Exploration et visualisation de la variation terminologique en corpus spécialisés complexes :

Aurélie Picton, Patrick Drouin, Julie Humbert-Droz

Exploration et visualisation de la variation terminologique en corpus spécialisés complexes : réflexions et propositions méthodologiques

Aurélie Picton (1) (2), Patrick Drouin (2), Julie Humbert-Droz (1) (3)

(1) TIM-FTI, Université de Genève, Unimail, Bd du Pont-d'Arve 40, 1211 Genève 4
(2) Observatoire de linguistique Sens-Texte (OLST), Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, Québec, Canada, H2C 3J7
(3) CLLE-ERSS, UMR 5263, CNRS et Université Toulouse – Jean Jaurès, Allée Antonio Machado 5, 31058 Toulouse

Résumé : L'analyse simultanée de plusieurs types de variation en corpus spécialisés comparables est une question encore peu traitée, malgré des besoins et des moyens méthodologiques grandissants pour les langues de spécialité. Pour discuter et alimenter la réflexion sur cette question, nous présentons deux exemples concrets de recherche : d'une part, l'étude de la déterminologisation, c'est-à-dire le passage de termes et concepts d'un domaine de spécialité vers la langue générale ; d'autre part, la création de l'Encyclopédie humanitaire dont l'un des objectifs principaux est d'analyser les différences d'usage des termes du domaine selon différentes perspectives. Sur la base d'une approche outillée en corpus, nous interrogeons différents axes et outils méthodologiques pour accompagner ce type d'analyse. Nous détaillons trois axes : le besoin de visualisation des données, de regroupement et de modélisation des phénomènes à observer, que nous illustrons sur la base d'exemples tirés de nos données.

Mots-clés : corpus spécialisés, diachronie, diastratie, corpus comparables, linguistique de corpus, variation terminologique, variation dialectale, analyse distributionnelle, visualisation

1 Introduction

Dans cet article, nous proposons d'alimenter la réflexion sur les outils et besoins inhérents à l'analyse simultanée de plusieurs types de variations en corpus spécialisés comparables. Depuis une vingtaine d'années, la prise en compte de la variation en langues de spécialité s'est largement développée sous l'impulsion des nouvelles théories de la terminologie, ainsi que du rapprochement de la discipline avec la linguistique et les approches en corpus. Si de plus en plus de travaux abordent la question de

la variation sous différents angles (méthodologiques, théoriques, descriptifs), encore peu de travaux traitent de plusieurs types de variations simultanément. Cependant, l'expression de nouveaux besoins pousse les langues de spécialité (LSP) à prendre en compte cette problématique. Nous proposons d'appuyer nos propos sur deux exemples concrets de recherche : d'une part, l'étude de la déterminologisation, c'est-à-dire le passage de termes et concepts d'un domaine de spécialité vers la langue générale, qui implique la prise en compte de la variation diachronique et diastratique ; d'autre part, la création de l'Encyclopédie humanitaire, dont l'un des objectifs principaux est d'analyser les différences d'usage des termes du domaine humanitaire selon différentes perspectives et dimensions relevant de la diatopie et de la diastratie.

Dans ce cadre, après avoir présenté le contexte général de la variation, et plus particulièrement dialectale, en terminologie (section 2), nous discutons les particularités de ces deux exemples de recherche et insistons sur les défis méthodologiques qu'ils créent (section 3). Notre démarche, basée sur une approche outillée en corpus spécialisés comparables (section 4), interroge différents axes et outils méthodologiques pour accompagner ce type d'analyses assez complexes. Nous détaillons en particulier trois axes : le besoin de visualisation des données, de regroupement et de modélisation des phénomènes à observer (section 5), que nous illustrons sur la base d'exemples tirés de nos données.

2 Contexte théorique

2.1 Variation en terminologie : constats

La variation terminologique est définie par exemple par Freixa et Fernández-Silva (2017 : 160) comme le phénomène comprenant « *all variation that affects the form or meaning of terms ; the former [...] denominative variation and the latter [...], conceptual variation (since terminology has traditionally emphasized concepts over meanings)* ». Suite au changement de paradigme connu par la terminologie depuis les années 1990 et son rapprochement avec la linguistique de corpus, la question de la variation est devenue l'une des pierres angulaires des recherches dans ce domaine. Depuis une trentaine d'années, en réponse à la perspective wüsterienne (ou Théorie générale de la terminologie (TGT), Wüster, 1981), qui préconisait en effet une vision prescriptive, et face aux réalités des langues de spécialité en discours, plusieurs propositions théoriques sont apparues (par exemple par Bourigault & Slodzian, 1999 ; Cabré, 1998 ; Desmet, 2007 ; Gaudin, 2003 ; Temmerman, 2000, etc.) ; toutes plaident en faveur d'un rapprochement avec la linguistique et de la prise en compte de la polysémie et de la variation des termes, en corpus. Du point de vue méthodologique également, plusieurs recherches ont vu le jour, qui proposent la mise en place d'outils de repérage automatisé de la variation (Daille, 2017 ; Jacquemin, 2001 ; Kerremans, 2017), ou des approches d'analyse en corpus de cette variation (par ex. León-Araúz & Reimerink, 2015 ; Picton, 2014). En résultent des descriptions assez nombreuses et documentées sur les types et les causes de variation (par ex. Dury, 2018 ; Freixa, 2002), l'analyse des phénomènes sémantiques en jeu (par ex. Fernández-Silva, 2016 ; León-Araúz, 2015) ou encore des typologies de la variation en langue de spécialité. Par exemple, Freixa (2006) en distingue cinq grands types : 1. la variation dialectale,

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

liée aux différences d'origines des auteurs ; 2. la variation fonctionnelle, liée aux différences de situation de communication ; 3. la variation discursive, liée aux différences de styles et de besoins d'expressivité des auteurs ; 4. la variation interlinguistique, liée aux contacts entre langues ; 5. la variation cognitive, liée aux différences de conceptualisation et de motivation. Dans cet article, nous nous intéressons à la variation dialectale.

2.2 Variation dialectale : définitions et besoins

La variation dialectale, largement travaillée en langue générale (par exemple par Coseriu, 1998 ; Gadet, 2003 ; Saussure (De), 1995) renvoie de manière générale aux changements de dénominations et/ou de concepts en fonction de différences de régions (zones géographiques) (diatopie), de l'évolution dans le temps (diachronie) ou de différences sociales, telles que l'appartenance à différentes sphères professionnelles des auteurs (diastratie) [1]. En langue de spécialité, ces types de variation sont régulièrement mentionnés (par exemple par la socioterminologie, la théorie variationniste, la terminologie textuelle, etc.) ; la diachronie fait en particulier l'objet de réflexions assez riches ces dernières années (voir Dury, 2018 ; Picton, 2018), tant en diachronie courte (moins de 20 ou 30 ans) qu'en diachronie longue (de plusieurs dizaines d'années à plusieurs siècles). Par contre, peu de travaux ont développé la réflexion sur l'axe de la diatopie (Drouin, 2017) ou de la diastratie (Auger, 2001 ; Bertaccini & Matteucci, 2005 ; Kacprzak & Goudaillier, 2014 ; Picton & Dury, 2017). Parallèlement, à l'instar des réflexions de Moreau (1997 : 284), notons que « [d']autres variables peuvent se révéler pertinentes pour rendre compte de la diversité à l'intérieur d'une langue : ainsi l'âge, le sexe, l'ethnie, la religion, la profession, le groupe et, de manière plus générale, toute variable sur laquelle les individus fondent leur identité (orientation sexuelle, appartenance à une congrégation religieuse, etc.) », qui toutes participent de la variation dialectale et ne sont pas (encore) bien décrites. De nombreuses pistes restent donc à explorer, la première relevant du besoin de description des phénomènes en jeu dans ces types de variation. Une deuxième piste, connexe à celle-ci, repose sur le fait que ces types de variation semblent étroitement liés, sans que ces liens soient connus ou décrits. Par exemple, un changement de strates de locuteurs (lorsqu'un terme d'un domaine intègre la langue générale par exemple) s'opère nécessairement dans le temps. Une hypothèse est donc bien que certains types de variation peuvent être sous-jacents à d'autres.

Ces quelques remarques reposent également sur un constat clair pour les langues de spécialité : de nombreux domaines, et très certainement la majorité, évoluent rapidement, sont partagés par différents groupes d'experts (ou de locuteurs), sont présents dans différentes régions du globe, et ce, simultanément. Ils sont donc sujets à tous ces types de variation, considérés ensemble. Pour ces différentes raisons, nous proposons dans cet article d'entamer la réflexion sur ces besoins et les questionnements méthodologiques qu'ils entraînent.

1. Coseriu (1966) introduit également la variation diaphasique, liée au style et registres de langue, et Gadet (2003) la variation diamésique, liée à la distinction oral/écrit. Nous ne prenons cependant pas ces variations en compte dans ce travail, ces questions n'étant traitées que de manière très secondaire à ce jour en LSP et n'étant notamment pas prises en compte par la typologie de Freixa que nous citons ici.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

3 Exemples de demandes pour l'analyse de différents types de variation

Notre réflexion repose ici sur deux exemples de demandes en cours : le développement des recherches sur la déterminologisation (thèse de Humbert-Droz, en cours) et la constitution de l'Encyclopédie humanitaire [2].

3.1 Analyse du phénomène de déterminologisation

La déterminologisation peut être définie comme le passage de termes et concepts d'un domaine de spécialité vers la langue générale (par ex. Meyer & Mackintosh, 2000 ; Ungureanu, 2003 ; Condamines & Picton, 2014 ; Galisson, 1978 ; Renouf, 2017). Ce passage de termes entre différentes « strates de locuteurs » (les experts et les « profanes ») s'effectue le plus souvent de manière progressive et continue, dans le temps et dans différents degrés de spécialisation, soit entre diachronie et diastratie. L'étude de la déterminologisation implique alors à la fois d'observer ce flux entre spécialisé et général, ainsi que la dimension diachronique. Ce projet repose sur une approche outillée en corpus (section 4.1). Il s'agit d'un projet de recherche fondamentale, qui vise la réflexion théorique et méthodologique sur ce phénomène, en français. Le domaine choisi pour cette étude est la physique des particules, et il s'agit d'un travail mené essentiellement par des linguistes, associés à deux experts du domaine. Ceci permet de contrôler de manière optimale les données linguistiques observées (corpus) et les variables souhaitées (ici des degrés de spécialité et des périodes différents).

3.2 Constitution de l'Encyclopédie humanitaire

Ce contexte de travail est une demande appliquée de la part des professionnels du monde humanitaire et initié par le Cerah [3]. L'un des objectifs principaux est d'analyser les différences d'usage des termes du domaine selon différentes perspectives dans ce domaine, en anglais. En effet, bien que les experts semblent s'accorder sur des principes et valeurs communes, beaucoup de termes et concepts montrent des divergences et nuances importantes selon l'origine géographique des acteurs de l'humanitaire, leur organisation, mais aussi leur discipline d'origine (par ex. Collinson & Elhawary, 2012). Ce projet est essentiellement un projet de recherche des acteurs de l'humanitaire, très impliqués, pour répondre à un besoin fort dans leur discipline. Les linguistes sont donc plutôt « consultés » et ce sont les experts qui contrôlent les variables souhaitées (en accord avec les besoins et pratiques linguistiques). Ce contexte permet donc une analyse plus proche des besoins et contextes réels des professionnels, mais entraîne des modalités différentes de contrôle des variables à explorer.

3.3 Défis communs

Bien que partant de fondements différents, ces deux contextes permettent d'entamer la réflexion sur un certain nombre de défis communs. Premièrement, du point de vue descriptif, ils permettent la prise en compte simultanée de plusieurs types de variation, qui répondent à la réalité de la langue de spécialité étudiée, ainsi que de leurs

2. https://humanitarianencyclopedia.org/, consulté le 01.06.2020.
3. https://www.cerahgeneve.ch/home/, consulté le 01.06.2020.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

liens éventuels. Deuxièmement, d'un point de vue théorique, ces contextes amènent à enrichir l'étude de différents types de variation en langue de spécialité, voire à réfléchir à la proposition de considérer les langues de spécialité comme de véritables *diasystèmes* (Coseriu, 1998), notion empruntée à la linguistique variationniste et renvoyant à une conception dynamique des langues comme des systèmes dont l'architecture se compose de différentes variétés linguistiques liées aux dimensions temporelles, géographiques et sociales (voir par exemple Dostie & Hadermann, 2016 ; Verjans, 2013). Enfin, et c'est ce que nous discutons dans la suite de cet article, ces contextes permettent de travailler l'approche de linguistique outillée de la construction de corpus « multi-comparables » (puisque prenant en compte plus d'une dimension) et la manière de manipuler et observer des données nécessairement nombreuses et hétérogènes.

4 Approche méthodologique

4.1 Linguistique outillée : aperçu général

L'approche globale dans laquelle s'inscrivent ces analyses est celle de la Terminologie textuelle (Bourigault & Slodzian, 1999) et de la « linguistique outillée ». Elle repose sur ce que l'on pourrait appeler un « trépied méthodologique » (Picton & Dury, 2015), à savoir : a. des textes (corpus à comparer), b. des outils et des indices (dont les termes sont les points d'entrée pour l'analyse) et c. des collaborations étroites avec les experts (afin de coconstruire une interprétation des données). Cette collaboration est établie tout au long du processus, de l'identification d'un besoin à la construction d'une interprétation finale, en passant par la compilation du corpus et l'analyse de données (Condamines & Picton, 2015). L'ensemble de ce processus est guidé par un besoin (par exemple Aussenac-Gilles *et al.*, 2002), le plus souvent appliqué, à l'origine de l'analyse (par exemple ici, la construction de l'Encyclopédie humanitaire).

Tout au long de ces étapes, plusieurs types d'outils sont mis en œuvre, tels que des extracteurs de termes (notamment TermoStat (Drouin, 2003)), des analyseurs syntaxiques (ici Talismane (Urieli, 2013)), des programmes *ad hoc*, des concordanciers (AntConc (Anthony, 2018)), des extracteurs de variantes (ici TermSuite (Daille, 2017)), ou encore des outils intégrant plusieurs de ces fonctionnalités (par ex. Sketchengine (Kilgarriff *et al.*, 2014)). Ces outils permettent en effet d'observer différents types d'indices dans les corpus, appartenant à quatre familles :

1. indices de types quantitatifs : calculs et comparaisons de fréquences, analyses statistiques variées sur l'absence, la présence, la répartition d'une unité ou d'un phénomène particulier dans les corpus ;
2. observation des variations de forme (variantes lexicales, graphiques, orthographiques ou syntaxiques, avec l'hypothèse qu'une variation de forme peut révéler une variation conceptuelle (voir Cabré, 1998 : 241 ou Tartier, 2004 sur cette question)) ;
3. observation de la distribution des unités et de sa variation pour analyser des comportements sémantiques spécifiques (ressemblances et différences de sens entre unités). Cet indice peut être mis en œuvre de deux manières :

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

a. sans interprétation *a priori*, c'est-à-dire un repérage de similarités ou divergences sémantiques sur la base de l'observation de la distribution des unités (syntaxique ou cooccurrence de n unités à gauche/droite ; voir par exemple Habert, 2005) ;

b. avec interprétation *a priori*, c'est-à-dire l'observation de cooccurrents définis en amont de l'analyse et spécifiquement ciblés pour cette analyse afin de servir de « marqueurs » (Meyer, 2001) de présence d'une information pertinente.

4.2 Constitution de corpus « multi-comparables »

L'approche de linguistique outillée repose en premier lieu sur la constitution de corpus, et en particulier dans cet article de corpus « multi-comparables ». Dans le cas de la déterminologisation cette « multi-comparabilité » se traduit par un double découpage des données, correspondant aux dimensions diachronique et diastratique, caractéristiques du phénomène (section 3.1). Nous proposons ainsi un corpus composé de cinq sous-corpus diastratiques chacun divisé en trois sous-corpus diachroniques, représentant différentes phases du processus d'intégration des termes dans la langue générale. Des textes relevant de différents genres et degrés de spécialisation constituent alors ces sous-corpus. Ils varient entre des textes très spécialisés (articles de recherche, thèses de doctorat, dans le sous-corpus *spécialisé*) et des textes non spécialisés (articles de presse généraliste, dans le sous-corpus *presse*), en passant par des textes d'un degré de spécialisation intermédiaire (ou d'experts à semi-/non-experts selon Bowker & Pearson, 2002), avec des communiqués de presse (sous-corpus *communiqués*), des rapports d'activités « grand public » provenant de divers laboratoires (sous-corpus *rapports*) ainsi que des sites et des articles de revues de vulgarisation (sous-corpus *vulgarisation*). Le découpage en sous-corpus diachroniques repose, quant à lui, essentiellement sur la définition d'événements-clés, susceptibles d'avoir été largement traités dans les médias et d'avoir ainsi contribué au processus d'intégration des termes de physique des particules dans la langue générale. Suite à des entretiens avec les experts, deux évènements importants pour la discipline et pertinents par rapport au sous-domaine sélectionné (le Modèle Standard de la physique des particules) ont été retenus : la mise en marche du Grand collisionneur de hadrons (LHC) au Cern (Organisation européenne pour la recherche nucléaire) en 2008 ainsi que l'annonce de la découverte du boson de Higgs en 2012. Ces deux évènements permettent de créer trois périodes homogènes, soit de 2003 à 2007, de 2008 à 2011 et de 2012 à 2016. Enfin, il convient de souligner que les textes ont été sélectionnés en collaboration avec les experts, et ce principalement pour s'assurer de leur cohérence vu le sous-domaine. Le corpus ainsi constitué compte un peu plus de 4 millions d'occurrences, dont la répartition dans les différents sous-corpus est illustrée dans la table ci-dessous (Humbert-Droz *et al.*, 2019).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Tableau 1 - Taille et composition du corpus de physique des particules

Sous-corpus	2003-2007	2008-2011	2012-2016	Total
Spécialisé	314 658	330 975	349 242	**994 875**
Communiqués	70 950	69 478	69 892	**210 320**
Rapports	516 820	302 552	322 501	**1 141 873**
Vulgarisation	216 969	194 675	208 401	**620 045**
Presse	367 378	365 650	365 680	**1 098 708**
Total	1 486 775	1 263 330	1 315 716	**4 065 821**

Dans le domaine de l'humanitaire, la constitution du corpus repose sur les principes suivants. Différents types de variation jugés pertinents par le consortium d'experts impliqués dans le projet doivent être intégrés, à savoir en priorité : le type d'organisations en jeu, l'origine géographique et le type de phénomènes/catastrophes/désastres gérés par les acteurs de l'humanitaire. Ce corpus se compose de rapports d'activité annuels, de documents généralistes d'organisations, de documents stratégiques, d'articles, etc. Au moment de la rédaction de cet article, il se compose d'un peu plus de 4 millions d'occurrences, 2,5 millions dans des documents dits « généraux » et 1,5 million dans des textes dits « stratégiques ». Les régions prises en compte à ce stade sont l'Afrique, l'Asie, l'Amérique centrale/Caraïbes, l'Europe, le Moyen-Orient/Afrique du Nord, l'Amérique du Nord et l'Océanie. Du point de vue des types d'organisations, sont considérés ici les corporations, les fédérations de NGO, les fondations, les organisations intergouvernementales, les réseaux, les organisations non gouvernementales, les instituts de recherche/éducation, les entités étatiques et les entités religieuses. Les sous-corpus comparés varient de 100 000 à 600 000 occurrences [4].

4.3 Principaux défis et enjeux méthodologiques

La constitution de ce type de corpus représente en elle-même un enjeu majeur de l'analyse dans ce type de contextes. Dans le cas de la physique des particules, le processus de constitution est mené par les linguistes, après discussion avec les experts sur les sources disponibles et pertinentes. La difficulté principale consiste alors à choisir finement ces sources et de les équilibrer en amont de l'analyse, à la fois du point de vue diachronique et du point de vue diastratique. Cela signifie ici un investissement en temps et énergie conséquent pour parvenir au meilleur équilibre possible, qui concerne tous les sous-corpus en diastratie pour toutes les années prises en compte, tout en tenant compte de la nature parfois très différente des genres de textes sélectionnés. Cet équilibre permet alors d'assurer une comparabilité forte pour les différents types de phénomènes observables.

Dans le cas de l'humanitaire, nous l'avons dit, la demande émane de professionnels du domaine, très nombreux (plus d'une vingtaine dans le consortium de discussion) et impliqués. La demande première consiste avant tout à composer avec les réalités d'un terrain complexe, tel que perçu par ces différents acteurs. La sélection de ces textes, bien que toujours discutée avec les linguistes, repose sur une répartition des rôles un peu différente, puisque ce sont plutôt, pour ainsi dire, les experts qui consultent les linguistes. Le nombre de dimensions souhaité étant particulièrement ambitieux

4. Au moment de la publication de cet article, le corpus est finalisé et compte 84 millions d'occurrences. Il peut être organisé en différents sous-corpus comparables variant de 1,5 à 26 millions d'occurrences.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

(*supra*), il est difficile de trouver un équilibre adéquat pour toutes ces dimensions en amont de l'analyse. Afin d'assurer une validité dans la comparaison des données, il convient alors de réfléchir à s'appuyer sur des statistiques pour pondérer les résultats des analyses dans les textes et projeter des schémas de description valides *a posteriori*.

Une fois les corpus constitués, le second défi est celui de la manipulation de ces données hétérogènes et nombreuses. Dans le cadre de l'approche outillée choisie ici, il s'agit avant tout d'identifier un point d'entrée pertinent dans les données pour commencer l'analyse. En effet, bien que le terme, comme chaîne de caractères, soit le point d'entrée des outils habituels dans ce type d'approche, cet élément ne suffit plus : il faut pouvoir décider quel(s) terme(s) ou groupe(s) de termes, dans quels sous-corpus, quels sous-corpus comparer en priorité pour optimiser l'analyse, etc. De plus, il n'est pas envisageable, cognitivement et ergonomiquement, de travailler à la lecture seule de tableaux ou concordances « nues » pour l'analyse. Pour cela, dans la suite de cet article, nous nous concentrons sur ce défi et, plus particulièrement, sur trois contraintes saillantes :

— premièrement, en diachronie, la nécessité de gérer la question du découpage temporel des données à comparer ;
— deuxièmement, le besoin de saisir les liens potentiels entre les différents types de variation observés ou, encore, le besoin d'observer simultanément ces différents types de variation ;
— enfin, le besoin d'interpréter des données aussi hétérogènes et complexes pour leur donner du sens.

5 Exploration de pistes

Afin de répondre au défi de la manipulation de données nombreuses, volumineuses et hétérogènes, issues de corpus, nous proposons d'explorer ici trois types de pistes qui permettent respectivement de visualiser les données, de les regrouper et de les modéliser. Plus particulièrement, nous focalisons notre attention sur les Motion Charts, le Clustering (et plus particulièrement l'approche VNC) et l'approche par vecteurs.

5.1 Visualisation des données

Les techniques de visualisation permettent de fournir à l'analyste (linguiste, expert) une représentation organisée et visuelle de différents types de données. L'objectif est donc avant tout de soutenir les analyses en limitant l'impact de la lourdeur des données à manipuler et en facilitant la mise en évidence de phénomènes pertinents difficilement identifiables « à l'œil nu ». Dans les contextes présentés, l'un des objectifs premiers de la visualisation est, d'une part, de pouvoir manier des ensembles de plusieurs termes, et d'autre part, de l'opérer sur plusieurs axes de variation simultanément. Pour ce faire, une piste est celle des Motion Charts (par exemple Gesmann & De Castillo, 2011), qui sont des graphiques dynamiques permettant d'observer plusieurs dimensions, dont l'une est la dimension temporelle. Ces graphiques sont notamment utilisés par certains chercheurs en diachronie en langue générale (par ex. Hilpert, 2011, dont nous nous inspirons ici). Ces graphiques peuvent être générés sous l'environnement R, à l'aide du package *GoogleVis* par exemple, open source. Afin d'illustrer ce poten-

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

tiel, nous proposons un cas d'analyse dans le contexte de la déterminologisation. La Figure 1, ci-dessous, présente un exemple de graphique dynamique : sur l'axe des abscisses sont représentées les fréquences relatives dans le sous-corpus de presse constitué pour l'étude (*Score_presse*). Sur l'axe des ordonnées sont représentées les fréquences relatives dans le sous-corpus spécialisé en physique des particules (*Score_spé*). Le tableau de droite indique la liste des termes observables dans ce corpus, ici l'ensemble des types de particules (soit plus de 90 termes). La case « Statut » permet de regrouper ces termes en fonction de caractéristiques définies en amont. Dans notre cas, nous faisons dans cet exemple l'hypothèse que les particules dites hypothétiques (c'est-à-dire dont les physiciens postulent l'existence, mais qui n'ont pas encore pu la prouver, en bleu ou gris foncé) ont potentiellement un comportement de fréquences différent de celui des particules prouvées expérimentalement (en vert ou gris clair). En bas de la figure se trouve un bouton « play » qui permet d'activer une frise chronologique de 2003 à 2016. En cliquant sur le bouton, le graphique dynamique s'anime et permet d'observer le film de l'évolution des fréquences de ces deux groupes de termes dans les deux sous-corpus. À l'aide de l'affichage (optionnel) des dénominations sur le graphique, l'analyste peut ainsi plus facilement voir des tendances de fonctionnements ou des fonctionnements isolés des termes à l'étude.

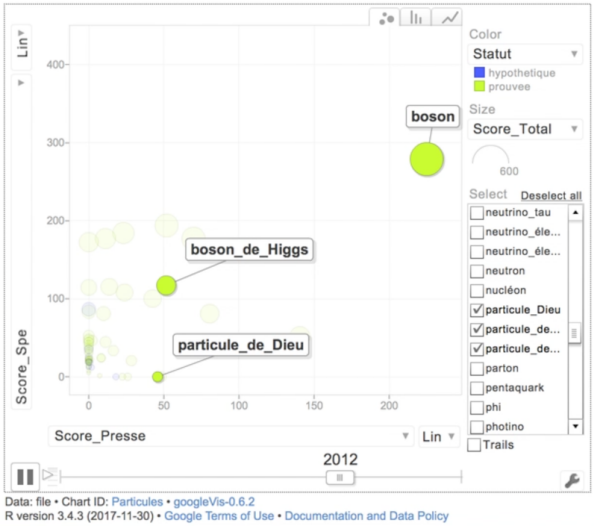

FIGURE 1: Exemple de Motion Chart (graphique dynamique)

La Figure 2 illustre un déroulé de ces graphiques dynamiques, qui montre la constance des particules hypothétiques : en effet, celles-ci restent majoritairement dans le corpus spécialisé et ne « passent » que de manière très anecdotique dans le corpus de presse. Rapporté à notre objectif d'analyse de la déterminologisation, ce type de visualisation permet de mettre en évidence des éléments en faveur d'une hypothèse selon laquelle certains types de concepts « passent » de manière privilégiée dans la langue générale par rapport à d'autres. Dans notre cas, nous avons ainsi une nouvelle piste à explo-

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

rer : celle de savoir si les termes renvoyant aux particules hypothétiques sont moins susceptibles de se déterminologiser que les termes renvoyant à des particules prouvées.

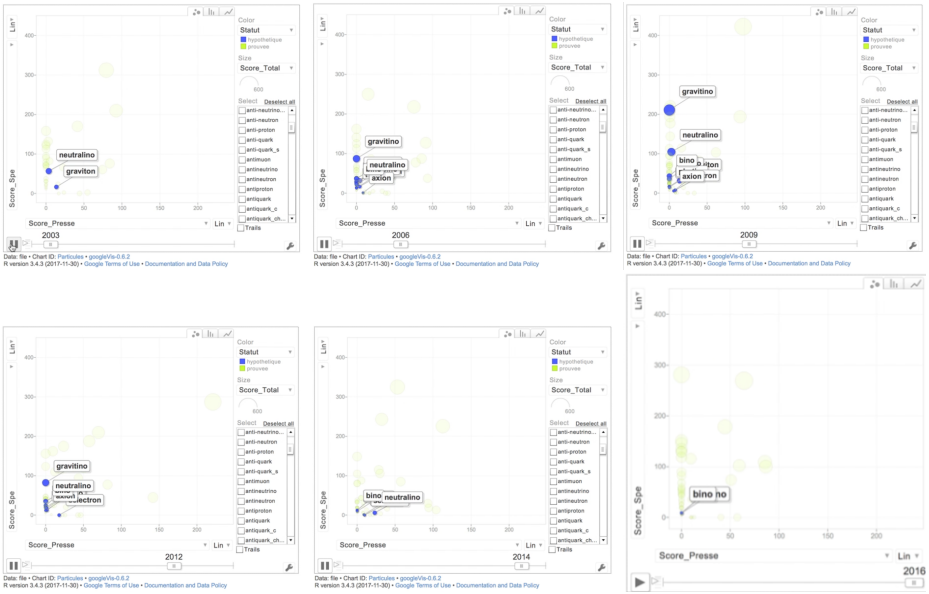

FIGURE 2: Motion Chart – Exemple de la dynamique des particules hypothétiques (points étiquetés) – 2003-2016

Afin d'explorer un peu plus avant cette hypothèse, la Figure 3 montre le déroulé de la dynamique du concept de BOSON DE HIGGS, particule hypothétique jusqu'en 2012, dont la preuve expérimentale a été apportée à cette date grâce à diverses expérimentations du Cern dans le LHC. Ce graphique dynamique révèle que, en 2003, l'unité *boson* (qui renvoie à une famille de particules dont certaines sont prouvées) est déjà utilisée dans la presse, mais que la dénomination *boson de Higgs*, l'est presque exclusivement dans le corpus spécialisé. En 2007, certaines dénominations renvoyant à ce concept passent dans la presse (*boson de Higgs* et *particule de Dieu* notamment), à la veille du lancement du LHC (en 2008), construit notamment pour tenter d'en démontrer l'existence. Le pic s'accentue en 2008-2009 pour *boson,* mais les autres dénominations de la particule hypothétique BOSON DE HIGGS restent stables (fréquentes dans les corpus spécialisés, mais quasi absentes dans la presse). Dès 2010, le passage semble se marquer un peu plus, et notamment en 2012 où un véritable pic de croissance pour *boson* et les autres variantes est marqué dans la presse (mais très peu dans le corpus spécialisé), date de la découverte du boson de Higgs, voir section 3.1. Dès 2013, ces pics redescendent dans la presse et les fréquences restent stables dans les corpus spécialisés. La situation de 2016 rejoint alors l'équilibre de 2003, avec une fréquence un petit peu plus grande de ces unités dans la presse.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

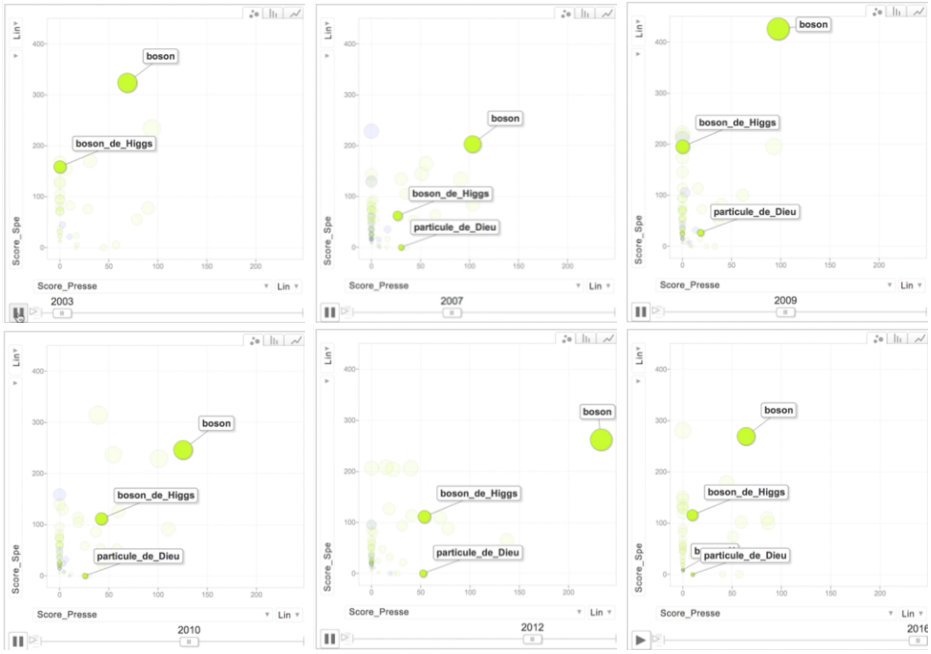

FIGURE 3: Motion Chart – Dynamique des particules en physique (cas de boson de Higgs et ses variantes, étiquetés)

À travers ces quelques exemples, le potentiel de ce type de graphique nous apparaît considérable : non seulement ces graphiques sont relativement simples à générer, mais ils permettent de manipuler une grande quantité d'informations en un « coup d'œil », de manière plutôt ergonomique et ludique. Il s'agit d'un outil efficace pour observer le lien entre plusieurs axes de variations et entre plusieurs sous-corpus simultanément. Si nous l'avons appliqué ici à une seule question, l'on perçoit également que ce type d'approche peut permettre d'analyser des groupes de termes (et potentiellement toutes les variantes pour un même concept), des termes isolés, mais également des catégories grammaticales, des types de constructions, etc. en prenant en compte différents types de variation. Dans notre cas, ces graphiques permettent de mettre en évidence la dynamique particulière des particules hypothétiques entre corpus spécialisé et presse, ainsi que la dynamique des termes dans des situations de « buzz » médiatique. Il s'agit donc d'un point d'entrée, certes limité et à compléter, mais extrêmement riche pour l'exploration d'hypothèses préliminaires sur des données complexes.

5.2 Analyse distributionnelle pour regrouper et modéliser les données

Depuis quelques années, on assiste à un retour en force de l'analyse distributionnelle dans le domaine du traitement automatique de la langue. Cette technique peut s'appliquer au dépistage de la variation selon au moins deux axes, que nous décrivons ici.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

5.2.1 Variability-based neighbor clustering, VNC

Le premier axe considéré ici consiste à étudier la variabilité du point de vue des documents afin de faire émerger des regroupements de ces derniers. Dans notre cas, nous proposons d'utiliser ce mode de regroupement pour l'analyse en diachronie et, plus spécifiquement, le choix du découpage temporel des sous-corpus. Ce choix, traditionnellement, repose sur deux stratégies dominantes : les sous-corpus sont découpés et organisés soit en fonction d'un événement extralinguistique (dans notre exemple de la physique des particules, il pourrait s'agir de la découverte du boson de Higgs, voir section 3.1), soit sur la base d'un découpage régulier (par exemple tous les cinq ans). Ces stratégies présentent cependant plusieurs limites : premièrement, elles impliquent une certaine arbitrarité dans les choix de découpage, qui peut générer du silence dans les résultats obtenus, si une évolution ne correspond pas au découpage choisi. En particulier, dans le premier cas, si l'un des objectifs de l'analyse est de chercher à saisir les liens entre dimension temporelle et passage du spécialisé au général, le passage de certains termes pourrait ne pas concorder avec le moment prédéfini arbitrairement en diachronie. Dans le second cas, un découpage en sous-corpus trop nombreux entraîne des limites techniques, ergonomiques et cognitives pour la gestion et l'observation de ces données.

Pour pallier ces limites, la classification ascendante hiérarchique par contiguïtés (variability-based neighbor clustering, VNC) de Gries & Hilpert (2008 ; 2012) permet, suite à une analyse du voisinage distributionnel des unités, d'identifier des regroupements de documents possédant des propriétés communes (des périodes chez les chercheurs cités). Les regroupements ainsi créés[5] peuvent ensuite faire l'objet d'analyses plus fines visant à décrire la variabilité observée. Cette proposition méthodologique est donc explorée par la linguistique diachronique qui travaille sur de très gros corpus et sous-corpus (plusieurs dizaines de millions d'occurrences), et sur des intervalles temporels généralement longs (plus de cent ans). Néanmoins, elle est applicable à des intervalles plus courts (par exemple Gries & Hilpert, 2010) et, dans une certaine mesure statistique, à des corpus moins volumineux pour dégager des périodes pendant lesquelles les distributions indiquent une certaine homogénéité de fonctionnement. Comme le décrivent les auteurs, ce calcul

> « *offers the best of both worlds : on the one hand, it provides* ***an objective, data-driven classificatory approach*** *for temporally-ordered data that avoids the above problems of inspecting the data visually and potentially losing important generalisations. On the other hand, it does not suffer from the problem of regular clustering algorithms that fail* ***to account for temporal ordering*** » (Gries & Hilpert, 2008 : 64).

Appliquée à nos données, toujours en physique des particules, cette approche permet d'obtenir des résultats extrêmement révélateurs. La Figure 4 montre le découpage obtenu pour le sous-corpus spécialisé, sur les fréquences relatives des différents types de particules.

5. Et qui reposent ici sur le coefficient de corrélation de Pearson (« Pearson product-moment correlation coefficient »).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

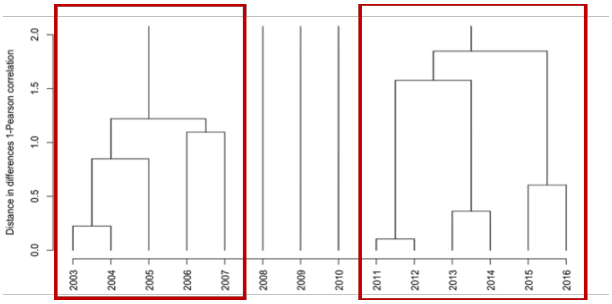

FIGURE 4: VNC - Regroupements en diachronie (corpus spécialisé)

Dans cette Figure 4, l'on voit clairement deux périodes se distinguer (dans les encadrés) : une période avant 2008 et la période 2011-2016. Ce découpage met en avant le fait que l'année 2008 a été déterminante pour l'évolution de la discipline et, *a fortiori*, de ces termes, dans les corpus spécialisés. Nous l'avons dit, cette année est celle du lancement du LHC au Cern. En revanche, l'année de la découverte du boson de Higgs, 2012, ne serait pas un événement pertinent pour un découpage dans ce cas (et bien que cela ait été notre hypothèse de départ pour cette analyse, section 3.1).

La Figure 5 montre la même approche, mais appliquée au sous-corpus de presse ; le découpage obtenu est tout à fait différent : ainsi, dans la presse, l'évolution des termes renvoyant aux différentes particules ne correspond pas à un découpage de type avant/après le lancement du LHC, mais requiert plutôt une observation des textes de 2007 à 2011 (soit « autour » du lancement du LHC) et à partir de 2011-2012 (soit la découverte du boson de Higgs).

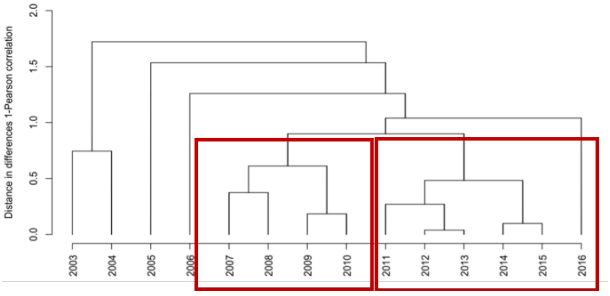

FIGURE 5: VNC - Regroupements en diachronie (corpus de presse)

Les deux exemples proposés mettent ainsi bien en évidence les dynamiques propres à chaque sous-corpus à observer pour un même groupe de termes, ou un même groupe de phénomènes. Ces figures sont également faciles à générer, bien que les regroupements par ascendance hiérarchique ne soient pas toujours faciles à lire ou interpréter. Cette approche par regroupement nous semble donc tout à fait pertinente pour les langues de spécialité, et permet une objectivisation claire des découpages en sous-corpus diachroniques. Si l'application sur de courts intervalles temporels semble entièrement pertinente, celle-ci doit cependant encore être testée pour être validée statistique-

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

ment sur des corpus de taille réduite (par rapport à ceux utilisés en linguistique diachronique).

5.2.2 Word embeddings (ou plongements lexicaux)

Une autre approche consiste à observer la variabilité avec une granularité plus petite et à s'intéresser aux phénomènes qui touchent directement les unités lexicales, les regroupements sont donc lexicaux dans un tel cas. Les travaux récents dans cette optique se fondent essentiellement sur l'exploitation des plongements de mots ou plongements lexicaux (ou *word embeddings*). Ces derniers ont pour but de capturer le sens en contexte des unités lexicales en les représentant sous forme de vecteurs numériques (Ferré, 2017; Mikolov, *et al.*, 2013). Les vecteurs sont *grosso modo* issus d'une analyse distributionnelle et peuvent être utilisés pour comparer le sens des unités sur la base des contextes qu'ils partagent.

Les vecteurs qui composent les plongements lexicaux peuvent être comparés entre eux afin de vérifier les unités qui sont les plus proches dans l'espace vectoriel. L'hypothèse est relativement simple et veut que ces proximités entre vecteurs soient représentatives d'une proximité sémantique. De plus, les plongements lexicaux permettent de consulter les unités qui contribuent à cette proximité (les voisins sémantiques).

Dans le contexte d'une étude sur un corpus pouvant être divisé en sous-corpus selon divers angles (diastratie, diachronie, diatopie, etc.), les outils nous permettent de construire un ensemble de plongements lexicaux pour chacun des sous-corpus du corpus. Il devient alors envisageable de comparer les vecteurs lexicaux pour une forme dans ces divers sous-corpus et d'illustrer la proximité dans un espace à deux dimensions. Nous avons généré des plongements lexicaux sur nos corpus afin de voir dans quelle mesure cette technologie nous donne accès à de l'information difficile à saisir à l'aide des approches plus traditionnelles. Les plongements ont été mis en place à l'aide de l'outil word2vec (Mikolov, *et al.*, 2013) avec un seuil de fréquence minimale de 2 et des vecteurs à 300 dimensions. En ce qui concerne les autres hyper-paramètres de l'algorithme, les valeurs par défaut ont été utilisées. La figure suivante est issue d'un tel procédé où les sous-corpus sont construits en fonction de critères diatopiques.

Comme on peut le constater dans la figure précédente, les sous-corpus sont relativement d'accord sur le sens contextuel du nom *protection* (en bas à droite). Dans la même région, on voit cependant apparaître la forme nominale *security* dont le sens diverge considérablement de celui observé dans les autres sous-corpus puisque ceux-ci l'associent plutôt à celui de *prevention*. L'avantage indéniable de la visualisation des plongements lexicaux est qu'elle permet d'observer des tendances qui ne seraient pas humainement observables dans un corpus textuel de grande taille.

On peut aussi voir dans l'exemple précédent que *prevention* en Asie et en Europe est plus souvent lié à *safety*, alors que dans le corpus d'Amérique du Nord et de l'Afrique, l'association est plutôt avec *respect*. Le retour au corpus, aux contextes et aux experts dans un tel contexte est inévitable afin de vérifier qu'il ne s'agisse pas d'un biais du corpus et bien d'une tendance généralisée liée à la variation. Ce type d'aller-retour n'est cependant pas propre à l'exploitation des plongements lexicaux et est, par ailleurs, valable pour toutes les études qui se fondent sur les corpus.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

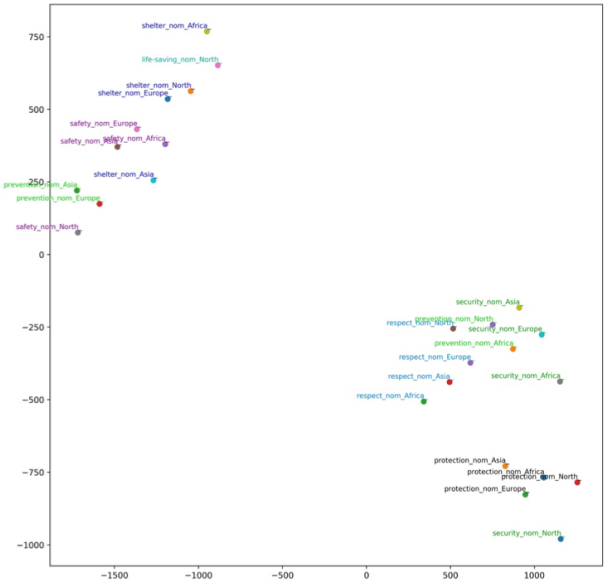

FIGURE 6: Visualisation t-sne (van der Maaten et Hinton, 2008) des vecteurs liés au concept *protection*

Bien que le recours à l'apprentissage profond et aux possibilités offertes par les plongements de mots soit intéressant, de telles techniques demeurent encore plutôt complexes à manipuler. Leur inclusion éventuelle dans les logiciels de textométrie comme TXM (Heiden 2010) ou iramuteq (Ratinaud 2009) facilitera leur utilisation. De plus, les plongements lexicaux ont des performances assez limitées lorsqu'on les utilise pour explorer des corpus de plus petite taille (moins de 1 million de mots). Pour des résultats optimaux, il faut donc de gros corpus et des machines relativement puissantes.

Pour le moment, les plongements de mots sont construits sur des sous-corpus constitués manuellement et *a priori*. Cependant, pour une étude sur la diachronie où la granularité peut varier d'une forme à une autre, les sous-corpus pourraient être construits dynamiquement à l'aide d'algorithmes de classification à la VNC (section 5.2.1). La convergence des technologies (logicielles et matérielle) et des algorithmes nous semble prometteuse et pouvoir conduire à une prise en charge outillée et semi-automatique de la variation dans un avenir rapproché.

6 Remarques conclusives

Dans cet article, nous avons initié une réflexion autour de contextes d'analyse de la variation dialectale en langues de spécialité. Plus précisément, dans ce cadre, nous avons examiné plusieurs possibilités d'outils pour l'exploration de ce type de variation, simultanément. Partant du principe que la variation s'observe sur des groupes de phénomènes (et non pas sur la terminologie en entier d'un domaine), nous avons testé trois approches différentes qui permettent de « brasser » de telles données complexes et volumineuses pour en entamer l'exploration, impossible « manuellement » ou sur la

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

base de tableaux et contextes seuls. Ces trois axes, la visualisation, le regroupement et la modélisation, s'ils ne sont pas parfaits, permettent en partie de s'affranchir de contraintes matérielles (telles que la lourdeur de tableaux à manipuler), cognitives (telles que l'impossibilité de traiter « humainement » de grosses quantités de données à comparer) et temporelles (puisque permettant d'optimiser et de faire ressortir des phénomènes saillants difficilement visibles à l'œil nu).

Malgré ce potentiel, ces types d'outils restent encore peu nombreux pour les LSP, qui présentent certaines spécificités. En particulier, les corpus utilisés ont souvent une taille plus réduite que ceux de langue générale, caractéristique qui a un impact direct sur la performance de certaines approches statistiques. Néanmoins, le potentiel de ces outils existe et est exploré par d'autres domaines ou points de vue complémentaires pour les langues de spécialité, tels que la linguistique diachronique. Nombre de ces outils sont librement disponibles et les LSP peuvent fortement s'inspirer de ces démarches. De plus, les besoins actuels, tant appliqués que fondamentaux, des langues de spécialité révèlent l'urgence de chercher à développer cet aspect.

Bibliographie

Anthony, Laurence (2018) « AntConc (Version 3.5.7) [Computer Software] ». Disponible sur http://www.laurenceanthony.net/software, Tokyo, Japan : Waseda University.

Auger, Pierre (2001), « Essai d'élaboration d'un modèle terminologique/terminographique variationniste ». *TradTerm*, 7, p. 183-224.

Aussenac-Gilles, Nathalie ; Condamines, Anne ; Szulman, Sylvie (2002) « Prise en compte de l'application dans la constitution de produits terminologiques ». *Actes des 2èmes assises nationales du groupe de recherche I3 (Information, Interaction, Intelligence)*, Claude Le Maître (dir.), Cépaduès, Nancy, France, p. 289-303.

Bertaccini, Franco ; Matteucci, Alessandra. (2005), « L'approche variationniste à la pratique terminologique d'entreprise. ». *Meta : le journal de traducteurs, Presses Universitaires de Montreal*, 50(4).

Bourigault, Didier ; Slodzian, Monique (1999), « Pour une terminologie textuelle ». *Terminologies Nouvelles*, 19, p. 29-32.

Bowker, Lynne ; Pearson, Jennifer, *Working with Specialized Language : a Practical Guide to Using Corpora*. Routledge, London/New York, 2002.

Cabré, Maria Teresa, *La terminologie : théories, méthodes et applications*. Armand Colin, Presses de l'Université d'Ottawa, Ottawa, 1998.

Condamines, Anne ; Picton, Aurélie (2015), « Terminologie outillée : analyse de corpus spécialisés dans différentes situations de néologie ». Conférence-hommage à John Humbley, « Quo vadis, Terminologia », Paris (France).

Coseriu, Eugenio (1966) « Structure lexicale et enseignement du vocabulaire. ». *Actes du premier colloque international de linguistique appliquée*, Nancy (France), p. 175-217.

Coseriu, Eugenio (1998), « Le double problème des unités "DIA-S" ». *Les cahiers DIA. Études sur la diachronie et la variation linguistique*, coll. « Communication et Cognition », p. 9-16.

Daille, Béatrice, *Term Variation in Specialised Corpora : Characterisation, automatic discovery and applications*, John Benjamins, Amsterdam, 2017.

Desmet, Isabel (2007), « Terminologie, culture et société. Éléments pour une théorie variationniste de la terminologie et des langues de spécialité ». *Cahiers du Rifal*, Terminologie, culture et société, p. 3-13.

Dostie, Gaétane ; Hadermann, Pascale (Éd.) (2016), *Diasystème et variation en français actuel : aspects sémantiques*. Carnets de lecture, 29.

Drouin, Patrick (2003), « Term Extraction Using Non-technical Corpora as a Point of Leverage ». *Terminology*, 9(1), p. 99-117.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Drouin, Patrick (2017) « Should we be looking for the needle in the haystack or in the straw poll? ». *in* Patrick Drouin ; Aline Francoeur ; John Humbley ; Aurélie Picton (dir.), *Multiple Perspectives in Terminological Variation*, John Benjamins, coll. « Terminology and Lexicography Research and Practice », Amsterdam/New York, p. 131–152.

Dury, Pascaline, *La dimension diachronique en anglais de spécialité : une approche terminologique*. note de synthèse pour l'Habilitation à Diriger des Recherches, Université Paris 7, 2018.

Fernández-Silva, Sabela (2016), « The cognitive and rhetorical role of term variation and its contribution to knowledge construction in research articles ». *Terminology*, 22(1), p. 52-79.

Ferré, Arnaud (2017) « Représentation de termes complexes dans un espace vectoriel relié à une ontologie pour une tâche de catégorisation ». *Actes Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA 2017)*, Caen (France).

Freixa, Judit (2002), *Anàlisi de la variació denomitiva en textos de different grau d'especialització de l'àrea de medi ambient*. Thèse de doctorat en linguistique, Université de Barcelone.

Freixa, Judit (2006), « Causes of Denominative Variation in Terminology – A Typology Proposal ». *Terminology*, 12(1), p. 51-77.

Freixa, Judit ; Fernández-Silva, Sabela (2017) « Terminological variation and the unsaturability of concepts ». *in* Patrick Drouin ; Aline Francoeur ; John Humbley ; Aurélie Picton (dir.), *Multiple Perspectives on Terminological Variation*, John Benjamins, coll. « Terminology and lexicography research and practice », Amsterdam/New York, p. 155-180.

Gadet, Françoise, *La variation sociale en français*. L'essentiel, Ophrys, Paris, 2003.

Gaudin, François, *Socioterminologie – Une approche sociolinguistique de la terminologie*. Champs linguistiques, De Boeck – Duculot, Bruxelles, 2003.

Gesmann, Markus ; De Castillo, Diego (2011), « Interface between R and the Google Visualisation API. GoogleVis package for R., http ://cran.r-project.org/Web/packages/googleVis/googleVis.pdf.

Gries, Stefan Th. ; Hilpert, Martin (2008), « The identification of stages in diachronic data : variability-based neighbour clustering ». *Corpora*, 3(1), p. 59-81.

Gries, Stefan Th. ; Hilpert, Martin (2012) « Variability-based neighbor clustering : a bottom-up approach to periodization in historical linguistics ». *in* Terttu Nevalainen ; Elizabeth. Traugott (dir.), *The Oxford handbook of the history of English*, Oxford University Press, Oxford, p. 134-144.

Gries, Stefan Th. ; Hilpert, Martin (2010), « Modeling diachronic change in the third person singular : a multifactorial, verb- and author-specific exploratory approach ». *English Language and Linguistics*, 14(3), p. 293-320.

Habert, Benoît ; Illouz, Gabriel ; Folch, Helka (2005) « Des décalages de distribution aux divergences d'acception ». *in* Anne Condamines (dir.), *Sémantique et Corpus*, Hermès, Paris.

Heiden, Serge ; Magué, Jean-Philippe ; Pincemin, Bénédicte. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (éd.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010) (vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.

Hilpert, Martin (2011), « Dynamic visualizations of language change : Motion charts on the basis of bivariate and multivariate data from diachronic corpora ». *International Journal of Corpus Linguistics*, 16(4), p. 435-461.

Humbert-Droz, Julie (en cours), *Circulation des termes entre langues de spécialité et langue générale : proposition d'un cadre théorique et méthodologique d'analyse du phénomène de la déterminologisation*. Thèse de doctorat en Traitement informatique multilingue/Sciences du langage, Université de Genève/Université de Toulouse – Jean Jaurès (cotutelle).

Humbert-Droz, J., Picton, A., & Condamines, A. (2019). How to build a corpus for a tool-based approach to determinologisation in the field of particle physics. Research in Corpus Linguistics, 7, 1-17. doi :10.32714/ricl.07.01

Jacquemin, Christian, *Spotting and Discovering Terms through NLP*. MIT Press, Cambridge MA, 2001.

Kacprzak, Alicia ; Goudaillier, Jean-Pierre (2014), « Dénominations des maladies en langue populaire et argotique (de la 'synonymite' des noms de maladies) ». *e-Scriptat Romanica*, 1, p. 1-8.

Kerremans, Koen (2017) « Towards a resource of semantically and contextually structured term variants and their translations ». *in* Patrick Drouin ; Aline Francoeur ; John Humbley ; Aurélie Picton (dir.), *Mul-*

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

tiple Perspectives on Terminological Variation, John Benjamins, coll. « Terminology and lexicography research and practice », Amsterdam/New York, p. 83-108.

Kilgarriff, Adam ; Baisa, Vít ; Bušta, Jan ; Jakubíček, Miloš ; Kovář, Vojtěch ; Michelfeit, Jan ; Rychlý, Pavel ; Suchomel, Vít (2014), « The Sketch Engine : ten years on ». *Lexicography*, 1(1), p. 7-36.

León-Araúz, Pilar ; Reimerink, Arianne (2015) « Signs and Symptoms in the Psychiatric Domain : A Corpus Analysis ». in Nieves Jiménez Carra ; Elisa Calvo ; Nuria Fernández-Quesada ; Alicia María López Márquez ; Alice Stender (dir.), *Procedia – Social and Behavioral Sciences, 32nd International Conference of the Spanish Association of Applied Linguistics (AESLA) : Language Industries and Social Change*, Elsevier, 173, p. 285-292.

León-Araúz, Pilar (2015) « Term variation in the psychiatric domain : transparency and multidimensionality ». *in* P. Hacken ; R. Panocová (dir.), *Word Formation and Transparency in Medical English*, Cambridge Scholars Publishing, Newcastle-upon-Tyne, p. 33-54.

Mikolov, Tomas ; Chen, Kai ; Corrado, Greg ; Dean, Jeffrey (2013), « Efficient estimation of word representations in vector space ». *arXiv :1301.3781*, http ://arxiv.org/abs/1301.3781.

Moreau, M.L., *Sociolinguistique : les concepts de base*. Mardaga, 1997.

Picton, Aurélie (2018), « Terminologie outillée et diachronie : éléments de réflexion autour d'une réconciliation ». *ASp [En ligne]*, 74, p. 27-52.

Picton, Aurélie ; Dury, Pascaline (2015) « Les discours d'expertise en langues de spécialité : le point de vue du terminologue ». *in* Céline Beaudet ; Véronique Rey (dir.), *Écritures expertes en questions*, Presses universitaires de Provence, Aix-en-Provence, p. 265-278.

Picton, Aurélie (2014) « The Dynamics of Terminology in Short-Term Diachrony : A Proposal for a Corpus-based Methodology to observe Knowledge Evolution ». *in* R. Temmerman ; M. Van Campenhoudt (dir.), *The Dynamics of Culture-bound Terminology in Monolingual and Multilingual Communication,* coll. « Terminology and Lexicography Research and Practice », John Benjamins, 16, Amsterdam/Philadelphie, p. 159–182.

Picton, Aurélie ; Dury, Pascaline (2017) « Diastratic Variation in Language for Specific Purposes. Observations from the Analysis of two Corpora ». *in* Patrick Drouin ; Aline Francoeur ; John Humbley ; Aurélie Picton (dir.), *Multiple Perspectives in Terminological Variation*, John Benjamins, coll. « Terminology and Lexicography Research and Practice », Amsterdam/New York, p. 57-79.

Ratinaud, P. (2009). *Iramuteq : Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires,* http ://www.iramuteq.org.

Saussure (De), Ferdinand, *Cours de Linguistique Générale*. (première édition 1916), Payot, Paris, 1995.

Tartier, Annie (2004), *Analyse automatique de l'évolution terminologique : variations et distances*. Thèse de doctorat en Informatique, Université de Nantes.

Temmerman, Rita, *Towards New Ways of Terminological Description. The Sociocognitive Approach*. John Benjamins, Amsterdam/Philadelphie, 2000.

Urieli, Assaf (2013), *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat en Sciences du langage, Université Toulouse – Jean Jaurès.

van der Maaten Laurens ; Hinton, Geoffrey (2008), « Visualizing High-Dimensional Data Using t-SNE », *Journal of Machine Learning Research*, 9, p. 2579–2605.

Verjans, Thomas (2013) « Les locutions conjonctives. Une hypothèse romane ». *in* Pascale Hadermann ; Marieke Van Acker ; Boutier Marie-Guy (dir.), *La variation et le changement en langue*, Société Néophilologique, Helsinki, p. 133-147.

Wüster, E. (1981) « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses ». *in* G. Rondeau et H. Felber (dir.), *Textes choisis de terminologie*, GIRSTERM, Groupe Interdisciplinaire de Recherche Scientifique et Appliquée en Terminologie, I. Fondements théoriques de la terminologie, sous la direction de V.I. Siforov, Québec, Canada.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454