Exploration d'un corpus de textes philosophiques québécois

Camille Demers, David Valentine, Sara-Maude Bergeron, Dominic Forest

▶ **To cite this version:**

Camille Demers, David Valentine, Sara-Maude Bergeron, Dominic Forest. Exploration d'un corpus de textes philosophiques québécois : enjeux, méthodes et résultats préliminaires. Humanistica 2024, Association francophone des humanités numériques, May 2024, Meknès, Maroc. hal-04559836

HAL Id: hal-04559836
https://hal.science/hal-04559836

Submitted on 25 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Exploration d'un corpus de textes philosophiques québécois: enjeux, méthodes et résultats préliminaires

Camille Demers, David Valentine, Sara-Maude Bergeron, Dominic Forest

Université de Montréal

{camille.demers,david.valentine,sara-maude.bergeron,dominic.forest}@umontreal.ca

Résumé

Cet article s'intéresse à l'application de méthodes de fouille de textes pour assister l'analyse d'un corpus de textes philosophiques québécois. Nous décrivons le traitement d'un corpus d'articles publiés au Québec entre 1945 et 2023 dans la revue *Laval théologique et philosophique*. Nous présentons les opérations menées sur ce corpus et les défis soulevés afin d'extraire les informations caractéristiques de la tradition philosophique de langue française au Québec et de cartographier de manière diachronique l'évolution des thématiques abordées sur une période de 79 ans.

1 Introduction

Cet article s'intéresse à l'application de méthodes de fouille de textes pour assister l'analyse d'un corpus de textes philosophiques québécois. Nous décrivons le traitement d'un corpus d'articles publiés au Québec entre 1945 et 2023 dans la revue *Laval théologique et philosophique* (LTP) (Dutron, 2021). L'objectif de cette recherche est d'analyser ce corpus historique à l'aide d'outils informatiques inspirés de la statistique textuelle et l'intelligence artificielle. Dans le cadre d'une analyse sémiotique de la textualité, nous cherchons à valider des technologies existantes et à concevoir de nouvelles approches méthodologiques pour appuyer le traitement d'un patrimoine numérisé trop vaste pour être maîtrisé par les méthodes d'analyse manuelle traditionnellement employées en philosophie. Plus spécifiquement, cet article présente les opérations menées et les défis soulevés afin : 1) d'extraire les informations caractéristiques de la tradition philosophique de langue française au Québec, qui sont autrement difficilement identifiables par des analyses manuelles et 2) de cartographier de manière diachronique l'évolution des thématiques abordées dans LTP sur une période de 79 ans.

2 Sources de données et méthodologie

2.1 Corpus

Nos travaux mobilisent un corpus dont la collecte et l'enregistrement ont été effectués en amont du projet. Les données proviennent d'un travail de numérisation rétroactive de la revue LTP par le Consortium Érudit. Il s'agit d'une revue comptant 79 volumes, publiés au rythme d'un volume de trois numéros par année, contenant un total de 3 824 documents, dont 1 553 articles publiés entre 1945 et 2023. Le corpus est rendu intégralement disponible dans les formats PDF et XML EruditArticle (Spina, 2014), et en HTML depuis 2002.

Seuls les 1 553 articles du corpus initial ont été retenus pour les analyses. Les 2 262 autres documents correspondent à des listes d'ouvrages rendus disponibles pour recension, des comptes rendus, des listes de mémoires et de thèses récemment publiés ou des textes d'opinion ; ces documents ont été considérés comme non pertinents et ont donc été exclus des analyses.

Parmi les 1 553 documents retenus, 280 articles rédigés dans d'autres langues que le français ont également été exclus (anglais, n=277 ; autres langues, n=5 [latin, espagnol]). Le corpus final regroupe donc 1 259 articles de philosophie en langue française, totalisant 7 080 221 occurrences de mots (tokens) pour 204 111 formes uniques (types).

Ce corpus est caractérisé par une distribution asymétrique du nombre d'articles publiés par année (voir figure 1), avec un nombre beaucoup plus important d'articles publiés entre les années 1990 et 2000. Cette distribution pose un certain nombre d'enjeux relatifs à la comparabilité des données à travers le temps, motivant le besoin d'analyses normalisées permettant de réduire d'éventuels biais associés à cette asymétrie.

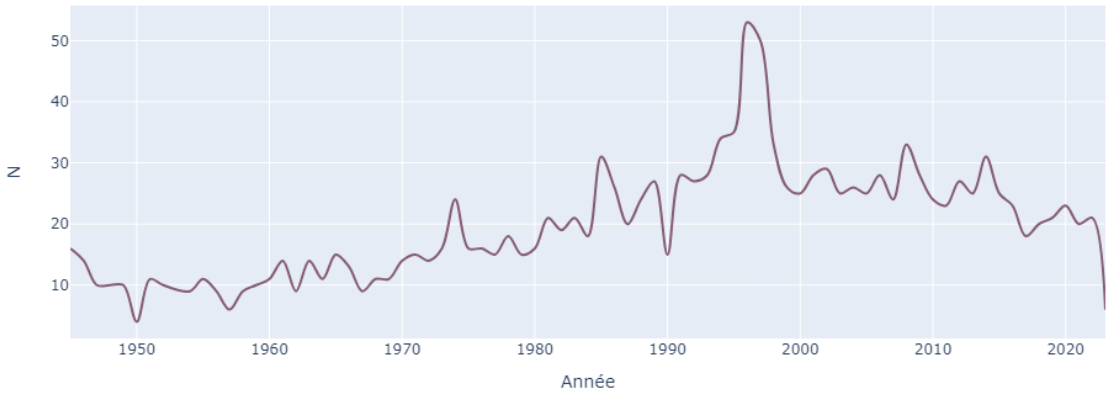

FIGURE 1 – Distribution du nombre d'articles publiés par année dans la revue *Laval théologique et philosophique*.

2.2 Nettoyage : formes de bruit rencontrées et solutions proposées

La structure et la qualité de la numérisation du corpus varient grandement dans le temps. En effet, plusieurs formes de bruit (Al Sharou et al., 2021) se manifestent. D'une part, les textes présentent un nombre substantiel d'erreurs d'océrisation et de problèmes d'encodage de caractères (voir figure 2). Un travail manuel d'annotation et de correction des métadonnées a été couplé à un nettoyage des textes en lot basé sur une combinaison variable de manipulations de la structure XML et sur l'emploi d'expressions régulières adaptées aux formes de bruit rencontrées. Ces approches complémentaires permettent d'augmenter considérablement la qualité des données.

D'autre part, certaines formes de bruit sont liées au processus de structuration des données effectué lors de numérisation rétroactive des documents. Bien que la numérisation ait été conduite en suivant un schéma XML documenté par Érudit (Spina, 2014), le corpus présente d'importantes différences de balisage à travers le temps, allant d'un balisage minimal à un balisage sémantique complet (voir figure 3). Pour pallier cette structuration différenciée des données, nous avons extrait du corpus au format XML un ensemble uniforme de métadonnées et d'annotations dans un format CSV plus flexible, mais moins riche, permettant un traitement simplifié des documents.

La présence de segments multilingues dans de nombreux articles du corpus (extraits cités en latin, en grec ancien, en allemand et en anglais) constitue également une source de bruit considérable. Nous avons conservé ces extraits en supprimant les mots

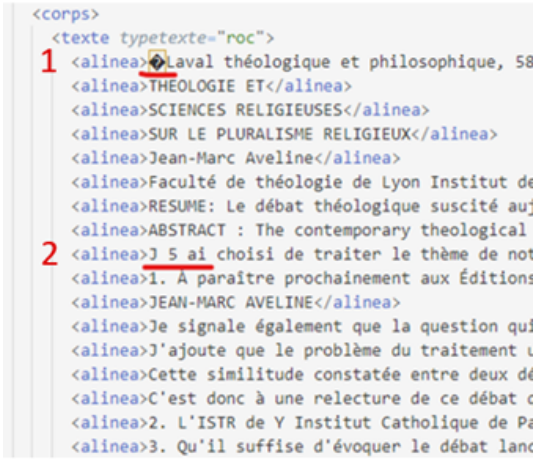

FIGURE 2 – Erreur d'encodage de caractères (1) et d'océrisation (2)

fonctionnels de l'ensemble des langues présentes dans le corpus en y appliquant une combinaison d'anti-dictionnaires spécifiques.

Finalement, la nature bidisciplinaire de LTP, laquelle découle historiquement d'une forte interaction entre la philosophie et la religion au Québec, pose une difficulté de fond relativement aux objectifs du projet. En effet, la présence de nombreux articles de théologie soulève la question de la frontière incertaine entre ces deux disciplines. Certains indicateurs textuels – qui prennent la forme d'annotations et de métadonnées telles que l'affiliation institutionnelle et départementale de l'auteur – ne permettent pas d'effectuer automatiquement cette distinction de manière satisfaisante puisque ces informations n'ont pas été systématiquement renseignées par la revue pendant plusieurs décennies. Il s'agit donc d'un problème de silence provoqué par les variations des processus éditoriaux.

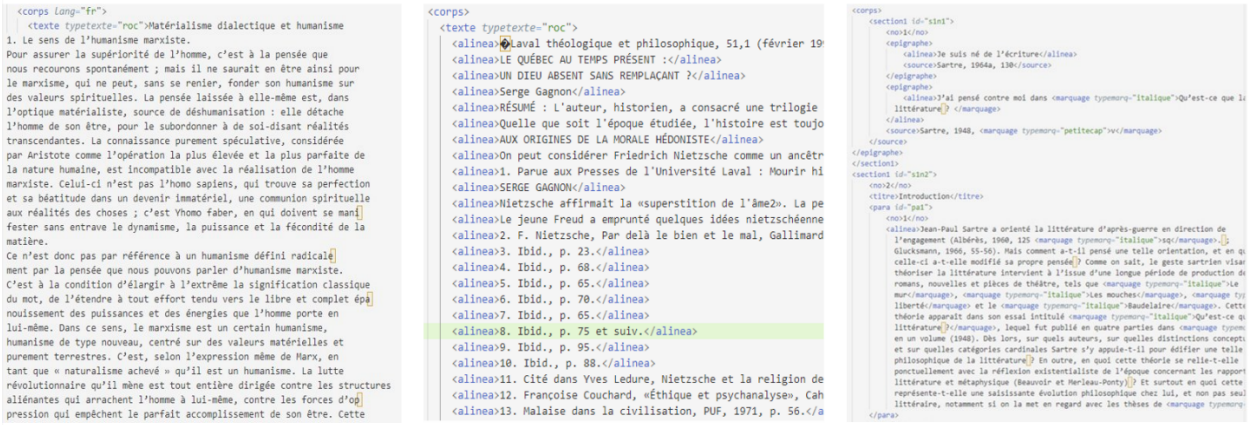

FIGURE 3 – Trois grandes phases dans le balisage XML du *Laval théologique et philosophique*. À gauche, aucun balisage dans le texte. Au centre, balisage en `<alinea>`. À droite, balisage sémantique complet.

3 Analyses

3.1 Extraction terminologique

Une première analyse a consisté à réaliser une tâche d'extraction terminologique visant à identifier les formes simples et les expressions complexes représentatives du corpus. Cette tâche a été réalisée en procédant à l'extraction des principaux syntagmes nominaux sur la base d'une pondération TF-IDF. Cette mesure de pondération permet d'ordonnancer les termes selon leur spécificité, en tenant compte à la fois de la fréquence de chaque terme dans chaque document du corpus (TF) et du nombre de documents dans lesquels se retrouve chaque terme (IDF), générant un score plus élevé pour les termes ayant une forte concentration dans un faible nombre de documents. Par la suite, nous avons procédé à l'extraction de séquences de mots (n-grammes) de longueur allant d'un (formes simples) à six mots, puis à la sélection d'un ensemble de termes candidats en fonction d'un filtrage basé sur les patrons syntaxiques des séquences extraites.

3.2 Analyse des termes au regard de vocabulaires d'indexation

Les termes extraits ont par la suite été mis en correspondance par une comparaison de chaînes de caractères avec les termes de deux vocabulaires d'indexation issus des bases de données bibliographiques Pascal et Francis (Khayari et al., 2021) :

— un vocabulaire de philosophie contenant 4 435 concepts bilingues (anglais/français) accompagnés de synonymes et d'une catégorie sémantique d'appartenance (par exemple « Courant de pensée », « Processus mental », « Titre de document/œuvre », etc.) ;

— un vocabulaire d'histoire et de sciences des religions, contenant 4 579 concepts bilingues et de leurs synonymes.

Parmi les termes de ces vocabulaires, 1 543 concepts se sont avérés communs aux deux ensembles.

Nous avons mené trois types d'analyses mobilisant ces vocabulaires : 1) une analyse globale basée sur le nombre de termes issus de chaque vocabulaire dans l'ensemble du corpus ; 2) une analyse chronologique basée sur la distribution du nombre de termes issus de chaque vocabulaire par année, et 3) une analyse cherchant à mobiliser les catégories sémantiques présentes dans ces vocabulaires pour explorer les termes identifiés.

3.2.1 Analyse globale

La mise en correspondance des deux vocabulaires et du corpus de textes philosophiques nous permet de constater une distribution équilibrée des termes issus de chacun d'eux au niveau global (pour toutes les années considérées). 3 725 termes du vocabulaire de philosophie s'y retrouvent, tandis que ce nombre s'élève à 3 472 pour le vocabulaire des sciences des religions (voir tableau 1).

Lexique	**Nombre de termes**	**%**
Philosophie	3 725	52%
Religions	3 472	48%
Total	7 197	100

TABLEAU 1 – Distribution des termes de deux vocabulaires d'indexation en philosophie et en sciences des religions au sein du corpus

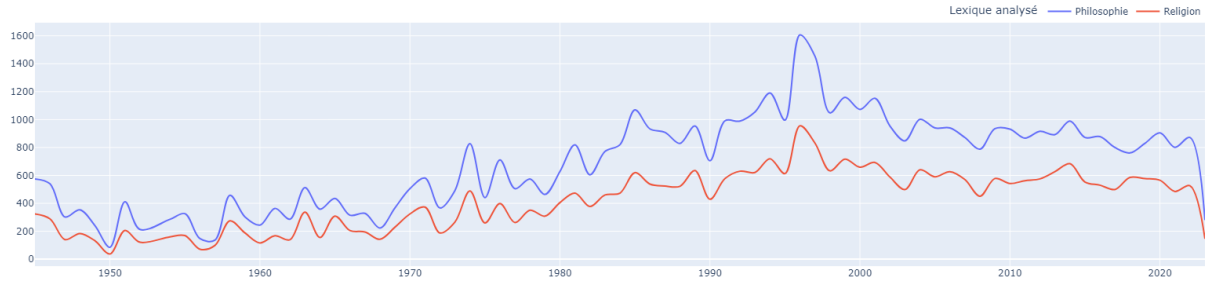

FIGURE 4 – Distribution du nombre de termes issus des vocabulaires de philosophie et de sciences des religions présents dans la revue LTP à travers le temps.

Cette distribution fait écho aux enjeux précédemment soulevés relatifs à la difficulté de distinguer entre philosophie et théologie au sein de la revue LTP.

3.2.2 Analyse chronologique

La distribution diachronique des deux vocabulaires en question dans le corpus illustre de légères variations des proportions à travers le temps (voir figure 4). Cette évolution de la distribution des termes à travers le temps demeure néanmoins conforme au portrait global.

3.2.3 Analyse par catégorie sémantique

Le tableau 2 illustre l'importance des catégories sémantiques du vocabulaire de philosophie dans le corpus. La tendance montre une prépondérance de termes liés à des anthroponymes (365 termes illustrants des noms de personnes comme Christ, Aristote, Thomas d'Aquin, Empédocle, Paul Ricœur, etc.), à des courants de pensée (263 termes à propos du christianisme, de l'islam, du cogito, du déterminisme, du nominalisme, etc.) et à des disciplines (220 termes relatifs à la théologie, à la métaphysique, à la dialectique, à la phénoménologie, à l'ontologie, etc.).

Catégorie sémantique	Nombre de termes
Anthroponyme	365
Courant de pensée	263
Discipline	220
Comportement	124
Idée / Abstraction	109
Toponyme	106
Activité	84
Langue / Langage	84
Processus mental	72
Titre de document / œuvre	63
Raisonnement logique	61
Activité / Comportement	59
Type de document / œuvre	58
Fonction	50

TABLEAU 2 – Distribution des termes issus de vocabulaires d'indexation de philosophie et de sciences des religions identifiés dans la revue LTP, par catégorie sémantique (n $\geq$ 50)

3.3 Regroupement automatique des termes

Une deuxième analyse a consisté à extraire des regroupements sémantiques entre les termes extraits à l'aide d'une méthode de fouille non supervisée basée sur l'emploi d'un algorithme de regroupement automatique (*clustering*). Cette méthode a été appliquée sur quatre périodes constitutives du corpus afin d'explorer l'évolution des principaux regroupements de termes à travers le temps.

Pour ce faire, nous avons extrait les 5000 termes (simples et complexes) les plus représentatifs des périodes 1940-1959, 1960-1979, 1980-1999 et 2000-2023, sur la base de leur score TF-IDF. Pour chaque période considérée, les 5 000 termes ayant le plus haut score TF-IDF ont été retenus pour les analyses. Un modèle de vectorisation (*embedding*) a été appliqué pour générer des représentations numériques permettant de calculer un score de similarité sémantique entre les termes (basé sur la distance cosinus entre les vecteurs) pour faire émerger des regroupements au sein des données. Le modèle employé pour la vectorisation est le modèle multilingue *paraphrase-multilingual-MiniLM-L12-v2* de la librairie Python *sentence-transformers* (Reimers et Gurevych, 2019). Le regroupement automatique des termes a été réalisé au moyen de l'algorithme *community detection*, également issu de la librairie *sentence-transformers*. Les figures 5, 6, 7 et 8 illustrent les regroupements identifiés entre les termes issus de chaque période chronologique. Ces illustrations révèlent certaines caractéristiques du discours

philosophique et religieux dans LTP. Ainsi, on note une distinction plus marquée entre le lexique philosophique et le lexique religieux dans les périodes 1980-1999 et 2000-2023 par opposition aux périodes précédentes. Dans le même ordre d'idées, ces figures mettent en lumière l'absence de termes relatifs à la philosophie morale dans la période vicennale 1940-1959.

4 Conclusion

Cette contribution présente les objectifs, les enjeux, les méthodes et les résultats préliminaires d'un projet cherchant à analyser et à visualiser un corpus de textes philosophiques à l'aide d'outils informatiques inspirés de la statistique textuelle et l'intelligence artificielle. Le corpus analysé est composé d'articles publiés au Québec entre 1945 et 2023 dans la revue *Laval théologique et philosophique*. Nous décrivons comment nous avons procédé au nettoyage des données textuelles et illustrons les variétés de bruit auxquelles nous avons été confrontés avant d'entamer nos traitements informatiques. Les résultats obtenus jusqu'à maintenant pour ce projet de recherche en cours sont encourageants. Ils nous permettent d'identifier globalement les termes les plus caractéristiques de ce corpus selon un découpage par période de 20 ans. Les prochaines étapes de notre projet consisteront à explorer de nouvelles méthodes afin d'extraire plus précisément les informations caractéristiques de ce corpus important dans le domaine de philosophie québécoise. Ainsi, nous souhaitons explorer les algorithmes de regroupement en tenant compte des différentes métadonnées ou annotations associées à ce corpus (nom d'auteurs, affiliations, etc.), et de certains algorithmes de classification supervisée.

Bibliographie

Khetam Al Sharou, Zhenhao Li, et Lucia Specia. 2021. Towards a Better Understanding of Noise in Natural Language Processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62. INCOMA Ltd.

Martin Dutron. 2021. La cartographie disciplinaire de vingt-cinq années de publication de « savoirs religieux » à Québec : le cas du périodique Laval théologique et philosophique (1945 à 1969). *Études d'histoire religieuse*, 87(1-2) :65.

Majid Khayari, Véronique Reszetko, Dominique Vachez, Nathalie Vedovotto, Jérémy Yon, et Sophie Aubin. 2021. De TermSciences à Loterre : comment l'Inist-CNRS a rendu les terminologies ouvertes plus conformes aux principes FAIR.

Francis Lareau. 2022. Approche computationnelle de l'analyse conceptuelle : présentation opérationnelle et approfondissement méthodologique de la détection d'un concept dans des extraits textuels. *Philosophiques*, 49(2) :413–431.

Nils Reimers et Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv :1908.10084 [cs].

Jean-Claude Simard. 2022. Un exemple d'humanités numériques : l'analyse de la revue Philosophiques. *Philosophiques*, 49(2) :395–412.

Isabelle Spina. 2014. Documentation du modèle XML Érudit (eruditarticle.xsd).

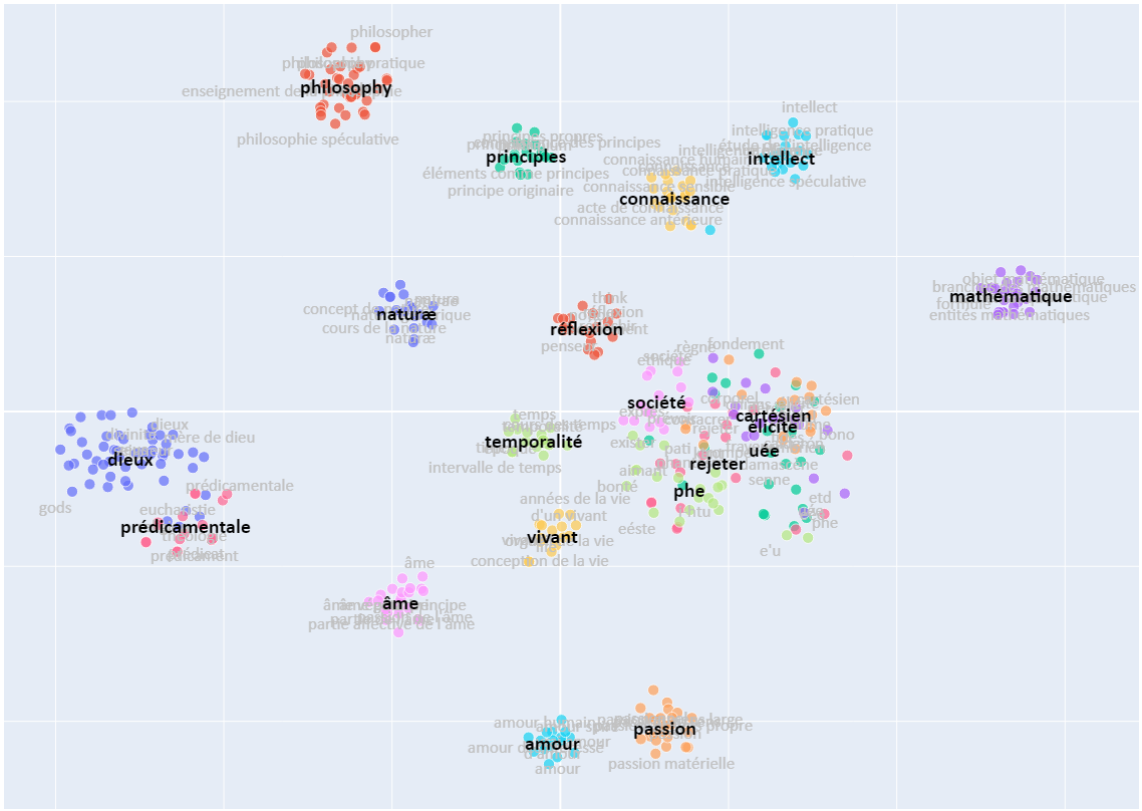

FIGURE 5 – Regroupement automatique des principaux termes extraits dans la revue LTP pour la période 1940-1959

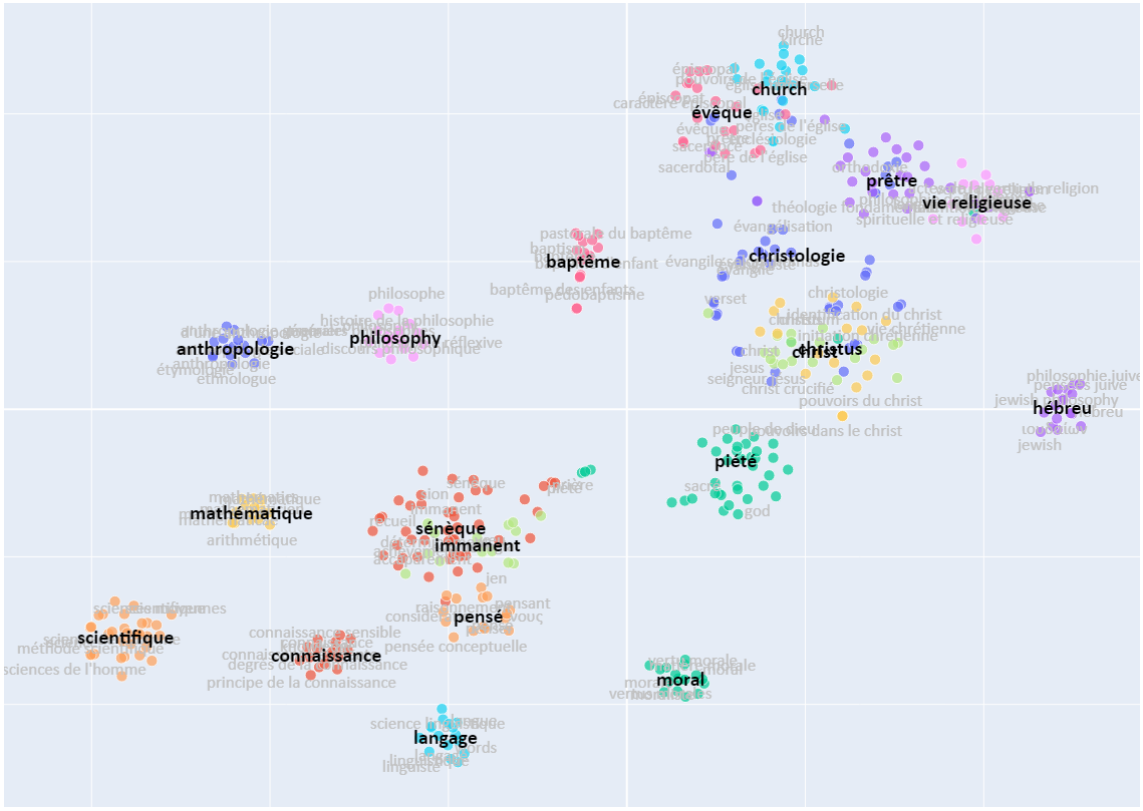

FIGURE 6 – Regroupement automatique des principaux termes extraits dans la revue LTP pour la période 1960-1979

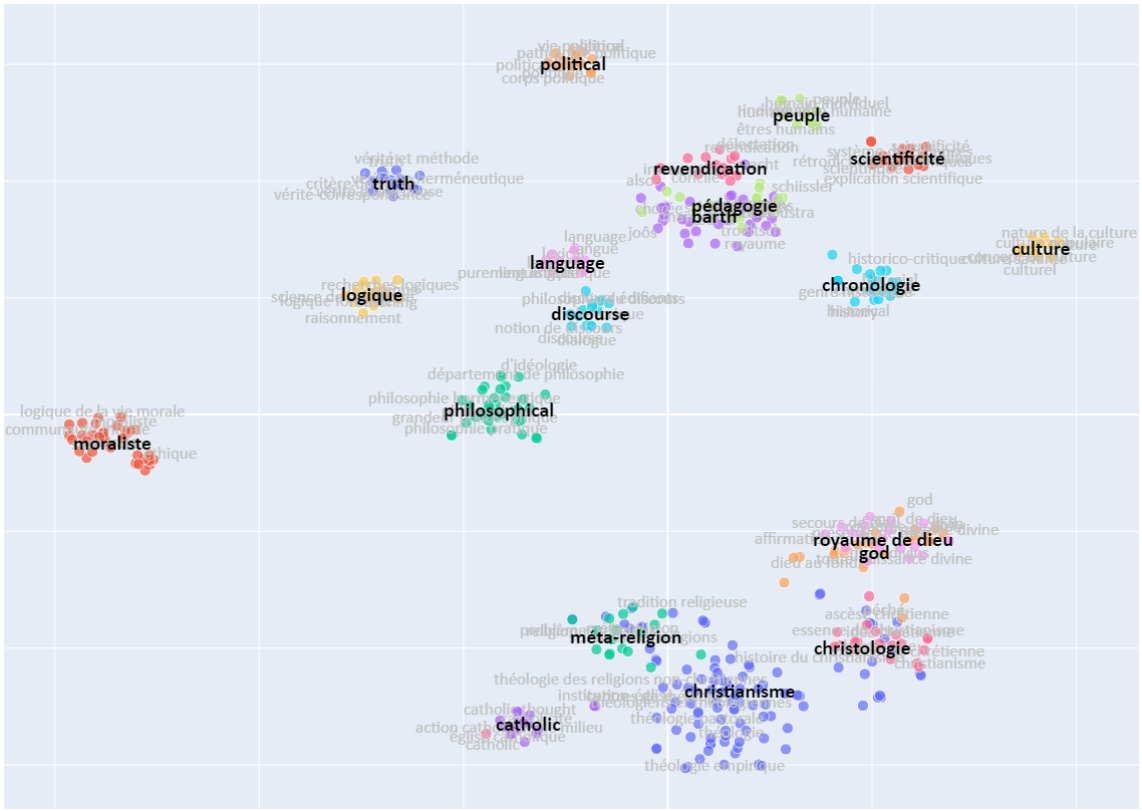

FIGURE 7 – Regroupement automatique des principaux termes extraits dans la revue LTP pour la période 1980-1999

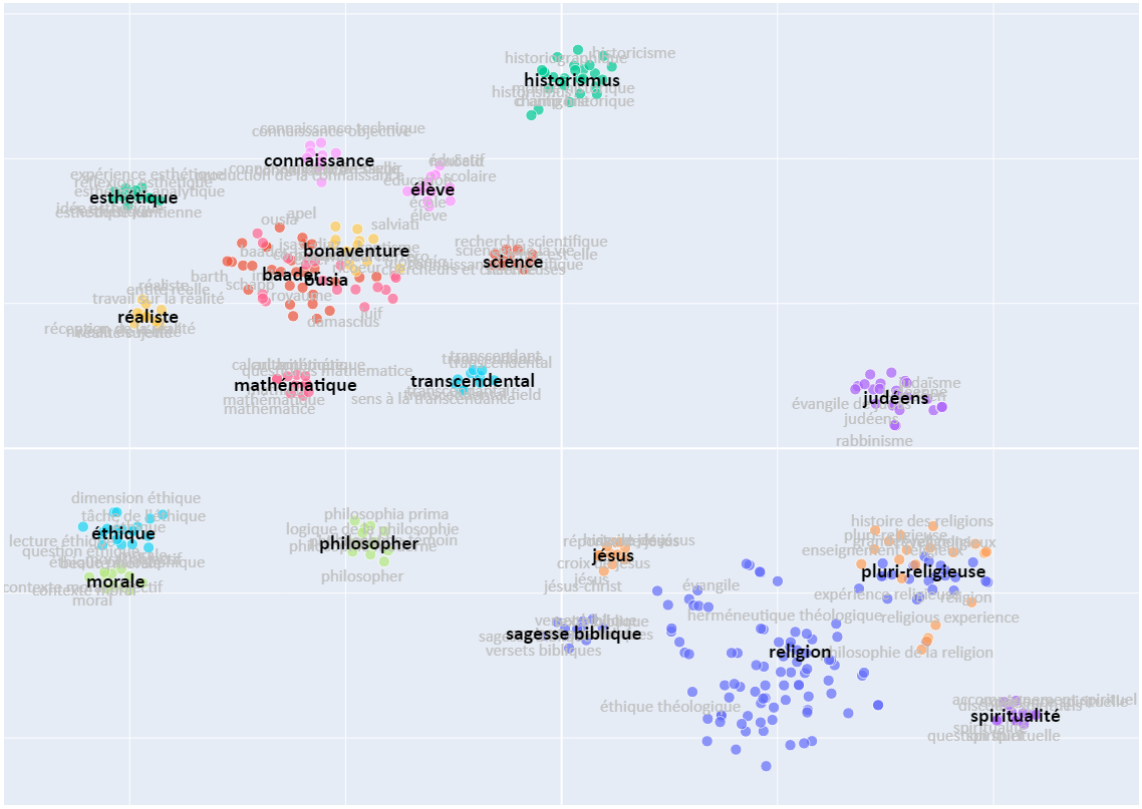

FIGURE 8 – Regroupement automatique des principaux termes extraits dans la revue LTP pour la période 2000-2023