 26 :

Le cluster 26 est divisé en deux regroupements. Nous avons attribué l'étiquette ESPACE au premier regroupement, bien que nous pourrions plutôt décrire cela par le milieu, la position ou la direction. Dans ce regroupement nous avons deux paires de quasi-synonymes avec d'un côté *contexte* et *cadre* et de l'autre, *ordre* et *orientation*. Il reste tout de même assez évident de faire le lien entre ces deux paires.

Concernant le deuxième regroupement, l'étiquette AVOIR UNE CERTAINE PROPRIETE semble bien le décrire. Le nom *unité* est un peu plus éloigné sémantiquement des autres membres du regroupement, mais nous l'entendons ici dans le sens d'élément.

L'ensemble des noms du cluster servent à décrire de façon précise une entité. Il reste cependant difficile d'attribuer une seule étiquette à ce cluster.

Cluster	**Etiquettes**	**Noms**	**Erreurs LST**	**Inclassables**
Cluster 27	FAIT ACTION **CARACTERISTIQUE**	*capacité* *défaut* *contraint* *effort* *risque* *besoin*		

Explications du cluster 27 :

Le cluster 27 est assez petit. On peut ainsi envisager de le regrouper avec un autre cluster.

Nous avons des mots qui semblent servir à représenter l'évaluation des risques, des besoins engendrés par un projet scientifique, ainsi que les efforts, contraintes que cela nécessite. Le lien sémantique avec le nom *défaut* est moins évident à réaliser. Nous l'entendons comme un

problème que peut générer un projet ou une activité scientifique.
Nous avons attribué l'étiquette CARACTERISTIQUE qui correspond le mieux mais qui est trop générale.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 28	ENTITE LIEU **ESPACE**	*axe* *plan* *voie* *centre* *limite*		*code* *formule*

Explications du cluster 28 :

Le cluster 28 est assez petit. Il regroupe des noms servant à décrire des graphiques ou des espaces comme avec *axe, plan* et *centre*. Au-delà de cela, il permet de situer un objet dans un espace. C'est pourquoi nous lui avons attribué l'étiquette ESPACE. Néanmoins, il se différencie des autres clusters « espaces », car ces derniers contiennent plutôt des espaces géographiques avec des noms tels que : *zone, environnement, terrain* etc.
Les liens sémantiques avec les noms *code* et *formule* sont plus difficiles à cerner, c'est pourquoi nous avons préféré les isoler.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 29	FAIT **PARAMETRE**	*degré* *dimension* *facteur* *composante* *paramètre*		

		point *poids* *trait* *élément* *fréquence* *force* *puissance* *résistance*		

Explications du cluster 29 :

Le cluster 29 est homogène, l'étiquette PARAMETRE s'applique bien pour décrire ces noms. Ils correspondent à tous les paramètres faisant varier une activité ou une expérimentation scientifique. Le lien avec le regroupement 2 du cluster 26 peut ainsi être fait.

Cluster	Etiquettes	Noms	Erreurs LST	Inclassables
Cluster 30	ENTITE **ENTITE INFORMATIONNELLE**	*connaissance* *donnée* *erreur* *information* *mesure* *précision* *ressource* *mémoire* *preuve*		

Explications du cluster 30 :

Le cluster 30 est intéressant car il est homogène, bien que les relations entre les mots ne soient pas de l'ordre de la quasi-synonymie. Nous avons attribué l'étiquette ENTITE

INFORMATIONNELLE. Celle-ci est un peu vague mais permet néanmoins de cerner le sens de ces noms, c'est-à-dire des noms renvoyant à tout ce sur quoi nous nous appuyer pour établir un raisonnement, ou la constitution d'un projet scientifique. En effet, nous avons des entités abstraites ou concrètes telles que : *connaissance*, *ressource* ou encore *preuve*.

Le nom *précision* est un peu plus difficile à relier aux autres, il permet d'améliorer le raisonnement ou le projet mais ne constitue pas réellement une base sur laquelle s'appuyer.

Discussion

Notre étude a fait émerger différents points que nous souhaitions discuter dans cette partie. Nous reprendrons l'analyse des clusters que nous avons obtenus. Cela nous permettra d'une part, de proposer des regroupements entre certains ensembles et d'autre part, nous expliquerons avec davantage de précision nos résultats, c'est-à-dire ce que nous pouvons penser des relations sémantiques entre les noms, que nous avons relevées. Enfin, nous évaluerons la qualité de ces résultats et ce que nous pouvons conclure de l'apport de la méthode distributionnelle et de la classification automatique à l'organisation d'une liste de noms brute.

1. Regroupement des clusters

Nous trouvons intéressant de proposer des regroupements entre certains clusters. En effet, quelques-uns possèdent des caractéristiques communes. Nous pensions qu'explorer cette similarité serait intéressante et que cela pourrait apporter une homogénéité aux classes sémantiques. Néanmoins, ces regroupements ne sont que des propositions et sont discutables. Nous sommes satisfaite de la qualité des distinctions que propose la méthode distributionnelle et sommes d'avis qu'il est aussi possible de conserver les clusters tels que nous les avons obtenus.

- Le cluster 9 et le cluster 10 :

personne homme population communauté individu sujet objet terrain territoire espace environnement monde pays catégorie classe groupe famille liste

Nous avions le cluster 9 : INDIVIDU ET ENSEMBLE D'INDIVIDUS et ESPACE ainsi que le cluster 10 : ENSEMBLE D'INDIVIDUS ET ENSEMBLE DE CHOSES REMARQUABLES. Nous obtenons ainsi, un cluster qui contient des noms faisant référence à des ensembles humains. Néanmoins, cela est aussi à nuancer car les mots tels que *groupe, espace, classe, liste* etc, sont polysémiques et ne font pas uniquement référence à des ensembles humains. Il est intéressant de constater que la méthode SGNS est capable de distinguer ces deux ensembles.

- Le cluster 17 et le cluster 20 :

absence biais complexité importance manque qualité stabilité efficacité apport dépendance différence impact influence rupture répartition distribution augmentation perte variation évolution différence taux écart séparation

Nous avions le cluster 17 : CARACTERE NEGATIF OU CARACTERE POSITIF et le cluster 20 contenant des noms permettant de décrire le comportement de données scientifiques.
Le regroupement que nous proposons nous semble intéressant. En effet, nous avons les noms décrivant des données scientifiques de manière précise et permettent également, de caractériser de façon négative ou positive ces données.

- Le cluster 26 et le cluster 29 :

degré dimension point facteur composante paramètre point poids trait élément fréquence force puissance résistance caractère caractéristique nature particularité propriété unité

Nous avions à l'origine, le regroupement 2 du cluster 26 : AVOIR UNE CERTAINE PROPRIETE et le cluster 29 : PARAMETRE.
Nous trouvions pertinent de les regrouper car ils contiennent tous les deux des noms faisant référence à des propriétés et permettant de caractériser un objet scientifique.

Ensuite, nous avons deux clusters qui ne nous semblent pas intéressants pour notre étude, il s'agit du 8 et du 16, qui contiennent peu de noms et il est difficile de les regrouper avec un autre cluster.

2. Bilan de l'analyse des clusters

Dans la majorité des cas, nous sommes satisfaite des clusters que nous avons obtenus. Ils ont quasiment tous des liens sémantiques intéressants. Ces derniers relèvent, pour la plupart, de la quasi-synonymie comme par exemple :
Dans le cluster 4 regroupement 2 :

Façon, manière, modalité et *mode*.

Nous obtenons également quelques cas où nous avons des relations d'antonymie comme, par exemple, dans le cluster 4 :

Conséquence et cause.

De plus, nous notons également des relations un peu différentes. Il s'agit de noms appartenant à un même champ sémantique, un même domaine, sans pour autant partager des liens paradigmatiques. Nous avons, par exemple, écarté le cluster 8 dans le cadre de notre étude, bien que ce dernier reste très intéressant. En effet il contenait des noms faisant référence au domaine du travail avec :

accord acte acteur action conflit demande emploi entreprise institution service société établissement pratique et situation.

En comparant nos clusters avec les classes sémantiques de Hatier et *al.* (2014), nous avons constaté des divergences entre les deux approches, manuelle et automatique. Tout d'abord, l'approche automatique que nous avons utilisée ne prend pas en compte un point essentiel : la polysémie. En effet, dans la classification manuelle, il arrive qu'un nom appartienne à deux classes sémantiques, parfois plus, s'il est polysémique. Par exemple : *discussion*, dans notre étude nous l'avons placé sous l'étiquette d'acte cognitif. Cependant, ce nom pourrait aussi se regrouper avec les textes, pouvant également être un texte écrit.

De nombreux travaux sur l'analyse distributionnelle ont quant à eux pris en compte la polysémie. C'est notamment le cas de Galop (2011) qui a utilisé l'analyse distributionnelle sur certains mots du LST. Il voulait observer si la position d'un mot dans un texte et son contexte syntaxique pouvaient faire varier son analyse sémantique. Les mots proches sémantiquement étaient regroupés par « cliques ». Un même mot pouvait appartenir à différentes cliques ce qui permettait de traiter les cas de polysémie. Par exemple, le nom *analyse* pouvait être regroupé avec les noms *approche* et *recherche* d'une part et *auteur, étude* et *travail* d'autre part.

Dans notre étude, nous avons été confrontée à ce problème où certains noms auraient pu tout à fait appartenir à d'autres classes et étaient donc par défaut, classés dans un seul cluster par l'algorithme de classification automatique que nous avons utilisé. Le traitement de la polysémie possède un réel intérêt lorsque l'on veut décrire le sens des mots. Davantage encore, si cette liste de mots est utilisée à des fins didactiques. En effet, un apprenant d'une langue étrangère a besoin de savoir exactement dans quels contextes le mot qu'il veut utiliser peut s'employer.

Bien que l'approche manuelle ne soit pas un « standard », elle reste tout de même plus fiable sur ce point, car un avis humain apporte toujours une méthodologie plus précise et fine. Effectivement, la plupart des clusters que nous avons obtenus sont assez homogènes. Néanmoins, nous avons eu plusieurs cas où nous avions des noms inclassables. Nous avons préféré les laisser dans les tableaux afin de conserver une certaine clarté dans les résultats que nous présentons. De plus, nous pourrons traiter ces noms inclassables dans une étape subséquente. Cependant, nous avons souhaité observer ce que cette quantité de noms inclassables représentait dans nos résultats en calculant la précision. Pour ce faire, nous avons comptabilisé le nombre de noms inclassables que nous avions. Nous comptons 25 noms sur 461 noms qui ne sont pas en lien avec le cluster dans lequel ils ont été placés, ce qui signifie que nous avons une précision de 94,577 %. La précision est haute, ce qui nous montre que nous avons peu d'erreurs et donc des clusters homogènes avec des liens sémantiques entre les mots riches et assez évidents à cerner.

Au-delà de cela, l'approche automatique permet de mettre en évidence des éléments auxquelles nous ne penserions pas avec une approche manuelle. Par exemple, il y avait des cas très intéressants, comme pour le cluster 8 où nous avions des noms ne partageant pas vraiment de liens paradigmatiques et pourtant regroupés ensemble. Nous pouvons ainsi voir que la méthode SGNS permet de regrouper ensemble des mots appartenant au même champ sémantique.

D'après nous, la méthode distributionnelle SGNS est avant tout une base, solide, permettant d'organiser sémantiquement une liste de noms brute.

Néanmoins, elle ne peut que difficilement fonctionner sans l'approbation d'un avis humain. Nous ne pouvons l'envisager seule et pensons qu'apporter une description sémantique manuelle est important et ajoute une véritable valeur à nos travaux. Nous avons notamment pu le voir dans les travaux de Morlane-Hondère & Fabre (2012, b). En effet, d'après eux la méthode distributionnelle est une véritable aide pour la constitution de lexique.

Conclusion

La méthode distributionnelle a su faire ses preuves concernant l'extraction des relations sémantiques ces dernières années. Dans notre étude, nous voulions observer en quoi celle-ci pouvait permettre d'organiser une liste de noms brute. Nous avons, exploité l'outil word2vec (la méthode SGNS), sur un corpus scientifique transdisciplinaire. Nous avons à la suite de cela, utilisé un algorithme de classification automatique afin de générer 30 clusters à partir d'une liste de 461 noms du Lexique Scientifique Transdisciplinaire (LST).

Après avoir analysé manuellement ces 30 clusters, nous avons pu, entre autres, observer le type de lien sémantique que nous avons entre les différents noms. Cette analyse sémantique nous a également permis d'appréhender la qualité de nos résultats. Ainsi, plusieurs points ont été mis en relief lors de cette analyse.

Dans un premier temps, nous sommes satisfaite des clusters que nous avons générés. En effet, ils sont pour la plupart homogènes et peu de noms sont inclassables. Nous avons notamment obtenu une précision de 94,577% qui nous permet de conclure qu'il y a eu peu de noms « mal classifiés » par l'algorithme.

Ensuite, les types de relations sémantiques que nous avons entre les mots sont assez variés et intéressantes. Nous avons effectivement obtenu dans la majorité des cas de la co-hyponymie, quelques relations de quasi-synonymie et aussi de l'antonymie. Ce qui a également particulièrement retenu notre attention est que la méthode SGNS permet aussi d'exploiter des relations qui ne sont pas d'ordre paradigmatique. Nous avons observé quelques cas où les noms étaient associés entre eux car ils appartenaient à un même domaine, comme par exemple le domaine du travail.

L'utilisation d'un algorithme de classification nous a permis d'améliorer cette liste de noms brute du LST. Tout d'abord, elle a permis de mettre en évidence les noms n'appartenant pas au LST. La plupart des erreurs de classification de noms que nous avions étaient dues au fait qu'ils n'étaient pas transdisciplinaires.

Néanmoins, l'utilisation seule de la méthode SGNS et de la classification automatique n'est pas suffisante. L'intérêt de nos travaux réside également dans le fait d'avoir apporté une réelle analyse manuelle à nos clusters. Nous avons pu ainsi décrire sémantiquement chacun d'entre eux afin d'observer ce que nous avions obtenu avec précision. Cela permet également de faciliter l'utilisation de ces classes et donc de cette liste de noms du LST. Effectivement, nous sommes convaincue qu'apporter une classification et une description sémantique aux

membres du LST est une véritable aide pour son utilisation. Sur le plan didactique, cela facilitera son emploi pour les locuteurs natifs ou les apprenants étrangers dans la rédaction et compréhension des méthodologies scientifiques.

Nous sommes satisfaite de l'utilisation de la méthode distributionnelle que nous avons exploitée, à savoir la méthode SGNS. Celle-ci, après les expérimentations que nous avons réalisées, s'est avérée être plus performante que les PMI et SVD. La qualité des résultats qu'elle propose est pertinente, bien que d'après nous, elle ne puisse se passer d'une analyse manuelle. Nous envisageons donc cette méthode, comme une base solide, permettant de générer des regroupements de qualité et de faciliter ce type de travaux, pouvant devenir rapidement fastidieux réalisés de façon manuelle. Dans notre étude, le problème le plus important a été le traitement de la polysémie qui n'était pas pris en compte par notre algorithme de classification.

Nous pouvons donc conclure que la méthode SGNS basée sur les réseaux de neurones artificiels est un véritable atout pour le traitement automatique des langues. Dans le cadre de notre étude, elle s'est prêtée de manière adéquate à l'organisation sémantique d'une liste de noms brute. Nous avons été convaincue de sa fiabilité reposant notamment sur un réglage des différents paramètres judicieux. Enfin, l'analyse sémantique nous paraît avoir été un complément essentiel dont nous n'aurions pu nous passer pour la compréhension de nos clusters.

Perspectives

À la suite de ces travaux, l'étape la plus immédiate serait le traitement de polysémie. En effet, il semble difficilement envisageable de proposer des classes sémantiques sans proposer les différents sens d'un nom. Cela apportera des regroupements avec une meilleure qualité et pourra permettre une utilisation didactique. Les apprenants étrangers et parfois même les locuteurs natifs ont besoin de savoir quels sont les usages exacts d'un mot et la polysémie est très présente dans le lexique scientifique transdisciplinaire. Nous avons notamment pu observer cela grâce aux classes de Hatier el al. (2016).

De plus, afin d'améliorer nos résultats, nous pensons qu'observer les noms en contextes serait un réel atout. En effet, cela aiderai à mieux comprendre les regroupements générés par l'analyse distributionnelle et faciliterai ainsi la description des classes et leur étiquetage. Des travaux sur le LST ont déjà montré qu'observer les cooccurrents d'un mot aidait à l'analyse sémantique. C'est notamment le cas des travaux de Galop (2011), qui consistaient à observer de quelle manière les relations syntaxiques et la position du mot dans un texte influençaient sur la proximité sémantique des mots.

À terme, nous pensons qu'il serait intéressant de pouvoir diffuser ces différentes classes sémantiques et leur description afin de pouvoir apporter un réel outil didactique. Cela facilitera l'usage des noms du LST pour la rédaction des méthodologies scientifiques.

Enfin, il semblerait également pertinent de réaliser le même type de traitement aux autres listes de mots du LST. Drouin (2007) avait également extrait les listes des verbes et des adjectifs du LST. Cela permettrait d'obtenir un lexique complet avec des descriptions fines. Néanmoins, les traitements seront certainement différents, notamment pour les verbes où nous avons appliqué un algorithme de classification automatique. Il semblait que les regroupements générés contenaient des verbes dont les liens sémantiques étaient difficiles à entrevoir. Il faudra donc trouver les traitements adaptés à ces deux classes du LST.

Bibliographie :

BERNIER-COLOBORNE, Gabriel. (2014). Analyse distributionnelle de corpus spécialisés pour l'identification de relations lexico-sémantiques. *21ème Traitement Automatique des Langues Naturelles*, Marseille.

BERNIER-COLOBORNE, Gabriel. (2015). Exploration de modèles distributionnels au moyen de graphes 1-PPV. *22ème Traitement Automatique des Langues Naturelles*. Caen.

COXHEAD, Averil. (2000). *A new academic word list*. TESOL Quarterly, 34(2):213–238.

COXHEAD, Averil. (2014). Corpus linguistics and vocabulary teaching: Perspectives from english for specific purposes. *In Corpus Analysis for Descriptive and Pedagogical Purposes : ESP Perspectives, Linguistic Insights*, Peter Lang Publishing, Incorporated, p. 289–301.

CRUSE, D. A. (1986). Lexical Semantics. Cambridge University Press.

FABRE, Cécile. & MORLANE-HONDERE, François. (2012). Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales. *CMLF 2012*, Jul 2012, France, p.1001 - 1015.

FABRE, Cécile, & LENCI, Alessandro. (2015). Distributional Semantics Today. *In Traitement automatique des langues, Sémantique distributionnelle*, 2015, Vol. 56- n°2.

DROUIN, Patrick. (2007). Identification automatique du lexique scientifique transdisciplinaire, *Revue française de linguistique appliquée* (Vol. XII), p. 45-64.

DROUIN, Patrick. (2010). From a bilingual transdisciplinary scientific lexicon to bilingual transdisciplinary scientific colloa- tions. In *Anne Dykstra et Tanneke Schoonheim*, éditeurs : Proceedings of the 14th EURALEX International Congress, pages 296–305, Leeuwarden/Ljouwert, Pays-Bas, Fryske Akademy.

GALOP, Mickaël. (2011). *Prise en compte de variables syntaxiques et textuelles dans l'analyse sémantique distributionnelle automatique. Mémoire de master 2 recherche, Industries de la langue*, ss. dir. Agnès tutin, , Université Stendhal-Grenoble3 : Grenoble.

GOUGENHEIM, G., MICHEA, R., RIVENC, P. et SAUVAGEOT, A. (1956), *L'élaboration du français élémentaire. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris : Didier.

GROUIN, Cyril. & FOREST, Dominic. (2012). *Expérimentations et évaluations en fouille de textes: un panorama des campagnes DEFT*. Paris, Hermès-Lavoisier. 248 pages.

HATIER, Sylvain. (2013). Extraction des mots simples du lexique scientifique transdisciplinaire dans les écrits de sciences humaines : une première expérimentation. *In Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues* (RECITAL'2013), pages 138–149, Les Sables d'Olonne, France. 00000.

HATIER, Sylvain., TUTIN, Agnès., JACQUES, Marie-Paule. JACQUEY, Évelyne. & KISTER, Laurence. (2014). Catégorisation sémantique des noms simples du lexique scientifique transdisciplinaire.

HATIER, Sylvain., TUTIN, Agnès., JACQUES, Marie-Paule., JACQUEY Évelyne. & KISTER, Laurence. (s.d.). Extraction et traitement sémantique des noms simples du lexique scientifique transdisciplinaire.

HATIER, S., & YAN, R. (2015). Comparaison de constructions verbales entre un corpus d'apprenants et un corpus d'articles de recherche. *8es Journées Internationales de Linguistique de Corpus (JLC2015)*, Orléans.

HATIER, S., AUGUSTYN, M., Yan, R., TRAN, T. T. H., TUTIN, A., & JACQUES, M.-P. (2016). French cross-disciplinary scientific lexicon: extraction and linguistic analysis. In G. Meladze (Éd.), *Proceedings of the XVII EURALEX International congress* (p. 355-365).

JANOD, Killian., MORCHID, Mohamed., DUFOUR, Richard. & LINARES, Georges. (2015). Apport de l'information temporelle des contextes pour la représentation vectorielle continue des mots, *22ème Traitement Automatique des Langues Naturelles,* Caen.

KRAIF, Olivier. & DIWERSY, Sascha. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. In *Actes de la conférence conjointe* JEP-TALN-RECITAL 2012, volume 2 : TALN, Grenoble, 4 au 8 juin 2012, pages 399–406.

LANDAUER, Thomas K. & DUMAIS, Susan T. (1997). A solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition. *Induction, and representation of Knowledge,* Vol. 104, No 2, p. 211-240.

LANDAUER, T. K., FOLTZ, P. W., & LAHAM, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes,* **25**, 259-284.

LAFON, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, 1, 128-165.

LEVY, Omar. & GOLDBERG, Yoav. (2014). Neural Word Embedding as Implicit Matrix Factorization.

LEVY, Omar., GOLDBERG, Yoav. & DAGAN, Ido. (2015). Improving distributionnal Similarity with lessons learned word embeddings.

MEL'CUK Igor. (1996). Lexical function : A tool for the description of lexical relations in the lexicon. In *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamin.

MIKOLOV, T., CORRADO, G., CHEN, K. & DEAN, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations, (ICLR 2013)*, p. 1–12.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. & DEAN, J. (2013b).

Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, p. 3111–3119.

MIKOLOV, T., YIH, W.-T. & ZWEIG, G. (2013c). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, p. 746–751.

MURPHY, L. (2003). Semantic relations and the lexicon. Cambridge University Press.

PERINET, Amandine. & HAMON, Thierry. (2014). Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité.

PERINET, Amandine. & HAMON, Thierry. (2015). Analyse distributionnelle appliquée aux textes de spécialité. Réduction de la dispersion des données par abstraction des contextes. In *Traitement automatique des langues*, Sémantique distributionnelle, Vol. 56- n°2.

PERINET, Amandine. (2015). Analyse distributionnelle appliquée aux textes de spécialité : réduction de la dispersion des données par abstraction des contextes. Computer Science [cs]. Université Paris 13; Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Sant´e. French.

PECMAN, Mojca. (2004). Phraséologie contrastive anglais-français : Analyse et traitement en vue de l'aide à la rédaction scientifique.

PICTON, Aurélie., FABRE Cécile. & BOURIGAULT Didier. (2008). « Méthodes linguistiques pour l'expansion de requêtes. Une expérience basée sur l'utilisation du voisinage distributionnel », *Revue française de linguistique appliquée* (Vol. XIII) , p. 83-95

PHAL, André. (1976). Vocabulaire général d'orientation scientifique (V.G.O.S.), part du lexique commun dans l'expression scientifique. *Linguistique appliquée*. Didier.

POLGUERE, Alain. (1998) La théorie Sens-Texte. *Dialangue*, Vol. 8-9, Université du Québec à Chicoutimi, pp. 9-30.

POLGUERE, Alain. (2002). Modélisation des liens lexicaux au moyen des fonctions lexicales, Nancy, 24-27.

TRAN, T-T-H. (2014). *Description de la phraséologie transdisciplinaire scientifique et réflexions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs* (Thèse de doctorat). Université de Grenoble, Grenoble.

TUTIN, Agnès. (2007). « Autour du lexique et de la phraséologie des écrits scientifiques », *Revue française de linguistique appliquée*, (Vol. XII), p. 5-14.

TUTIN Agnès. (2014). La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico- rhétorique. *In L'écrit scientifique, du lexique au discours, Collection Rivages linguistiques*, pages 24–44. Presses de l'Université de Rennes.

TUTIN, Agnès., THI THU, Hoai Tran., KRAIF, Olivier. & HATIER, Sylvain. (s.d). French collocations of cross-disciplinary scientific lexicon.

WANNER, Leo. (1996). Lexical functions *in lexicography and natural language processing*.

John Benjamins.

HARRIS, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Sitographie

DUBOIS, J., DUBOIS-CHARLIER, F. (2010). Dictionnaire Électronique des Mots. [En ligne] http://rali.iro.umontreal.ca/rali/?q=fr/dem consulté août 2016

DUBOIS, J., DUBOIS-CHARLIER, F. (1997). Les Verbes du Français. [En ligne] http://rali.iro.umontreal.ca/rali/?q=fr/lvf consulté août 2016

FOLGERT, Kardsorp. (2015). Word2Vec: an introduction. [En ligne] http://www.folgertkarsdorp.nl/word2vec-an-introduction/ consulté décembre 2015.

LAFOURCADE, M . (2008). Jeux de Mots. [En ligne] https://linuxfr.org/news/jeuxdemots-un-jeu-en-ligne-pour-produire-des-données-lexicales consulté août 2016

LEVY, Omar. (2015). How does work Word2vec? [En ligne] https://www.quora.com/How-does-word2vec-work consulté décembre 2015.

LEVY, Omar. (2015). Hyperwords. [En ligne] https://bitbucket.org/omerlevy/hyperwords consulté janvier 2015.

WIKIPÉDIA. [En ligne] https://fr.wikipedia.org/wiki/Accueil consulté août 2016.

Liste des tableaux

Table des annexes

Annexe 1 : Classes sémantiques des noms du LST de Hatier et al. (2014)

Classe	Sous-Classe	Noms
	Orientation : sens – être sur un N	axe 02, direction 01, orientation, sens 03, voie 01
etat_humain :, -expérienceur humain, **-**faire preuve de N		caractère 02, préoccupation, souci 01, concentration 03, expérience 02
etat_qualité : , **-** état ou qualité d'une entité abstraite ou concrète, **-** avoir Dét N ?,	**Axiologique-négatif** : Jugement négatif de valeur sur la qualité/état	absence 02, biais 04, défaut, manque 01
	Axiologique-positif : Jugement positif de valeur sur la qualité/état **-** Trouver que ça présente Det N	avantage 01, cohérence 02, compétence 01, intérêt 03, maîtrise 02, mérite, pertinence 01, pertinence 02, précision, rigueur 01, utilité, validité
	Caractéristique:manière dont une personne ou une chose se présente à la vue ou à l'esprit **-** présenter Dét N	apparence 02, aspect 02, caractère 01, caractéristique 02, état 01, forme 01, identité 02, nature 03, particularité, profil 02, propriété 02, qualité 01, trait U3-01
	Certitude : Expression de l'attitude du locuteur par rapport au contenu propositionnel de son énoncé	possibilité 01, impossibilité 01, incertitude
	Changement-positif : Présence de variation, modification **-** Dét N débute	dynamique 04, tendance 01, rupture 01
	Changement-négatif : Absence de variation, modification **-** assurer Dét N	continuité 01, stabilité 01
	Complexité : présenter un dégré de N	ambiguïté 01, complexité 01, difficulté, obstacle, problème 02,
	Composition : Nature d'une entité en tant qu'unité **-** une entité est caractérisée par sa/son N	composition U3-01, configuration 02, distribution 01, diversité 02, répartition, Structuration U3-01, unité 02, variété, version 02
	Évaluation/capacité : qualité d'une entité à évaluer **-** évaluer Dét N	aptitude 01, capacité 01, faculté 01, performance 04, qualité 02, valeur 04
	Importance - accorder Dét N	fonction 03, importance, poids 04, portée 04, préférence 01, priorité 01, rôle 02, statut
	Manière - Procéder à la N de	façon 02, manière 01, modalité 02, mode 01

Classe	Sous-Classe	Noms
	Nécessité	exigence, nécessité 01, nécessité 02
	Nouveauté-positif : Degré d'originalité	originalité 01, singularité 01, spécificité
groupe/partie de :, -Dénote un groupe ou une partie , -diviser en plusieurs N	**partie**: Groupe rassemblant plusieurs éléments - faire partie de Dét N	échantillon 03, élément, extrait 01, partie 01, section 02, segment 02, unité 01
	groupe: Élément membre d'une entité plus grande - extraire Dét N d'un tout	catégorie 01, classe 01, ensemble 04, groupe 01, série 02, totalité 01
Objet_scientifique :, -Renvoie aux observables et objets construits par l'activité scientifique,	**artefact** :Objet abstrait, élément construit pour l'activité scientifique	composante, état 05, genre 01, image 02, norme
	équivalence: Lien d'équivalence entre des observables - X est le N de Y ?,	équivalent 02, figure 01, indice 01, marque 03, référence, repère, signe 01, symbole 01
	explicatif : Élément de l'argumentation scientifique	analyse U3-1, argument 01, code 02, conclusion 02, contenu 02, critère, dynamique 02, exemple 02, facteur 02, hypothèse 02, indicateur 04, loi 02, modèle 03, modèle 01, niveau, notion, paradigme, paramètre 01, plan 04, principe, rang, règle 01, réponse 02, réponse 01, représentation 01, résultat 01, sens 04, signification, solution 02, témoignage, théorie 01, thèse 01, type 01, typologie, variable 03
	Instrument : Instrument - à l'aide de N	instrument 01, mesure 02, moyen 01, outil 04, outil 01, support
	méthode : Objet abstrait renvoyant à la méthode de recherche scientifique	approche 03, convention 01, démarche 01, dispositif 02, formule 05, logique 03, matériel 02, méthode, méthodologie, plan 03, positionnement, posture U3-01, procédé 02, procédure 02, programme 02, programme 05, questionnaire, recherche U3-01, réseau 01, schéma 02, stratégie, structure, système 03, technique 03, technique 04, technologie, vision 03, voie 02
	objectif: Objet abstrait renvoyant au but - Nous visons tel N	but 04, cible, enjeu 02, finalité 02, motif 01, objectif 04, objet 03, projet 01
	objet-mathématiques : Objet abstrait du domaine des mathématiques	axe 04, fonction 04, formule U3-1, indice 02, logique 02
	Observable : Renvoie aux observables	contenu 01, corpus 01, couple 01, donnée, entité, erreur 01, flux 01, information, instance U3-1, marque

Classe	Sous-Classe	Noms
		01, matériau 02, matériau 01, objet 01, point 01, population 02, réel 02, vocabulaire
	Situation : Situation, cas	alternative 01, cas 01, circonstance, condition 02, événement, exception 01, fait 01, option 01, phénomène 02, situation 02
	thème : Objet abstrait renvoyant au thème, au sujet -Le N de l'article est X	discipline 02, problème 03, science, angle u3-1, connaissance 01, idée, idéologie, perspective 02, point 03, problématique 02, projet 03, question 01, savoir 01, sujet 01, thématique 02, thème 01
Personne :, -Renvoie à une personne ou un groupe de personnes, -+humain	**Collectif** : Groupe de personnes	classe 02, collectif 02, communauté 02, faculté 02, groupe 02, institut, laboratoire, milieu 02, mouvement 02, organisation 02, organisme 02, population 01, université 02
	Individu	auteur 03, chercheur 03, expert 01, individu, membre 02, observateur 02, pair U3-01
processus_cognitif :, -Renvoie aux processus cognitifs, -,	**Choix**	adoption U3-1, appropriation 02, choix, décision 02, rejet 01, sélection 01
	Classement	catégorisation, classement, classification, distribution 04, organisation 01, structuration
	Constat	constat 02, découverte 01, description, identification, observation 01, perception 01, traduction
	Évaluation : Processus d'évaluation - émettre un N sur X	appréciation 01, appréhension 02, compréhension 01, considération 01, définition 01, détermination 01, estimation 02, évaluation, interprétation 01, jugement 02, lecture U3-01, localisation, opinion, reconnaissance 03, mesure 01
	Explication : Processus d'explication - proposer un N	analyse 01, attribution 01, critique 06, étude 01, examen 01, expertise 01, explication 02, généralisation U3-01, illustration 02, interrogation, justification 01, présentation 01, questionnement, raisonnement, réflexion 01, résolution 03, synthèse 01, démonstration 01
	Inclure_séparer : Processus groupant ou	différenciation, distinction 01, division 01,

Classe	Sous-Classe	Noms
	séparant les objets	intégration 01, introduction 01
	Réalisation: Processus de réalisation	composition 02, conception 01, constitution 01, construction 01, création 01, élaboration, préparation, production 01, réalisation 01, Recueil U3-01
processus_evolutif :, -Renvoie aux processus dénotant une évolution, un changement, -constater un N depuis des années	**Amélioration_augmentation**: Processus dénotant une évolution positive	accroissement, affirmation U3-1, amélioration, augmentation, avancée 01, croissance 02, développement 02, expansion 04, extension, gain, généralisation, multiplication 01, progrès 02, progression, renforcement
	Baisse: Processus dénotant une évolution négative	baisse 01, déclin, diminution, perte 01, réduction 01
	Changement: Processus dénotant une évolution	changement 02, communication 01, déplacement 01, diffusion 02, évolution 02, évolution 01, modification, mouvement 01, mutation 02, tendance 02, transfert 01, transformation 01, transition, transmission 01, variation
	Manifestation :	apparition 01, émergence, expression U3-1, genèse, manifestation 01
processus_gen : Renvoie aux processus de fonctionnement **-** qc a 1 N / le N de qc ?	**Action**	activité 01, apport 01, comportement, contribution 01, fonctionnement, opération 01, processus, tâche 01, traitement 04, travail 05,
processus_humain : Renvoie aux processus impliquant un sujet humain	**Interaction :** Processus d'interaction entre humains -Un N entre X+hum et Y+hum	collaboration 01, communication 02, compromis, consensus, interaction 01, publication 01
	Usage : Processus générique	acquisition, application 02, échec 01, épreuve 01, implication 01, maîtrise 01, manipulation, pratique 01, pratique 02, recours 01, tentative, tradition, usage 01, utilisation
processus_scientifique :, -Renvoie aux processus typiques de l'activité scientifique, **-** Le chercheur effectue un N	**Proc_mathématiques** : Processus typique de l'activité scientifique du domaine des mathématiques **-** Un N mathématique	calcul 01, division 02, multiplication 02
	Méthodologie : Processus typique de la méthodologie de l'activité scientifique **-** L'auteur a mené un N	enquête 01, entretien 02, essai 01, expérience 01, expérimentation, investigation, recherche 01, test 02,

Classe	Sous-Classe	Noms
Quantité :Renvoie à la notion de quantité	**Grandeur** : Grandeur - le N total	ampleur, capacité 02, concentration 01, dimension 01, échantillon 01, effectif 02, faiblesse 01, fréquence 03, intensité 01, note 04, quantité 02, seuil 02, taille 01, volume 04
	Rapport : Rapport entre plusieurs quantités - calculer N	bilan 02, différence 03, distance 01, distance 02, écart 01, échelle 04, égalité 03, indice 03, majorité 01, maximum, minimum 01, minorité 03, moyenne, pourcentage, proportion 01, somme 02, statistique 02, taux 01, tiers U3-01, total 02
	Unité : Unité de mesure, de calcul - ajouter un N	chiffre 01, chiffre 03, degré 02, nombre 01, point 05
Relation :, -Dénote une relation entre entités, -Il existe une relation de N entre ... ?,	**Association**: Relation de structuration d'une entité permettant sa description - Dét N entre deux éléments	appartenance 02, articulation 04, association 01, combinaison 01, corrélation, équilibre 01, indépendance 01, interaction 02, liaison 01, lien 02, loi 03, ordre 02, relation 02
	Correspondance : Correspondance entre entités - Un N est établi entre X et Y	comparaison 01, convergence 02, correspondance 01, identité 01, proximité 01, rapprochement
	Implication : Renvoie à une relation d'implication - X a/est un N pour Y	base 01, cause 01, condition 03, conséquence 01, contrainte, dépendance 01, effet, explication 01, fondement 01, impact U3-01, implication 02, incidence, indication, influence 02, mécanisme 02, rapport 04
	Opposition : Renvoie à une opposition entre entités - Un N est établi entre X et Y	avantage 02, clivage, confrontation 01, contradiction 01, contraste, différence 01, divergence 01, divergence 02, division 05, opposition 01, paradoxe
Temporalité :, -nom renvoyant au temps, -Avoir lieu pendant/à N,	**Chronologie** : nom renvoyant à un instant donné - a lieu à Dét N	approche 02, date, étape 02, horizon 03, origine 01, terme 01
	Durée : nom renvoyant à une durée, - pendant Dét N	année 01, cycle 02, durée 01, période 01, phase 02, siècle 02
	Fréquence	fréquence 01
Temporalité/espace ? : Temps et à l'espace		absence 01, défaut U3-01, existence, présence 01

Annexe 2 : Clusters méthode PMI, clustering K-means:

0 reprise transition sou clé intérieur

1 milieu monde mouvement face famille tête temps esprit main maison société côté coeur communauté commune corps culture homme association région vie ville origine groupe dernier droit partie pays personne population premier valeur raison élément forme trait structure lien logique nature composante contrainte puissance aspect vue dimension dynamique

2 représentant force majorité membre sein charge contrôle sécurité protection puissance aide autorité décision direction plupart prise réseau monde terrain territoire emploi espace matière secteur lieu zone cas centre champ contexte histoire projet objet domaine

3 regard reste résistance foi espèce sorte loi compte considération attention réaction disposition part peine place figure erreur exception extérieur sens limite côté caractéristique distribution
4 énergie équipe établissement faute totalité tour essentiel extérieur manque masse source seul soin chemin clé contribution absence appel apport voie grâce dehors disparition donnée intérieur preuve écart énergie minimum mois qualité quantité masse surface somme unité production
5 modalité tâche texte titre mécanisme méthode stratégie support contenu article objet opération ouvrage document image procédure rapport échelle moyen facteur terme statut niveau cadre condition critère profil système origine degré donnée plan
6 figure tableau type ensemble suite liste nom catégorie classe série époque étape fois maison section catégorie classe cours heure jour an année départ date période phase
7 fond terrain trace est exemple surface ligne cas chose contact propos angle auteur axe reconnaissance extension mémoire main compréhension compte contact contrôle dépendance détail détermination identification identité intégration prise
8 réseau territoire entreprise environnement espace marché structure secteur zone centre champ contexte organisation domaine institution ressource technique ensemble environnement support ligne code connaissance contenu produit programme objectif outil dispositif image indice information
9 échelle économie équilibre état proximité distance idéal variation événement évolution modification trace transformation changement amélioration déplacement
10 moment fait tentative travail effet effort lieu but choix comportement conflit contraire acte discours discussion doute phénomène point économie équilibre équipe société communauté autorité
11 ressource matière mesure service connaissance produit programme projet action activité outil information instrument intervention production représentant famille tête majorité membre sein couche région ville direction personne
12 recherche représentation étude théorie lecture base comparaison conception analyse approche réflexion référence vision observation définition démarche description interprétation perspective présentation résistance modalité question faute tendance effet effort manque souci succès liberté besoin biais nécessité coeur compétence conflit considération constat contraire courant propriété proximité absence avantage règle réaction réalité volonté opposition décision dehors différence difficulté doute particularité partie peine phénomène précision problème problématique
13 résultat remarque formule thèse expression section conclusion constat hypothèse proposition argument idée moitié moment fin siècle long naissance début
14 tendance exigence situation nature norme composante condition contrainte orientation dimension position principe problématique établissement milieu mouvement face enjeu entreprise sujet service commune communication corps homme acteur appartenance appel association auteur organisation groupe individu institution pays population pratique premier
15 élément événement mot expérience mémoire signe littérature caractéristique cause compétence courant histoire propriété accord appartenance règle réalité identité répartition exercice vie
16 valeur être moyen façon fonction emploi manière pouvoir risque fréquence taille temps transfert marché nombre charge coût profit proportion demande distance durée paramètre perte poids pression prix

17 rôle rapport relation modèle mode question facteur forme technique terme trait enjeu statut sujet schéma lien logique notion cadre caractère concept profil acteur aspect système vue objectif ordre dispositif dynamique indice individu plan pratique problème processus respect espèce conservation sécurité protection aide gestion
18 moitié fin siècle cours année début être est pouvoir
19 époque étape sortie naissance départ date issue période phase remarque formule taux conclusion proposition
20 évidence exception lumière oeuvre défaut inverse parallèle recherche étude thèse théorie traitement travail type méthode lecture littérature base comparaison conception analyse approche réflexion observation démarche description distinction interprétation intervention perspective présentation
21 minimum mois sou code heure jour an relation représentation façon fait fonction expression manière situation solution caractère chose action activité référence vision définition
22 taux couche évidence fond lumière séparation angle oeuvre issue parallèle preuve
23 fois sens chapitre cité composition synthèse rôle force efficacité existence stabilité complexité apport attention augmentation orientation impact importance influence intérêt part place point position présence
24 limite niveau critère degré paramètre reste mot tableau texte titre exemple source nom cité propos acte argument article synthèse ouvrage discussion document
25 répartition qualité quantité fréquence taille erreur somme long nombre coût profit proportion unité demande distribution durée prix regard modèle mode exigence mécanisme schéma loi norme notion cause concept accord ordre droit instrument principe procédure processus
26 recours échange fonctionnement formation traitement exercice communication construction jeu usage utilisation accès gestion développement interaction introduction recours retour fonctionnement seul signe choix comportement composition culture jeu usage utilisation accès grâce discours disposition idéal interaction
27 reconnaissance reprise respect retour transfert transition existence extension maintien compréhension conservation constitution création séparation application réalisation opposition ouverture diffusion distinction identification intégration présence résultat réponse tâche tentative esprit expérience mesure sorte stratégie soin liaison but capacité chemin conséquence hypothèse axe voie défaut dernier idée possibilité
28 risque écart efficacité stabilité succès liaison biais complexité amélioration augmentation dépendance différence impact importance influence intérêt perte poids pression foi sortie liste contribution apparition disparition introduction inverse passage
29 raison réponse souci solution liberté besoin nécessité capacité conséquence avantage volonté difficulté particularité possibilité rupture échange état mise formation totalité tour essentiel maintien suite chapitre constitution construction création série application réalisation opération ouverture développement diffusion plupart
30 variation rupture évolution mise modification transformation changement apparition déplacement détail détermination passage précision

Annexe 3 : Clusters méthode PMI, clustering Hclust :

1 milieu monde mouvement face famille tête temps esprit main maison société côté coeur communauté commune corps culture homme association région vie ville origine groupe dernier droit partie pays personne population premier
2 représentant force majorité membre sein charge contrôle sécurité protection puissance aide autorité décision direction plupart prise
3 regard reste résistance foi espèce sorte loi compte considération attention réaction disposition part peine place
4 énergie équipe établissement faute totalité tour essentiel extérieur manque masse source seul soin chemin clé contribution absence appel apport voie grâce dehors disparition donnée intérieur preuve
5 modalité tâche texte titre mécanisme méthode stratégie support contenu article objet opération ouvrage document image procédure
6 figure tableau type ensemble suite liste nom catégorie classe série
7 fond terrain trace est exemple surface ligne cas chose contact propos angle auteur axe
8 réseau territoire entreprise environnement espace marché structure secteur zone centre champ contexte organisation domaine institution
9 échelle économie équilibre état proximité distance idéal
10 moment fait tentative travail effet effort lieu but choix comportement conflit contraire acte discours discussion doute phénomène point
11 ressource matière mesure service connaissance produit programme projet action activité outil information instrument intervention production
12 recherche représentation étude théorie lecture base comparaison conception analyse approche réflexion référence vision observation définition démarche description interprétation perspective présentation
13 résultat remarque formule thèse expression section conclusion constat hypothèse proposition argument idée
14 tendance exigence situation nature norme composante condition contrainte orientation dimension position principe problématique
15 élément événement mot expérience mémoire signe littérature caractéristique cause compétence courant histoire propriété accord appartenance règle réalité identité
16 valeur être moyen façon fonction emploi manière pouvoir
17 rôle rapport relation modèle mode question facteur forme technique terme trait enjeu statut sujet schéma lien logique notion cadre caractère concept profil acteur aspect système vue objectif ordre dispositif dynamique indice individu plan pratique problème processus
18 moitié fin siècle cours année début
19 époque étape sortie naissance départ date issue période phase
20 évidence exception lumière oeuvre défaut inverse parallèle
21 minimum mois sou code heure jour an
22 taux couche
23 fois sens chapitre cité composition synthèse
24 limite niveau critère degré paramètre
25 répartition qualité quantité fréquence taille erreur somme long nombre coût profit proportion unité demande distribution durée prix
26 recours échange fonctionnement formation traitement exercice communication construction jeu usage utilisation accès gestion développement interaction introduction
27 reconnaissance reprise respect retour transfert transition existence extension maintien compréhension conservation constitution création séparation application réalisation opposition ouverture diffusion distinction identification intégration présence
28 risque écart efficacité stabilité succès liaison biais complexité amélioration augmentation dépendance différence impact importance influence intérêt perte poids pression
29 raison réponse souci solution liberté besoin nécessité capacité conséquence avantage volonté difficulté particularité possibilité

30 variation rupture évolution mise modification transformation changement apparition déplacement détail détermination passage précision

Annexe 4 : Clusters méthode SGNS, clustering Hclust :

1 variation rupture évolution modification transformation existence stabilité limite caractéristique complexité séparation utilisation amélioration apport avantage opposition dépendance détermination différence difficulté distinction impact importance influence particularité précision présence
2 recours respect retour échange mise maintien manque succès biais absence accès appel départ déplacement passage prise
3 répartition minimum transfert profit augmentation distribution perte poids pression
4 raison énergie qualité quantité faute masse surface somme coût conséquence unité défaut demande donnée partie peine preuve prix
5 équipe établissement fonctionnement formation traitement exercice service soin composition conservation usage application gestion développement diffusion production
6 relation trace contact contribution observation discussion interaction
7 efficacité base nature charge communication compétence connaissance sécurité protection action association attention organisation orientation direction disposition information institution intervention position
8 reprise évidence mois fois taux exception sou suite sens lumière compte considération cours oeuvre introduction inverse issue
9 reste face fin tête temps tour ensemble essentiel extérieur main seul siècle liste naissance côté clé série origine début dehors dernier intérieur part place premier
10 extension sortie constitution construction création apparition réalisation ouverture disparition intégration
11 reconnaissance résistance foi totalité liberté contrôle réaction décision identification
12 valeur représentation réponse façon tendance expression manière propriété réalité référence vision image
13 ressource question fonction technique matière mesure source situation solution cause condition aide grâce pratique
14 économie époque maison société cité communauté commune culture région ville pays
15 force espèce esprit expérience mémoire catégorie histoire puissance appartenance autorité vie idéal identité
16 réseau terrain territoire entreprise environnement espace marché secteur zone centre champ contexte domaine
17 rôle être milieu monde mouvement membre statut sujet sein nom code coeur comportement corps profil acteur groupe droit individu pouvoir
18 représentant écart moitié famille fréquence taille majorité liaison nombre classe comparaison couche proportion proximité distance personne plupart population
19 long heure jour an année date durée période
20 échelle étape transition phase
21 modalité tâche tentative méthode stratégie activité voie opération démarche procédure
22 recherche remarque étude figure formule tableau thèse théorie section lecture littérature compréhension conception conclusion proposition analyse approche réflexion synthèse définition détail description interprétation perspective présentation problématique
23 rapport regard risque mot moyen fait fond terme titre emploi erreur est exemple signe ligne cas chemin chose contraire courant homme doute intérêt parallèle
24 chapitre propos acte auteur discours
25 texte travail effort support but contenu produit programme projet article objectif objet outil ouvrage dispositif document instrument
26 résultat loi constat hypothèse argument règle idée
27 élément événement facteur trait enjeu caractère aspect axe indice phénomène problème processus
28 moment sorte lieu niveau critère degré paramètre point
29 équilibre état effet exigence souci lien besoin nécessité capacité changement conflit accord volonté possibilité

30 modèle mode forme type mécanisme structure schéma logique norme notion cadre choix composante concept contrainte jeu angle système vue ordre dimension dynamique plan principe

Annexe 5 : Clusters méthode SGNS, clustering K-means :

0 expérience mémoire histoire vie observation
1 variation rupture évolution modification transformation existence stabilité limite caractéristique complexité séparation utilisation amélioration apport avantage opposition dépendance détermination différence difficulté distinction impact importance influence particularité précision présence recherche étude formation lecture littérature compétence compréhension connaissance réflexion discussion
2 recours respect retour échange mise maintien manque succès biais absence accès appel départ déplacement passage prise échelle économie équilibre société communauté unité identité
3 répartition minimum transfert profit augmentation distribution perte poids pression rôle rapport recours fonctionnement statut succès caractère contenu usage accès accord apport objectif développement intérêt poids
4 raison énergie qualité quantité faute masse surface somme coût conséquence unité défaut demande donnée partie peine preuve prix variation respect écart modification moyen erreur exigence existence maintien stabilité liaison limite base biais but comparaison complexité contrôle séparation amélioration augmentation déplacement détermination identification impact précision
5 équipe établissement fonctionnement formation traitement exercice service soin composition conservation usage application gestion développement diffusion production élément état événement enjeu support changement conflit acte objet intérieur
6 relation trace contact contribution observation discussion interaction milieu moitié fin sein siècle début
7 efficacité base nature charge communication compétence connaissance sécurité protection action association attention organisation orientation direction disposition information institution intervention position représentant reste retour échange époque équipe établissement être mise mois moment face fait fois fond tête taux temps totalité tour emploi espèce essentiel est exception exercice extérieur membre sou suite sujet sens seul lieu long niveau nom côté chapitre chemin choix cité clé coeur contraire courant cours heure homme jour proximité acteur année appel référence oeuvre origine grâce départ dehors dernier donnée individu inverse parallèle part passage personne pouvoir premier prise
8 reprise évidence mois fois taux exception sou suite sens lumière compte considération cours oeuvre introduction inverse issue rupture mouvement effort esprit souci nécessité capacité projet volonté idéal idée
9 reste face fin tête temps tour ensemble essentiel extérieur main seul siècle liste naissance côté clé série origine début dehors dernier intérieur part place premier résultat remarque mécanisme comportement conclusion constat critère profil argument indice paramètre phénomène point preuve
10 extension sortie constitution construction création apparition réalisation ouverture disparition intégration terme notion concept principe
11 reconnaissance résistance foi totalité liberté contrôle réaction décision identification ressource risque énergie titre entreprise manque masse matière service soin besoin sécurité aide attention information
12 valeur représentation réponse façon tendance expression manière propriété réalité référence vision image reprise sortie naissance constitution construction création apparition réalisation ouverture disparition issue
13 ressource question fonction technique matière mesure source situation solution cause condition aide grâce pratique regard réseau modèle monde terrain territoire environnement espace secteur cadre centre champ contexte corps système vue ordre groupe discours domaine plan
14 économie époque maison société cité communauté commune culture région ville pays texte travail contribution article auteur ouvrage document
15 force espèce esprit expérience mémoire catégorie histoire puissance appartenance autorité vie idéal identité norme composante condition contrainte pression
16 réseau terrain territoire entreprise environnement espace marché secteur zone centre champ contexte domaine question faute tendance trace chose conséquence avantage réalité défaut difficulté doute importance particularité peine possibilité problème
17 rôle être milieu monde mouvement membre statut sujet sein nom code coeur comportement corps

profil acteur groupe droit individu pouvoir représentation réponse thèse théorie traitement expression mesure conception hypothèse utilisation analyse approche vision définition description interprétation
18 représentant écart moitié famille fréquence taille majorité liaison nombre classe comparaison couche proportion proximité distance personne plupart population mot facteur figure trait signe angle aspect axe dimension
19 long heure jour an année date durée période famille maison zone classe commune couche région ville pays population
20 échelle étape transition phase relation situation contact action interaction
21 modalité tâche tentative méthode stratégie activité voie opération démarche procédure lien caractéristique opposition différence distinction image influence
22 recherche remarque étude figure formule tableau thèse théorie section lecture littérature compréhension conception conclusion proposition analyse approche réflexion synthèse définition détail description interprétation perspective présentation problématique répartition minimum qualité quantité fréquence taille transfert efficacité marché surface somme nombre coût produit profit proportion demande distance distribution durée partie perte prix production
23 rapport regard risque mot moyen fait fond terme titre emploi erreur est exemple signe ligne cas chemin chose contraire courant homme doute intérêt parallèle mode type absence voie degré
24 chapitre propos acte auteur discours valeur raison reconnaissance résistance foi force main liberté catégorie cause charge code culture propriété puissance appartenance association autorité réaction organisation orientation décision direction disposition droit position
25 texte travail effort support but contenu produit programme projet article objectif objet outil ouvrage dispositif document instrument étape façon tâche tentative manière sorte schéma activité opération période problématique processus
26 résultat loi constat hypothèse argument règle idée modalité forme formule technique méthode source stratégie structure solution logique loi programme proposition application règle outil démarche date dispositif instrument perspective phase pratique procédure
27 élément événement facteur trait enjeu caractère aspect axe indice phénomène problème processus transformation extension nature communication composition conservation protection gestion diffusion institution intégration intervention
28 moment sorte lieu niveau critère degré paramètre point transition an dépendance
29 équilibre état effet exigence souci lien besoin nécessité capacité changement conflit accord volonté possibilité évidence évolution fonction tableau effet ensemble exemple majorité section ligne liste lumière cas compte considération jeu série propos synthèse détail dynamique introduction place plupart présence présentation
30 modèle mode forme type mécanisme structure schéma logique norme notion cadre choix composante concept contrainte jeu angle système vue ordre dimension dynamique plan principe

Annexe 6 : Clusters méthode SVD, clustering Hclust :

1 texte support contenu usage outil document information instrument

2 ressource réseau mode technique matière norme champ contexte programme projet système dispositif domaine pratique processus

3 valeur fait fonction tendance effet choix comportement objectif dynamique idée phénomène problème

4 modèle forme structure situation logique cadre dimension discours position principe
5 rôle relation enjeu statut lien notion composante concept aspect vue ordre interaction plan point
6 représentation façon théorie manière lecture nature caractère conception approche réalité vision définition description interprétation perspective
7 étude tâche expérience méthode stratégie schéma analyse réflexion règle observation opération démarche distinction problématique procédure
8 résultat élément facteur trait constat critère profil apport argument indice
9 risque effort exigence mesure besoin condition contrainte absence degré
10 réponse solution loi base connaissance contribution propriété unité attention référence orientation discussion disposition
11 raison tentative mécanisme source limite biais code conséquence courant acte angle avantage idéal
12 variation minimum quantité transfert somme coût augmentation perte poids
13 équilibre moyen qualité marché niveau produit profit demande prix production
14 reconnaissance échange composition constitution création accès réalisation ouverture diffusion identification présentation
15 énergie répartition fréquence taille erreur masse surface long nombre proportion distribution donnée durée paramètre
16 écart extension maintien stabilité liaison complexité compréhension séparation amélioration déplacement détermination différence précision
17 traitement communication utilisation application
18 respect rupture évolution modification fonctionnement transformation efficacité existence souci liberté nécessité capacité changement compétence contrôle volonté décision difficulté impact importance influence possibilité
19 échelle état événement fond force trace comparaison conflit contact proximité accord opposition dépendance distance intérêt présence pression
20 recours résistance mise face faute manque succès soin charge compte considération appel réaction défaut peine preuve prise
21 ligne cause chemin jeu axe voie place
22 remarque époque étape modalité formule thèse section caractéristique catégorie conclusion hypothèse série proposition année date période particularité phase
23 représentant reprise reste retour évidence mois moitié fin foi fois tête taux totalité tour transition espèce essentiel exception extérieur main maison majorité membre sortie sou sens siècle liste lumière but naissance chapitre cité clé coeur conservation cours heure jour an apparition synthèse oeuvre début départ détail dehors dernier disparition intérieur introduction inverse issue parallèle passage plupart premier
24 société littérature communauté construction sécurité protection aide gestion développement intégration
25 équipe établissement formation esprit exercice corps autorité direction droit
26 économie environnement culture histoire puissance action activité appartenance vie organisation identité institution intervention
27 milieu monde mouvement territoire espace zone acteur groupe
28 famille entreprise secteur service centre classe commune couche association région ville origine pays population

29 recherche figure tableau terrain est exemple article auteur ouvrage
30 rapport regard être moment mot question temps terme titre travail type emploi ensemble expression mémoire sorte suite sujet sein seul signe lieu nom côté cas chose contraire homme propos objet grâce doute image individu part partie personne pouvoir

Annexe 7 : Clusters méthode SVD, clustering K-means :

0 résultat rôle élément état événement fait fond temps tentative terme trace effet enjeu essentiel mécanisme sorte signe lieu caractère caractéristique cause concept conséquence constat propriété absence apport argument règle réalité objectif degré doute passage phénomène point problème
1 texte support contenu usage outil document information instrument recours risque rupture modification force tendance succès limite biais changement conflit avantage différence difficulté distance impact importance influence intérêt peine poids présence pression
2 ressource réseau mode technique matière norme champ contexte programme projet système dispositif domaine pratique processus remarque étape modalité mois fois espèce section catégorie chapitre série année opération date période particularité phase
3 valeur fait fonction tendance effet choix comportement objectif dynamique idée phénomène problème économie esprit société littérature communauté culture histoire puissance vie identité image
4 modèle forme structure situation logique cadre dimension discours position principe heure jour an
5 rôle relation enjeu statut lien notion composante concept aspect vue ordre interaction plan point représentation modèle façon théorie méthode manière stratégie solution lecture conception analyse approche réflexion vision vue définition démarche description interprétation perspective problématique procédure
6 représentation façon théorie manière lecture nature caractère conception approche réalité vision définition description interprétation perspective ressource faute totalité effort exercice exigence manque souci besoin but nécessité capacité code contraire unité volonté idéal possibilité pouvoir
7 étude tâche expérience méthode stratégie schéma analyse réflexion règle observation opération démarche distinction problématique procédure réponse foi tâche tableau taux tour mémoire liste loi naissance chose clé coeur contact accord réaction référence ouvrage départ donnée présentation premier preuve
8 résultat élément facteur trait constat critère profil apport argument indice équilibre qualité fonction taille marché masse produit profit part prix
9 risque effort exigence mesure besoin condition contrainte absence degré être mot est membre sujet sein seul nom classe homme acteur groupe dernier individu personne
10 réponse solution loi base connaissance contribution propriété unité attention référence orientation discussion disposition facteur trait type expression comportement composante condition critère profil angle aspect axe indice
11 raison tentative mécanisme source limite biais code conséquence courant acte angle avantage idéal moitié fin siècle cours début
12 variation minimum quantité transfert somme coût augmentation perte poids réseau environnement extension centre chemin conservation création voie développement
13 équilibre moyen qualité marché niveau produit profit demande prix production valeur relation mode forme technique statut structure schéma situation lien logique nature norme notion cadre champ choix contexte contrainte jeu système ordre organisation orientation dimension discours dispositif disposition domaine dynamique plan position pratique principe processus
14 reconnaissance échange composition constitution création accès réalisation ouverture diffusion identification présentation recherche regard étude figure terrain travail expérience comparaison conclusion connaissance contribution proposition article attention observation discussion
15 énergie répartition fréquence taille erreur masse surface long nombre proportion distribution donnée durée paramètre formation matière sécurité protection action aide application gestion direction institution intervention
16 écart extension maintien stabilité liaison complexité compréhension séparation amélioration déplacement détermination différence précision évidence mise sou sens lumière oeuvre parallèle prise
17 traitement communication utilisation application représentant équipe établissement majorité
18 respect rupture évolution modification fonctionnement transformation efficacité existence souci liberté nécessité capacité changement compétence contrôle volonté décision difficulté impact importance influence possibilité apparition synthèse déplacement diffusion production
19 échelle état événement fond force trace comparaison conflit contact proximité accord opposition

dépendance distance intérêt présence pression échange texte traitement source support ligne base communication contenu usage utilisation accès objet outil document information instrument interaction
20 recours résistance mise face faute manque succès soin charge compte considération appel réaction défaut peine preuve prise variation écart énergie répartition minimum moyen quantité fréquence transfert erreur surface somme long niveau nombre coût couche proportion augmentation demande distribution durée paramètre perte
21 ligne cause chemin jeu axe voie place rapport retour époque moment face titre exception exemple suite côté cas cité propos appel grâce défaut intérieur inverse issue
22 remarque époque étape modalité formule thèse section caractéristique catégorie conclusion hypothèse série proposition année date période particularité phase raison question formule thèse existence hypothèse auteur distinction idée
23 représentant reprise reste retour évidence mois moitié fin foi fois tête taux totalité tour transition espèce essentiel exception extérieur main maison majorité membre sortie sou sens siècle liste lumière but naissance chapitre cité clé coeur conservation cours heure jour an apparition synthèse oeuvre début départ détail dehors dernier disparition intérieur introduction inverse issue parallèle passage plupart premier respect résistance fonctionnement efficacité maintien stabilité liaison liberté compétence complexité compréhension amélioration dépendance détermination précision
24 société littérature communauté construction sécurité protection aide gestion développement intégration monde mouvement corps courant programme projet opposition origine droit
25 équipe établissement formation esprit exercice corps autorité direction droit famille emploi ensemble entreprise maison secteur service activité association ville partie
26 économie environnement culture histoire puissance action activité appartenance vie organisation identité institution intervention milieu territoire espace zone commune région pays population
27 milieu monde mouvement territoire espace zone acteur groupe reconnaissance reprise échelle évolution transformation transition sortie composition constitution construction séparation proximité appartenance réalisation ouverture détail disparition identification intégration introduction
28 famille entreprise secteur service centre classe commune couche association région ville origine pays population reste tête extérieur dehors plupart
29 recherche figure tableau terrain est exemple article auteur ouvrage main mesure soin charge compte considération contrôle acte autorité décision place
30 rapport regard être moment mot question temps terme titre travail type emploi ensemble expression mémoire sorte suite sujet sein seul signe lieu nom côté cas chose contraire homme propos objet grâce doute image individu part partie personne pouvoir

TABLE DES MATIERE

MOTS-CLÉS :

Méthode distributionnelle, Réseaux de neurones artificiels, Classification automatique, Lexique Scientifique Transdisciplinaire, Description sémantique.

RÉSUMÉ

L'hypothèse distributionnelle (Harris 1954) montre que deux mots proches sémantiquement ont tendance à apparaître dans les mêmes contextes. Reprise par la suite sous forme de méthode, nous la mettons à l'épreuve afin d'organiser sémantiquement, une liste de noms brute du Lexique Scientifique Transdisciplinaire. Ce dernier est le lexique commun à différentes disciplines scientifiques.
Pour ce faire, nous testons trois méthodes distributionnelles : deux « classiques » et une basée sur les réseaux de neurones artificiels, au travers d'un outil appelé word2vec. Cette dernière méthode retient notre attention et nous l'utilisons pour le reste de nos travaux. Dans un second temps, nous appliquons à la liste de noms, un algorithme de classification automatique afin d'obtenir des regroupements entre les noms les plus proches sémantiquement. Enfin, la dernière phase de nos travaux consiste à apporter une description sémantique à chaque groupe généré. Cela nous permet d'analyser la qualité de nos résultats et de faciliter l'utilisation de cette liste, par exemple pour les usages didactiques.

KEYWORDS :
Distributional method, Artificial neural networks, Automatic classification, Transdisciplinary Scientific Lexicon, Semantic description.

ABSTRACT

The distributional hypothesis (Harris 1954) shows that two semantically close words tend to appear in the same contexts. Subsequently recovery in the form of a method, we test it in order to organize semantically, a raw list of names of the Transdisciplinary Scientific Lexicon. This last one, is the common lexicon of different scientific disciplines.
To do this, we test three distributional methods: two « standard » and one based on artificial neural networks, by using a tool called word2vec. The last method holds our attention and we use it for the rest of our work. Secondly, we apply to the list of names, an automatic classification algorithm to obtain clusters among the closest names semantically. The last phase of our work, is to bring a semantic description for each generated cluster. This allows us to analyze the quality of our results and facilitate the use of this list, for example, for educational uses.