**Analyse des corpus spécialisés "multi-comparables" à l'aide d'une approche outillée : quelles limites et quelles pistes de travail?**

Patrick Drouin, Aurélie Picton & Julie Humbert-Droz

## 1. Contexte de la recherche

Dans cette communication, 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'année, 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 sous différents angles la question de la variation (méthodologiques, théoriques, descriptifs), peu de travaux encore 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 (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. L'étude de la déterminologisation implique alors à la fois d'observer ce "flux" entre spécialisé et général, ainsi que la dimension diachronique,
- d'autre part, la création de la Humanitarian Encyclopaedia (https://humanitarianencyclopedia.org/) dont l'un des objectifs principaux est d'analyser les différences d'usage des termes du domaine selon différentes perspectives dans ce domaine. 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 contexte impose aux LSP de développer des méthodes qui permettent d'explorer ces dimensions simultanément, autant que possible.

## 2. Mise en place d'une approche outillée

### 2.1. Constitution des corpus

Dans un premier temps, nous décrirons en détail les deux corpus "multi-comparables" construits dans ces études, ainsi que les principaux défis rencontrés dans cette tâche.

Le corpus de déterminologisation regroupe des textes dans le domaine de la physique des particules de 2003 à 2016, répartis en 5 catégories pour approcher le continuum entre spécialisé et général (Humbert-Droz, 2017). Le corpus humanitaire regroupe des textes du domaine, provenant d'organisations, disciplines et régions du monde différentes[1].

[1] À propos d'un corpus préliminaire, voir : Humanitarian Encylopedia, "Evidence 3 : 100 most frequently used terms by humanitarian actors", Humanitarianencyclopedia.org, November 2017, https://humanitarianencyclopedia.org/events/hew-humanitarian-evidence-week/evidence-3-100-most-frequently-used-terms-by-humanitarian-actors/ (12.02.2018)

### 2.2. Approche outillée, indices "classiques" et nouveaux défis pour les LSP

L'approche choisie dans cette étude s'appuie sur les méthodes et outils de la Terminologie textuelle (par ex. Condamines et Picton, 2015). Cette approche se base sur des outils classiques tels que des concordanciers, des analyseurs syntaxiques, etc., ainsi que sur trois familles d'indices tels que 1. la comparaison de fréquences et de comportements statistiques, 2. l'observation de variations de formes et 3. l'analyse des cooccurrents.

Néanmoins, appliquée à des corpus comparables complexes tels que ceux construits pour les deux contextes mentionnés, cette approche montre différents défis et limites. Nous soulignerons en particulier le besoin de saisir le "lien" entre différents types de variations (par exemple la chronologie inhérente à la déterminologisation), ainsi que les difficultés techniques de travail sur des données de taille modeste pour l'analyse des cooccurrents. En réponse à ces limites, nous discuterons en particulier deux axes : les besoins de modélisation des comportements lexicaux et les besoins de visualisation des données.

## 3. Axes de développement : propositions

### 3.1. Visualisation

Afin de permettre de faire émerger les phénomènes à l'oeuvre, nous avons exploré les possibilités offertes par les MotionCharts (par ex. Gesmann & de Castillo 2011) et déjà utilisés en diachronie en langue générale (par ex. Hilpert 2011). Cet outil permet en effet d'observer deux types de dimensions simultanément, l'une d'entre elles étant nécessairement la dimension diachronique. Nous montrerons la manière dont cet outil permet notamment d'observer certains mécanismes (ou certains rythmes) de déterminologisation dans le cas de l'analyse de la terminologie des particules en physiques des particules.

### 3.2. Analyse distributionnelle pour la modélisation

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 deux axes. Un premier axe consiste à étudier la variabilité du point de vue des documents afin de faire émerger de ces derniers des regroupements. 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 mots, d'identifier des regroupements de documents possédant des propriétés communes (des périodes chez les chercheurs cités). Les regroupements ainsi créés peuvent ensuite faire l'objet d'analyses plus fines visant à décrire la variabilité observée. Une autre approche consiste à observer la variabilité avec une granularité plus petite et à s'intéresser aux phénomènes qui touchent directement les mots, 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 (*word embeddings*) qui ont pour but de capturer le sens en contexte des unités lexicales en les représentant sous forme de vecteurs numériques (Mikolov et al. 2013, Ferré 2017). Nous explorerons ces deux techniques afin de faire émerger des divisions thématiques au sein de nos corpus et d'illustrer les variations de sens associées aux formes au sein de ces mêmes corpus comparables.

**4. Références citées**


Collinson, S. et Elhawary, S. (2012). « Humanitarian space: a review of trends and issues »,| Overseas Development Institute (ODI). London.

Condamines, A. & Picton, A. (2014). « Étude du fonctionnement des nominalisations déverbales dans un contexte de déspécialisation ». In: Congrès Mondial de Linguistique Française. Berlin (Germany). [s.l.] : [s.n.], 2014. p. 697-711.

Condamines, A. & Picton, A. (2015) « Terminologie outillée : analyse de corpus spécialisés dans différentes situations de néologie ». *In* conférence-hommage à John Humbley, « Quo vadis, Terminologia », Paris, France. conférence-hommage à J. Humbley, « Quo vadis, Terminologia ».

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

Galisson, R. (1978). Recherches de lexicologie descriptive : la banalisation lexicale, Nathan, collection « Université, Information, Formation », Paris.

Gesmann M. & De Castillo D. (2011). « Interface between R and the Google Visualisation API. GoogleVis package for R. » http://cran.r-project.org/web/packages/googleVis/googleVis.pdf, Consulté le 12 mars 2018.

Gries S.T. & Hilpert M. (2008). « The identification of stages in diachronic data: variability-based neighbour clustering », Corpora 3 (1), 59-81.

Gries S.T. & Hilpert M. (2012). « Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics », in T. Nevalainen & E. Traugott (eds), The Oxford handbook of the history of English, Oxford : Oxford University Press, 134-144.

Hilpert, M. (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), 435–461.

Humbert-Droz, J. (2017). « Définition d'un corpus comparable pertinent pour l'étude de la déterminologisation ». *Colloque Jeunes Chercheurs – DyLis 2017, Diversité de la constitution des données : sur quoi travaille-t-on en Sciences du langage ?* Université de Rouen, 23-24 novembre.

Meyer, I., & Mackintosh, K. (2000). « L'"étirement" du sens terminologique : aperçu du phénomène de la déterminologisation ». In H. Béjoint & P. Thoiron (Eds.), *Le sens en terminologie* (pp. 198–217). Lyon: Les Presses Universitaires de Lyon.

Mikolov, T., Chen, K., Corrado, G. et J. Dean. (2013). « Efficient estimation of word representations in vector space ». arXiv:1301.3781. http://arxiv.org/abs/1301.3781, Consulté le 12 mars 2018.

Renouf, A. (2017). « Some Corpus-Based Observations on Determinologisation ». *Neologica*, 11, 21–48.

Ungureanu, L. (2003). *L'interpénétration langue générale-langue spécialisée dans le discours d'internet*. Thèse de doctorat en Sciences du Langage. Université Paris 13/Université Technique de Moldova.



**Mots clés :** corpus comparables, linguistique de corpus, variation terminologique, diachronie, diastratie, analyse distributionnelle, visualisation