Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2024

Université Savoie Mont Blanc

6 & 7 juin 2024

Les ouvrages TOTh précédents sont disponibles :

- sur le site des Presses universitaires Savoie Mont Blanc (https://btk.univ-smb.fr/categorie-produit/livres)
- sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr)
- ou auprès de : contact@toth.condillac.org

27, rue Marcoz
BP 1104
73011 Chambéry cedex
www.univ-smb.fr

R. Giannadakis
Réalisation : C. Roche, M. Papadopoulou, ~~C. Djambian~~, L. Vachon
Collection « Terminologica »
ISBN : 978-2-37741-107-8
ISSN : 2607-5008
Dépôt légal : décembre 2025

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2024

Université Savoie Mont Blanc

6 & 7 juin 2024

http://toth.condillac.org

avec le soutien de :

- Université de Crète (Grèce) – TALOS funded by the European Union under Horizon Europe (project TALOS-AI4SSH 101087269)
- Université Savoie Mont Blanc, École d'ingénieurs Polytech Annecy Chambéry
- Ministère de la Culture. La publication des actes est soutenue financièrement par le ministère de la Culture - Délégation générale à la Langue Française et aux Langues de France

MINISTÈRE
DE LA CULTURE
Liberté
Égalité
Fraternité

Presses universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique

Christophe Roche	University of Crete (Greece) Université Savoie Mont Blanc (France)

Comité de programme 2024

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http://toth.condillac.org/committees.

Manuel Alcántara-Plá	Universidad Autónoma de Madrid – Spain
Amparo Alcina	Universitat Jaume I – Spain
Albina Auksoriute	Institute of the Lithuanian Language – Lithuania
Bruno Bachimont	Université Technologie de Compiègne – France
Harry Bunt	Tilburg University – Netherlands
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Dardo De Vecchi	Kedge Business School – France
Valérie Delavigne	Université Paris 3 – France
Giorgio-Maria Di Nunzio	Università degli Studi di Padova – Italy
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Caroline Djambian	Université Grenoble Alpes – France
Antoine Doucet	Université de la Rochelle – France
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Cécile Frérot	Université Stendhal Grenoble 3 – France
Francesca Frontini	Institute for Computational Linguistics «A. Zampolli» Italy
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Gernot Hebenstreit	University of Graz – Austria
Frédérique Henry	ATILF – France
John Humbley	Université Paris 7 – France
Yangli Jia	University of Liaocheng – China
Olha Kanishcheva	University of Jena – Germany
Barbara Inge Karsch	BIK Terminology – USA
Hendrik Kockaert	University of Leuven – Belgium
Héba Lecocq	Université Sorbonne nouvelle – France
Hélène Ledouble	Université de Toulon – France

Avant-propos

La dix-huitième édition de la conférence TOTh s'est tenue, comme à son habitude, le premier jeudi et vendredi du mois de juin. TOTh s'est déroulée à la fois en présentiel et à distance, permettant au plus grand nombre de participer et à ceux qui ont eu l'opportunité de venir de pouvoir échanger de vive voix et de profiter de moments de convivialité dont le dîner de gala dans la ville historique de Chambéry est le point culminant. Nous continuerons à adopter ce mode hybride devenu aujourd'hui une pratique courante.

Notre collègue Antoine Doucet de l'université de La Rochelle a ouvert la Conférence TOTh 2024 avec une présentation des dernières avancées sur l'extraction automatique de termes.

Quatre-vingt-neuf personnes se sont inscrites à la conférence et trente-quatre à la première partie d'une formation dédiée à l'intelligence artificielle pour la terminologie, cette année étant dédiée aux ontologies et aux graphes de connaissances, la formation 2025 portera sur les grands modèles de langage (LLM) pour la terminologie. Plus de soixante personnes ont suivi de manière assidue les présentations. Vingt-deux pays étaient représentés : Allemagne, Belgique, Brésil, Cameroun, Canada, Chine, Espagne, États-Unis, France, Grèce, Italie, Lituanie, Pologne, Portugal, Royaume-Uni, Roumanie, Fédération de Russie, Rwanda, Slovénie, Suisse, Turquie, Ukraine.

Vingt-deux communications ont été retenues pour publication après une sélection rigoureuse par un comité de programme international issu de dix-neuf pays. La diversité des sujets abordés, tant théoriques que pratiques, illustre la richesse et le dynamisme de notre discipline.

Les actes sont publiés aux Presses universitaires Savoie Mont Blanc dans la collection *Terminologica*. Ceux des années précédentes sont accessibles à partir du site de la conférence et des Presses universitaires Savoie Mont Blanc.

Je terminerai en remerciant le ministère de la Culture, la Délégation générale à la langue française et aux langues de France, l'université de Crète (Grèce) et le projet Européen TALOS, l'université Savoie Mont Blanc et l'école Polytech Annecy-Chambéry pour leur support et leur aide financière à l'organisation de la conférence et à la publication des actes.

Christophe Roche, président du Comité scientifique

Sommaire

Posters

Conférence d'ouverture

An overview of automatic term extraction

Antoine Doucet

L3i Laboratory, La Rochelle Université, France
antoine.doucet@univ-lr.fr

1. Introduction

Automatic term extraction (ATE) is a natural language processing (NLP) task that is meant to ease the effort of manually identifying terms from domain-specific corpora by providing a list of candidate terms. As defined by ISO 1087, a term is "the designation of a defined concept in a special language by a linguistic expression". As units of knowledge in a specific field of expertise, extracted terms are not only beneficial for several terminographical tasks but also support and improve several complex downstream tasks, *e.g.*, information retrieval, machine translation, topic detection, and sentiment analysis. ATE systems, along with annotated datasets, have been studied and developed widely for decades, but recently we observed a surge in novel neural systems to address this task. This keynote talk presented an overview of recent ATE approaches, notably deep learning-based ones, with a focus on Transformer-based neural models. Full details on the opportunities and risks involved with using large language models were provided. Neural methods were also compared to the previous ATE approaches, which were mainly based on feature engineering and non-neural supervised learning algorithms. This paper provides a summary of the keynote talk at the 2024 TOTh conference.

2. Artificial Intelligence and Generative Models

Artificial intelligence (AI) is a broad and often misinterpreted term that encompasses a wide array of methods and programs aimed at enabling computer systems to perform tasks that typically require human intelligence. Within this spectrum, generative language models like GPT have demonstrated exceptional capabilities by predicting and generating coherent text based on preceding input. Such models have proven highly effective in numerous NLP tasks, including term extraction, thanks to their ability to capture linguistic nuances and context.

Nonetheless, it is essential to be aware of the main weaknesses of using generative language models and AI in general. Among these issues lies the challenge of explainability, which follows a blackbox model in which the errors (and successes) of AI models are very hard to explain, which even triggered a whole subfield of research around the notion of explainable AI. Difficulties to explain results are problematic when such models are known for producing hallucinations amidst what can otherwise appear as reasonable content. This questions the reliability of such AI models. Next, copyright is another concern since copyrighted material sent over online tools is at a clear risk of being used to feed the model. A solution is then to run models locally, although this requires facing another major issue posed by AI: considerable computing power is needed to experiment with the models, which requires significant budget and has a clear impact on the use of resources and the environment (to be added to the cost of training the models). This critical issue in fact fostered another subfield of AI, called frugal (or green) AI.

3. Methods of Automatic Term Extraction

The current landscape of ATE methods can be categorized into three primary approaches: machine learning-based methods, neural-based methods, and prompt engineering. This overview is largely based on a recent paper, which includes further details and full bibliography (Tran *et al.*, 2023).

3.1. Machine Learning Methods

Machine learning methods for ATE generally follow a structured approach. First, preprocessing tasks such as sentence segmentation, word segmentation, and part-of-speech tagging are applied to prepare the text for further analysis. Next, feature engineering is performed to extract lexical, syntactic, and semantic characteristics that enhance the learning process. Finally, classification algorithms such as Naive Bayes, Support Vector Machines (SVM), and XGBoost are employed to distinguish terms from non-terms based on these features.

3.2. Neural-Based Methods

Neural methods have transformed ATE by enabling the extraction of complex patterns from large datasets. Traditional approaches include non-contextual embeddings like Word2Vec, GloVe, and FastText, which represent words based on their statistical co-occurrence in text. More recent advancements have introduced contextual embeddings such as BERT and FLAIR, which

incorporate surrounding context to generate more accurate representations. In addition, sequence classification and token classification methods have been widely used to label entire sequences or individual tokens as terms. Span-based models further refine this approach by predicting the start and end positions of terms within a sentence, ensuring more precise extraction.

3.3. Prompt Engineering

Prompt engineering takes advantage of the capabilities of large language models (LLMs) such as GPT-3.5 and Llama2-chat. By designing effective prompts, researchers can guide these models to produce relevant outputs for term extraction. This method has been particularly useful in few-shot learning scenarios, where models generate accurate term predictions with minimal labeled data. Moreover, prompt engineering reduces the need for extensive fine-tuning, making it a cost-effective solution for term extraction in various domains (Tran *et al*., 2024).

4. Current Status and Challenges

Despite significant progress, ATE still faces several challenges. One major issue is domain variability, where different subject areas exhibit distinct terminologies, making it difficult to create a universal extraction model. Another challenge is the gradual relevance of terms, as their importance can shift depending on context. Additionally, high-quality, well-annotated corpora remain scarce, limiting the effectiveness of supervised learning approaches. Finally, explainability in term extraction models is a growing concern, as researchers seek to understand the decision-making process behind automated term identification.

5. Future Directions

Future research in ATE should aim to address these challenges. Fine-tuning large language models for domain-specific term extraction is a promising direction that could enhance accuracy across different fields. The use of prompt-based extraction with chain-of-thought reasoning could improve model performance by encouraging deeper linguistic analysis. Unsupervised methods may also play a key role in mitigating data scarcity by enabling term extraction with minimal labeled resources. Finally, increasing the transparency and interpretability of ATE models will be essential for fostering trust and usability in real-world applications.

6. Conclusion

Automatic term extraction represents a vital intersection of linguistic expertise and technological innovation. By refining existing methods and addressing current limitations, future research can further enhance the accuracy, efficiency, and applicability of ATE systems across diverse domains. Continued advancements in AI and NLP will play a crucial role in shaping the future of term extraction, ultimately contributing to improved information management and knowledge discovery.

This work has been supported by the TERMITRAD (2020-2019-8510010) project funded by the Nouvelle-Aquitaine Region, France. Use of artifical intelligence tools. Automated tools were used to improve the grammar and style of this paper.

References

Tran, H.T.H., Gonzalez-Gallardo, C.E., Delaunay, J., Doucet, A., Pollak, S., 2024. "Is prompting what term extraction needs?", in: *Text, Speech, and Dialogue: 27th International Conference*, TSD 2024, Brno, Czech Republic, September 9–13, 2024, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg. p. 17–29. doi.org/10.1007/978-3-031-70563-2_2.

Tran, H.T.H., Martinc, M., Caporusso, J., Doucet, A., Pollak, S., 2023. "The recent advances in automatic term extraction: A survey". arXiv:2301.06767

Articles

Terminologie pratique : fondements généraux, finalités et outils

Samuel Benga*, Julien-Hervé Mbappé**, Jérémie Fifen Moluh***

samuel.benga@outlook.com, julienmbappe@outlook.fr, fifenmoluh@gmail.com

Résumé. Cela fait déjà plusieurs décennies que le travail terminologique ne s'effectue plus sur du papier avec des crayons comme ça l'était certainement à l'époque où paraissait la thèse de Wüster (1898-1977), fondateur d'une théorie de la terminologie et figure de proue de l'École de Vienne (Roche, 2005). En effet, depuis les années 1990 et la prolifération des « humanités numériques » (Abiteboul & Hachez-Leroy, 2015), une avalanche d'outils de précision s'offre plus que jamais aux praticiens de la terminologie classique dont l'objet consiste à créer des produits sémantiques accessibles à des utilisateurs humains autant qu'à des « agents artificiels » (Carsenty, 2020). C'est fort de ce contexte que dans le cadre de cet article, nous avons voulu non seulement revisiter les principes majeurs sur lesquelles s'appuie cette pratique, mais également présenter quelques-uns de sa panoplie d'outils indispensables qui participent de sa chaîne de travail vers un produit final en adéquation avec ces principes majeurs.

Abstract. *For several decades now, terminology work has no longer been carried out on paper with pencils, as was certainly the case at the time when the thesis of Wüster (1898–1977), founder of a theory of terminology and leading figure of the Vienna School, was published (Roche, 2005). Indeed, since the 1990s and the proliferation of "digital humanities" (Abiteboul & Hachez-Leroy, 2015), a flood of precision tools has become available to practitioners of classical terminology, whose aim is to create semantic products accessible to both human users and "artificial agents" (Carsenty, 2020). It is against this backdrop that, in this article, we wanted not only to revisit the major principles underpinning this practice, but also to present some of the essential tools that contribute to the workflow towards a final product that complies with these major principles.*

1. Fondements d'une pratique

« Pour étudier des principes, on doit d'abord décrire leur visée, en l'occurrence ici la théorie de la terminologie », disait Budin (2007, p. 12). La clé de la réussite de chaque travail scientifique se trouvant en priorité dans le respect de ses principes, ses méthodes et ses solides bases théoriques, c'est le lieu d'indiquer les fondements sur lesquels s'appuie le travail terminologique en tant que discipline scientifique à part entière.

1.1. Travail terminologique

Tel qu'on la reconnaît aujourd'hui au-delà de ses mutations et ses conflits idéologiques, la pratique du travail terminologique repose sur des bases scientifiques solides postulées au sein du Cercle de Vienne dont le chantre est Eugen Wüster, fondateur d'une théorie de la terminologie. « À partir des années 1930, les travaux d'un ingénieur autrichien contribuent à mettre en place une réflexion scientifique sur cette pratique. Présentée en 1931, sa thèse de doctorat jette les bases de ce qui pourrait être une discipline propre à l'analyse des termes employés dans les sciences et les techniques ». (Depecker, Constitution de la terminologie en tant que discipline, s.d.).

Ingénieur de son état, Wüster est un individu qui baigne dans la technique, et ses travaux sur la terminologie, qui portent sur les langues de spécialité de « toutes les sciences fondamentales de la nature, qu'il s'agisse des sciences du vivant ou de la matière », ont inspiré de nombreux travaux ultérieurs (Candel, Samain, & Savatovsky, 2022, p. 15). Nous conviendrons toutefois que « même si nous réservons une place particulière aux travaux de Wüster, nous n'avons pas voulu réduire la terminologie à une source unique, celle de l'École de Vienne, qui elle-même s'alimente à d'autres sources et se développe parallèlement aux écoles tchèque et soviétique » (Savatovsky & Candel, 2007, p. 3). Il ressort donc que selon toute filiation wüsterienne, la pratique du travail terminologique dite « classique » s'appuie sur cinq fondements ou principes majeurs (Wüster, 1931) (Felber, 1984) (Condamines, 1994) (Candel, 2004) (Condamines, 2005) (Roche, 2005) (Van Campenhoudt, 2006) (Van Campenhoudt, 2006) (Roche, 2007) (Budin, 2007) (Depecker & Roche, 2007) (Depecker, s.d.) (Roche, 2008) (ISO 704, 2009) (Piccini, 2015) (Humbley, 2018) (ISO 704) (ISO 1087).

1.1.1. Onomasiologie et sémasiologie

La perspective onomasiologique conforte la théorie du concept et suppose que contrairement aux apparences, la terminologie n'est pas la science des termes, mais celle des concepts, le concept étant entendu comme « ensemble de caractères communs » (Roche, 2005, p. 55), ou « unité de connaissance créée par une combinaison unique de caractères » (ISO 1087, 2019). Wüster postule donc la primauté de la démarche onomasiologique. Autrement dit, dans un travail terminologique, les concepts devront toujours être étudiés ou pris en compte avant les termes. Or, de nos jours, loin d'être battue en brèche du point de vue de la pratique (Roche, 2005, p. 54), l'onomasiologie est plutôt complétée par la sémasiologie, une démarche plus linguistique intégrant la manière dont les experts s'expriment en discours sur des sources textuelles et documentaires, notamment en s'appuyant sur un traitement linguistique automatique de ces sources (corpus textuels) à des fins d'extraction de candidats termes ou de « reconnaissance d'entités nommées » (Schaeffer, Interdonato, & Roche, 2020, p. 104-112). » Longtemps invoquée comme méthode spécialement adaptée à la terminologie, et qui distinguait celle-ci de la linguistique générale à orientation plutôt sémasiologique, l'onomasiologie a fini par se trouver minorisée dans ce contexte, à la faveur de méthodes fondées sur l'exploitation de corpus textuels » (Humbley, 2018, p. 2).

Dans les pays de langue française, les critiques faites à la terminologie classique wüsterienne – certains l'ayant d'ailleurs complètement rejetée – s'attaquent à ses principes qu'elles considèrent parfois comme des idéaux sans une once de réalité, « idéaux de la monosémie stable, de l'univocité, de la démarche onomasiologique, de la précision des définitions, du terme comme *étiquette* apposée de manière immuable et quasi immanente sur la chose qu'il désigne, de la standardisation ou normalisation, bref d'une langue fabriquée de toutes pièces et contrôlée par la communauté linguistique pour façonner le monde » (Thoiron & Béjoint, 2010, p. 106).

Mais comme nous le révèle Candel (« Wüster par lui-même », 2004), la démarche sémasiologique n'est pas ignorée par Wüster lui-même qui est réaliste, mais qui tient bien compte du cheminement qui caractérise cette dernière : « *Zum Identifizieren und Fixieren eines Begriffes ist eine Benennung oder ein anderes Zeichen unentbehrlich. Geht man umgekehrt vom Zeichen für den Begriff aus, so wird der Begriff die Bedeutung des Zeichens oder dessen Sinn genannt* ». [Pour identifier ou fixer un concept, une dénomination ou un autre signe est indispensable. Si l'on part inversement du signe vers le

concept, alors le concept sera appelé la signification du signe ou son sens] (cité dans Candel, 2004, p. 21). C'est une concession bien maigre venant de Wüster et des praticiens wüsteriens, qui savent que « même si la terminologie a tout intérêt à s'approprier le signifié », le signifié n'est pas le concept, comme s'en interroge Roche : « Le fait qu'un terme puisse être mobilisé au sein de discours de façon similaire à un signe linguistique l'identifie-t-il pour autant à un tel signe réduisant du même coup le concept à un signifié ? ».

Toutefois, dans une perspective pratique, l'onomasiologie qui « hante l'inconscient culturel du terminologue » et porte en elle-même un « paradoxe » (Petit, p. 275), ne saurait prospérer en faisant chemin seule. Cela suppose que les deux approches onomasiologique et sémasiologique ont vocation à être prises en compte et à cohabiter dans un même projet de travail terminologique (Piccini, p. 319). Et de toutes les manières, quelle que soit la perspective avec laquelle l'on abordera ce travail, certains préliminaires restent intangibles, étant donné que « l'analyse linguistique de corpus constitue une des premières étapes, quasiment incontournable, de la démarche terminologique » (Roche, 2005, p. 53).

1.1.2. Principe de la délimitation des concepts

Ce principe stipule qu'en travail terminologique, les termes qui sont des unités linguistiques ne servent qu'à marquer des limitations entre concepts d'un domaine de spécialité, et l'étymologie même du mot *terme* aide à cette compréhension de la terminologie wüsterienne, lorsque l'on sait que « terme » provient du latin « *terminus* », qui signifie littéralement en français « limite », « frontière », « démarcation », etc. Les termes servent donc à indiquer la limite entre chaque concept et leur place exacte dans un système de concepts. La délimitation des concepts, certes matérialisée dans le système des concepts à l'aide de désignations monosémiques, se construit elle-même en amont par la prise en compte des relations classiques qu'entretiennent les concepts d'un domaine entre eux, et dont le métalangage nous semble parfois lui-même polysémique. Mais il faut toutefois reconnaître que la typologie habituelle des relations conceptuelles est la relation générique, la relation partitive et la relation associative.

1.1.3. Principe de la définition

Ce principe stipule que la définition ou la description du concept (et non celle du terme, qui n'est que leurs désignations) obéit à une logique qui tient compte

de leur position particulière dans un système de concepts, et des relations qu'ils peuvent y entretenir les uns avec les autres (ISO 860, 2007, p. 2). Chaque concept étant une « unité de connaissance créée par une combinaison unique de caractères » (ISO 1087, p. 3), définir ou décrire un concept consiste donc à présenter l'ensemble de ses caractéristiques essentielles, celles qui marquent sa relation d'appartenance à un genre prochain, et de l'autre sa différentiation spécifique. C'est le principe wüsterien de la définition intensionnelle et extensionnelle (ISO 704, p. 6). « La mise en avant du caractère sémantique de l'inclusion conduit, tout naturellement, à postuler qu'il s'agit de l'inclusion d'un sens d'un terme dans celui d'un autre, d'un rapport d'inclusion entre signifiés lexicaux » (Kleiber & Tamba, 1990, p. 8). C'est la démarche à adopter pour régler la question de la représentation des relations entre concepts au sein d'un domaine de connaissance (Van Campenhoudt, p. 164).

1.1.4. Principe de la biunivocité

Le terme est l'aspect le plus visible du travail terminologique, il se décline en unité(s) linguistique(s) ou sémiotique(s), et désigne un concept qui « existe ». Son aspect invisible, le concept, relève de l'« extralinguistique » (Roche, 2005, p. 53), c'est le concept. et entretient avec ce dernier une relation de « biunivocité », principe qui signifie qu'il n'y aura en fin de compte qu'un seul terme pour désigner un concept, la synonymie, tout comme l'homonymie et la polysémie, n'étant pas des options fiables en terminologique classique. « C'est une relation stable et biunivoque entre le terme et le concept » (Candel, 2004, p. 21). La biunivocité ne faisant pas bon ménage avec la synonymie « L'essentiel du travail, c'est l'élimination de la synonymie, pour arriver à un rapport plus direct entre concept et dénomination, dans l'idéal une relation de la biunivocité, la fameuse *Eineindeutigkeit* » (Humbley, 2015, p. 282-283).

L'*eineindeutigkeit* wüstérienne renvoie à une normalisation rigide qui a valu plusieurs critiques au « père fondateur » qui transparaissent dans divers travaux explicatifs de son œuvre ; « on reconnaitra bien là qu'on est chez un normalisateur, dont le but est de désambiguïser, de clarifier, de toucher du doigt et de voir s'établir, autant que faire se peut, la relation de biunivocité » (Candel, 2004, p. 20).

1.1.5. Principe de la synchronisation : immuabilité du concept

Les termes et les concepts sont étudiés de manière synchronisée. La notion de « synchronie » intervient dans la linguistique saussurienne comme une

dimension d'objet d'étude différente, opposée à une démarche « diachronique ». Ainsi, dans un travail terminologique « synchronisé », le concept est immuable, fixé dans un temps, et le terme qui le désigne est explicitement indicatif de l'état du concept et toutes ses caractéristiques essentielles à un moment donné du temps. « Le terme révèle ainsi sa nature de signe fonctionnant dans un système linguistique spécifique à une société, une culture, une époque donnée » (Piccini, Bellandi, & Abrate, 2019, p. 57). Cela impliquerait qu'en terminologie wüsterienne un concept quel qu'il soit (générique, intégrant ou pragmatique) ne saurait se percevoir dans tout son cheminement historique. « Le terme – fondamentalement monosémique – est conçu comme une étiquette associée à un concept – essentiellement unidimensionnel – d'une manière immuable et indifférente à toute variation culturelle et temporelle. En d'autres termes, les structures conceptuelles sont universelles et statiques ainsi que leurs termes et leurs référents » (Piccini, Bellandi, & Abrate, 2019, p. 55).

Or, cette immuabilité du concept est souvent battue en brèche par ce qu'il convient d'appeler la terminologie textuelle et la socioterminologie, des approches qui postulent que le concept ne saurait être une donnée immuable, il est façonné par les discours à travers le temps, et à travers la structuration de la société (Boulanger, 1992) (Gaudin, 2003) (Petit, 2012, p. 252).

Dans le même ordre d'idée, les approches pratiques modernes et les outils qu'elles suscitent admettent des expressions synonymiques et rejettent l'attitude prescriptive de la terminologie classique dans laquelle un concept ne donne lieu qu'à un seul terme. « *By being studied in the context of communicative situations, terms are no longer seen as separate items in dictionaries or part of a semi-artificial language deliberately devoid of the functions of other lexical items* » (cité dans Lavagnino, 2009, p. 242).

1.2. Renaissance de l'ontologie

En cette ère numérique, une pratique régulière du travail terminologique rapprochera automatiquement tout praticien du concept de l'ontologie. Mais d'où provient ce concept et surtout que vient-il faire dans la terminologie pratique ? Il convient de présenter ses origines philosophiques et d'observer sa mutation moderne vers l'ingénierie des connaissances.

1.2.1. L'ontologie philosophique

« Le monde de la technique a un soubassement philosophique », disait Ferry (2017), et nombreux sont ceux qui ont entendu le mot *ontologie* pour la

première fois en classe de terminale « philosophie », plus précisément dans les leçons sur le thème de la « Connaissance ». Nous ne nous doutions alors pas que l'ontologie viendrait renaître dans notre futur, en pleine société numérique. L'ontologie est présentée dans ce cadre philosophique par Aristote comme une branche de la métaphysique qui traite de la signification du mot « être ». Le philosophe grec antique postule qu'il y a « une science qui étudie l'être en tant qu'être ainsi que les attributs qui lui appartiennent de par sa nature propre. Elle ne se confond avec aucune des sciences dites particulières. En effet, aucune de celles-ci n'étudie de manière générale l'être en tant qu'être. » (Cité dans Keiji, 2008, p. 306). Dans son discours préliminaire de l'*Encyclopédie*, D'Alembert rappelait que « les êtres, tant spirituels que matériels ayant quelques propriétés générales comme l'existence, la possibilité, la durée, l'examen de ces propriétés forme d'abord une branche de la philosophie dont toutes les autres empruntent en partie leurs principes ; on la nomme l'ontologie, ou science de l'être, ou métaphysique générale » (cité dans Lalande, 1926, p. 715).

Roche réitère ce principe et ancre davantage l'ontologie dans la philosophie aristotélicienne, en d'autres termes « La définition du concept comme expression d'un ensemble de caractères communs est une démarche naturelle et bénéficie d'une longue histoire. Cette première théorie, de tradition aristotélicienne, repose sur l'*axiome des relations internes* qui considère le monde comme un ensemble de substrats supports d'attributs » (Roche, 2005, p. 55).

> Créé au moyen du grec *to on* à partir du participe présent substantivé du verbe être *einai* et des mots *logia* (théorie) et *logos* (discours), l'ontologie est la *science de l'être en tant qu'être* (Roche, 2005, p. 58).

Aussi, il nous est cher de rappeler que le mot *ontologie* ne se retrouve pas mentionné une seule fois dans des manuels qui nous ont initié dans la pratique du travail terminologique tels que Dubuc (1990), Pavel & Nolet (2001) et bien d'autres, tandis que Wüster lui-même, déjà au début des années trente, semble avoir élargi le champ du travail jusqu'à l'ontologie, lorsqu'il précise que la terminologie doit faire des emprunts à la logique et à l'ontologie. Wüster a donc lui-même lié la terminologie à l'ontologie lorsqu'il dit que :

> Ein wesentlicher Unterschied zwischen Terminologielehre und Sprachwissenschaft ist darin zu suchen, dass die Terminologielehre Anleihen bei der Logik und bei der Ontologie machen muss und sich mit einer dritten formalen Wissenschaft überschneidet, nämlich der Informationswissenschaft.

> *Il y a une différence essentielle entre la terminologie et la linguistique, c'est que la terminologie doit faire des emprunts à la logique et à l'ontologie et qu'elle se recoupe avec une troisième science formelle, à savoir la science de l'information (cité et traduit dans (Candel, 2004, p. 16).*

1.2.2. Les ontologies de l'ingénierie des connaissances

« L'ontologie, d'essence philosophique, s'est trouvée dispersée en ontologies, objets modélisés » (Depecker & Roche, 2007, p. 111). Du point de vue de l'informatique, l'ontologie est une représentation formelle des connaissances, représentation définie de manière lisible par une machine. C'est ainsi que Gruber la définit comme une description explicite d'une conceptualisation, l'expression d'une logique : « *an ontology is an explicit specification of a conceptualization…formally, an ontology is a statement of a logical theory* » (1995, p. 908).

L'ontologie dans ce cadre particulier décrit les concepts et les relations entre ces concepts. Elle est *formelle* parce que tant au niveau de la sémantique que de la syntaxe, sa compréhension se veut sans ambiguïté, et les connaissances d'un domaine de spécialité ou d'une communauté de travail qui y sont représentées peuvent être déchiffrées par des machines (Cimiano, Völker, & Buitelaar, 2010). C'est dans cette perspective que la terminologie et l'ontologie semblent poursuivre le même but.

> Les relations entre terminologie et ontologie se renforcent régulièrement, ainsi qu'en témoigne le développement des conférences TOTh, qui ont pour objectif de rassembler industriels, chercheurs, utilisateurs et formateurs dont les préoccupations relèvent de ces deux domaines et, de façon plus générale, de la langue et de l'ingénierie des connaissances (cité dans Thoiron & Béjoint, p. 112).

Roche réitère les similitudes de fond entre la terminologie et l'ontologie, en précisant que cette dernière « vise des objectifs similaires à ceux de la terminologie classique », c'est-à-dire « permettre la communication et l'échange d'information entre différentes communautés de pratique. Pour cela elle s'appuie sur une conceptualisation partagée d'un domaine sur laquelle repose la signification des termes » (Roche, 2007, p. 9).

Il ne serait pas extravagant de postuler que Wüster est une pièce charnière essentielle entre l'ontologie purement philosophique – depuis Aristote (*La Métaphysique*) jusqu'à Heidegger (*Qu'est-ce qu'une chose*) – et l'ontologie informatique fortement incarnée dans des travaux plus contemporains à l'instar

de l'« ontoterminologie » dans laquelle cette mutation de l'ontologie vers l'ingénierie des connaissances ne saurait être mieux perçue dans la perspective, en fin de compte, d'une opérationnalisation du travail terminologique.

> Nous introduisons le néologisme ontoterminologie pour désigner cette approche qui place l'ontologie au centre de la terminologie. Une approche où l'ontologie joue un rôle fondamental à double titre : pour la construction du système notionnel et pour l'opérationnalisation de la terminologie (Roche, 2007, p. 14).

2. Finalités: quel produit pour quel usage ?

Une typologie de produits et applications possibles du travail terminologique ne serait pas extravagante. Mais d'emblée, et selon l'objectif que l'on souhaite atteindre à travers la pratique (base terminologique, données liées destinées à une publication dans la toile sémantique, etc.), il va de soi que le choix des outils de pratique sera déterminant au bon accomplissement d'une mission, étant donné qu'aujourd'hui, la terminologie a plusieurs applications et peut se mouvoir dans moult domaines de connaissances, ce qui marque bien la diversité et la richesse de cette discipline (Avant-propos, 2020).

Toujours est-il que Cabré (1998) distingue trois types d'orientations finales que peut prendre le travail terminologique selon le type d'utilisateur du produit qui en découlera :

- La terminologie orientée vers le système linguistique ;
- la terminologie visant à un aménagement linguistique, et
- la terminologie dont la gestion est orientée vers la traduction.

Mais avec l'apport de l'ontologie (§ 3.1.) et notamment l'introduction de l'ontoterminologie (Roche, 2007, p. 14), le produit terminologique prend une tout autre ampleur et ne peut plus être séparé de l'ingénierie des connaissances.

> Les études et les applications de la terminologie contemporaine sont caractérisées par l'élaboration de cadres de représentation des connaissances formalisées qui visent à concilier les exigences opposées des utilisateurs humains et des agents artificiels, car ceux-ci sont souvent la double cible des Bases de Connaissances Terminologiques (Carsenty, 2020, p. 193).

Il en ressort que le produit terminologique participe au premier chef d'un « cadre de représentation des connaissances formalisées » (onomasiologie). Et surtout, c'est un produit qui se conçoit pour être utilisé non seulement par des personnes (« utilisateurs humains »), mais aussi par des machines ou des « agents artificiels ».

Le produit terminologique pourrait donc s'adresser à l'une ou l'autre de ces deux cibles ou aux deux cibles en même temps, selon le domaine de spécialité, l'objectif et l'usage que l'on souhaite en faire et qui déterminent le choix des outils de pratique susceptibles d'être mobilisés.

3. Fondements et boîte à outils

Parmi la panoplie des « humanités numériques » aujourd'hui figurent de nombreux outils qui sont mis à la disposition des praticiens dans leur effort de créer des produits terminologiques qui permettent de satisfaire les fondements de pratique évoqués plus haut. Pavel et Nolet (2001) notent d'ailleurs que :

> […] ceci est particulièrement vrai pour les terminologues travaillant dans une entreprise, un organisme gouvernemental ou un service de traduction où ils doivent créer, tenir à jour et exploiter d'importants fichiers terminologiques informatisés conçus à l'intention de nombreux utilisateurs, et tenant compte de besoins de communication clairement définis (Pavel & Nolet, 2001).

Si le choix et l'exploitation des logiciels disponibles sur le marché sont loin d'être anodines, la première question que nous nous sommes posée face aux outils de pratique que nous glanions ci et là, était de savoir si ces derniers nous permettraient de respecter les fondements et reproduire les principes sur lesquels repose le travail terminologique tel que nous l'avions appréhendé dans des manuels tels que (Dubuc, 1990), (Pavel & Nolet, 2001) ou des articles tels que (Roche, 2005) principalement.

Quels sont ces types d'outils et que permettent-ils de pouvoir faire précisément dans la chaîne du travail terminologique ? Dans cette section, nous examinons leur typologie et nous pencherons sur la pertinence des fonctions qu'ils contiennent, et notamment sur les principes de pratique qu'ils permettent de reproduire dans le processus qui vise à mettre en œuvre un produit terminologique.

3.1. Outils de collecte documentaire

En terminologie sémasiologique ou descriptive, les praticiens ont besoin de réunir une documentation ciblée et équilibrée au sein d'un domaine de connaissance, ou au sein d'une communauté de travail ; c'est cette documentation qui fait l'objet d'un traitement linguistique pour une recherche terminologique systématique. Ce deuxième cas échéant, même s'il convient d'admettre qu'un groupe de travail interne consacré au pilotage d'un travail terminologique sera toujours d'une grande utilité collaborative pour chaque

praticien qui en dispose ne serait-ce que pour ce qui est de la collecte documentaire, il existe néanmoins des outils numériques capables d'accéder directement à des banques documentaires distantes et organisées à cet effet, permettant aux praticiens d'y extraire des documents ciblés en usant de mots-clés. Cette classe de logiciels se décline généralement sous la dénomination « CorGen » (de l'anglais *Corpus Generation)* (Anthony, AntCorGen, 2021).

3.2. Outils de référencement bibliographique

En terminologie pratique, tous les termes d'un index alphabétique et toutes les « justifications textuelles » (Pavel & Nolet, 2001) qu'ils contiennent sont référencés par des contextes dans des sources documentaires dont il convient d'obtenir des versions codifiées. Les styles bibliographiques les plus répandus tels que l'APA, l'IEEE, l'ISO 690 ou le VANCOUVER qui se trouvent généralement dans l'onglet *Références* de l'outil Microsoft *Word* fournissent pour chaque source documentaire un code qui peut être réutilisé pour les besoins d'une quelconque publication numérique. Ce sont les versions codifiées de ces sources que l'on mentionne pour référer l'onomasiologie et les justifications textuelles qui sont les contenus de chaque catégorie de données présente dans la fiche terminographique.

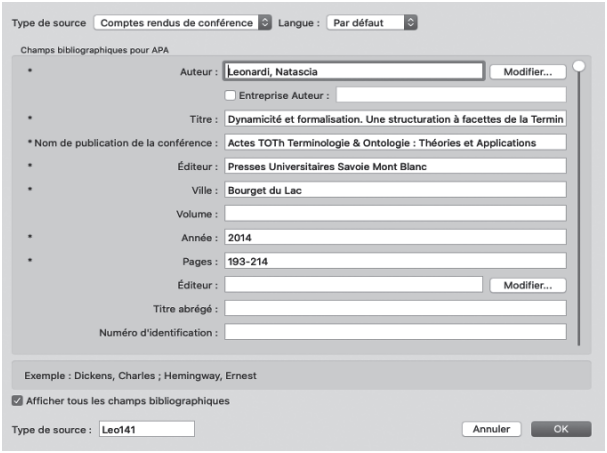


Figure 1. Indication d'une source codée utilisable au référencement bibliographique d'un concept (ex :Type de source : Leo141).

Dans une base terminologique, la source codée peut faire l'objet d'un lien hypertexte vers la source véritable.

3.3. Outils de constitution et de traitement de corpus

En travail terminologique, la mise ensemble et le traitement automatique d'un corpus textuel (corpus de travail) nécessiteront une batterie d'outils indispensables à la pratique. Parmi ces outils, une liste non exhaustive inclurait des convertisseurs de documents, des logiciels de lecture optique de caractères (OCR), des dépouilleurs automatiques, des concordanciers et des aligneurs de bitextes pour n'en citer que quelques-uns.

3.3.1. Les convertisseurs de documents

La plupart des outils numériques de traitement de corpus textuels n'admettent que des documents d'un certain format, notamment le format texte simple TXT. Dans les opérations de constitution de corpus, le rôle des convertisseurs de documents est justement de mettre ces derniers au format adéquat requis par l'application pour le traitement automatique de corpus.

3.3.2. La lecture optique de caractères (OCR)

La lecture optique de caractères est un recours fréquent des praticiens du travail terminologique dans le souci de n'avoir qu'un corpus entièrement numérique pour objet de traitement, de manière à pouvoir traiter l'ensemble du corpus de projet dans un logiciel de concordance (§ 3.1.3.3.). La lecture optique de caractères s'applique au corpus matériel fait de documents papiers qui parfois peuvent être abondants, voire dominants par rapport à l'ensemble du corpus. Certes, le traitement du corpus matériel pourrait faire l'objet d'une extraction manuelle en usant notamment du surlignage lorsque l'on a affaire à une petite quantité de documents (Pavel & Nolet, 2001, p. 39). Mais dans notre environnement où les centres de documentation et d'archives avancent vers la numérisation à pas lents, les praticiens d'Afrique subsaharienne devront encore s'attendre à recevoir une bonne quantité de documents papiers lors de leurs missions. Les lecteurs optiques de caractères procèdent par numérisation de papiers contenant des caractères langagiers. L'on obtient ensuite du texte éditable pouvant être versé dans un concordancier afin de subir un traitement automatique.

3.3.3. Les dépouilleurs automatiques

Les dépouilleurs automatiques effectuent des dépouillements-machines de textes unilingues et dressent automatiquement une liste qui contiendra

toujours beaucoup d'unités pseudo-terminologiques que l'on devra éliminer manuellement pour parvenir à une liste finale de termes réels. Les dépouillements automatiques livrent, en plus d'unités terminologiques, beaucoup de « bruit », c'est-à-dire, des découpages pseudo-terminologiques ou des éléments fortuitement regroupés dans le discours, qui ne désignent pas des concepts particuliers. Un bref examen des contextes permettra dans ce cas d'éliminer le bruit, d'écarter les termes appartenant à d'autres domaines et d'intégrer les concepts absents à une représentation conceptuelle plus complète (Pavel et Nolet, 2001, p. 43). Les pièces d'un corpus sont téléversées dans le dépouilleur automatique qui génère une feuille Excel que l'on peut traiter manuellement pour obtenir des candidats termes.

3.3.4. Les logiciels de concordance linguistique ou « concordanciers »

Lors de sa communication présentée dans le cadre de la huitième université d'automne en terminologie, le professeur Marc van Campenhoudt signalait déjà que depuis plusieurs décennies, les lexicologues et lexicographes avaient recours à l'exploitation de corpus électroniques qu'ils dépouillaient à l'aide de logiciels dénommés concordanciers, des logiciels jadis réservés aux centres de recherche. Bourigault définit un logiciel concordancier comme « un programme informatique qui recevra en entrée un corpus de textes techniques portant sur un domaine quelconque et qui devra en sortie proposer des mots ou des séquences de mots, extraits de corpus, qui pourront être retenus par un analyste humain en charge de la construction d'un produit terminologique (lui aussi quelconque) ». Les concordanciers sont des programmes informatiques qui, lorsque l'on veut bien s'en servir, remplissent un rôle primordial dans la chaine du travail terminologique, notamment en phase de traitement des supports d'information, avec pour objectif la constitution d'un corpus numérique, c'est-à-dire un corpus pouvant justement rentrer dans un concordancier.

Utilisés en symbiose avec les convertisseurs de documents, les concordanciers sont des applications complètes dotées d'un certain nombre de fonctions puissantes telles que la prise en charge des contextes définissables par l'utilisateur, l'affichage en plusieurs volets, la possibilité d'analyser statistiquement des textes sélectionnés et d'exporter les résultats de la concordance sous forme de texte (Anthony, 2005) (Folger Institute, 2019) (Berberich & Kleiber, 2023).

Dans un concordancier, une des fonctions de dépouillement les plus utiles pour la construction progressive du champ conceptuel (notamment en ce qui concerne l'identification de concepts manquants) est la fonction KWIC (*Key Word In Context*, en français littéralement « mot-clé en contexte »), qui donne

par exemple au praticien la possibilité de saisir un concept superordonné (mot-clé) et d'apercevoir toute la kyrielle de ses concepts subordonnés en contexte, observant ainsi l'horizon d'une relation hiérarchique générique.

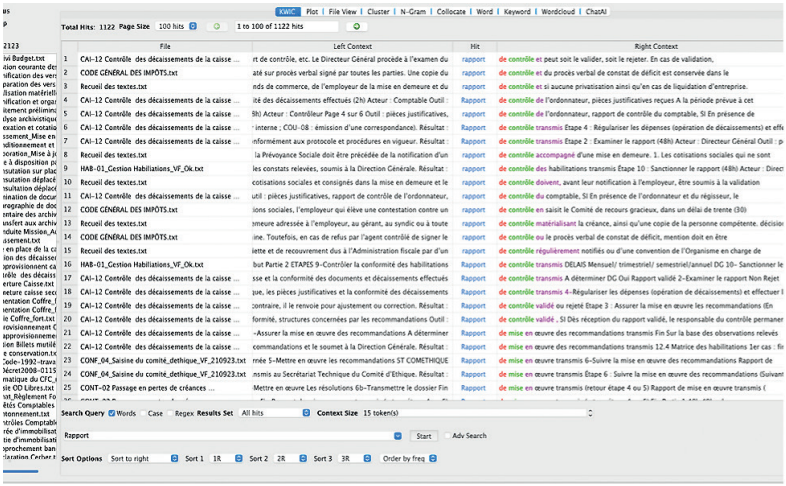


Figure 2. KWIC pour <Rapport>. Le logiciel utilisé ici pour le traitement s'appelle AntConc (Anthony, 2005).

Les concordanciers offrent une approche de fouille de texte qui donne au praticien la possibilité d'observer la sémantique d'un corpus et d'y identifier des unités terminologiques ; ils mettent notamment en évidence « la sémantique distributionnelle – c'est-à-dire l'analyse de la distribution de mots sur la base de contextes – aidant à construire la signification des termes en demeurant au plus près des textes. Les champs lexicaux de la sémantique distributionnelle apportent une aide précieuse pour la définition et l'organisation des concepts » (Roche, 2005, p. 52).

3.3.5. Les aligneurs de bitextes

Existant parfois comme modules dans certaines suites logicielles d'aide à la traduction, parfois comme solution autonome, les aligneurs de bitextes sont des outils puissants notamment en terminologie bilingue ou multilingue, car ils donnent aux praticiens la possibilité de constituer des corpus de bitextes dans une multitude de formats de fichiers, et présentent l'avantage de pouvoir extraire les concepts et leurs équivalences linguistiques en même temps.

Dans un contexte de projet de base terminologique interne d'entreprise, la constitution de vastes corpus de bitextes requiert la collaboration d'une Cellule de traduction lorsque cette dernière existe au sein de l'entreprise, les traducteurs étant ceux qui détiennent les bitextes.

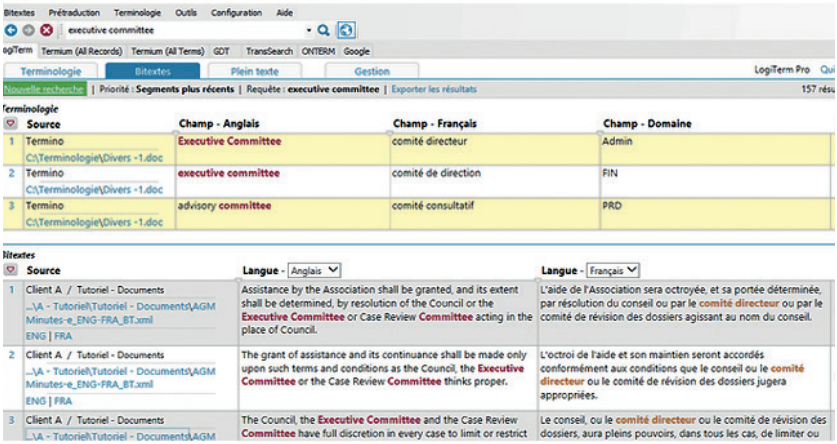


Figure 3. Aperçu d'un aligneur de bitextes.

Le dépouillement des bitextes est effectué simultanément dans les deux langues : d'abord la langue originale (ici, l'anglais) puis la traduction française. « Il est recommandé de dépouiller les textes originaux en premier (langue de départ), puis les textes traduits (langue d'arrivée) » (Pavel & Nolet, p. 39).

3.4. Outils de modélisation sémantique

La finalité de la plupart des travaux terminologiques de nos jours étant numérique en perspective, cette forme de modélisation consiste à organiser les dossiers terminologiques uninotionnels et produire des fiches terminologiques pour une publication numérique (Pavel & Nolet, p. 69). Ce travail de modélisation s'effectue sur des éditeurs d'ontologies, littéralement des environnements d'ingénierie sémantique à l'instar du *Tedi* ou du *Protégé* pour les plus populaires.

Tels qu'ils sont conçus (Musen, 2015) (Papadopoulou & Roche, 2018), les éditeurs d'ontologies permettent aux praticiens rompus aux principes de

mettre plus ou moins en application plus ou moins les fondements généraux énumérés plus haut ; ils offrent un cadre doté de moult fonctions qui favorisent la délimitation et la description de concepts et d'unités terminologiques, notamment en leur affectant systématiquement des URI (*Universal Resource Identifier*, Identificateur universel de ressources), véritables signes distinctifs qui, nous semble-t-il, donnent finalement satisfaction au grand principe de biunivocité si cher au Cercle de Vienne.

Ces outils de l'ingénierie des connaissances sont dépositaires de l'« ontoterminologie » de Christophe Roche, car ce sont eux qui permettent de pratiquer en même temps « Terminologie et Ontologie », offrant au praticien un cadre de représentation de relations conceptuelles plus ou moins sophistiqué. Du point de vue terminographique, les outils de modélisation sémantique offrent parmi leurs propriétés d'annotations des catégories de données qui sont des vocabulaires homologués pour l'édition de définitions en langue naturelle, d'observations, d'exemples d'usage et d'indication de contextes explicatifs.

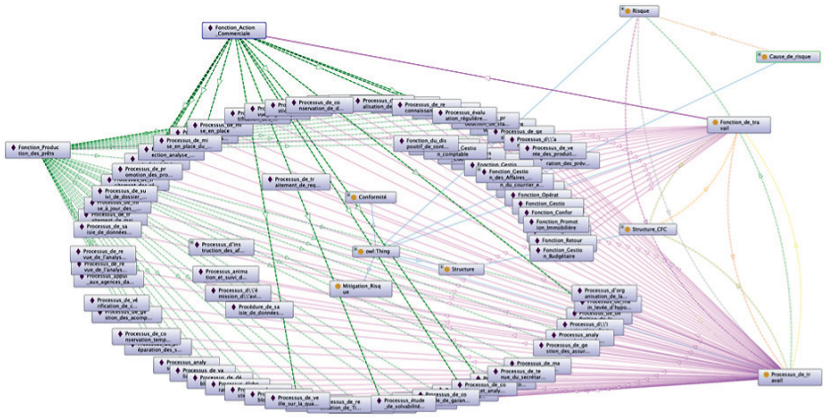


Figure 4.

Graphe de relations conceptuelles construit à partir de l'éditeur d'ontologies « Protégé » qui permet de représenter aisément toutes sortes de relations génériques, partitives et associatives, ces dernières étant souvent suggérées par inférence lorsque cette dernière est activée dans l'éditeur.

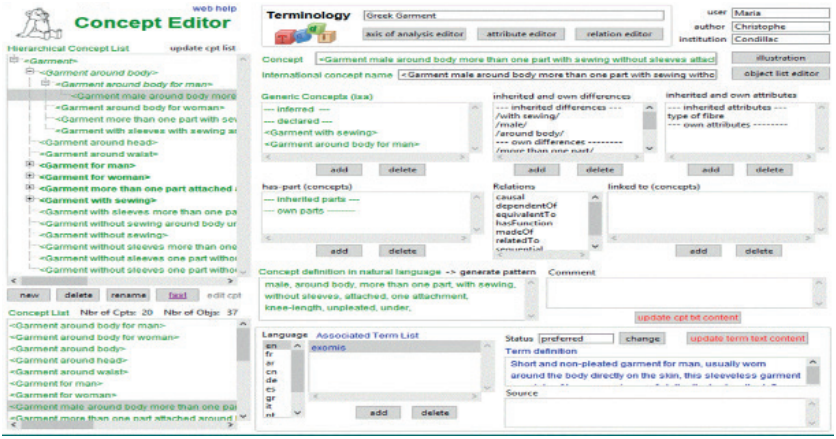


Figure 5. L'éditeur de concepts « Tedi » mis au point par le professeur Christophe Roche de l'Université Savoie Mont Blanc.

Papadopoulou & Roche, 2018.

4. Conclusion

Nous avons abordé les fondements majeurs du travail terminologique au sens de la terminologie wüsterienne et des outils de pratique indispensables pour sa production et sa mise en service à des utilisateurs humains et artificiels. Il ressort que la terminologie pratique peut être abordée avec plusieurs approches théoriques en même temps, notamment l'approche onomasiologique et l'approche sémasiologique, qui sont quasiment devenues complémentaires. Tandis que l'approche onomasiologique s'intéressera à « des notions (concepts) et aux expressions les nommant », l'approche sémasiologique mettra l'accent sur « des expressions métiers et les connaissances qu'elles désignent » (Roche, 2005, p. 52). Toutefois, la structure de représentation des concepts en désignations monosémiques reste au cœur d'une production terminologique normée. Et que dire des ontologies de l'ingénierie des connaissances et leurs apports inestimables au travail terminologique en pleine aube de ce millénaire ? Le cadre de description logique qu'elles procurent aux praticiens contemporains par le biais de leurs outils pointilleux, leur permet de satisfaire les exigences parmi les plus intransigibles de la terminologie classique. Enfin, la typologie des outils de pratique présentée dans cet article ne prétend pas à une quelconque exhaustivité, mais participe du désir de partage de nos modestes parcours de pratique faits de tribulations et d'enseignements.

Références

Abiteboul, S., & Hachez-Leroy, F. (2015). « Humanités numériques », *Bulletin de la société informatique de France*, nº 6, p. 41-57.

Anthony, L. (2005). « AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom », *International Professional Communication Conference Proceedings*, IEEE, p. 729-737.

Anthony, L. (2021). *AntCorGen*, Center for English Language Education in Science and Engineering, School of Science and Engineering, Waseda University, Tokyo. En ligne : <https://www.laurenceanthony.net/software/antcorgen/releases/AntCorGen120/help.pdf>

Berberich, K., & Kleiber, I. (dir.). (2023). *Tools for Corpus Linguistics*. En ligne : <https://corpus-analysis.com/>.

Boulanger, J.-C. (1992). « Compte-rendu de SAGER, Juan C. (1990) : *A Practical Course in Terminology Processing* », *Meta,* vol. 37, nº 3, p. 391-573.

Budin, G. (2007). « L'apport de la philosophie autrichienne au développement de la théorie de la terminologie : ontologie, théories de la connaissance et de l'objet », *Langages,* vol. 4, nº 168, p. 11-23.

Cabré, M. T., & Nazar, R. (2011). « Terminus: A Workstation for terminology and corpus management », *Actes de la cinquième conférence TOTh*. Annecy : Institut Porphyre, Savoir et Connaissance.

Cabré, T. M. (1998). *Terminology: Theory, methods and applications.* Philadelphia : John Benjamins.

Candel, D. (2004). « Wüster par lui-même ». *Cahier du CIEL,* p. 15-31.

Candel, D. (2007). « Terminologie de la terminologie. Métalangage et reformulation dans l'*Introduction à la terminologie générale et à la lexicographie terminologique* d'E. Wüster », *Langages,* vol. 4, nº 168, p. 66-81.

Candel, D. (2012). « Eugen Wüster (1898-1977), terminologie et variation ». *Actes TOTh Terminologie & Ontologie : Théories et Applications*, Chambéry : Université Savoie Mont Blanc, p. 21-35.

Candel, D., Samain, D., & Savatovsky, D. (2022). *Eugen Wüster et la terminologie de l'école de Vienne.* Paris : SHESL HEL Livres.

Candel-Mora, M. A. (2017). « Criteria for the Integration of Term Banks in the Professional Translation Environment ». *SENDEBAR* , nº 28, p. 243-260.

Carsenty, S. (2020). « Première approche de la modélisation ontoterminologique de la balance des paiements ». *Actes TOTh Terminologie & Ontologie : Théories et Applications. Collection* « *Terminologica* », Chambéry : Presses Universitaires Savoie Mont Blanc, p. 81.

Cimiano, P., Völker, J., & Buitelaar, P. (2010). « Ontology Construction » , dans N. Indurkhya, F. J. Damerau, R. Herbrich, & T. Graepel (dir.), *Handbook of Natural Language Processing* (p. 577-604). Boca Raton, FL : Taylor and Francis Group, LLC.

Condamines, A. (1994). « Terminologie et représentation des connaissances ». *Didaskalia*, nº 5, p. 35-51.

Condamines, A. (2005). « Sémantique et corpus, quelles rencontres possibles ? », *Langages*, nº 157, p. 36-47.

Depecker, L. (s.d.). *Constitution de la terminologie en tant que discipline*. En ligne : <https://www.universalis.fr/encyclopedie/terminologie/2-constitution-de-la-terminologie-en-tant-que-discipline/>.

Depecker, L., & Roche, C. (2007). « Entre idée et concept : vers l'ontologie », *Langages*, nº 168, p. 106-114.

Djambian, C. (2011). « Le métier : son savoir, son parler ». *TOTh Terminologie & Ontologie : Théories et Applications*. Annecy : Institut Porphyre, Savoir et Connaissance, p. 75-90.

Dubuc, R. (1990). *Manuel pratique de Terminologie*. Conseil international de la langue française.

Ferry, L. (2017). « Heidegger : l'œuvre philosophique expliquée », *Heidegger : un cours particulier de Luc Ferry*.

Folger Institute. (2019). *Glossary of digital humanities terms*. En ligne sur Folgerpedia : <https://folgerpedia.folger.edu/Glossary_of_digital_humanities_terms>.

Gruber, T. R. (1995). « Towards principles for the design of ontologies used for knowledge sharing », *International Journal of Human-Computer Studies*, nº 43, p. 907-928.

Humbley, J. (2004). « La réception de l'œuvre d'Eugen Wüster dans les pays de langue française », dans C. Cortès, dir., *Les Cahiers du CIEL*, p. 33-51.

Humbley, J. (2015). « L'encyclopédisme du dix-huitième siècle : prémisses de la terminologie et de la langue de spécialité », *Actes TOTh : Terminologie & Ontologie : Théories et Applications*, Chambéry : Presses universitaires Savoie Mont Blanc, p. 265-290.

Humbley, J. (2018). « L'onomasiologie comme principe constituant de la néonymie diachronique », *NEOLEX*, nº 1, p. 1-18.

ISO 1087. (2019). *Travail terminologique et science de la terminologie. Vocabulaire*. Genève, Suisse : International Standards Organisation.

ISO 704. (2009). *Travail terminologique, principes et méthodes*. Genève, Suisse : International Standard Organisation.

Keiji , N. (2008). « Le problème de l'être et la question ontologique », *Laval théologique et philosophique*, vol. 2, nº 64, p. 305-325.

Kleiber, G., & Tamba, I. (1990). « L'hyponymie revisitée : inclusion et hiérarchie », *Langages*, vol. 25, nº 98, p. 7-32.

Lalande, A. (1926). *Vocabulaire technique et critique de la philosophie*. Paris : Presses universitaires de France.

Lavagnino, E. (2009). « De l'agriculture biologique aux espaces naturels : une étude des syntagmes terminologiques à l'intérieur des textes de spécialité », dans C. Roche (dir.), *Actes TOTh Terminologie & Ontologie : Théories et Applications* , Annecy : Institut Porphyre, Savoir et Connaissance, p. 235-252.

Leonardi, N. (2014). « Dynamicité et formalisation. Une structuration à facettes de la Terminologie », *Actes TOTh Terminologie & Ontologie : Théories et Applications*, Chambéry : Presses universitaires Savoie Mont Blanc, p. 193-214.

Musen, M. A. (2015). The Protégé Project: A Look Back and a Look Forward. *AI MATTERS*, vol. 1, nº 4, p. 4-12.

Papadopoulou, M., & Roche, C. (2018). « Tedi: A Platform for Ontologisation of Multilingual Terminologies for the Semantic Web: A Use Case from the Domain of Ancient Greek Cultural Heritage », *SEMAPRO The Twelfth International Conference on Advances in Semantic Processing*, IARIA.

Pavel, S., & Nolet, D. (2001). *Précis de Terminologie*. Canada : Ministère des Travaux publics et Services gouvernementaux Canada.

Petit, G. (2012). « La standardisation terminologique dans la perspective de la dénomination : un rêve adamique ? », dans C. Roche (dir.), *Actes TOTh Terminologie & Ontologie : Théories et Applications* , Chambéry : Presses Universitaires Savoie Mont Blanc, p. 245-280.

Piccini, S. (2015). « PLOTITERM : modélisation de la terminologie de Plotin en OWL », dans C. Roche (dir.), *Actes TOTh Terminologie & Ontologie : Théories et Applications*, Chambéry : Presses universitaires Savoie Mont Blanc, p. 313-343.

Piccini, S., Bellandi, A., & Abrate, M. (2019). « Diaterm : un modèle pour représenter l'évolution diachronique des terminologies dans le web sémantique », dans C. Roche (dir.), *Actes TOTh Terminologie & Ontologie : Théories et Applications*, Chambéry, Presses universitaires Savoie Mont Blanc, p. 55-68.

Roche, C. (2005). « Terminologie et Ontologie ». Langages, vol. 1, nº 157, p. 48-62.

Roche, C. (2007). « Le terme et le concept : fondements d'une ontoterminologie », dans C. Roche (dir.), *Actes TOTh :* « *Terminologie & Ontologie : Théories et Applications* » , Annecy : Institut Porphyre Savoir et Connaissance, p. 1-22.

Roche, C. (2008). « Faut-il revisiter les Principes terminologiques ? », *Actes TOTh 2008 Terminologie & Ontologie : Théories et Applications*, Annecy : Institut Porphyre, Savoir et Connaissance, p. 51-71.

Roche, C. (2020). « Avant-propos ». *Actes TOTh Terminologie & Ontologie : Théories et Applications.* Chambéry, Presses universitaires Savoie Mont Blanc.

Rossi, M. (2014). « Systèmes conceptuels et isotopies métaphoriques dans les langues de spécialité : analyse contrastive de quelques domaines en français et en italien », *Actes TOTh Terminologie & Ontologie : Théories et Applications*, Chambéry, Presses universitaires Savoie Mont Blanc, p. 47-67.

Roxin, I., Tajariol, F., Hosu, I., & Pélissier, N. (2019). « Information, Communication et Humanités numériques. Enjeux et défis pour un enrichissement épistémologique ». *Actes du 23e colloque bilatéral Franco-Roumain en sciences de l'information et de la communication.* Cluj-Napoca : Accent.

Ruimy, N., Piccini, S., & Giovannetti, E. (2012). « Les outils informatiques au service de la terminologie saussurienne ». *CMLF 2012 Congrès mondial de linguistique française*, EDP Sciences, p. 1043-1056.

Savatovsky, D., & Candel, D. (2007). « Présentation », *Langages*, vol. 4, nº 168, p. 3-10.

Schaeffer, C., Interdonato, R., & Roche, M. (2020). « Construction d'un corpus sur la problématique de la sécurité alimentaire guidée par un lexique et des approches de fouille de textes », dans C. Roche (dir.), *Actes TOTh Terminologie & Ontologie : Théories et Applications*, Chambéry : Presses universitaires Savoie Mont Blanc, p. 111-112.

Thoiron, P., & Béjoint, H. (2010). « La terminologie, une question de termes ? », *Meta*, vol. 55, nº 1, p. 105-118.

Van Campenhoudt, M. (2006). « What remains of Eugen Wüster? Terminology against (or with) Linguistics », dans D. Candel, D. Samain, & D. Savatovsky (dir.), *Actes du colloque de la Société d'histoire et d'épistémologie des sciences du langage et du laboratoire Histoire des théories linguistiques*, p. 89-107.

Van Campenhoudt, M. (1999). « Terminologie descriptive: petite initiation à l'exploitation de corpus », Huitième université d'automne en terminologie,

Université Rennes II. En ligne sur TERMISTI : <https://termisti.ulb.ac.be/archive/rennes99.pdf>.

Varga, C. (2012). « "Terminologie en nuage". La virtualisation de la recherche terminologique », *Synergies Roumanie*, nº 7, p. 205-216.

Wüster, E. (1931). « Internationale Sprachnormung in der Technik, Besonders in Der Elektrotechnik ». Berlin : VDI-Verl.

Un formulaire informatisé pour la validation de données terminologiques juridiques

Beatriz Curti-Contessoto*, Gercélia Batista de Oliveira Mendes*

*Laboratoire CeRLA, Université de Lille
bfcurti@gmail.com, gercelia.mendes@gmail.com

Résumé. Cet article explore la terminologie juridique et bioéthique liée à la fin de vie et l'équivalence diachronique dans la traduction juridique, fondée sur une thèse doctorale et une recherche postdoctorale. La thèse, soutenue en 2024, a produit un glossaire transdisciplinaire et une base de données trilingue (allemand, français, portugais brésilien), tandis que la recherche postdoctorale s'est concentrée sur l'évolution historique des termes juridiques. Les experts consultés pour valider les données terminologiques ont souvent été limités par le volume d'informations, ce qui a révélé la nécessité d'outils adaptés. Un formulaire de validation est proposé pour faciliter la collaboration entre terminologues et spécialistes du droit. Cette initiative vise à standardiser et optimiser la validation des données, tout en renforçant la précision et la cohérence des traductions juridiques.

Abstract. *This communication explores legal and bioethical terminology related to end-of-life and diachronic equivalence in legal translation, based on a doctoral thesis and postdoctoral research. The 2024 thesis produced a transdisciplinary glossary and a trilingual database (German, French, Brazilian Portuguese), while the postdoctoral research focused on the historical evolution of legal terms. Experts consulted to validate terminological data were often constrained by the volume of information, revealing the need for tailored tools. A validation form is proposed to facilitate collaboration between terminologists and legal specialists. This initiative aims to standardize and optimize data validation while enhancing the accuracy and consistency of complex legal translations.*

1. Introduction

Cet article propose un formulaire informatisé pour la validation des données terminologiques, spécifiquement conçu pour les spécialistes-valideurs impliqués dans un travail terminographique juridique. L'objectif principal est d'améliorer l'assertivité des consultations auprès des experts juridiques. Cette initiative résulte de la combinaison de deux expériences de recherche terminologique : un doctorat mené en France et un postdoctorat réalisé dans le cadre d'une collaboration entre une faculté brésilienne et un centre de recherche français.

La première expérience provient d'une thèse de doctorat intitulée « Dire la mort : terminologie juridique et bioéthique de la fin de vie en allemand, français et portugais du Brésil », soutenue en janvier 2024 à l'Université de Strasbourg. Cette recherche visait la création de deux « ressources terminologiques » (ISO 2021), à savoir un glossaire transdisciplinaire et une base de données regroupant les termes juridiques et bioéthiques dans trois langues : allemand d'Allemagne, français de France et portugais du Brésil. Le travail s'appuyait sur un corpus de textes juridiques (législatifs, jurisprudentiels et doctrinaires) publiés entre 2000 et 2020, représentant environ 1 200 000 mots, et un corpus de référence constitué de textes de presse publiés entre 2022 et 2023, comptant environ 19 000 mots. Les termes inclus dans les ressources terminographiques proviennent principalement de ce corpus, enrichi par des recherches menées dans le cadre de cette thèse.

La seconde expérience découle d'une recherche postdoctorale intitulée « L'équivalence terminologique du point de vue diachronique : le divorce, la séparation de corps et les régimes matrimoniaux brésiliens et français », menée à l'Université de São Paulo (2020-2024) et au Centre de recherche en linguistique appliquée de l'Université Lumière Lyon 2 (2022-2023), avec le soutien financier de la *FAPESP* (*Fundação de Apoio à Pesquisa do Estado de São Paulo*). L'objectif était de démontrer l'importance de l'information diachronique pour les traducteurs assermentés. Les données terminologiques en français et en portugais brésilien proviennent du *DroitCivilCorpusFR* et du *DroitCivilCorpusBR*. Ces corpus diachroniques s'articulent autour de l'institution du mariage civil, introduite en 1791 en France et en 1890 au Brésil. Les termes ont été extraits de deux corpus de travail dans chaque langue : des textes législatifs (46 textes français et 30 textes brésiliens) et des actes de mariage (102 français et 333 brésiliens). Un corpus complémentaire, constitué de dictionnaires, bases terminologiques, glossaires, ainsi que d'articles et

ouvrages de juristes sur l'histoire du mariage civil, le divorce et le droit des femmes, a été utilisé pour valider les données.

En s'appuyant sur l'exemple du traitement terminologique des termes via la plateforme de gestion de données *Heurist*, l'étude explore les types de données à valider, les critères de validation et les différents profils des spécialistes sollicités pour cette tâche. L'objectif final est de proposer un formulaire informatique destiné à faciliter la validation des données par les experts, tout en optimisant leur intégration par le terminologue dans la ressource terminologique concernée.

L'article souligne l'importance de la collaboration entre juristes et terminologues, en analysant le traitement des données sur la plateforme Heurist. Il propose un formulaire informatique pour simplifier la validation par les experts et optimiser son intégration dans les ressources terminologiques.

2. Jurilinguistique : quelques notions

L'étude du langage juridique et de ses dimensions socioculturelles a attiré l'attention de nombreux chercheurs à travers le monde. Il est largement reconnu que les langages juridiques utilisés dans les législations des différents pays et à diverses époques, présentent des caractéristiques distinctes par rapport au langage écrit quotidien (Mattila, 2006 ; Lerat, 2021). Un des aspects fondamentaux de ces langages est constitué par la terminologie juridique, qui exprime les concepts en droit et présente des spécificités de différentes natures.

La première particularité des termes juridiques réside dans l'influence du système juridique dans lequel ils sont utilisés (Mendes 2024), ce qui entraîne une diversité terminologique. Ainsi, les systèmes juridiques, qu'ils relèvent de la tradition romano-germanique ou de la *common law*, exercent une influence considérable sur la terminologie employée dans chaque pays. Ainsi, par exemple, les termes *tutelle* en français et *tutela* en portugais brésilien ne correspondent pas, malgré leur similitude morphologique et conceptuelle. En effet, en droit français, la tutelle s'applique aux mineurs non émancipés et aux majeurs protégés, tandis qu'en droit brésilien, la *tutela* concerne uniquement les mineurs. Cette différence illustre l'influence du cadre juridique national sur les concepts nommés par ces termes dans ce contexte spécialisé.

Une autre caractéristique importante du langage juridique est la précision et la clarté des textes législatifs (Mendes, 2024), reflétant toutes deux un impératif du droit. La précision est un principe fondamental du droit, notamment en vertu du principe général de clarté de la loi. La loi doit être claire

pour être compréhensible par les justiciables. Par conséquent, la terminologie juridique est censée être particulièrement soignée pour éviter toute ambiguïté. Prenons les termes « tutelle » et « curatelle » en droit français : malgré leur proximité, ces deux mesures de protection des majeurs sont soumises à des conditions d'application distinctes, détaillées de façon rigoureuse dans l'article 440 du Code civil. Une rigueur qui illustre l'importance de la précision dans la définition légale ou jurisprudentielle.

Bien que la précision et la clarté soient des aspects essentiels, notamment dans les textes législatifs, le langage juridique se caractérise également par sa grande technicité. Il s'agit d'un langage à haut degré de spécialisation, souvent incompréhensible pour les non-initiés. Certains termes, comme « sauvegarde de justice » ou « fidéicommis », sont employés presque exclusivement dans le domaine juridique. Ainsi, le fidéicommis désigne « une disposition testamentaire par laquelle le stipulant transmet un bien, ou tout ou partie de son patrimoine à un bénéficiaire apparent, en le chargeant de retransmettre ce ou ces biens à une tierce personne spécifiquement désignée dans l'acte[1] ».

D'ailleurs, dans le domaine juridique, un même terme peut avoir des significations différentes selon le contexte ou le domaine de droit concerné. Prenons l'exemple du terme « responsabilité » : en droit pénal, il fait référence à la culpabilité pour un crime, tandis qu'en droit civil, il s'agit de l'obligation de réparer un préjudice. Cette polysémie intradomaine impose une analyse minutieuse de chaque contexte d'utilisation.

En outre, la terminologie juridique évolue avec le temps. Bien que l'on puisse penser que le droit, en cherchant à garantir la sécurité juridique, impose un certain figement terminologique, il est indéniable que sa terminologie doit également s'adapter aux réalités sociales et économiques en constante évolution. Par exemple, le terme *suicide assisté* a progressivement été remplacé par « mort médicalement assistée », une évolution terminologique qui reflète les changements dans le discours et les perceptions sociales. Ce nouveau terme allège la charge morale que *suicide* peut évoquer en relation avec ce type d'assistance en fin de vie (Mendes, 2024).

Enfin, la dernière caractéristique du langage juridique que nous souhaitons souligner dans cet article concerne la persistance de termes d'origine latine (Mendes, 2024), tels que « prima facie » pour désigner une présomption. Cette utilisation confère à la terminologie juridique un caractère distinctif. Bien que

1 Dictionnaire juridique Serge Braudo en ligne, disponible sur : https://www.dictionnaire-juridique.com/definition/fideicommis.php.

ces latinismes puissent sembler obscurs pour le grand public, ils jouent un rôle crucial dans la précision et la rigueur des discours juridiques. En effet, ces termes sont souvent ancrés dans une tradition juridique millénaire, reflétant une profondeur historique et une autorité. Leur maintien dans le langage juridique souligne également l'importance de la continuité et de la stabilité des concepts juridiques à travers le temps. De plus, ces latinismes permettent d'exprimer des idées complexes de manière concise, facilitant ainsi la communication entre professionnels du droit.

En raison des particularités du langage juridique et des enjeux que sa traduction représente, une discipline scientifique dédiée à l'étude de ce langage et de ses applications a vu le jour. Cette discipline se caractérise par une diversité d'approches théoriques et méthodologiques, ce qui explique les différentes dénominations qui lui sont attribuées, telles que la linguistique juridique (Cornu, 2005), la jurilinguistique (Gémar, 1979) ou encore la juritraductologie (Monjean-Decaudin, 2022). Spécifiquement dans le domaine de la juritraductologie, le défi de la traduction des termes se double d'une difficulté supplémentaire qui est celle du besoin de mettre constamment en parallèle les systèmes juridiques impliqués. La « prise de décision terminologique » (Prieto Ramos, 2021, 278) dans ce cadre exige une mise en contexte des concepts en fonction de paramètres juridiques tel le système juridique et le sous-domaine du droit. La gestion des différents niveaux d'asymétrie dans la traduction intersystémique est également essentielle, en tenant compte des priorités communicatives. Une telle approche exige du terminologue non seulement des compétences méthodologiques spécifiques, notamment en matière de recherche et gestion de l'information, mais implique aussi très souvent une analyse juridique comparative, soulignant l'importance des compétences dans l'extraction d'informations stratégiques et juridiques (*ibid.*). La section suivante présente le sujet central de cet article : la validation des informations enregistrées dans notre base de données.

3. Rôle de l'expert dans la validation des données juridiques : deux expériences

L'introduction de cet article a précisé que notre objectif est d'examiner le rôle de l'expert dans le cadre de recherches en terminologie juridique, comme celles que nous menons. Cette réflexion s'est imposée en raison de la nécessaire convergence des expertises, entre celle des juristes et la nôtre en tant que terminologues spécialisés en droit. De plus, répondre aux questions

que se pose le terminologue requiert le recours à des ressources dépassant le domaine purement linguistique, afin d'acquérir une compréhension approfondie de la réalité sous-jacente (Roche, 2007, 5). Dans ce cadre théorique, où la terminologie est à la fois science des signes et science des réalités (*ibid*., 3), l'étape préliminaire la plus cruciale du travail terminologique est la structuration notionnelle du domaine concerné, laquelle ne peut être menée de manière satisfaisante sans cette convergence des savoirs (Mendes, 2024).

Avant de poursuivre nos réflexions, une clarification conceptuelle s'impose. Dans cet article, *expert* désigne exclusivement les juristes, dont l'expertise découle de leur formation et de leur pratique professionnelle dans le domaine du droit. Les traducteurs et interprètes assermentés ne sont pas considérés comme des experts spécifiquement dans le cadre du présent article, bien que dans le système judiciaire français ces professionnels soient effectivement considérés comme des experts judiciaires (Driesen, 2016).

Cela dit, bien que les terminologues juridiques possèdent une expertise dans le domaine du langage juridique, pour les fins de cette étude, ils sont considérés comme des experts en linguistique. En tant qu'experts en linguistique, et plus spécifiquement en terminologie, les terminologues — qu'ils aient une formation en droit ou non — ils doivent posséder des compétences spécifiques afin d'exercer leur profession de manière efficace. L'une des compétences essentielles pour les « travailleurs en terminologie », c'est de « comprendre le principe de l'orientation conceptuelle dans le cadre du travail terminologique et (d')être capables de l'appliquer » (ISO 2021, 28). Toutefois, il convient de reconnaître que cette compétence est rarement maîtrisée dans toute son étendue, particulièrement lorsque le terminologue doit jongler entre plusieurs systèmes juridiques, distincts par essence. En effet, si « le droit est exprimé de multiples façons au sein d'une même langue » (Gémar, 2011, 133), et si les concepts juridiques sont liés à un système conceptuellement unique au sein d'une société, il n'y a pas de véritable corrélation dans une autre société (L'Homme 2020). Face à ces lacunes potentielles dans la connaissance des systèmes conceptuels en jeu, la meilleure approche demeure le recours aux experts, dont le rôle dans le travail terminologique se revêt d'une importance majeure (Chiocchetti *et al*., 2015).

Traditionnellement, dans le cadre d'un travail terminologique, le spécialiste est avant tout un expert du domaine de la connaissance concerné. Son rôle ne se limite pas à la simple validation des termes, mais s'étend à plusieurs aspects clés de la gestion terminologique. Dans de nombreux travaux,

> les références, directes ou indirectes, à l'avis des experts sont fréquentes, que ce soit pour la constitution des corpus, la validation des termes candidats ou encore la structuration des terminologies. Plusieurs auteurs soulignent la pluridisciplinarité de cette tâche ainsi que son caractère collaboratif (Bourigault, 2000 ; Després, 2001 ; Hamon et Nazarenko, 2001 ; Després et Szulman, 2008). (Thoiron et Béjoint, 2010, 12)

Certains travaux mettent en avant l'importance d'une coopération étroite et complémentaire entre différents types d'experts, notamment les spécialistes du domaine concerné, les terminologues, les lexicologues et les experts en ingénierie des connaissances (Calberg-Challot *et al.*, 2007, 199). Cette approche collaborative souligne que la précision et la richesse des terminologies résultent non seulement de l'expertise technique, mais aussi de la synergie entre ces disciplines. La création et la validation des terminologies doivent ainsi être vues comme un processus dynamique, où chaque expert apporte une contribution essentielle, plutôt que comme un effort isolé.

Dans tous les domaines spécialisés, l'intervention d'un expert est cruciale pour garantir le bon déroulement et la rigueur d'un projet de terminologie. Cela est également vrai pour la terminologie juridique, où il est recommandé de solliciter un expert juridique afin d'assurer la précision et la pertinence des termes utilisés (Fähndrich, 2005). Cela soulève la question de savoir comment cet avis est formulé et pris en compte dans les travaux terminologiques.

Chaque projet terminologique poursuit un objectif spécifique, ce qui influence la procédure collaborative adoptée. En général, l'expert du domaine joue un rôle clé dans une étape cruciale du processus terminologique : la validation des données. Cette étape peut être réalisée de différentes manières, selon des méthodologies adaptées au projet. Une des approches consiste à diviser le travail de validation en trois phases :

i. une phase d'explicitation des choix de validation à partir d'un ensemble restreint de termes candidats ;
ii. une phase de validation par les terminologues ;
iii. une phase de double validation avec l'expert juridique (Szulman *et al.*, 2013, 104).

L'article présente ensuite comment cette démarche a été mise en œuvre dans les deux études scientifiques évoquées en introduction.

Dans le cadre de la thèse mentionnée, six experts (deux en médecine/bioéthique et trois juristes) ont été sollicités pour valider des données terminologiques sur la fin de vie. Deux ont décliné en raison d'un manque de temps,

tandis que quatre ont reçu un échantillon de données centré sur le terme « directives anticipées de volonté ». Parmi eux, trois ont répondu, évoquant soit des contraintes de temps face au volume de données, soit une intention de traiter les informations ultérieurement. Le quatrième n'a pas répondu.

Dans la recherche postdoctorale mentionnée, six experts ont participé à la validation des données : deux juristes enseignants-chercheurs en histoire du droit, deux avocats spécialisés en droit familial (mariage, divorce, séparation) et deux officiers de l'état civil. Ils ont reçu un échantillon incluant le terme, sa définition, la période de validité légale, les informations législatives associées et des notes diachroniques. Les retours ont mis en évidence l'impossibilité de vérifier l'ensemble des données, menant à la proposition de les synthétiser et de prioriser les questions essentielles, selon les suggestions de la responsable terminologique.

4. Formulaire informatisé : une proposition de réponse

À partir de nos expériences dans le cadre de ces deux projets de recherche en terminologie juridique, nous avons identifié trois difficultés majeures, à la fois perçues par nous et ressenties par les experts juridiques que nous avons consultés. Ces obstacles ayant eu un impact significatif sur le processus de validation des données et par conséquent les résultats obtenus au niveau du traitement de la terminologie, ils méritent une réflexion approfondie.

La recherche postdoctorale a mis en évidence trois défis majeurs liés à la validation des données terminologiques juridiques. Le premier défi réside dans l'impossibilité de valider un volume conséquent d'informations. Sur les 119 termes identifiés (56 en français et 63 en portugais brésilien) relatifs au divorce, à la séparation de corps et aux régimes matrimoniaux, chacun présentait plusieurs états sémantiques selon les périodes étudiées : 10 pour le terme *divorce*, 11 pour *séparation de corps* et 13 pour *régime matrimonial*. Cette surcharge a souligné la nécessité de prioriser les données à valider et de mettre en place des stratégies pour optimiser le temps des experts.

Le second défi concerne le besoin d'une expertise spécifique pour chaque sous-domaine juridique. Les concepts juridiques évolutifs, tels que ceux mentionnés, ont exigé l'intervention de spécialistes en histoire du droit civil capables de comprendre les nuances historiques et législatives associées à chaque état sémantique. Les experts généralistes se sont avérés insuffisants pour répondre à ces exigences.

Enfin, le troisième défi porte sur la nécessité d'inclure des experts en droit comparé. La comparaison des législations nationales est cruciale pour garantir une compréhension précise des concepts et de leurs équivalences dans d'autres systèmes juridiques. L'absence de cette expertise a révélé des lacunes dans la validation de termes nécessitant une analyse comparative internationale. Pour illustrer cette problématique à travers un exemple concret du travail terminologique trilingue susmentionné, examinons le concept de *directives anticipées de volonté*. Ce terme désigne un document juridique par lequel une personne exprime ses volontés concernant les soins médicaux à recevoir ou à refuser dans des situations où elle ne serait plus en mesure de prendre des décisions par elle-même. Un concept simple en apparence, mais qui révèle une grande complexité lorsqu'il est examiné dans plusieurs systèmes juridiques.

En Allemagne, ce document est appelé *Patientenverfügung* et est encadré par une législation claire depuis 2009. Le document doit être mis par écrit et peut couvrir des situations spécifiques dans lesquelles la personne souhaite refuser des traitements médicaux. En France, on parle de *directives anticipées de volonté*, introduites par la loi Léonetti en 2005, modifiée en 2016 pour rendre les directives contraignantes pour les médecins. Au Brésil, le terme employé est *diretivas antecipadas de vontade*, mais contrairement à l'Allemagne ou à la France, il n'existe pas de législation nationale spécifique concernant la rédaction de ce document. Toutefois, il est reconnu dans certaines pratiques médicales et dans la doctrine juridique, avec des orientations pour les médecins et les établissements de santé.

Ici, la complexité du travail terminologique trilingue réside dans la comparaison des termes et des notions juridiques sous-jacentes. Même si les trois termes semblent correspondre à des réalités similaires, leurs conditions et implications varient. Ainsi, en France, les directives anticipées peuvent être modifiées à tout moment, alors qu'en Allemagne, une rigidité plus importante s'impose une fois le document signé. Au Brésil, l'absence de cadre législatif clair peut poser des problèmes d'interprétation et d'application, surtout en ce qui concerne le caractère contraignant du document.

Un expert en droit comparé est indispensable pour garantir une équivalence terminologique précise entre langues et systèmes juridiques. Cette expertise permet d'analyser les nuances conceptuelles entre législations nationales et d'identifier similitudes et divergences. En son absence, des erreurs peuvent survenir, comme une mauvaise interprétation des *diretivas antecipadas de vontade* brésiliennes, risquant de créer des malentendus dans

des cadres médicaux ou juridiques français ou allemands où les pratiques sont plus strictes (Mendes, 2024).

Ces défis soulèvent une question centrale : comment optimiser la validation des données terminologiques juridiques ? La réponse passe par une gestion efficace des volumes de données, l'implication de spécialistes qualifiés et l'intégration de l'expertise en droit comparé. Cela nécessite de rationaliser la collaboration entre terminologues et juristes, tout en développant des méthodologies adaptées pour garantir la rigueur et la précision des informations validées.

Nous avons donc envisagé la création d'un formulaire de validation informatisé, qui permettrait de rationaliser et d'optimiser le processus de validation des données terminologiques. Cependant, avant de procéder à la mise en place de cet outil, deux étapes préliminaires ont été considérées :

- **Première étape** : Organisation préalable du système conceptuel, réalisée par le terminologue, l'expert juridique, ou de préférence les deux, en collaboration. Cette phase établit une structure claire et cohérente pour guider le processus de validation.
- **Deuxième étape** : Sélection par le terminologue des champs de données à inclure dans le formulaire de validation. Ce tri est crucial pour garantir que seules les informations pertinentes seront soumises aux experts. Il doit être basé sur la complexité des termes, leur importance législative, et leur évolution historique et comparée.

Une question importante s'est posée lors de la deuxième étape : comment sélectionner les données de manière à répondre aux exigences des recherches en terminologie juridique, tout en facilitant le travail des experts impliqués dans le processus de validation ? Pour y répondre, nous proposons un système de catégorisation des données terminologiques juridiques, structuré autour de trois niveaux de validation : vert, jaune et rouge. Cette classification permettrait de hiérarchiser les besoins en validation, optimisant ainsi le processus de vérification par les experts.

- **Niveau vert :** Ce niveau concerne les données terminologiques qui ne nécessitent aucune validation externe, car elles sont directement validées par des sources fiables (textes législatifs, jurisprudence, doctrine). Par exemple, le terme « mariage civil » en français et son équivalent « *casamento civil* » en portugais brésilien, dont l'équivalence est évidente et largement attestée, ne requièrent pas de validation supplémentaire.

- **Niveau jaune :** Ce niveau concerne les données qui présentent des ambiguïtés ou des incertitudes. Ces doutes peuvent découler de concepts mal définis dans les dictionnaires juridiques ou de désaccords entre spécialistes. Par exemple, l'évolution du concept de séparation dans le droit brésilien, suite à la simplification du divorce en 2010, a rendu la définition de la séparation conjugale incertaine, nécessitant une clarification par un expert.
- **Niveau rouge :** Ce niveau regroupe les données pour lesquelles le terminologue n'a aucune certitude, nécessitant une validation prioritaire par un expert. Un exemple est l'utilisation du terme «*divórcio*» dans un décret brésilien de 1890. Il n'était pas clair si ce terme désignait réellement une séparation de corps entre 1890 et 1916, et la validation par un expert en histoire du droit civil était indispensable pour confirmer cette interprétation cette interprétation.

Après avoir classé les données terminologiques dans la catégorie rouge, l'étape suivante consiste à créer un formulaire informatisé pour faciliter la validation par les experts. Dans ce cadre, *Google Forms* a été choisi pour sa gratuité, son accessibilité et sa capacité de personnalisation. Le formulaire proposé est divisé en deux sections principales. La première section, illustrée par la figure 1, se concentre sur l'identification de l'expert consulté, en recueillant des informations telles que le domaine d'expertise, les qualifications professionnelles et d'autres détails pertinents concernant le spécialiste chargé de valider les données terminologiques.

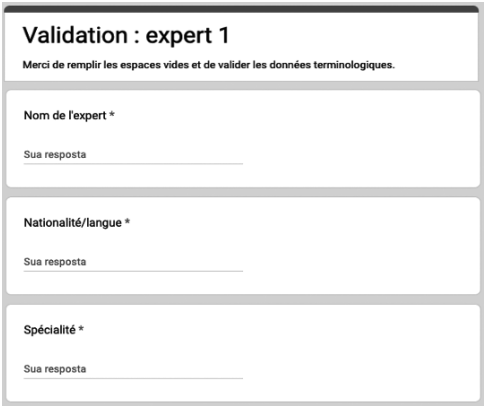


Figure 1. Première section macrostructurale du formulaire de validation.

Dans ce volet introductif, l'objectif est d'identifier le profil de l'expert, ce qui est crucial pour garantir la pertinence des évaluations fournies. Tout d'abord, le champ relatif à la nationalité et à la langue du spécialiste a pour fonction de déterminer sa capacité à évaluer non seulement les termes dans sa langue maternelle, mais aussi d'émettre des avis éclairés sur les équivalences terminologiques dans d'autres langues, le cas échéant. Cette information est capitale dans le cadre des recherches terminologiques multilingues. Ensuite, le champ relatif à la spécialité permet de préciser les sous-domaines du droit dans lesquels l'expert se sent le plus compétent. Cette information nous aide à cibler avec précision les contributions de chaque spécialiste et à répartir les tâches de validation selon leurs compétences respectives. Cela permet d'assurer que les termes les plus techniques et spécialisés sont validés par des professionnels ayant une maîtrise approfondie des concepts en question.

La deuxième section du formulaire est conçue pour traiter directement les articles terminologiques issus de la base de données. Chaque article correspond à une entrée terminologique nécessitant validation. Cette section est construite de manière à offrir à l'expert un aperçu complet des termes à valider, y compris leur définition, leur usage historique et juridique, ainsi que toute autre information importante pour faciliter une évaluation éclairée. L'expert est ainsi invité à fournir ses observations, ses corrections éventuelles ou à confirmer la validité des données sélectionnées. La figure 2 ci-après montre un autre exemple extrait de notre proposition de formulaire également relatif au terme *curatelle*, mais cette fois-ci centré sur la validation de son équivalence terminologique :

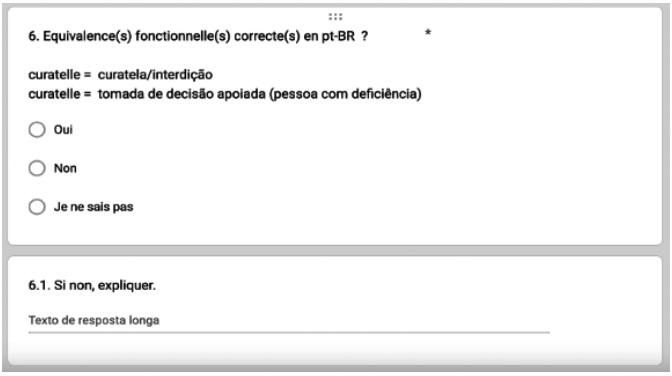


Figure 2. Seconde section macro-structurale de notre formulaire. L'équivalence faisant partie du niveau rouge.

Dans ce cas spécifique, l'équivalence proposée est classée au niveau rouge, signifiant que son exactitude est incertaine et nécessite une validation prioritaire par l'expert. Comme pour la validation de la définition, l'expert est amené à indiquer si l'équivalence est correcte[2] par un oui ou par un non. Si l'expert coche l'option négative, il/elle doit obligatoirement justifier cette réponse en fournissant une explication détaillée. Cette étape permet de mieux comprendre les divergences ou les inexactitudes relevées par l'expert, tout en apportant des pistes d'amélioration pour le choix d'équivalences plus pertinentes. Par ailleurs, nous avons intégré une troisième option : « je ne sais pas ». Cette option permet à l'expert de signaler son incertitude ou son manque de connaissances sur l'équivalence en question, sans pour autant interrompre le processus de validation. Cette flexibilité permet de minimiser les blocages tout en assurant que l'expert se prononce uniquement sur des domaines où il/elle se sent suffisamment compétent(e).

5. En guise de conclusion

Les défis rencontrés au cours de deux recherches terminologiques précédemment mentionnées ont démontré que le spécialiste dispose au moins potentiellement de la clé qui peut ouvrir la porte de la réalité sous-jacente, son domaine de connaissance, au terminologue, dont l'objet de travail est la langue. Néanmoins, la mise en œuvre d'une coopération efficace entre terminologue et spécialiste se heurte souvent aux contraintes de disponibilité de la part de ces derniers. En effet, l'analyse et la validation d'un seul terme et encore plus de toute la terminologie d'un domaine ou sous-domaine, aussi restreint soit-il, peut s'avérer une opération complexe et chronophage, surtout lorsqu'il s'agit d'un travail terminologique multilingue. Multiplication des langues, multiplication d'éléments socioculturels, multiplication des systèmes juridiques concernés : la tâche n'est pas simple.

Dans le présent travail collaboratif, l'objectif était donc de jeter les bases d'un modèle conceptuel de formulaire de consultation destiné aux spécialistes en vue de la validation de termes juridiques dans le cadre d'un travail terminographique d'élaboration d'une base de données terminologiques juridiques impliquant deux langues ou plus. Dans le but d'optimiser davantage le travail du spécialiste ainsi que celui du terminologue, il a été envisagé

2 Dans ce contexte, il est souhaitable de vérifier si cette équivalence fonctionnelle est correcte en procédant à une comparaison des états sémantiques actuels des termes concernés.

d'intégrer ce formulaire de validation directement à la base de données terminologique créée, en l'associant aux articles nécessitant une validation. Cette fonctionnalité est disponible sur la plateforme Heurist. Cela reste une perspective prometteuse pour l'évolution de notre proposition, qui pourrait être réalisée dans un avenir proche. Une telle intégration renforcerait encore la synergie entre terminologues et juristes, facilitant une prise de décision terminologique plus éclairée et garantissant une plus grande fiabilité des ressources terminologiques.

Références

Cabré, M. T. (2008). *La terminologie : théorie, méthode et applications*. Ottawa/Paris : Presses de l'Université d'Ottawa/A. Colin.

Calberg-Challot, M., Candel, D., & Roche, C. (2007). « De la variation des usages au consensus terminologique : vers un dictionnaire de l'ingénierie nucléaire ». Dans *Actes de la première conférence TOTh* (Annecy, 1er juin 2007) (p. 119-141). Annecy : Institut Porphyre.

Chiocchetti, E., Ralli, N., & Wissik, T. (2015). « The role of the domain expert in legal/administrative terminology work ». *Language and Law in Social Practice*.

Conceição, M.C. (1999). « Terminologie et transmission du savoir : (re)construction(s) de concepts ». Dans V. Delavigne & M. Bouveret (dir.), *Sémantique des termes spécialisés* (p. 33-42). Rouen : Presses Universitaires de Rouen.

Cornu, G. (2005). *Linguistique juridique* (3e éd.). Montchrestien.

Driesen, C. (2016). « L'interprétation juridique : surmonter une apparente complexité. *Revue française de linguistique appliquée* », vol. 21, no 1, p. 91-110. doi.org/10.3917/rfla.211.0091

Fähndrich, U. (2005). « Terminology project management ». *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 11, no 2, p. 225-260.

Gémar, J.-C. (2011). « Aux sources de la "jurilinguistique" : texte juridique, langues et cultures ». *Revue française de linguistique appliquée*, no 16, p. 9-16. doi.org/10.3917/rfla.161.0009

Gémar, J.-C. (dir.). (1979). *The legal translation*. Édition spéciale *Meta*, vol. 24, no 1.

Humbley, J. (2018). *La néologie terminologique*. Limoges : Lambert-Lucas.

ISO. (2021). *NF ISO 12616-1:2021. Travail terminologique appuyant la communication multilingue. Partie 1 : Principes fondamentaux de la terminographie axée sur la traduction*. ISO.

Lerat, P. (2021). « La terminologie juridique ». *International Journal for the Semiotics of Law*, no 34, p. 1173-1213.doi.org/10.1007/s11196-020-09794-7

L'Homme, M.-C. (2020). *Lexical Semantics for Terminology: An introduction*. Amsterdam : John Benjamins Publishing Company.

Mac Aodha, M. (2008). Review of J.-C. Gémar et N. Kasirer (dir.) (2005), « *Jurilinguistique : entre langues et droits. Jurilinguistics: Between Law and Language* » (Montréal, Thémis/Bruylant), dans *Meta : Journal des*

traducteurs, vol. 53, nº 3, p. 679-684. En ligne sur :<http://id.erudit.org/iderudit/019247ar>.

Mattila, H. E. S. (2006). *Comparative legal linguistics*. Ashgate Publishing Limited.

Mendes, Gercélia B. de O. (2024). « Dire la mort : terminologie juridique et bioéthique de la fin de vie en allemand, français et portugais du Brésil ». Thèse de doctorat en Sciences du langage, Université de Strasbourg.

Monjean-Decaudin, S. (2022). *Traité de juritraductologie : épistémologie et méthodologie de la traduction juridique*. Presses universitaires du Septentrion.

Møller, B. (1998). « À la recherche d'une terminochronie », *Meta*, vol. 43, nº 3, p. 426-438. doi.org/10.7202/003655ar

Prieto Ramos, F. (2021). « The use of resources for legal terminological decision-making: patterns and profile variations among institutional translators ». *Perspectives*, vol. 29, nº 2, p. 278-310.doi.org/10.1080/0907676X.2020.1803376

Roche, C. (2007a). « Le terme et le concept : Fondements d'une ontoterminologie ». Dans *TOTh 2007 : Terminologie et Ontologie : Théories et Applications* (p. 1-22).En ligne sur : <https://hal.archives-ouvertes.fr/hal-00202645>.

Szulman, S., Zargayouna, H., & Paul, E. (2013). « Retour d'expérience sur la création d'une ressource termino-ontologique (RTO) juridique ». *Proceedings 10th International Conference on Terminology and Artificial Intelligence TIA 2013*, p. 103-106.

Thoiron, P., & Béjoint, H. (2010). « La terminologie : une question de termes ? », *Meta : Journal des traducteurs*, vol. 55, nº 1, p. 106-118.

Do Terminologists Have a Role to Play Regarding Climate Change? Aligning Terminological Work with Climate Work through a Variationist Approach

Pauline Bureau

Pauline.bureau18@gmail.com

Abstract. Through this paper, our purpose is to show that terminology as a discipline has a role to play in addressing climate change and that this role can notably be carried out by studying terminological variation. After highlighting the difficulties that terminologists might encounter when studying *climate terms*, we define the domain of climate change which might be used as a framework for terminological study in the latter. Finally, we demonstrate that approaching this terminology from a variationist perspective can yield relevant results regarding the intentionality (van der Yeught, 2016) of this domain, which we define as encompassing three broad objectives: the reduction in gaps in climate knowledge, the diffusion of that knowledge, and climate action. We notably rely on tools from textual terminology for the extraction and analysis of key climate terms and show how data on terminological variation can help to apprehend these three objectives.

Résumé. *À travers cet article, notre objectif est de montrer que la terminologie en tant que discipline a un rôle à jouer dans la lutte contre le changement climatique et que ce rôle peut notamment être exercé par l'étude de la variation terminologique. Après avoir souligné les difficultés auxquelles les terminologues peuvent être confrontés lorsqu'ils étudient les termes liés au climat, nous définissons le domaine du changement climatique de façon à offrir un cadre d'analyse pour une étude terminologique de ce dernier. Enfin, nous démontrons qu'une approche variationniste de cette terminologie peut produire des résultats pertinents concernant l'intentionnalité (van der Yeught, 2016) de ce domaine, que nous définissons comme englobant trois grands objectifs : la réduction des lacunes dans les connaissances sur le climat, la diffusion de ces connaissances et l'action climatique. En nous appuyant sur des outils de la terminologie textuelle pour*

l'extraction et l'analyse des termes clés du domaine, nous montrons comment les données sur la variation terminologique peuvent aider à appréhender ces trois objectifs.

1. Introduction

As a global phenomenon concerning society at large, climate change has become a transdisciplinary issue, being studied across a variety of disciplines which have dedicated parts of their work to better understanding and addressing it. Through this paper, our purpose is to show that terminology as a discipline has also a role to play in addressing climate change and that this role can notably be carried out by studying terminological variation across various types of discourses on this topic. As is common in terminology studies, we begin by defining the domain under investigation – climate change –, which is addressed in the first section. We demonstrate that, as a complex socio-environmental issue, climate change constitutes a non-prototypical domain, which then leads us to introduce a theoretical and methodological framework tailored to studying this issue (second section). Finally, we apply this framework through two case studies to reinforce our argument.

1.1. What is climate work? Defining the non-prototypical domain of climate change through its purpose.

1.1.1. On the difficulties of outlining a *domain of climate change*

Climate change is a societal issue that involves a great variety of actors, sectors and disciplines. As such, what would be the *specialised domain of climate change* cannot be defined in terms of strict boundaries (Bordet, 2013). These "fuzzy" boundaries (*ibid*.) thus make it difficult to study the terminology of climate change and even to talk of a "terminology of climate change" at all, since terms are identified in relation to a domain which should be clearly outlined (L'Homme, 2004). Thus, probably as a result of this difficulty and despite analysis of specific units or genres in the literature (Cabezas-Garcia & León-Araúz, 2022; Biros *et al.*, 2018, 2020; Koteyko *et al.*, 2010) as well as dictionaries dedicated to the broader field of the environment that integrates terms pertaining to climate change[1], studies explicitly focusing on the terminology of this contemporary phenomenon remain scarce.

1 See for example the DicoEnviro and the EcoLexion respectively developed by the Observatoire Linguistique Sens-Texte at the University of Montréal and the

At the same time, there is evidence of a form of linguistic specialisation around the topic of climate change, a major one being that there are glossaries dedicated to the latter[2]. Drawing on this evidence as a first form of justification for studying what would be the *terminology of climate change*, we dedicate the next section to defining the eponymous domain.

1.2. Defining the domain through its underlying purpose

Since the field of climate change tends to resist the prototypical definition of a domain, whereby the latter refers to a sector or a field of activity which occupies a recognisable place in a given society (Petit, 2010, § 20), we define it through what Van der Yeught (2016) refers to as the *intentionality* of the agents involved in it. In other words, rather than being defined by its actors or the activities they perform, specialised domains are "sets of knowledge and practice which transcend their originators and are harnessed to the service of one particular purpose" (*ibid*., § 40). In the case of climate change, this "particular purpose" can be defined as threefold. The first one is the reduction of gaps in knowledge on that very phenomenon, its potential consequences, and on ways to address it. This objective is best exemplified by the very existence of a group of scientific experts dedicated to climate change such as the International Panel on Climate Change (IPCC). In the more recent history of the domain, it has also been tackled by social sciences and humanities, whose more specific aim has been to account for the social consequences of climate change and for the socio-cultural factors that might constitute obstacles or adjuvants to its apprehension.

Besides, because of the societal scope of climate change, which mobilises a diversity of actors with different levels and forms of specialisation, the second purpose of the eponymous domain is the transfer of the knowledge thus produced. Being composed of experts, whose purpose is, by definition, to transfer specialised knowledge to non-specialists, the IPCC also illustrates this aspect of the domain's intentionality, along with other actors such as non-governmental organisations, think tanks, the media or vulgarisers.

Lexicon Research Group at the University of Granada

2 Those which accompany each report by the Intergovernmental Panel on Climate Change (IPCC) being one example, along with the glossary published by the international Union for Conservation of Nature (IUCN) and entitled "Terminologies Used in Climate Change" (Adhikari *et al.*, 2011).

The ultimate objective of the domain of climate change is *climate action*, which corresponds to initiatives and measures that participate in reducing the threat of climate change (Bureau, 2023, 77). These might take many forms and materialises at a great variety of scales, from the global scale with legal frameworks aimed at constraining countries to reduce emissions under a certain threshold, through that of a state or a territory deploying adaptation measures to protect its coastlines from sea-level increase, to that of individuals who might decide to reduce their meat consumption. As we shall see in the following sections, climate action might also take more subtle forms and materialise through specific discursive strategies.

These three objectives thus provide us with a reference for identifying *climate terms*, which are terminological units that either participate in the description of climate change or refer to strategies, measures or tools for addressing it. As such, this framework should allow for considering as *terms* units that originally belonged to other domains (*e.g.*, *sea-level, energy efficiency,* etc.), providing that they are used in climate discourses to apprehend the objectives of the domain. In other words, it is with this threefold purpose in mind – rather than merely because of their use in discourses from a particular sector of society – that the terminology of climate change is defined and identified. It is also in relation to this threefold purpose that one must demonstrate the role terminology can play in addressing climate change.

2. The variationist approach: addressing the transdisciplinary nature of climate terminology and the purpose of the domain

2.1. Variation is inevitable in the field of climate change

The threefold purpose that defines the field of climate change is tackled by a variety of actors and discourse communities, which may pertain to different degrees of specialisation. Besides, despite the overarching purpose of the domain, these communities do not necessarily share the same worldviews and, as such, they may not have the same ways of actually tackling climate objectives: *e.g.*, while both NGOs and policymakers take part in the annual Conference of the parties (COPs), one can imagine that the former's demands are not always in line with the agendas of policymakers. Thus, according to the type of discourse from which climate terms are extracted, their denominations and meanings might differ. Studying the translation of

global warming from English to Spanish, Cabezas-Garcia & León Araúz, 2009 have thus shown that variation is frequent around the latter, occurring at the semantic level, at that of the communicative situation, or both. Variation has also been shown to affect the conceptualisation of *climate justice* (Biros *et al.*, 2018) and *decarbonisation* (Bureau, 2023) in specialised discourses, or to occur diachronically, through the emergence of neologisms on climate change (Francoeur, 2022; Zollo, 2022; Koteyko *et al.*, 2010).

Terminological variation seems therefore inevitable, a conclusion that is in line with the post-Wüsterian literature in terminology, where terminological variation is considered a necessary – if not a beneficial – phenomenon. In socioterminology for example, documenting variation is envisioned as a way to better understand the social groups between which it occurs and improve communication between them, conferring a very pragmatic dimension to the study of variation (Delavigne & Gaudin, 2022). For a textual terminologist, variation is a form of data to be observed to account for the evolution of knowledge (Picton, 2009) or its despecialisation (Humbert-Droz, 2021). Finally, from a sociocognitive perspective, Fernández-Silva (2022) shows that variation can be used by speakers to highlight conceptual relationships between various terms, a knowledge transmission strategy increasing literacy around the terms at stake.

2.2. Participating in the domain's objectives through the study of variation: a framework

Drawing on these theoretical perspectives, we formulate two principles for the study of variation in this paper: 1) studying terminological variation can uncover information, dynamics and patterns that might be relevant to the specialised domain or situation under study and 2) the study of variation is a necessary step for improving a communicative situation. As it gathers information about both the domain and its language users, the study of variation is indeed a way to contextualise a given linguistic situation and to identify potential terminological obstacles and deficits in communication. As such, it allows for a relatively rich description of the given situation, which is crucial for making relevant suggestions to improve the latter.

Applying these two principles to the domain of climate change, we thus argue that studying terminological variation can enhance our understanding of this non-prototypical domain and the dynamics between various discursive communities, as well as identify communication obstacles. Specifically, we

will demonstrate that studying climate-term variation can provide data relevant to actualising the domain's three objectives, as outlined in the following subsections.

2.2.1. Apprehending objective 1: Reducing gaps in knowledge

Participating in objective 1 from the perspective of terminology requires making use of specific theoretical principles in the field, such as the idea that terms index knowledge. Doing so, terminologists can rely on the study of terminological diachronic variation to document the evolution of expert knowledge (Picton, 2009), and thus be able to provide experts with an overview of the types of knowledge gaps that have narrowed over time and of those that persist (see 3.3). Thus, studying diachronic variation in the domain of particle physics, Picton (2009) was able to identify various forms of evolution such as the new centrality of a particular theme at a given period or the lasting implantation of a concept in the field. Besides, drawing on a second exception of the notion of *gaps* in climate knowledge, whereby they would refer to the "poor connectivity" of the already existing knowledge on climate change (Hulme, 2018, 332), terminology can participate in objective 1 by relying on diastratic variation to identify key concepts specific to different sectors or discourses and that would be valuable to others. For instance, a study comparing key terms used in the IPCC's third Working Group reports – which focus on climate change mitigation – with those in the scientific literature on mitigation could reveal terms or concepts overlooked by IPCC experts, thereby enriching future reports.

2.2.2. Apprehending objective 2: Transferring and communicating climate knowledge

A first obvious contribution of variationist terminology to climate knowledge transfer lies in documenting terminological instability and inaccuracies, which can affect communication between experts and from experts to non-specialists. This is notably the contribution offered in Cabezas-Garcia & León Araúz (2009, 434), where the authors identify inaccuracies in some of the Spanish variants used to translate *global warming.* What is more, terminologists could rely on the study of diastratic variation between expert and terminographic discourses to pinpoint potential limitations in the latter, for instance if key semantic features that were identified in expert discourses are not mentioned in the definition of the corresponding term.

Finally, studying terminological variation might be a way to document knowledge transfer between climate discourse communities such as experts and the press, notably by comparing the terms they used over a certain period. It is the approach taken by Biros & Peynaud (2019), who compared the semantic categories identified in the nouns and verbs used by the United Nations, the press, and the Earth Negotiation Bulletins between two time periods to account for the dissemination of climate knowledge. This approach might allow for the identification of categories of concepts that would be over- or under-represented in the press in contrast with expert discourses, and as such provide guidance for a more exhaustive and balanced mediatic coverage of climate change.

2.2.3. Apprehending objective 3: Climate action

Outlining the potential contributions of terminology to climate action requires us to go beyond the referential function of terms and instead consider them as communicative units embedded in discourse (Cabré, 2001). In that context, they have the potential to be the cornerstones of communicative strategies which aim at exerting an influence on the representations of the co-speaker, and as such on the way the latter might choose to act in the world.

This propensity might be realised in the cotext of a term, which can crystallise various representations of the referent it denotes. As such, studying how the cotexts of a given term vary across different discourses might uncover these varying representations. Relying on the study of the predication strategies attributed to key terms in the field of gene editing, Nikitina (2020) thus showed that this technology can either be described as a form of progress or as a dangerous method depending on the newspapers, two representations which could respectively promote positive actions toward the latter or on the contrary lead to further regulation around its use.

Besides, terms themselves can contribute to the emergence and transfer of specific representations of their referent. This function becomes especially apparent with the introduction of neological variants of a "neutral" term, as illustrated by *climate emergency* or *climate crisis,* which are negative variants of *climate change.* This phenomenon yet prompts us to question the classification of such units as *terms,* this metalinguistic category being prototypically defined as having primarily a referential purpose[3]. Considering this defining

3 In contrast, expressive purposes – which typically motivate the use of connoted units – are "usually quite rare in terminological discourse" (Cabré 1998, 112).

feature, our approach calls upon the notion of *degrees of termhood* (Humbley, 2018, 79): rather than defining a term in the absolute, this notion accounts for the fact that not all terms correspond to the same extent to the prototype of this metalinguistic category. This idea is also in line with later developments in Cabré's work, where she acknowledged that terms are in fact socio-communicative units which as such may fulfil other purposes than a purely referential one (2001). Thus, a potential role of terminologists in that respect could be to 1) rely on diachronic and/ or diastratic variation to identify connotated climate units, drawing on the hypothesis that they tend to be neologisms and that they might be more frequent in less specialised climate discourses and 2) rely on their knowledge of what a term is to show how these units present a lesser degree of termhood than prototypical terms. Doing so, they might be able to 3) bring to the fore the connotations associated with such units and as such, highlight their propensity to constitute persuasive strategies in discourse, in service of specific actions and attitudes.

Summarising the three previous sub-sections, Figure 1 below illustrates how the study of terminological variation can align with each of the three purposes of the domain of climate change.

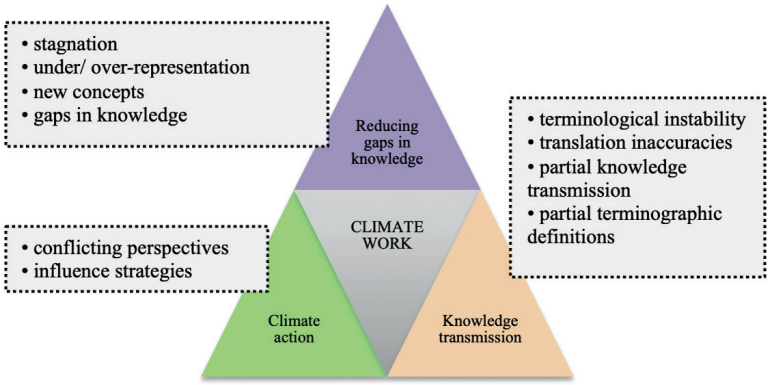


Figure 1. Examples of data that can be gathered by terminologists based on the objectives of the domain of climate change.

Each greyed rectangle presents various types of data (the lists are not exhaustive) that can be obtained by studying variation and that are relevant to the objective the rectangle is aligned with. All in all, it synthesises and highlights

data that might be an object of study for the terminologist who seeks to take part in climate work through their speciality.

3. The climate terminologist at work: two case studies

In this section, we illustrate more concretely how terminological work can align with climate work through two case studies where the study of variation brings information that can be relevant regarding one, two or the three objectives of the domain. We first account for cotextual variation around the concept of *carbon capture and storage* between climate expertise and the press, before analysing diastratic and diachronic variation between these two communities at the scale of the whole set of terms used in their discourses. The final step of each case study is to highlight the implications of the observations made in the latter regarding climate work, as defined *supra*.

3.1. Description of the corpora

The analyses were carried out on a comparable corpus[4] involving three sub-corpora representing three different discursive communities playing a key role in the domain of climate change: intergovernmental organizations[5] (IGOs), non-governmental organizations[6] (NGOs), and the British and US press[7]. The three corpora, each totalizing about 1.3 million tokens, are respectively composed of expert reports (IGOs and NGOs corpora) and articles on climate change (press corpus), published between 2007 and 2022.

4 While access to the press corpus is restricted due to copyright, the two expert corpora can be accessed at https://hdl.handle.net/11403/climate-discourses, in the "Corpus IGOs" and "Corpus NGOs" sections (Bureau *et al*., 2024).

5 UNFCCC (United Nations Framework Convention on Climate Change), IPCC (Intergovernmental Panel on Climate Change), UNDP (United Nations Development Program), UNEP (United Nations Environment Program), UN (United Nations), UNCDF (United Nations Capital Development Fund), UN-REDD (United Nations Collaborative Programme on Reducing Emissions from Deforestation and Forest Degradation in Developing Countries).

6 Greenpeace, Oxfam, Friends of the Earth, WWF (World Wildlife Fund), EDF (Environmental Defense Fund), EJF (Environmental Justice Foundation), NRDC (Natural Resources Defense Council), The Climate Group, The Climate Institute.

7 *The Guardian*, *The Telegraph* (UK), *Financial Times*, *New York Times*, and *USA Today* (US).

They can each be subdivided into three diachronic sub-corpora organised around different major Conferences of the parties (COPs): COP 15 (2009), COP 21 (2015), and COPs 25 and 26 (2019 and 2021).

3.2. Case study nº 1: *carbon capture and storage* in expert discourses

3.2.1. Methodology

Our analysis of cotextual variation around the term *carbon capture and storage* (CCS) is motivated by the distributional hypothesis, whereby the meaning of a word is reflected in its cooccurrents (Harris, 1954). Applying this principle, we compared the thirty most specific lemmatised cooccurrents of *carbon capture*[8] in the press to those extracted from the IGO sub-corpus on the one hand and the NGO sub-corpus on the other[9]. This approach was used as a first step to identify genre-specific patterns in the discursive treatment of *carbon capture*, which were then explored further through an analysis of concordance lines around the cooccurrents at stake, and of the extended cotext (the whole paragraph or article) when needed. More specifically, we relied on categories of analysis from Critical Discourse Analysis (Baker *et al.*, 2008) such as predication strategies ("What qualities are attributed to the referent across the corpora?") and prosody ("Do the collocates of the term convey specific connotations?") to decipher the meaning conveyed by the combination of a given term and its collocates in discourse.

8 We shortened our search to *carbon capture* to account for denominative variants of the full compound such as *carbon capture, carbon capture technology* and *carbon capture and storage technology.* Note that in the IPCC's glossary, the term is lexicalised as *carbon dioxide capture and storage*, a variant which we excluded from our search as it had a relative frequency of about 3 in the expert corpus (*vs* 39 for *carbon capture*) and was absent from the list of terms extracted from the press corpus.

9 We used the "cooccurrence" function in TXM (Heiden *et al*., 2010) and set the window to 10 words on each side of the node and a specificity threshold of 2. Grammatical words (prepositions, articles) and punctuation signs were deleted from the resulting lists, while numbers and units of measurement were kept. The full lists can be accessed at https://hdl.handle.net/11403/climate-discourses, in the section "Tableurs_Données_Corpora", file "COOC_Presse_OIG_ONG".

3.2.2. Results and implications regarding the purpose of the domain

This approach allowed us to identify differences in the discursive treatment of the concept of carbon capture and storage between climate expertise and the press. Thus, we identified potentially negatively connotated cooccurrents in NGOs' and the press corpora ("speculation" and "unproven", respectively), as well as units underscoring the recency of carbon capture technologies ("nascent", "new"), which are absent from the list representing IGOs. An analysis of the cotexts around those cooccurrents revealed various debate points on CCS in the press and NGOs.

First, conflicting stances have emerged in the press regarding the readiness and reliability of CCS. Some articles depict it as "unproven technologies far in the future" (*The Guardian*, 26 June 2020), stating that it "ha[sn't] been developed yet" (*Financial Times*, 5 March 2021) and "is yet to be demonstrated at full scale" (*The Telegraph*, 23 October 2010). In contrast, others argue that CCS technologies "can be brought to the market place in the near future" (*Financial Times*, 15 April 2008) and are "ready to go immediately" (*The Telegraph,* 12 September 2016).

These contradictions may reflect a relative uncertainty around CCS that transferred from expert discourse[10] to the press, but they may also signal a broader debate on the relevance of promoting CCS to address climate change. This idea is in fact apparent in the cotexts surrounding the node term *carbon capture* and its cooccurrents: in some articles, CCS is mentioned in cotexts presenting it as an opportunity, as illustrated by *The Telegraph*'s quotations of the former non-executive president of Shell Lord Oxburgh ("Carbon capture project 'can revive parts of the UK' ", "We can dramatically reduce our CO_2 emissions, create ten thousands of jobs […], by making use of technologies which are now well understood and fully proved" (12 September 2016), or *The New York Time*'s reference to the government strategy of using CCS "to solve global warming without sacrificing coal, gas or oil" (21 February 2020). Conversely in *The Guardian*, an expert presents CSS as part of a "net-zero" strategy which is "used as a smokescreen to avoid actual transition away from fossil fuels" (16 November 2021). This critical view, framing CCS as an "anti-transition" strategy, is echoed in expert reports arguing for instance that CSS "cannot substitute for a global, long-term wind down of coal, oil and gas"

10 The IPCC itself both describes CSS as an option that has been "technically proven at various scales" but whose deployment "may be limited by economic, financial, human capacity and institutional constraints" (IPCC 2018a).

(UNEP 2021, VII). Ultimately, since CCS targets CO_2-emitting fossil fuel and energy industries – an inherent aspect of its definition[11] – it raises questions about the appropriateness of deploying this technology, as it could potentially prove counterproductive.

As such, this debate raises the broader question of the energetic mix on which to rely for energy production in a context of dangerous ecological changes: should we aim toward a negative emission objective ("net zero") and thus potentially authorise fossil fuels, whose emissions would be compensated by relying on CCS? On the contrary, should we move away from fossil fuels and promote an energetic system relying completely on renewable energies? In that respect, the various stances observed in discourse appear as means to influence public opinion on these questions, and by extension, decision-making regarding CCS technologies.

Thus, starting from a term and its cooccurrents – *carbon capture* – to zoom into its use in various discursive contexts allowed us to get some insight into the representations that might emerge in public opinion around its referent. As such, working with tools from Discourse Analysis and concepts and units from terminology (the term), terminologists can have a unique perspective on a given topic – here, carbon capture technologies – and bring to the fore information that might be relevant regarding climate work. In that case, we were able to pinpoint points of debate and conflicting perspectives in public opinion as represented in discourse, which could influence the actions taken regarding carbon capture technologies.

3.3. Case study n° 2: climate terminology in the media

3.3.1. Methodology

For this case study, we relied on the study of diachronic and diastratic variation to account for the transmission of expert knowledge to a broader and less specialised audience, as represented by the British and US press. To that aim, we extracted three lists of the most specific terms in the press corpus (one for each diachronic sub-corpus) and compared it with a list of terms extracted

11 "A process in which a relatively pure stream of *carbon dioxide (CO2)* from industrial and energy-related sources is separated (captured), conditioned, compressed and transported to a storage location for long-term isolation from the *atmosphere*. Sometimes referred to as carbon capture and storage" (IPCC 2018b, 544).

from a corpus composed of the two expert corpora (IGOs and NGOs), thus representing climate expertise. We relied on the measure of the specificity score (Drouin 2003) in the software TXM (Heiden *et al.*, 2010) for this extraction procedure and used the criteria defined in Bureau (2023, 132-137) for validating the terminological status of the units thus extracted. Comparisons between the lists of terms extracted from the three press sub-corpora and that representing climate expertise were based on relative frequency, so as to avoid corpus-size effect (the expert corpus being almost six times larger than each press sub-corpus). Besides, we measured the chi-square goodness of fit to determine whether the variation observed was significant or not (Picton, 2009, 112-113), considering as significant only the cases where P = 0,01 at most (with 2 degrees of freedom). To facilitate comparisons between the different lists of terms, we categorised the most frequent units[12] based on their semantic categories and represented the latter in the form of comparable tree-diagrams[13]: as such, we could see which categories of terms were more or less represented in each diachronic press sub-corpus and in comparison with those in expert discourses[14].

3.3.2. Results and implications regarding the purpose of the domain

Comparing the three diachronic press tree diagrams with the one representing climate expertise, we first observed a relative imbalance in the categories of terms displayed in the press, with a greater representation of terms pertaining

12 We chose a threshold of 30 for the expert corpus and 15 for the press. Note that the choice of a threshold tends to be relatively subjective and thus to vary depending on the authors: Nikitina (2020) relies for instance on a threshold of 30 for 110.000 words (or about 0,03%), while in Breeze (2013), this threshold is of 0,005%. In our case, the difference between the two threshold is motivated by the significant difference in the size of the press sub-corpora compared to the expert corpus: a higher threshold for the press would not have allowed us to represent enough terms for the comparisons to be telling, while a smaller threshold in the expert corpus would have implied considering too many terms for it to be feasible in the context the study. Thus, we were able to categorise and represent 24%, 23% and 21,5% of the terms for the three press sub-corpora respectively, which is comparable with the share of terms considered in the expert corpus (23% of the total list of terms extracted was included in the corresponding tree diagram).

13 See Bureau (2023, 138-141) for a detailed description of the procedure for building the tree-diagrams.

14 The tree-diagrams can be accessed at https://hdl.handle.net/11403/climate-discourses.

to the description of climate change (left part of the diagram) than to ways of addressing it (right part of the diagram): a majority of the categories in the left part of the diagram have terms representing them while it is a minority in the right part, a pattern which is true across all three figures. This distribution is striking because no such imbalance was observed in the expert tree diagram, where both parts appeared to be relatively equally represented. While indicating that the physical consequences of climate change and its anthropogenic origin are relatively well accounted for in the press, this pattern also reveals a certain gap in the mediatic coverage of climate knowledge (and therefore in the actualisation of objective 2 of the domain), where possible measures to addressing climate change are underrepresented. Besides, considering that citizens would be more likely to take action if they are aware of possible ways of addressing climate change than by knowing the reality of this phenomenon and its consequences alone (American Psychological Association [APA], 2009, 24), this imbalance could also have consequences regarding climate action (objective 3).

A closer look at the categories represented in the right part of the tree also indicates a certain focus on renewable energies (*renewable energy, solar power, wind power, solar energy*, etc.). This pattern is more and more obvious over the period, as new terms pertaining to that category appeared around COP 21, and a statistically significant increase ($x2 = 38{,}25$) in the relative frequency of *renewable energy* can be observed between COP 15 and COP 21. By contrast, other forms of responses such as resilience and climate justice do not appear before the second tree diagram and are associated with a smaller variety of terms and lower relative frequencies in each tree: for example, in the most recent diagram the category "energy transition" is represented by 12 different terms, which amounts to a total relative frequency of 1237, while the categories "developing adaptation and resilience capacities" and "promoting equity" are both represented by only 3 different terms whose total relative frequencies respectively amount to 90 and 97.

Relying on the idea that terms encapsulate knowledge and on the possibility of tracking their circulation with tools from corpus linguistics, we have thus been able to identify a relative imbalance in the mediatic coverage of climate change. This observation may be relevant regarding objective 2 of the domain (ie. the diffusion of climate knowledge), with the tree diagrams working as tools for the press to become aware of the areas of climate knowledge which might be insufficiently covered and terminologists as mediators of such data.

4. Conclusion

Through this article, we have defined the non-prototypical domain of climate change through its underlying objectives (*i.e.* reducing gaps in knowledge, transmitting this knowledge, and taking climate action) and sought to demonstrate that the study of terminological variation around climate terms could bring to the fore data that are relevant regarding the actualisation of this very purpose. Through two case studies, we thus identified potential obstacles to knowledge transmission regarding carbon capture and storage, as well as a certain imbalance in the mediatic coverage of climate change in comparison to what is being put forward by climate experts: the press appears to focus more on the meteorological manifestations of climate change and its causes than on ways to address it. While further case studies would be necessary to illustrate the full range of data that terminologists could gather to inform climate work, we hope to have demonstrated that studying climate change terminology is possible despite the non-prototypicality of the eponymous domain, if not necessary considering the potential implications of terminological variation regarding climate objectives. Because of these implications, we believe that terminologists can participate in the global effort toward the actualisation of these objectives. They notably do so through their role of latent information gatherers, collecting data pertaining to the social apprehension of climate change using terms as an entry point, and as such potentially helping to understand the gap that exists between the analysis of climate change as a societal threat and the difficulties in taking actions for minimising this threat.

I would like to thank Camille Biros and Caroline Rossi, who supervised and mentored me during the PhD that made this article possible. I am also grateful to the reviewers for their careful rereading of this article and their kind suggestions for improvement.

References

Adhikari, Anu, Shah Racchya *et al.* 2011. *Terminologies Used in Climate Change*. Kathmandu: IUCN Nepal, 2011.

American Psychological Association. 2009. "Psychology and Global Climate Change: Addressing a Multi-Faceted Phenomenon and Set of Challenges, A Report by the Task Force on the Interface between Psychology and Global Climate Change." Washington, D.C.: APA. https://www.apa.org/science/about/publications/climate-change

Biros, Camille, Rossi Caroline and Sahakyan Inesa. 2018. "Discourse on climate and energy justice: a comparative study of Do It Yourself and Bootstrapped corpora". *Corpus* 18 (July). http://journals.openedition.org/corpus/3376

Biros, Camille, and Peynaud Caroline. 2019. "Disseminating Climate Change Knowledge. Representation of the International Panel on Climate Change in Three Types of Specialized Discourse." *Lingue e Linguaggio*, Bologna: Il Mulino, 29 (Special Issue): 179–204. doi.org/10.1285/I22390359V29P179.

Biros, Camille, Rossi, Caroline & Talbot, Aurélien. 2020. "Translating the International Panel on Climate Change Reports: Standardisation of Terminology in Synthesis Reports from 1990 to 2014". *Perspectives* 29(2), 231–44. doi.org/10.1080/0907676X.2020.1800059.

Bordet, Geneviève. 2013. "Brouillage des frontières, rencontres des domaines : quelles conséquences pour l'enseignement de la terminologie et de la traduction spécialisée". *ASp. La revue du GERAS* [Online] 64, 95–115. doi.org/10.4000/asp.3851.

Breeze, Ruth. 2013. "Lexical Bundles across Four Legal Genres". *International Journal of Corpus Linguistics* 18(2), 229–53. doi.org/10.1075/ijcl.18.2.03bre

Bureau, Pauline. 2023. "Variation terminologique et néologie dans le domaine du changement climatique". *PhD thesis in English for Specific Purposes*, Grenoble: Université Grenoble Alpes. https://theses.hal.science/tel-04353683

Bureau, Pauline, Biros, Camille, Peynaud, Caroline, *et al.* 2024. Climate discourses [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/climate-discourses.

Cabezas-García, Melania, and Pilar Araúz. 2022. "Term and Concept Variation in Climate Change Communication". The Translator 28(4): 429–31. doi.org/10.1080/13556509.2023.2182168.

Cabré Castellví, M. Teresa. 1998. "Terminology: Theory, Methods and Applications". Translated by Janet Ann DeCesaris. *Terminology and Lexicography Research and Practice 1.* Amsterdam Philadelphia: John Benjamins Publishing Company.

Cabré Castellví, M. Teresa. 2001. "Elements for a Theory of Terminology: Towards an Alternative Paradigm." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 6(1): 35–57. doi.org/10.1075/term.6.1.03cab

Delavigne, Valérie & Gaudin, François. 2022. "Founding Principles of Socioterminology". In *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge. Terminology and Lexicography Research and Practice* (23), edited by Pamela Faber & Marie-Claude L'Homme. Amsterdam: John Benjamins Publishing Company, 177–96. doi.org/10.1075/tlrp.23.08del

Fernández-Silva, Sabela. 2022. "Chapter 20. Cognitive Approaches to the Study of Term Variation". In *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge. Terminology and Lexicography Research and Practice* (23), edited by Pamela Faber & Marie-Claude L'Homme. Amsterdam: John Benjamins Publishing Company, 435–56. doi.org/10.1075/tlrp.23.20fer

Francoeur, Aline. 2022. "Entre climato-alarmistes et climato-dénégateurs : une saga néologique de notre temps". In *Neologica. Néologie et Environnement* (16), edited by Vincent Balnat and Christophe Gérard, 151–71. doi.org/10.48611/isbn.978-2-406-13219-6.p.0151

Harris, Zellig S. 1954. "Distributional Structure." *WORD* 10(2–3), 146–62. doi.org/10.1080/00437956.1954.11659520

Heiden, Serge, Magué Jean-Philippe, and Pincemin Bénédicte. 2010. "TXM : Une plateforme logicielle open-source pour la textométrie. Conception et développement." *Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data. JADT 2010* (2), 1021. Roma: Edizioni Universitarie di Lettere Economia Diritto. https://halshs.archives-ouvertes.fr/halshs-00549779.

Hulme, Mike. 2018. "'Gaps' in Climate Change Knowledge: Do They Exist? Can They Be Filled?". *Environmental Humanities* 10(1): 330–37. doi.org/10.1215/22011919-4385599.

Humbert-Droz, Julie. 2021. "Définir la déterminologisation : approche outillée en corpus comparable dans le domaine de la physique des particules". PhD thesis in Multilingual Communication Technology and Language

Sciences, University of Geneva. doi.org/10.13097/ARCHIVE-OUVERTE/UNIGE:157351

IPCC. 2018a. "Summary for Policymakers". In *Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*, edited by Valérie Masson-Delmotte *et al.* Cambridge and New York: Cambridge University Press, 3-24. doi.org/10.1017/9781009157940.001

IPCC. 2018b. "Annex I: Glossary." In *Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-Industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*, edited by Valérie Masson-Delmotte *et al.* Cambridge and New York: Cambridge University Press, 541–62. DOI: doi.org/10.1017/9781009157940

Koteyko, Nelya, Thelwall Mike & Nerlich Brigitte. 2010. "From Carbon Markets to Carbon Morality: Creative Compounds as Framing Devices in Online Discourses on Climate Change Mitigation". *Science Communication* [Online] 32(1), 25–54. doi.org/10.1177/1075547009340421.

L'Homme, Marie-Claude. 2004. "2. Le terme dans le texte spécialisé". In *La Terminologie : Principes et Techniques* [Online]. Montréal: Presses de l'Université de Montréal, 52–82. doi.org/10.4000/books.pum.10705

Nikitina, Jekaterina. 2020. "Representation of Gene-Editing in British and Italian Newspapers. A Cross-Linguistic Corpus-Assisted Discourse Study." *Lingue e Linguaggi* 34 (Special Issue), 51–75. doi.org/10.1285/I22390359V34P51

Petit, Michel. 2010. "Le discours spécialisé et le spécialisé du discours : repères pour l'analyse du discours en anglais de spécialité". *E-rea. Revue électronique d'études sur le monde anglophone* [Online] 8.1. doi.org/10.4000/erea.1400

Picton, Aurélie. 2009. "Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial". PhD thesis in linguistics, University of Toulouse le Mirail – Toulouse II.

Van der Yeught, Michel. 2016. "A Proposal to Establish Epistemological Foundations for the Study of Specialised Languages". *ASp. La revue du GERAS* [Online] (69), 41–63. doi.org/10.4000/asp.4788.

Zollo, Silvia Domenica. 2022. "Les néologismes de Glenn Albrecht face au changement écologique : entre créativité lexicale et bouleversement émo-

tionnel." In *Neologica - Néologie et Environnement* (16), edited by Vincent Balnat and Christophe Gérard, 203–21. doi.org/10.48611/ISBN.978-2-406-13219-6.P.0203.

Corpus references

The documents below are the excerpts of our press and expert corpora used as examples in the article.

Financial Times, 5 March 2021. "Bank of England's Climate Mandate Has Global Resonance".

Financial Times, 15 April 2008. "LETTERS TO THE EDITOR – Misguided case against climate change action".

Greenpeace. 2015. "Energy [r]evolution a sustainable world energy outlook". <https://www.greenpeace.de/sites/default/files/publications/energyrevolutionreport_engl_0.pdf>

The Guardian, 18 February 2014. "Comment: Let the money talk: Disinvesting in firms that don't curb emissions shows the way forward for climate campaigners".

The Guardian,26 June 2020. "Government Climate Advisers Running Scared of Change, Says Leading Scientist".

The Guardian, 16 November 2021. "'A death sentence': Indigenous climate activists denounce Cop26 deal".

The New York Times, 21 February 2020. "Here to Help; The Facts About Alcohol and Climate Change".

The Telegraph, 23 October 2010. "WILL THE LIGHTS GO OUT?".

The Telegraph, 12 September 2016. "Carbon capture project 'can revive parts of UK'".

UNEP. 2021. "The Production Gap Report: Governments' planned fossil fuel production remains dangerously out of sync with Paris Agreement limits" <https://www.unep.org/resources/report/production-gap-report-2021>

Crowdsourcing Net Rights

Harry Halpin*, Vincent Puig**

*Nym Technologies, Neuchâtel
hhalpin@ibiblio.org

**Institut de recherche et d'innovation, Paris
Vincent.Puig@iri-centrepompidou.fr

Abstract. Our study engages with the question of whether or not it is possible to crowdsource a new class of rights – net rights – using various tools for annotation, terminologies, and formal ontologies. Net rights are a possible and new class of rights brought into being by the widespread usage of the Internet. They reflect new techno-social goals around the right to Internet access and the right to privacy online. Various initiatives have been made by different countries to crowdsource net rights, and we review these efforts. Building on these documents, we explore the various debates around net rights. While there is not a clear and unified set of net rights due to open terminological and ontological questions, our efforts show that there is an emerging consensus on what may compose net rights.

Résumé. *Notre étude s'intéresse à la question de savoir s'il est possible ou non de produire de manière participative («* crowdsourced *») une nouvelle classe de droits, les droits du net, en utilisant divers outils d'annotation, de terminologie et d'ontologies formelles. Les droits du net sont une nouvelle classe de droits, engendrée par l'utilisation généralisée d'Internet, qui reflète de nouveaux objectifs techno-sociaux autour du droit d'accès à Internet et du droit à la vie privée en ligne. Diverses initiatives ont été prises par différents pays pour produire de manière participative des droits du net, et nous passons en revue ces efforts. À travers des outils d'annotation, nous explorons les différents débats relatifs aux droits du net. Même s'il n'existe pas d'ensemble unifié clair des droits du net et que diverses questions terminologiques et ontologiques restent inexplorées, nos travaux montrent qu'un consensus émerge sur les droits du net.*

1. Introduction

Given the threat to the Internet posed by centralization by both large platforms and increasingly authoritarian governments, the inventor of the Web, Tim Berners-Lee, has called for a kind of constitution for the Web to restore existing fundamental human rights and extend a new class of "net rights" to the Web. Net rights are a possible new class of rights brought into being by the widespread usage of the Internet. They reflect new techno-social realities around the right to Internet access and the right to privacy online. Key to this concept of new "net rights" on the Internet is – as put by Snowden – that fundamental rights have to be inscribed into the Internet protocols themselves: "a Magna Carta for the Internet is exactly what we need: we need to encode our values not just in writing but in the structure of the Internet" (Pitts-Brooke, 2014). Yet there is widespread terminological disagreement over whether or not there are net rights, including debate over whether they are a new class of rights at all (Cerf, 2012; Berners-Lee and Halpin, 2012).

We first present the concept of net rights for the Internet in Section 2, and then review various crowd-sourced initiatives from France, Brazil, the UK, and Italy to describe net rights in Section 3. Moving to the study itself in Section 4, we give a brief overview a set of hypertext tools used by participants in our study to annotate and discuss net rights in a number of in-person workshops. Then, we present our initial results in creating a terminology and formal ontology of net rights in Section 5, and discuss potential problems in Section 6. Finally, we conclude with a philosophical analysis of the importance of net rights and the underlying ontological reasons for the lack of coherency in the current state of net rights.

2. The Concept of Net Rights

Traditionally, the framework of rights was constructed to preserve the fundamental rights of the individual in the face of the coercive power of the nation-state, from the American "Bill of Rights" to the "Declaration of the Rights of Man and of the Citizen" in France. These declarations of rights were thought to be universal, embedded in human nature itself. After the devastation of the Holocaust, these rights were inscribed in the 1948 "Universal Declaration of Human Rights." All of these declarations of fundamental rights made twin assumptions: first, that these rights would be formulated and enforced by the nation-state, and second, that the rights would be focused on the individual human. Today, in the era of the Internet, nation-state jurisdictional boundaries

are increasingly blurred and nation-states themselves seem unable to effectively enforce these rights on the Internet. More importantly, these fundamental rights do not take into account the changing role of computational media in defining the locus of rights as not just being the individual human, but also the complex web of data that is now constitutive of being human in the era of the Internet. For example, does one's ability to interact and flourish in society depend on access to the Internet and access to your personal data, even if it is stored on the computers of a large platform such as Google? One example of this slow incorporation of data into the notion of being human is the right to "data protection" given by the General Data Protection Regulation (GDPR) in Europe, including the controversial "right to be forgotten." However, the ability of nation-states to actually enforce these rights on large online platforms has been rather limited, as shown by the failure of the "Safe Harbour" and "Privacy Shield" legislation meant to protect European citizen rights over data stored in the United States.

The debate about whether the Internet confers new rights was first provoked by the Internet shutdown during the 2011 Egyptian revolution. The United Nations released a report calling for access to the Internet as a human right, with Frank LaRue (2011) stating: "There should be as little restriction as possible to the flow of information *via* the Internet" (4). This led to a response from the co-inventor of the Internet, Vint Cerf (2012), provocatively entitled "Internet Access is Not a Human Right," in which he argued that rights such as privacy and the right to Internet access were civil rather than human rights, as human rights were established by the United Nations and must be implemented by every state regardless of technology. However, Cerf (2012) also argued that it is "the responsibility of technology creators themselves to support human and civil rights," as "it is engineers – and our professional associations and standards-setting bodies […] – that create and maintain these new capabilities" (26). However, Cerf's narrow understanding of rights as an unchanging set of rights given by the United Nations led to a response by Berners-Lee and Halpin (2012) in "Defend the Web" where they argue that only if "we start to talk about access to the capabilities of the Internet in relation to human rights" then "we can imagine that various governments might uphold them as rights" (p. 6). Basing their argument on the Extended Mind Hypothesis (A. Clark & D. Chalmers 1998), they assert that if the Net is an extension of a person's mind, then a person's rights to control their data over the Net should be equivalent to the rights given to the person outside the Net, leading to the conclusion that access to the Internet could be a right (Berners-Lee & Halpin 2012). This line

of argument by Berners-Lee led to the amorphous idea of net rights, a new class of rights adhering to humans extending their cognitive capabilities *via* the Internet. While there was initial thinking that such rights would include unfettered access to the Internet and the right to communicate privately, the exact formulation of such net rights remained unclear at their inception.

3. Initiatives for Net Rights

Although the inventor of the Web, Berners-Lee, conceived of the idea of net rights as a program for guaranteeing the rights of users of the Internet, it seemed apropos that the "crowd" of Internet users themselves should be the ultimate source of the content of net rights. Could net rights be crowdsourced? By "crowdsourcing," we mean that the general public in some sense, as opposed to only lawmakers or experts, were to be involved in the drafting of, or at least allowed to comment on, the proposed net rights via the new public agora of the Internet. The guiding idea was that rights regarding the Internet, as a universal space of what Habermas would term as "communicative rationality" via online discussion and deliberation, should come from the Internet (1984). The implicit hypothesis behind crowdsourcing net rights was that this process of commentary would lead to a more coherent and ultimately better set of net rights, and also set an example of the kind of increased democratic engagement imagined into the law itself made possible by the Internet.

After the Snowden revelations of mass surveillance on the Internet, a number of attempts to crowdsource net rights happened on the national-level, following the successfully crowdsourced *Marco Civil* of the Internet[1] in Brazil and the more traditionally created GDPR in Europe. Italy produced a Declaration of Internet Rights, published online by the Italian Chamber of Deputies in 2014.[2] The Internet Governance Forum (IGF) of the United Nations also started a process to determine if there were new rights on the Internet via forming the Internet Rights and Principles Coalition. In 2015, the Digital Committee of the French National Assembly began their own proposals for net rights, laying out a hundred recommendations, with over 4,000 comments, including the right to Internet access, the preservation of net neutrality, and the production of a commons. While it eventually found its way into French national law as Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, this particular law had several gaps; for example, it

1 https://www.cgi.br/pagina/marco-civil-law-of-the-internet-in-brazil/180.
2 https://www.camera.it/leg17/1179.

did not address the issue of mass surveillance. In Germany, journalists, philosophers (such as Jürgen Habermas), and politicians announced a digital rights charter, but the effort never succeeded in becoming a nationally-binding law.[3] An effort that was supposed to lead to a world-wide revolution in net rights, the Magna Carta Initiative was launched in 2015 under the "Web We Want" slogan by the Web Foundation, following the lead of its founder Berners-Lee. The Magna Carta for the Web languished, being handed off to Consumers International, and eventually failing to be launched as a global process for the creation of a set of unified net rights. There were efforts such as a crowdsourcing campaign aimed at young people in 2015 by the British Library to propose and debate a list of ten top net rights.[4] The final iteration was a survey of 600 people who gave a numerical ranking to a list of nine net rights in the "Contract for the Web" endorsed by Berners-Lee and the Web Foundation.[5]

4. Tools for Crowdsourcing Net Rights Terminologies

The focus of our experiments was whether or not agreement over the terminology and underlying concepts of net rights is even possible. This would involve "crowdsourcing" via a group of participants whether or not they agreed on the net rights outlined in the set of documents as described in Section 3. As in the larger scales of national and international initiatives, ordinary people would be allowed to comment on and discuss various proposed net rights. Over the course of 2017 to 2020, we ran nine workshops on net rights with 45 participants to discuss net rights at Institut de recherche et d'innovation du Centre Pompidou in Salle Triangle. In order to stimulate discussion over net rights, workshop participants could annotate key documents from Section 3 to determine a common basis that could, at least in principle, lead to a unified terminology for net rights.

We adapted the Hypothesis annotation tool for the workshops and included "meta-tags" and "action-tags" for the user to help them clarify terms. Hypothesis is an implementation of the W3C Web Annotation standard that allows anyone to add annotations and keywords to a shared set of hypertext documents. Each document pertaining to net rights was loaded into Hypothesis for annotation and discussion, as shown in Figure 1.[6]

3 https://digitalcharta.eu/.
4 https://www.bbc.com/news/technology-33132662.
5 https://contractfortheweb.org/.
6 https://hypothes.is/.

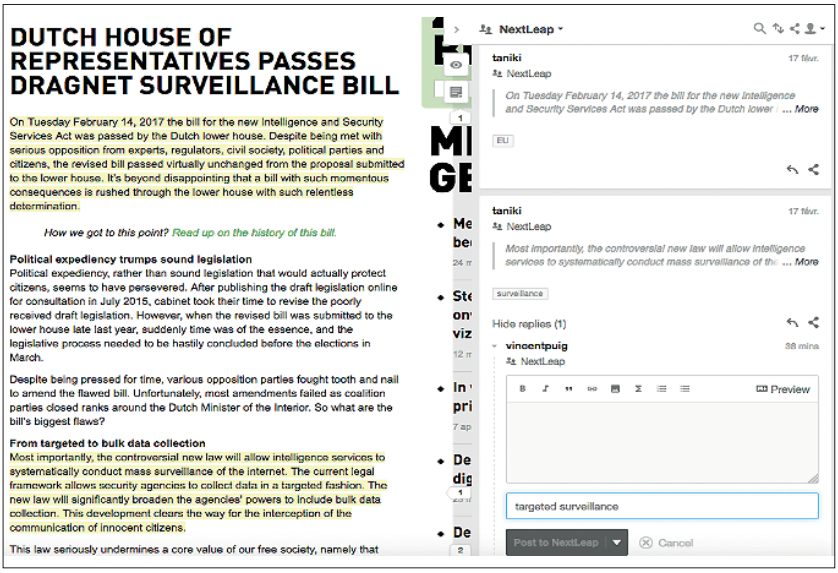


Figure 1. Hypothesis Interface for annotating documents about Net Rights.

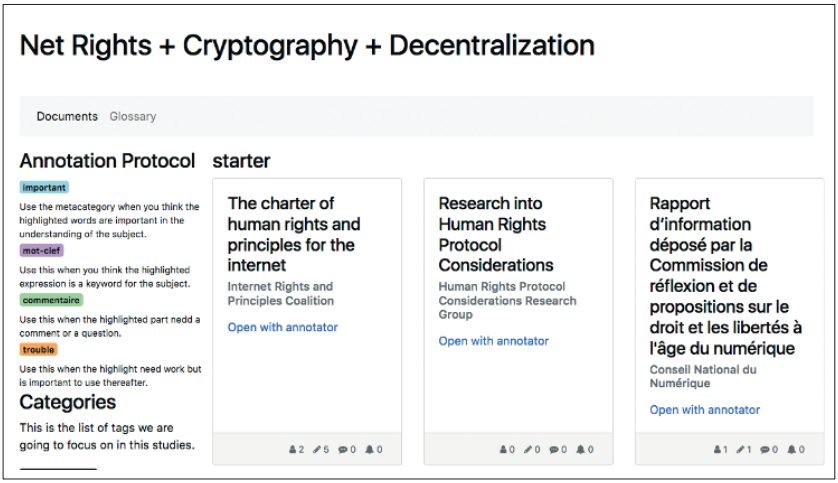


Figure 2. Interface for annotating documents discussing net rights using meta-tags.

Note that these tools were chosen to go beyond the limitations of the previous approaches to crowdsourcing net rights which were limited primarily to textual comments in the context of a discussion forum or, even worse, simple numerical rankings of proposed net rights. The interface for crowdsourcing net rights and the results are online.[7]

A two-step annotation protocol was used in the workshops for creating a terminology of net rights. The first step of the protocol consists in the use of a semi-controlled vocabulary of categories for tagging terms during the process of annotation, including but not limited to: accessibility, anonymity, decentralization, encryption, governance, open data, industrial lobby, surveillance, and so on. This allowed various terms related to net rights to be grouped together under one or more categories, with new terms and categories being added as more documents were annotated. Thus, the categories indexed diverse terms related to net rights even when the different lexical terms related to a single category were in different languages. During the initial round of annotation, the annotator could add meta-categories about their understanding of a term and its importance via "meta-tags," as shown in Figure 2. The following meta-tags that were already present in Hypothesis were used in our workshop: *Important*, *Comment*, *Keyword*, and *Confusion*. In addition, new meta-tags were added to Hypothesis: *Reference*, *Approval*, *Disapproval*, and *Question*. The *Reference* tag, in particular, was used to create a glossary of various documents relating to a term outside the initial corpus. Then the following metacategories were added for signaling to other annotators that more information or a discussion was required via using the following "action-tags": *Request for expertise*, *Request for discussion*, and *Request for collective action*. The goal was to build consensus on the definition and use of a term. Going beyond definitions, collective action could also entail informally agreeing that the proposed net right was either to be accepted or rejected by the workshop participants, which in practice often meant being subsumed by a pre-existing human right.

In the second phase of annotation, a glossary for net rights was created automatically from the annotations of the various key lexical terms. For each keyword, a discussion forum was created. This was followed by a discussion by the annotators over the terms in the glossary in order to find a unified definition, as well as clarify their relationship to any proposed net rights, as shown in Figure 2. The automatic generation was done via the *keyword*

7 https://netrights.iri-research.org/.

meta-tag ("*mot-clef*") that indexed highlighted terms by the number of occurrences (given by annotated *references*) across documents and languages as shown in Figure 3, showing in the interface the additional context given by the document needed for discussion. Although the goal of the discussion and references was to create a single definition, the interface supported multiple definitions for each term.

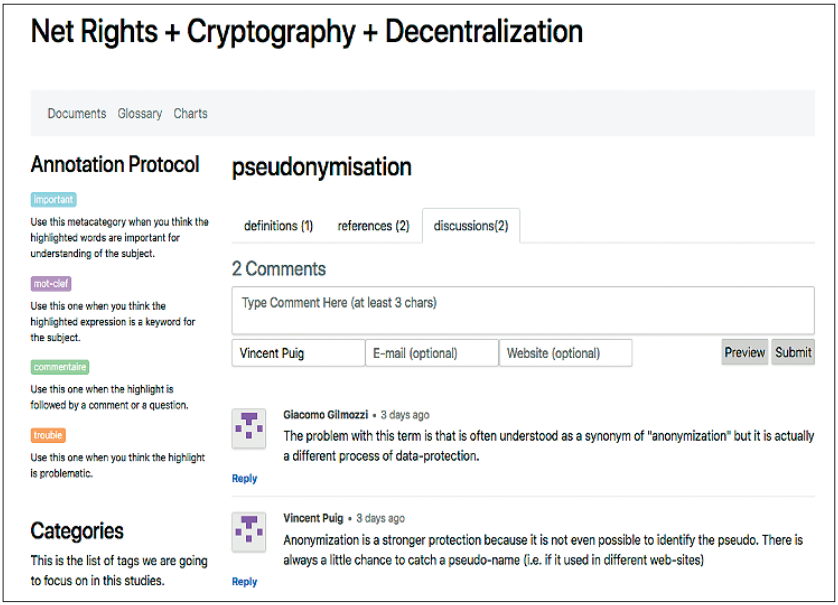


Figure. 3. An example of a discussion over a definition in the Net Rights glossary.

5. Results

Before the workshops, we collected a set of 39 documents for annotation. Of these, fifteen documents were related to European law and regulation, and three documents were related to French law. This included proposals by the French Parliament Digital Commission (FPDC), with the 100 propositions (recommendations) of the FPDC alone. Annotated documents ranged from

a draft of "Magna Charta for the Web" proposed by Ars Industrialis[8] for Berners-Lee to the finalized "The charter of human rights and principles for the Internet" by the IGF's Internet Rights and Principles Coalition,[9] as well as related articles such as "Degrees of Freedom, Dimensions of Power" by Yochai Benkler (2016) and historic documents such as the 1996 "A Declaration of the Independence of Cyberspace" by John Perry Barlow.[10] During the workshops, 13 documents were annotated. The distribution of annotations is given in Figure 4. The majority of annotations were for discussing keywords (with some disagreement) and the most common topics were surveillance and decentralization.

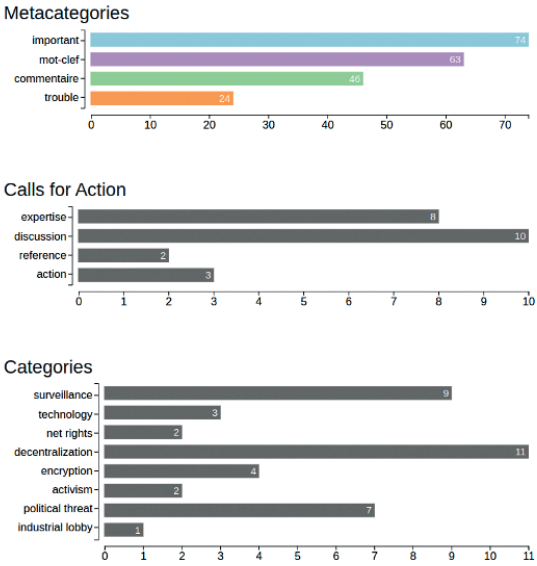


Figure. 4. The distribution of various annotations across net rights documents.

The glossary created for net rights via the process of annotation had 63 terms. However, only 14 terms managed to receive definitions, although the rest of

8 https://nextleap.eu/netrights/articles/Magna_Charta_constitution.pdf.
9 https://internetrightsandprinciples.org/charter/.
10 https://www.eff.org/cyberspace-independence.

the terms were discussed and had references. This demonstrates the lack of emerging consensus and the expert knowledge required to define certain terms related to net rights. Note that as each participant could see the other participants' annotations, there was a strong effect from the interface to encourage discussion, if not agreement, between annotators. Although there was a wide range of disagreement and agreement, standardized statistical measures of inter-rater agreement such as Krippendorff's alpha do not really make sense given the aforementioned effects of the interface. Therefore, various statistical measures are not reported at this stage of the experiments, but are left for future work.

Right #	**10**
Type of right (legal category)	Individual autonomy (III, B)
Title	Right for individuals to control their personal data (R58) including control over technologies (R49)
Source	FPDC, R58
Description	FPDC recommends that each individual has a right of self-control its personal data (auto-determination) but without exploitation by of their personal data. Prior consent (opt in) for each transaction is impossible, so authorization to collect personal data may be given for all transactions on a given site but some cases (balance of power, frequent requests, impact on relatives) should be excluded from this scheme (1) and require specific consent (2) and certain authorizations should be prohibited for instance on DNA information (3)
References	(1) Y. Poullet et A. Rouvroy, « Le droit à l'autodétermination informationnelle et la valeur du développement personnel. Une réévaluation de l'importance de la vie privée pour la démocratie. », in *État de droit et virtualité*, K. Benyekhlef et P. Trudel (dir.), Montréal : Thémis, 2009. (2) Ruth R. Faden et Tom L. Beauchamp, A history and theory of informed consent, OUP USA, 1986. (3) Frederik J. Zuiderveen Borgesius, « Consent to behavioural targeting in european law - What are the policy implications of insights from behavioural economics? », 2013.
Other comments	These recommendations raise the need for standardization of the information related to personal data.
Categories	governance, accessibility

Table 1. An example of an annotated net right from the proposal for net rights from the French Parliament Digital Commission (FPDC).

Via the process of categorization and discussion, the experiments in annotation produced a large number of proposed net rights that were given distinct titles, types, categories, and description, and references and related comments were included as shown in Table 1. In summary, 33 net rights were extracted from the documents in this format, with the full list of net rights online. Of these 33 net rights, there was a convergence on the following net rights: 1) Right to participate, 2) Right to individual autonomy (See Table 1), 3) Right to access public data, 4) Right to access the Internet (net neutrality), 5) Right to own personal data, 6) Right to creative expression (commons-based licensing), 7) Right to anonymity. The following more traditional rights also had consensus: 8) Right to privacy, 9) Right to freedom of expression, and 10) Right of free association. Other proposed rights around social justice, diversity, and governance were more controversial due to political differences, and the contradiction between a right to security and the rights to privacy and anonymity.

While a lexicon exhibits both descriptive linguistic and conceptual relationships, the conceptual relationships can be modelled by a formal ontology (Roche, 2012). Thus, net rights were considered the ontological conceptualization of the various lexical expressions of these rights in the annotated documents. So each of these rights was formalized into a Semantic Web ontology, with each net right being a class and each of the properties being derived from the final description as exemplified by Table 1. The Semantic Web OWL ontology allowed each net right to be formally connected to documents, where the mention of this right in their own set of lexical terms is an instance of this class. This is illustrated in Figure 5.[11]

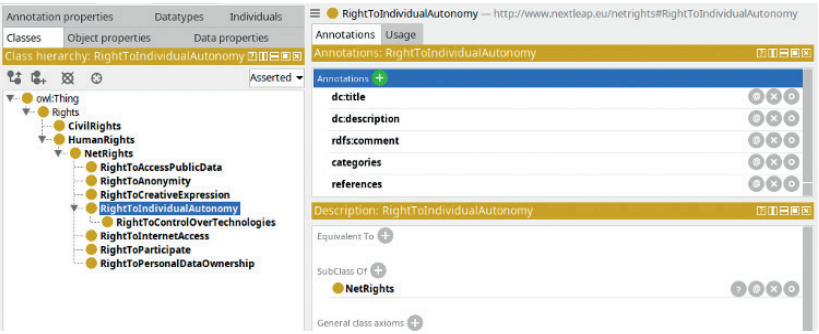


Figure 5. An example of the formal Semantic Web ontology for net rights.

11 http://nextleap.eu/.

6. Discussion

Although a coherent consensus on net rights emerged from our workshops, there remain a number of thorny problems. There were problems of linguistic terminology and ontological expertise, as well as a larger problem of enacting these rights in political practice.

Many terminological issues that became apparent in the deliberations of workshop participants had their origin in linguistic and cultural issues. Although the Internet puts itself as a universal space of information, there persists deeper historical and cultural differences revealed by even minor terminological differences. The workshop participants were primarily native French and English-speakers and so one key difference that appeared was that in French there is not a clear term for "privacy," with the closest term being "*la protection de la vie privée*," although another term, "*confidentialité*" resonates with the meaning of privacy in terms of the protection of secrets outside of the individual's life. Technical terms also proved to be difficult, such as the translation of "encryption" being rendered often as "*chiffrement*," which is related to the transformation of text into numerical signs (numerals are "les chiffres") as required by computer processing. On the other hand, English lacks a clear term for the broad European notion of "data protection" as it relates to the "*autodétermination*" of data, as the lexical term "self-determination" caused considerable confusion in English speakers. Further participants from Germany, Greece, Russia, Lebanon, and Guatemala also had their own particular terminological inflections. Although the project separated annotated linguistic lexical relationships from the conceptual relationships as formalized in an ontology in Figure 5 (Roche, 2012), there is nonetheless a trace of ontological difference in many terminological differences.

The question of genuine differences and contradictions on the level of ontology also arose in the course of crowdsourcing net rights. On the one hand, even though it is possible that the currently rather vague concepts of net rights – around internet access, data protection, net neutrality, online anonymity, and so forth – map onto traditional human rights inscribed by the United Nations, the exact formulation of the mapping is unclear (Redeker *et al*., 2018). On the other hand, although it is possible that there are entirely new net rights that do not map onto any existing human rights, there is no widespread agreement on what this new category of genuinely new "post-human" rights would entail. Various attempts to design or crowdsource net rights have remained incoherent and even contradictory: How can the right to remain free of online

abuse, for instance, be squared with the right for online anonymity? Further work in order to systematize these rights and their possible contradictions may require not just the work of ordinary humans, but various legal experts who can outline the contradictions and edge cases inherent in different rights. Therefore, future work will require legal experts working with Semantic Web experts in order to create a durable formal ontology of net rights.

There is also the practical question of how these net rights can be enforced. The Internet's lack of a legal body capable of enacting binding legislation makes international jurisdiction for net rights difficult, and the fact that most nation-states do not want to curtail their own sovereignty, including the surveillance activities of their own intelligence agencies, makes preserving these rights within traditional legal frameworks difficult. The idea brought forward by Snowden to encode these rights into the fundamental standards of the Internet may ultimately be a more feasible path to enforcement for rights such as the right to anonymity (Pitts-Brooke, 2014), yet even then many of these net rights may be not be capable of being enforced on the level of code, such as the right of participation or internet access.

7. Conclusions

Even if the practical aspects proved difficult, the project of crowdsourcing net rights from the collective intelligence of Internet users is a compelling proposition. Ultimately, the very concept of human rights has undoubtedly been altered by the advent of the Internet, and the invention of a new concept of net rights is coherent with both various popular notions of post-humanism and the extended mind in cognitive science (A. Clark & D. Chalmers, 1998). As put by the philosopher Bernard Stiegler, the essence of being human is the interiorization and exteriorization of media, as shown by the interpretative process of reading, writing, and discussion (Stiegler, 2018). In this vein, this particular terminological experiment was meant to "build and implement systems dedicated to the individual and collective interpretation of traces… with the goal of rebuilding the architecture of the World Wide Web, and as a way of responding to the call issued by Tim Berners-Lee" (Stiegler, 2018, 49-50). The key is that net rights cannot be reduced to a set of transcendental net rights, at least at this moment in history. Net rights are a way of thinking about the politics inherent in the historicized and dynamic process immanent in our evolving usage of the Web. Therefore, it should be no surprise that both the number and kinds of net rights will continue to be in flux over time. What meta-stability exists in terms of agreement over net rights is a witness to the

stabilization of certain social expectations over the Internet, but even these expectations are not ahistorical, much less transcendental. In final analysis, the goal of our efforts in crowdsourcing net rights is a new form of hermeneutics, not a unified terminology with a clear ontological foundation. There is thus a natural tendency towards both a consensus and a countervailing dissensus, a logomachy that is ultimately productive of new concepts of rights. Our particular experiment is just one step in provoking the collective deliberation necessary to transform our own relationship to the Internet.

The authors are supported by NEXTLEAP (EU H2020 ref: 688722). We would also like to thank Giacomo Gilmozzi and the many seminar participants at IRI for their discussions over net rights.

References

Benkler, Yochai. 2016. “Degrees of Freedom, Dimensions of Power.” *Daedalus*, 145(1), 18-32.

Berners-Lee, Tim, and Halpin, Harry. 2012. “Defend the Web.” In *Digital Enlightenment Yearbook*, edited by Jacques Bus, Michael Crompton, Mireille Hildebrandt, and Metakides, George, 3–7. Amsterdam: IOS Press.

Cerf, Vinton. 2012. “Internet Access is Not a Human Right.” *New York Times,* 25–26. https://www.nytimes.com/2012/01/05/opinion/internet-access-is-not-a-human-right.html

Clark, Andy, and Chalmers, David. 1998. “The Extended Mind.” *Analysis* 58(1): 7–19.

Habermas, Jürgen. 1984. *Theory of Communicative Action, Volume One: Reason and the Rationalization of Society*. Boston: Beacon Press.

LaRue, Frank. 2011. “Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. United Nations General Assembly.” https://www.ohchr.org/sites/default/files/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf

Pitts-Brooke, Jack. 2014. “Tim Berners-Lee Joins Edward Snowden Onstage for Virtual High-Five and Privacy Chat.” *The Independent.* https://www.independent.co.uk/tech/tim-bernerslee-joins-edward-snowden-onstage-for-virtual-highfive-and-privacy-chat-9203165.html

Redeker, Dennis, Gill, Lex, and Gasser, Urs. 2018. “Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights.” *International Communication Gazette* 80(4): 302–319.

Roche, Christophe. 2012. “Ontoterminology: How to unify terminology and ontology into a single paradigm.” In *LREC International Conference on Language Resources and Evaluation,* 2626–2630. Istanbul: Turkey.

Stiegler, Bernard. 2018. *The Neganthropocene*. London: Open Humanities Press. Benkler.

Représentation computationnelle des données terminologiques en diachronie : le cas des fibres artificielles

Andrea Bellandi*, Silvia Calvi**, Klara Dankova***, Silvia Piccini****

* Istituto di Linguistica Computazionale "A. Zampolli", Pisa
andrea.bellandi@ilc.cnr.it

**Università Cattolica del Sacro Cuore, Milano
silvia.calvi1@unicatt.it

*** Università Cattolica del Sacro Cuore, Milano
klara.dankova@unicatt.it

**** Istituto di Linguistica Computazionale "A. Zampolli", Pisa
silvia.piccini@ilc.cnr.it

Résumé. L'article porte sur la création d'une base de données diachronique des fibres textiles (FR-IT-EN) selon les principes du web sémantique – un projet né de la collaboration entre l'Istituto di Linguistica Computazionale « Antonio Zampolli » (ILC-CNR) et l'Osservatorio di Terminologie e Politiche Linguistiche (OTPL – Università Cattolica del Sacro Cuore). L'accent est mis sur la représentation formelle des données du sous-domaine des fibres artificielles.
Après avoir introduit les caractéristiques spécifiques de ce type de fibres textiles, notamment leur impact sur l'environnement, la nature des données formalisées jusqu'à présent est illustrée. Dans l'étape suivante, les principes du modèle DIATERM, adopté dans ce projet, sont expliqués, suivis d'une description détaillée de la modélisation de la ressource des fibres textiles, qui repose sur deux niveaux d'analyse : linguistique et conceptuel. Enfin, sa capacité à donner des réponses précises à des requêtes formelles est démontrée au moyen d'une série d'interrogations diachroniques.

Abstract. *The article focuses on the creation of a diachronic database of textile fibres (FR-IT-EN) based on the principles of the Semantic*

Web – a project born from the collaboration between the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) and the Osservatorio di Terminologie e Politiche Linguistiche (OTPL - Università Cattolica del Sacro Cuore). Emphasis is placed on the formal representation of data in the subdomain of artificial fibres.

After introducing the specific features of this type of textile fibres, especially their environmental impact, the nature of the data formalised so far is illustrated. Then, the principles of the DIATERM model, adopted in this project, are explained, followed by a detailed description of the modelling of our terminological resource, which includes two levels of analysis: linguistic and conceptual. Finally, its effectiveness in providing precise answers to formal queries is demonstrated through a series of diachronic queries.

1. Introduction[1]

Cette contribution vise à présenter les premières phases de construction d'une ressource terminologique computationnelle diachronique dédiée aux fibres textiles. Ce domaine s'avère en effet particulièrement propice à une étude diachronique. L'industrie des fibres textiles a connu une évolution remarquable au cours du XX^e^ siècle : les avancées technologiques, les transformations dans les méthodes de production, les innovations matérielles et les changements dans les préférences des consommateurs ont sculpté de manière significative ce domaine industriel, en ouvrant de nouvelles perspectives en matière de durabilité, de performance et d'applications diverses.

Parmi les avancées les plus marquantes, l'émergence des fibres manufacturées vers la fin du XIX^e^ siècle avec la mise au point des *fibres artificielles* a profondément transformé le paysage de ce secteur, les matières textiles pouvant être obtenues, pour la première fois, indépendamment des ressources naturelles traditionnelles. Pour bien comprendre les enjeux liés à la diffusion de définitions précises des termes de ce sous-domaine, il faut souligner qu'en fonction de leur typologie, les fibres artificielles présentent un impact sur l'environnement très varié, dépendant notamment de la technologie de leur production ; considérons, entre autres, la consommation et la pollution de l'eau, la consommation d'énergie, la pollution des sols et les émissions de CO_2.

1 Andrea Bellandi a rédigé les §§ 4 et 6, Silvia Calvi les §§ 3 et 5.1, Klara Dankova les §§ 1 et 2 et Silvia Piccini les §§ 4.1, 4.1.1, 4.1.2, 4.1.3 et 5.

Actuellement, les impératifs écologiques de l'industrie textile sont de plus en plus partagés à un niveau individuel – par les consommateurs changeant leurs habitudes d'achat et de gestion des produits textiles – aussi bien qu'institutionnel : de nouvelles mesures visant à rendre le cycle de vie des produits textiles plus durable sont adoptées à l'échelle nationale et internationale. À titre d'exemple, le Pacte vert de l'UE, qui a pour objectif une transition vers une économie circulaire d'ici 2050, prévoit des plans d'action en vue de réduire l'empreinte environnementale de l'industrie textile : ces plans portent notamment sur des limites plus strictes d'emploi de certains produits chimiques, la gestion durable des déchets et la promotion du réemploi et du recyclage des matières textiles[2].

Dans ce contexte, une définition claire et précise des fibres textiles s'impose, en vue de distinguer nettement les matières polluantes de celles qui sont plus éco-responsables. En plus, l'adoption de la perspective diachronique permettant de retracer l'évolution à la fois conceptuelle et linguistique du domaine peut être d'intérêt non seulement pour les consommateurs, mais aussi pour les experts du secteur et ceux du service marketing : la base de données peut constituer un précieux stimulant de créativité pour le développement de nouvelles matières et la création de leurs dénominations, tout en contribuant à la conservation et à la transmission des savoir-faire des générations précédentes.

C'est dans cette optique que la construction de la ressource terminologique – fruit de l'effort conjoint de l'Istituto di Linguistica Computazionale « Antonio Zampolli » (ILC – CNR) et de l'Osservatorio di Terminologie e Politiche Linguistiche (OTPL – Università Cattolica del Sacro Cuore) – a débuté avec une focalisation initiale sur le sous-domaine des fibres artificielles.

Formaliser de manière à rendre les évolutions dans ce domaine calculables par la machine représente une démarche intrigante, d'autant plus lorsqu'elle s'inscrit dans le paradigme du web sémantique (Berners-Lee *et al.*, 2001) et des données liées ouvertes (Bizer *et al.*, 2011). De nos jours, la communauté des terminologues manifeste un intérêt croissant pour ces technologies qui facilitent la construction de ressources linguistiques (terminologies, lexiques, dictionnaires et ontologies) réutilisables, partageables et interopérables

2 Parlement européen. Comment parvenir à une économie circulaire d'ici 2050 ?. Disponible sur : https://www.europarl.europa.eu/topics/fr/article/20210128STO96607/comment-parvenir-a-une-economie-circulaire-d-ici-2050 [consulté le 16 novembre 2024].

(Chiarchos *et al.*, 2013). Toutefois, leur utilisation pour représenter l'évolution diachronique des données terminologiques, dans leurs dimensions linguistique et conceptuelle, constitue un défi intéressant, en raison du caractère statique et monotone de l'OWL (Web Ontology Language), le langage standard du W3C développé pour représenter les ontologies dans le Web sémantique.

La présente étude se concentre dans un premier temps sur l'analyse du sous-domaine des fibres artificielles, en considérant notamment leur empreinte écologique (§ 2), avant d'aborder la méthodologie adoptée pour l'extraction des données terminologiques (§ 3). La construction de la ressource computationnelle selon le modèle DIATERM est ensuite décrite en détail (§ 4), avec une attention particulière portée aux dimensions linguistique et conceptuelle. Des requêtes SPARQL (§ 5) sont ensuite illustrées pour montrer comment interroger la base de données et extraire des informations ciblées sur les termes et les concepts. Enfin, nous revenons sur les principaux résultats (§ 6) afin d'en tirer des conclusions et des perspectives.

2. Les fibres artificielles dans le paysage des matières textiles

Pendant de longs siècles, l'industrie textile reposait essentiellement sur des matières fibreuses disponibles dans la nature : d'origine animale (exemple : la laine), végétale (exemple : le lin) ou minérale (exemple : les fils métalliques d'or), les différentes typologies de fibres naturelles ne pouvaient être obtenues que dans des territoires circonscrits, présentant des contextes géologiques spécifiques et/ou des conditions climatiques appropriées pour la culture de différentes espèces des plantes ou pour l'élevage de certains animaux. Le lien étroit entre une fibre et le lieu de son origine se traduit dans la production de matières présentant des caractéristiques physiques très variées : la longueur d'une fibre de coton, par exemple, peut aller de 1 à 4 cm (Weidmann, 2010, 42), ce qui entraîne des conséquences, entre autres, sur la fabrication et la qualité du fil, aussi bien que sur son prix de vente.

Ainsi l'arrivée des *fibres manufacturées* ou *chimiques*, c'est-à-dire matières textiles produites par l'homme par un processus de transformation chimique (Weidmann, 2010, 14 ; Baum et Boyeldieu, 2018, 257-258), a-t-elle révolutionné le secteur textile : dorénavant, les fibres peuvent être fabriquées en grande quantité, indépendamment du lieu de production ou des facteurs externes, comme des intempéries ou l'émergence de maladies, pouvant compromettre gravement l'exploitation des fibres végétales ou animales.

La première fibre manufacturée a été créée en 1884 dans le but précis de pallier la pénurie de la soie naturelle : appelée *soie Chardonnet* d'après son inventeur, cette nouvelle matière à base de nitrate de cellulose donne naissance aux fibres artificielles, ou bien celles obtenues par traitement chimique des polymères naturels (Baum et Boyeldieu, 2018, 257). Au fil du temps, plusieurs typologies des fibres cellulosiques ont été développées, dont la soie au cuivre ou *cupro* (1890), la viscose (1898) ou l'acétate de cellulose (1899) (Agulhon, 1962, 9-12), et d'autres substances organiques ont commencé à être utilisées dans la fabrication de ces fibres, telles que les polysaccharides (exemple : les fibres d'alginate obtenues à partir de l'acide alginique extrait des algues brunes, inventées dans les années 1910, Miraftab *et al.*, 2001, 165) et les protéines végétales (exemple : les fibres produites à partir des protéines du soja ou du maïs, mises au point vers la fin des années 1930, Agulhon; 1962, 14). Les différentes typologies de fibres artificielles sont représentées dans la carte conceptuelle de la Figure 1.

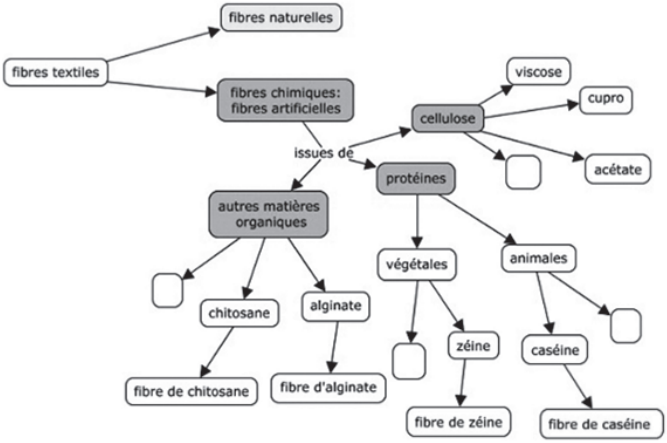


Figure 1. Le sous-domaine des fibres artificielles (Dankova 2023, 109).

L'évolution technologique du secteur aux XX^e^ et XXI^e^ siècles a porté au développement d'autres catégories de fibres textiles, dont l'organisation est illustrée dans la Figure 2, transformant celles-ci en matières premières incontournables non seulement dans la production textile, mais aussi dans d'autres secteurs industriels, parmi lesquels la santé, l'agriculture et le bâtiment.

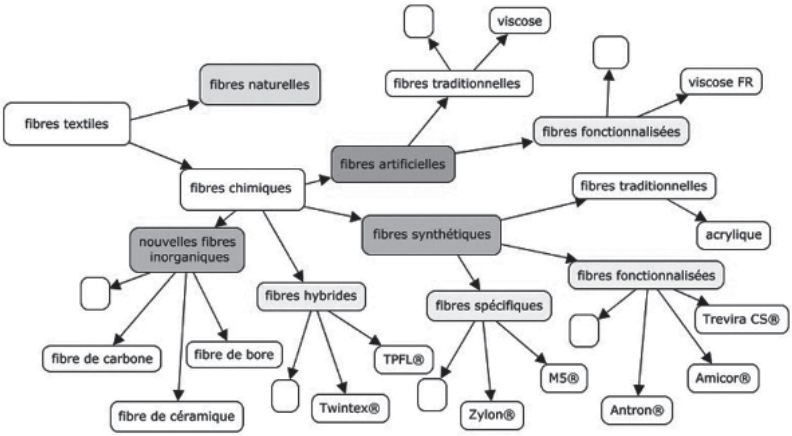


Figure 2. Les fibres textiles au XXIe siècle (Dankova, 2023, 113).

Les différents procédés de fabrication, exploitant des matières premières très variées, sont responsables non seulement des caractéristiques techniques (exemple la finesse, le point de fusion) et fonctionnelles (exemple la résistance aux agents chimiques ou aux micro-organismes) spécifiques de ces matières, mais aussi de leur impact sur l'environnement, évalué en considérant leur cycle de vie dans son intégralité, de la production jusqu'au traitement des déchets. Afin de comprendre pleinement les enjeux en matière de durabilité des fibres artificielles, il faut les distinguer, tout d'abord, de la deuxième grande catégorie des fibres manufacturées, les fibres synthétiques. Développés vers la fin des années 1930, ces produits textiles, tels que le polyamide, le polyester ou l'acrylique, sont créés par synthèse chimique, en utilisant, dans la plupart des cas, des matières premières dérivées du pétrole (Baum et Boyeldieu, 2018, 257-259). Contrairement aux fibres artificielles exploitant des ressources naturelles durables, comme la cellulose ou les polysaccharides, les fibres synthétiques consistent dans le traitement des matières non renouvelables. En plus, leur potentiel de biodégradabilité est, en général[3], très limité et leur recyclage, dont la pratique pour certaines matières comme le polyester est désormais assez diffusée, n'est pas sans interrogations.

3 Ces dernières années, la recherche dans le domaine des fibres synthétiques a porté au développement de nouveaux produits écoresponsables, dont la fibre Amni Soul Eco® de la société italienne Fulgar, un polymère en polyamide 6.6 biodégradable, utilisé notamment dans la fabrication de vêtements de sport. Voir le site : https://www.fulgar.com/prodotti/60/amni-soul-eco [consulté le 4 décembre 2024].

C'est pour cette raison qu'actuellement, les fibres artificielles sont beaucoup appréciées pour leurs caractéristiques écologiques, notamment, l'emploi des matières premières renouvelables et leur caractère biodégradable : des études récentes ont montré que quelques fibres artificielles, comme la viscose, se décomposent même plus vite que certaines fibres naturelles, dont le coton ou la ramie (Mehta, 2023 ; Niu *et al*., 2012). D'où le développement et le succès de certains produits innovants, comme Orange Fiber[4], une fibre textile fabriquée à partir de déchets d'agrumes. Cependant, même au sein de cette catégorie, l'empreinte écologique des fibres peut varier considérablement, en fonction du procédé de fabrication : les fibres artificielles cellulosiques de troisième génération, telles que le lyocell, mis au point dans les années 1970 et commercialisé en Europe depuis les années 1990, sont beaucoup plus écologiques que celles de première génération, fabriquées par le procédé viscose, breveté, pour la première fois, déjà en 1892 (Dankova, 2023, 163-164 ; 176-177).

Compte tenu de la prise de conscience de plus en plus forte des questions d'environnement, qui se reflète, dans le domaine textile, dans le changement progressif des habitudes et des comportements d'achat ainsi que dans l'augmentation de certaines pratiques, telles que le recyclage ou le réemploi, il est essentiel de fournir aux consommateurs des définitions précises des termes utilisés afin de leur permettre de prendre des décisions d'achat plus informées, dans le respect des valeurs qu'ils partagent. D'où l'importance de la clarté terminologique dans le domaine des fibres textiles, à laquelle notre ressource terminologique vise à contribuer.

3. La collecte des données terminologiques

La construction de notre ressource computationnelle a comme point de départ les données terminologiques du domaine des fibres chimiques, recueillies dans le cadre de l'étude de Dankova (2023), couvrant la période allant de la naissance de ce secteur industriel vers la fin du XIX^e^ siècle jusqu'à présent. Celles-ci ont été rassemblées au moyen de l'extraction terminologique semi-automatique effectuée à partir d'un corpus en français créé *ad hoc* et en s'appuyant sur un corpus de documentation contenant différentes typologies de sources, entre autres, des normes industrielles et juridiques et des

4 Voir le site : https://orangefiber.it/it/ [consulté le 4 décembre 2024].

documents institutionnels[5]. Le corpus textuel pour l'extraction des termes en français compte 112 136 occurrences, provenant de quatre types de textes :

- un catalogue d'un salon professionnel : Première Vision Yarns (12-14 février 2019) ;
- un document institutionnel : DGE/ UBIFRANCE 2006. *Textiles Techniques. Le futur se tisse en France*. Paris : Ministère de l'Economie, de l'industrie et de l'emploi.
- un ouvrage de vulgarisation : Fauque, Claude et Sophie Bramel. 1999. *Une seconde peau : fibres et textiles d'aujourd'hui*. Paris : Éditions Alternatives ;
- un manuel technique : Weidmann, Daniel. 2010. *Aide-mémoire Textiles techniques*. Paris : Dunod.

Ces sources ont été choisies dans le but d'obtenir un corpus représentatif contenant les principaux termes nécessaires pour acquérir une connaissance solide du domaine pendant toute la période de son existence. Bien que publiés dans la période récente (1999-2019), ces documents contiennent des termes utilisés à l'époque actuelle (Première Vision Yarns, 2019, DGE/ UBIFRANCE, 2006, Weidmann, 2010) aussi bien que dans le passé (surtout Fauque et Bramel, 1999 ; Weidmann, 2010). En plus, considérant que la terminologie change beaucoup en fonction du contexte d'emploi, notamment pour ce qui est des phrases de la production et de la commercialisation des produits (Zanola, 2016, 2018, 2019), les différentes typologies des sources ont été sélectionnées également en vue de couvrir cette variation de nature diaphasique.

Dans la première étape de la construction de notre ressource terminologique, nous nous sommes concentrés sur les données concernant le sous-domaine des fibres artificielles. Pour ce qui est de sa dimension conceptuelle, 55 concepts identifiés lors de l'étude de Dankova (2023) ont été complétés par 11 autres en donnant un ensemble de 66 concepts. Du côté linguistique, la ressource enregistre 123 termes en français, dont 77 noms génériques et 46 noms de marque, caractérisés par une variation diachronique aussi bien que fonctionnelle (exemple : code, dénominations chimiques, etc.), d'intérêt pour différentes catégories d'utilisateurs. À part les termes en français, la ressource fournit aussi leurs équivalents en anglais et en italien.

5 Pour plus d'information sur la structure du corpus de documentation, nous renvoyons à Dankova (2023, 87-94).

4. La représentation diachronique de la terminologie : le modèle DIATERM

Comme mentionné dans l'introduction, l'application des technologies du web sémantique, et en particulier du langage OWL-DL, pose des défis significatifs lorsqu'il s'agit de représenter l'évolution diachronique de l'information. Ce langage, fondé sur le modèle de données RDF, repose sur une structure en triplets (Sujet - Prédicat - Objet) et se base sur des fragments décidables de la logique de description (DL), permettant de raisonner sur des axiomes logiques complexes portant sur des prédicats unaires (classes) et binaires (relations). L'introduction d'une troisième dimension, à savoir la dimension temporelle, transforme une relation binaire en relation ternaire, modifiant ainsi la nature des prédicats. De plus, la nature monotone de l'OWL permet uniquement de représenter des scénarios statiques et cohérents. Toute nouvelle assertion, qu'elle soit explicitée dans la base de connaissances ou déduite par des mécanismes de raisonnement, ne peut en aucun cas contredire ce qui a été précédemment affirmé ou déduit.

Pour relever ce défi, diverses approches ont été proposées dans la littérature (voir notamment Flouris, 2008, Makri, 2011), parmi lesquelles figurent les logiques descriptives temporelles (Artale and Franconi, 2000; Lutz *et al.*, 2008), le versioning (Klein and Fensel, 2001), le modèle perdurantiste ou 4-D fluents (Welty *et al.*, 2006) ainsi que la réification (Kanmani *et al.*, 2016).

Dans le but de représenter formellement l'évolution diachronique des concepts et des termes liés au domaine des fibres textiles, décrit dans la section précédente, notre choix s'est porté sur le mécanisme de réification des relations N-aires, qui consiste à transformer un prédicat en objet, autrement dit à représenter une relation sous la forme d'une classe.

Plus précisément, nous avons adopté le modèle DIATERM (Piccini *et al.*, 2021), qui exprime les relations d'arité supérieure à deux à l'aide d'un modèle de conception d'ontologie (*Ontology Design Pattern*) basé sur la réification (Noy *et al.*, 2006).

Dans le modèle DIATERM, l'approche N-aire a été privilégiée pour trois raisons principales. Tout d'abord, cette méthode permet de gérer la complexité inhérente à la modélisation diachronique d'un système conceptuel tout en maintenant des performances élevées. Contrairement au modèle perdurantiste, elle limite efficacement la prolifération des entités, qu'il s'agisse de classes, de relations ou encore d'assertions sous forme d'axiomes (Batsakis *et al.*, 2017). De plus, l'approche N-aire s'inscrit pleinement dans les bonnes

pratiques reconnues par la communauté du web sémantique. Elle est fortement recommandée par le groupe de travail *Semantic Web Best Practices and Deployment* (SWBPD) qui fait partie de l'*Ontology Engineering and Patterns Task Force* (OEP) et qui se consacre à définir des solutions standardisées et éprouvées, afin d'assurer à la fois l'interopérabilité et la cohérence des modèles ontologiques. Enfin, cette approche présente un avantage technologique majeur : elle est largement compatible avec les principaux entrepôts de triplets actuellement disponibles, qu'ils soient commerciaux ou *open-source*. Parmi ces outils, on peut notamment citer GRAPHDB et Apache Jena.

Il convient toutefois de souligner que, bien que l'adoption de l'approche *N*-aire présente de nombreux avantages, elle comporte également certaines limitations. L'un des aspects les plus complexes concerne les capacités de raisonnement en OWL, qui deviennent plus limitées et ne peuvent être maintenues qu'au prix de stratégies chronophages. En effet, la transformation de la propriété réifiée en classe fait que les caractéristiques habituellement associées directement à une propriété en OWL, telles que la fonctionnalité, la fonctionnalité inverse, la transitivité ou la symétrie, ne sont plus applicables de manière directe à la propriété elle-même. En d'autres termes, l'introduction d'un objet intermédiaire, correspondant à une instance de la classe réifiée, fragmente la relation initiale, entraînant une redéfinition de son domaine et de son codomaine.

Plusieurs techniques peuvent néanmoins être envisagées pour tenter de préserver ces capacités de raisonnement malgré les contraintes imposées par la réification. Parmi celles-ci, on peut citer : i) l'utilisation de règles SWRL (*Semantic Web Rule Language*), qui permettent de traduire les caractéristiques des propriétés en règles déductives explicites ; ii) l'emploi du *punning*, une fonctionnalité introduite avec OWL 2, qui permet de représenter des individus à la fois comme des classes et des propriétés, tout en laissant le contexte préciser la nature exacte de l'utilisation.

L'exploration de ces techniques, bien que prometteuse, dépasse les objectifs de cet article. Dans notre cas spécifique, nous avons adopté une approche pragmatique en opérant des choix méthodologiques qui reflètent les compromis nécessaires pour atteindre les objectifs de notre étude. Ces décisions, tout en limitant la portée du modèle, visent à maintenir un équilibre entre la faisabilité pratique et la cohérence des résultats obtenus.

En particulier, nous avons choisi de ne pas inclure les caractéristiques des propriétés temporelles, telles que la symétrie ou la transitivité. Cependant, ces informations pourraient être récupérées par l'utilisateur lors de la consultation

de la ressource grâce à des techniques de visualisation de l'information. Une application logicielle dédiée pourrait, par exemple, représenter ces relations sous forme de frises chronologiques ou de lignes de temps, permettant ainsi à l'utilisateur de reconstruire visuellement les caractéristiques d'une propriété sur différents intervalles temporels.

En ce qui concerne la représentation temporelle (dates, relations topologiques entre instants et intervalles, durée des intervalles), DIATERM réutilise le vocabulaire OWL-Time, développé par le W3C pour décrire la dimension temporelle du contenu des pages web. Les règles d'Allen permettent de préciser les relations entre les entités temporelles, telles que les chevauchements ou les commencements d'intervalles.

L'encodage des informations est actuellement effectué manuellement avec Protégé, un processus laborieux qui consiste à créer les individus, à les relier selon le modèle et à instancier les processus de réification. À ce jour, 62 concepts et 101 termes, incluant des marques et des codes de fibres, ont été formalisés dans la ressource.

Dans les sections suivantes, nous illustrerons l'application du modèle DIATERM au domaine des fibres textiles, en nous concentrant sur les dimensions linguistique et conceptuelle.

4.1. Le modèle DIATERM et la ressource des fibres textiles

Le modèle DIATERM est caractérisé par une architecture à trois niveaux, avec l'axe diachronique traversant les trois dimensions où le changement peut se produire, à savoir le texte, le lexique et la conceptualisation. Dans le cadre de ce travail, nous avons restreint notre analyse aux plans linguistique et conceptuel, qui, bien que séparés, sont reliés par la relation `ontolex-:reference`, laquelle associe chaque sens d'un terme à un concept formalisé dans l'ontologie. Il convient de préciser à ce propos que, dans le présent travail, nous avons posé l'hypothèse théorique selon laquelle un concept est considéré comme existant pendant un intervalle temporel *T* si, et seulement si, il est désigné par un sens lexical dans *T*. En d'autres termes, l'introduction d'un nouveau concept est contextuellement liée à sa lexicalisation et peut être inférée à partir des valeurs temporelles associées à la propriété `ontolex-:reference`, qui relie un terme au concept qu'il dénote. Notons par ailleurs que, d'un point de vue théorique général, cette hypothèse pourrait sembler discutable. En effet, il est bien établi que certains concepts peuvent faire partie d'un système conceptuel établi par des experts, sans pour autant trouver

de correspondance lexicale explicite dans une langue donnée. Toutefois, cette hypothèse est conforme à la démarche sémasiologique adoptée dans le présent article et s'inscrit ainsi en cohérence avec le cadre méthodologique choisi.

4.1.1. Le niveau linguistique

En ce qui concerne le niveau linguistique, la ressource propose un répertoire trilingue de termes (français, italien et anglais), le français étant la langue principale d'investigation. Les données terminologiques ont été formalisées en suivant le modèle OntoLex-Lemon (McCrae *et al.*, 2017), reconnu comme la norme *de facto* pour la représentation de la composante lexicale des ontologies dans le web sémantique. Chaque terme est représenté comme une instance de la classe `LexicalEntry`. Plus précisément, un terme simple correspond à une instance de la classe `Word`, tandis qu'un terme composé est représenté par une instance de la classe `Multiword`. Les variantes morphologiques associées à une entrée lexicale, modélisées comme instances de la classe `Form`, sont décrites et associées à une représentation écrite via l'attribut `writtenRep`. Le sens spécifique d'une entrée lexicale, instance de la classe `LexicalSense`, est relié à une entité conceptuelle (classe, instance ou relation), qui reçoit une description formelle et explicite dans une ontologie externe, selon le principe de la « sémantique par référence » (Buitelaar, 2010).

Conformément au modèle DIATERM, une série d'autres classes caractérise le niveau linguistique de notre ressource, découlant du processus de réification des relations définies dans le modèle OntoLex-Lemon. Parmi celles-ci on trouve : la classe `senseEvent`, qui représente la réification de la propriété `ontolex:sense`, reliant un individu de la classe `LexicalEntry` à un sens spécifique ; la classe `referenceEvent`, qui réifie la relation `ontolex-:reference`, associant le sens d'une entrée lexicale à un concept de l'ontologie ; ainsi que les classes `equivalentEvent`, `incompatibleEvent`, `broaderEvent` et `narrowerEvent` qui représentent respectivement la réification des relations `lexinfo:synonym`, `lexinfo:antonym`, `lexinfo:hypernym` et `lexinfo:hyponym`. Les instances de ces classes sont reliées aux individus de la classe `ontolex:LexicalSense` par de nouvelles relations, telles que `temporalHypernym`, `temporalHyponym`, `temporalSynonym` et `temporalAntonym`, toutes définies comme sous-propriétés de `temporalProperty`.

Toutes les classes issues de la réification sont formalisées comme des sous-classes de `dcmtype:Event`, qui représente de manière générale la réification d'une propriété dans un intervalle donné *T*. Ces classes, ainsi que

leur superclasse `dcmitype:Event`, sont reliées à la classe `time:ProperInterval` *via* la propriété `during`, qui est modélisée comme une sous-propriété de `dcterms:date`[6] et spécifie l'intervalle de temps pendant lequel l'événement en question se produit.

Nous présenterons dans les pages suivantes (§ 4.1.3.) un exemple de modélisation de la variation temporelle du terme *fibre manufacturée* selon le modèle DIATERM.

4.1.2. Le niveau conceptuel

Bien que le modèle DIATERM permette de temporiser les propriétés qui caractérisent un concept en utilisant le même principe de réification, le domaine des fibres textiles s'est révélé relativement stable sous cet aspect. En effet, dans le cadre de nos études de cas, nous n'avons pas constaté de redéfinition de l'intension des concepts existants. La dynamique observée a plutôt consisté en l'introduction de nouveaux concepts et en la disparition progressive de ceux devenus obsolètes. Ce phénomène a été modélisé, comme mentionné précédemment, à travers la réification de la propriété `ontolex:reference`. Concrètement, cela signifie que chaque concept introduit dans le domaine concerné possède également une lexicalisation, du moins en français, qui est la langue du corpus de référence. À l'inverse, un concept « cesse d'exister » dès lors qu'il n'est plus associé à un terme au niveau linguistique.

L'architecture de l'ontologie de domaine s'inspire du modèle SIMPLE (Lenci *et al.*, 2000). Conçu à l'origine pour la lexicographie, ce modèle s'est également révélé particulièrement efficace dans la structuration des lexiques spécialisés (Piccini *et al.*, 2016). Il repose sur les principes fondamentaux de la théorie du Lexique Génératif développée par Pustejovsky (1995) qui offre un cadre solide pour capturer et formaliser la multidimensionalité des concepts. L'un des éléments clés du modèle SIMPLE est en effet la *Structure Qualia*, qui permet de représenter les différentes dimensions d'un concept à travers quatre rôles fondamentaux : *formel*, *constitutif*, *télique* et *agentif*. Ce cadre dépasse

6 Conformément à la philosophie des données liées, dans DIATERM les classes et les propriétés des vocabulaires considérés comme standard et/ou largement utilisés dans la communauté scientifique ont été empruntées dans la mesure du possible. Ainsi, par exemple, certaines classes et propriétés ont été tirées du vocabulaire de la *Dublin Core Metadata Initiative* (DCMI) ou, comme mentionné précédemment, de l'ontologie OWL-Time.

les simples relations hiérarchiques de subsomption, offrant une modélisation plus riche des concepts, particulièrement adaptée à des domaines spécialisés, comme celui des fibres textiles.

Dans le cadre de notre ressource, nous avons intégré les concepts généraux du modèle SIMPLE pertinents pour notre domaine d'application, en les formalisant comme les nœuds les plus élevés de la hiérarchie taxonomique. Une approche essentiellement descendante a été adoptée pour affiner et étendre le modèle, en tenant compte des problématiques spécifiques liées aux données du domaine des fibres textiles.

De plus, une série de relations, inspirées également de la *Structure Qualia*, a été intégrée, classée en relations formelles (`is-A`), constitutives (`is PartOf`, `hasAsPart`, `madeOf`, etc.), téliques (`usedFor`, `usedBy`, etc.), et agentives (`resultOf`, `causedBy`, `createdBy`, `derivedFrom`, etc.). Cette approche a permis de définir de manière fine et exhaustive tous les aspects qui caractérisent les concepts les plus complexes.

4.1.3. La formalisation d'une entrée lexicale : *fibre manufacturée*

À titre illustratif, nous présentons dans les Figures 3 et 4 l'application du modèle DIATERM au cas de la *fibre manufacturée*. Le terme a été utilisé au fil du temps pour désigner deux concepts différents, illustrés dans la partie (b) de la Figure 3. Il s'agit, d'une part, du concept de « fibre textile créée par l'homme, obtenue par traitement chimique des polymères naturels » (classe <FIBRE ARTIFICIELLE>[7]), et, d'autre part, du concept de « fibre textile produite par l'homme par un processus de transformation chimique » (classe <FIBRE CHIMIQUE>).

Le terme *fibre manufacturée* a été introduit pour la première fois en 1884, comme synonyme de *fibre artificielle*. Alors que ce dernier terme continue de désigner le concept de <FIBRE ARTIFICIELLE>, le terme *fibre manufacturée* a évolué. À partir de la fin des années 1930, avec l'émergence des fibres synthétiques, il a progressivement cessé d'être utilisé pour désigner ce concept et il a acquis une signification plus large, désignant simplement les fibres textiles produites par l'homme, englobant ainsi à la fois les fibres artificielles et synthétiques.

7 Dans cet article, les concepts sont encadrés de crochets angulaires pour les distinguer des termes. La définition des concepts est, quant à elle, placée entre guillemets "…".

La représentation temporelle des termes associés à ces concepts est illustrée dans la Figure 3(a). Selon le modèle OntoLex-Lemon, chaque terme (représenté par un nœud vert) a un ou plusieurs sens (représentés par des nœuds jaunes), chacun dénotant un concept décrit dans une ontologie externe. Les relations entre les nœuds sont dotées d'une valence temporelle, mise en œuvre par le modèle DIATERM. Les axiomes formalisant les informations de la Figure 3(a) sont présentés dans la Figure 4(a) ; en particulier, les axiomes liés au modèle DIATERM, qui intègre la valence temporelle des relations *via* le modèle de réification, sont mis en gras.

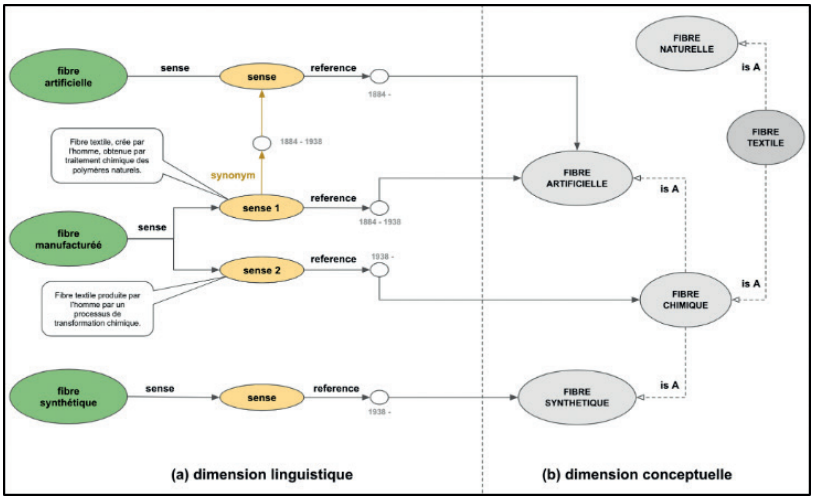


Figure 3. Extrait de la hiérarchie taxonomique relative aux fibres textiles (b) et leurs dénominations en français selon le modèle DIATERM (a).

Enfin, dans la Figure 4(b) les caractéristiques essentielles qui définissent les concepts de la Figure 3(b) ont été formalisées à l'aide de l'éditeur Protégé. Par exemple, le concept <FIBRE SYNTHETIQUE> est défini comme une fibre textile (sous-classe de <FIBRE _ TEXILE>), créée par l'homme (createdBy <HUMAN>), résultant du processus de synthèse chimique (resultOf <SYNTHESE _ CHIMIQUE>) et dérivée des matières non renouvelables (derivedFrom <MATIERE _ NON _ RENOUVELABLE>). Par contre, le concept <FIBRE _ ARTIFICIELLE> est défini comme une fibre textile (sous-classe de <FIBRE _ TEXILE>), créée par l'homme (createdBy <HUMAN>), résultant du traitement chimique des polymères naturels

(resultOf <TRAITEMENT_CHIMIQUE_DES_POLYMERES_NATURELS>) et dérivée des matières renouvelables (derivedFrom <MATIERE_RENOUVELABLE>).

Les deux classes sont définies comme disjointes, ce qui signifie qu'une instance ne peut appartenir simultanément à l'une et à l'autre de ces classes.

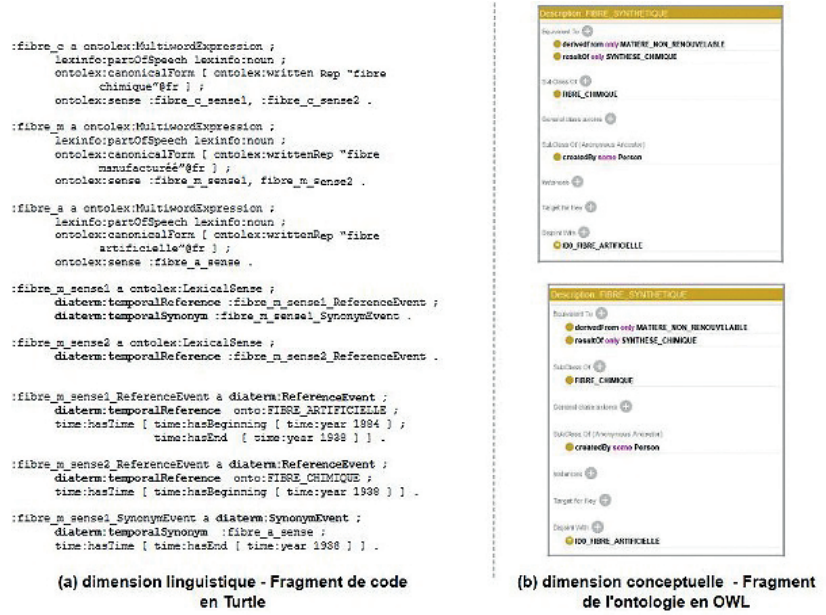


Figure 4. Formalisation du terme fibre manufacturée selon le modèle DIATERM (a) et représentation des concepts <FIBRE SYNTHÉTIQUE> et <FIBRE ARTIFICIELLE> à l'aide du logiciel Protégé.

5. Interrogation de la base de connaissance

Dans le cadre de ce travail, la validation de l'ontologie a été réalisée à travers l'approche des questions de compétence, qui ont joué un rôle fondamental tout au long du processus. Ces questions, formulées en langue naturelle par les experts du domaine, ont permis de guider la construction même de l'ontologie en définissant les concepts, les relations et les informations clés qu'elle

devait représenter. Une fois l'ontologie du sous-domaine des fibres artificielles élaborée, ces questions de compétence ont été traduites en requêtes formelles – notamment en SPARQL – et utilisées pour évaluer la capacité du modèle à fournir des réponses précises et cohérentes. Ce processus a permis de vérifier de manière rigoureuse si la formalisation adoptée répondait aux besoins identifiés par les experts, tout en assurant une cohérence théorique et pratique. De plus, il a favorisé une interaction directe avec les experts, renforçant ainsi la pertinence et l'exhaustivité de la modélisation.

Pour systématiser cette analyse, huit questions de compétence ont été identifiées et classées selon deux critères principaux : le « niveau », distinguant les requêtes qui se concentrent soit sur la dimension linguistique, soit sur la dimension conceptuelle, ou qui portent sur les deux dimensions à la fois ; et l'« axe temporel », qui sépare les requêtes synchroniques des requêtes diachroniques. Ces questions sont présentées dans le tableau 1 (*infra*), où elles sont également classées par ordre de complexité.

Les deux premières requêtes, en vert, concernent le plan linguistique synchronique. Elles permettent à l'utilisateur, par exemple, de connaître les marques associées à un type de fibre ou encore les équivalents en italien et anglais d'un terme français donné. Pour ce qui est de la dimension conceptuelle, les requêtes 3 et 4 en vert permettent respectivement de comprendre les propriétés d'un concept en synchronie, ou encore de déterminer le moment où un concept est apparu pour la première fois. Il convient de souligner que les réponses à cette dernière requête reposent sur la temporalisation de la propriété `ontolex:reference` et sur le postulat théorique selon lequel lorsqu'un concept est introduit, il est simultanément nommé.

Les requêtes en jaune, quant à elles, concernent uniquement la dimension diachronique linguistique. Elles permettent, entre autres, de savoir quels sont les synonymes d'un terme donné à un moment précis ou sur une période définie par l'utilisateur, ou encore d'identifier les termes complexes dérivés d'un terme simple utilisés pendant une période donnée (*T*). Enfin, les deux dernières requêtes, en rouge, se distinguent par leur complexité, liée à la nécessité d'interroger simultanément plusieurs dimensions (linguistique et conceptuelle) tout en intégrant des aspects temporels. Elles permettent de répondre à des questions telles que : *Quels termes ont été utilisés pour désigner un concept donné (*C*) pendant une période (*T*) ?* ou encore : *Quels sont les termes associés à un concept donné (*C*), et pendant quelles périodes ont-ils été utilisés* ?

ID	Question	Niveau	Axe temporel
1	Quels sont les noms de marque de *C*?	linguistique	synchronique
2	Quels sont les termes italiens et anglais correspondant au terme français *t*?	linguistique	synchronique
3	Quelles sont les propriétés du concept *C*?	conceptuel	synchronique
4	Quand le concept *C* a-t-il été introduit?	conceptuel	diachronique
5	Quels sont les synonymes du terme t dans une période *T*?	linguistique	diachronique
6	Quels termes complexes créés à partir d'un terme simple sont utilisés dans une période *T*?	linguistique	diachronique
7	Quels sont les termes utilisés pour désigner le concept *C* dans une période *T*?	linguistique et conceptuel	diachronique
8	Quels sont les termes désignant le concept *C* et à quelle époque ont-ils été utilisés?	linguistique et conceptuel	diachronique

Tableau 1. Questions de compétence portant sur les niveaux linguistique, ontologique et temporel de la ressource.

5.1. Des exemples de requêtes SPARQL

En nous focalisant sur les interrogations de type diachronique, nous allons présenter, à titre illustratif, quatre cas de requêtes spécifiques.

En interrogeant le niveau linguistique de la ressource, l'utilisateur peut obtenir, par exemple, les variantes synonymiques d'un terme *t* pendant la période *T*. La Figure 5 illustre une requête SPARQL de ce type, visant à repérer les variantes synonymiques de *triacétate* dans les années 1970. L'interrogation fournit comme résultats deux termes, *fibre de triacétate* et *triacétate de cellulose*.

Figure 5. Requête SPARQL 1 : variantes synonymiques de triacétate dans les années 1970.

```
SELECT ?synonym
WHERE {
    ?term ontolex:canonicalForm [ ontolex:writtenRep "triacétate"@fr ] ;
                                  ontolex:sense ?sense .
    ?sense diaterm:temporalSynonym ?synEvent .
    ?synEvent diaterm:temporalSynonym [ ontolex:isSenseOf
            [ ontolex:canonicalForm
                [ ontolex:writtenRep ?synonym ]
            ]
    ] ;
                time:hasTime ?t .
    OPTIONAL { ?t time:hasBeginning [ time:year ?start ] }
    OPTIONAL { ?t time:hasEnd [ time:year ?end ] }
    FILTER((1970<COALESCE(?start, 0) && 1979>COALESCE(?start, 0)) ||
           (1970>COALESCE(?start, 0) && 1979<COALESCE(?end, 0)) ||
           (1970<COALESCE(?end, 0) && 1979>COALESCE(?end, 0)))
```

Une autre requête portant sur la dimension linguistique peut explorer, par exemple, tous les termes complexes contenant un terme simple ou un autre élément, utilisés pendant la période *T*. Celle de la Figure 6, par exemple, a pour objectif d'identifier les termes complexes contenant l'adjectif *artificiel*, en limitant la recherche à la période entre 1883 et 1940 :

```
SELECT DISTINCT ?term
WHERE {
    ?le ontolex:canonicalForm [ ontolex:writtenRep ?term ] .
    ?le decomp:constituent [ decomp:correspondsTo
            [ ontolex:canonicalForm
                  [ ontolex:writtenRep "artificielle"@fr]
            ]
    ] .
    ?le ontolex:sense [ diaterm:temporalReference [ time:hasTime ?t ] ]
    OPTIONAL { ?t time:hasBeginning [ time:year ?start ] }
    OPTIONAL { ?t time:hasEnd [ time:year ?end ] }
    FILTER((1883<COALESCE(?start, 0) && 1940>COALESCE(?start, 0)) ||
           (1883>COALESCE(?start, 0) && 1940<COALESCE(?end, 0)) ||
           (1883<COALESCE(?end, 0) && 1940>COALESCE(?end, 0)))
}
```

Figure 6. Requête SPARQL 2 : termes complexes contenant artificiel dans la période 1884-1940.

L'interrogation donne cinq termes :

- terme 1 : *fibre artificielle*_1 « fibre manufacturée » ;
- terme 2 : *fibre artificielle*_2 « fibre manufacturée issue du traitement chimique des polymères naturels » ;
- terme 3 : *soie artificielle*_1 « fibre manufacturée issue du traitement chimique des polymères naturels » ;
- terme 4 : *soie artificielle*_2 « fibre manufacturée cellulosique » ;
- terme 5 : *soie artificielle*_3 « fibre manufacturée cellulosique sous forme de filaments continus ».

Exploitant les deux niveaux de la ressource, linguistique et conceptuel, l'utilisateur peut chercher les termes désignant un concept *C* au cours de la période *T*. Ainsi la requête SPARQL représentée dans la Figure 7 vise-t-elle à obtenir les termes utilisés pour désigner le concept « fibre manufacturée cellulosique obtenue à partir d'une solution cupro-ammoniacale » dans deux périodes différentes, avant 1940 et après 1940.

A. t= ? (C 'fibre manufacturée cellulosique obtenue à partir d'une solution cupro-ammoniacale', T = 1884 – 1940)

```
SELECT ?period
(GROUP_CONCAT(concat(str(?term), "@", lang(?term));SEPARATOR="\n") AS ?terms)
WHERE {
    BIND("before 1940" as ?period)
    ?le ontolex:canonicalForm [ ontolex:writtenRep ?term ] .
    ?le ontolex:sense [ diaterm:temporalReference ?refEvent ] .
    ?refEvent diaterm:temporalReference [ rdf:type fibreOnto:ID15_CUPRO] ;
          time:hasTime ?t .
    OPTIONAL { ?t time:hasBeginning [ time:year ?start ] }
    OPTIONAL { ?t time:hasEnd [ time:year ?end ] }
    FILTER((0<COALESCE(?start, 0) && 1940>COALESCE(?start, 0)) ||
           (0>COALESCE(?start, 0) && 1940<COALESCE(?end, 0)) ||
           (0<COALESCE(?end, 0) && 1940>COALESCE(?end, 0)))
} GROUP BY ?period
```

B. t = ? (C 'fibre manufacturée cellulosique obtenue à partir d'une solution cupro-ammoniacale', T = 1940 – présent)

```
SELECT ?period
(GROUP_CONCAT(concat(str(?term), "@", lang(?term));SEPARATOR="\n") AS ?terms)
WHERE {
    BIND("after 1940" as ?period)
    ?le ontolex:canonicalForm [ ontolex:writtenRep ?term ] .
    ?le ontolex:sense [ diaterm:temporalReference ?refEvent ] .
    ?refEvent diaterm:temporalReference [ rdf:type fibreOnto:ID15_CUPRO] ;
          time:hasTime ?t .
    OPTIONAL { ?t time:hasBeginning [ time:year ?start ] }
    OPTIONAL { ?t time:hasEnd [ time:year ?end ] }
    FILTER((1940<COALESCE(?start, 0) && 2024>COALESCE(?start, 0)) ||
           (1940>COALESCE(?start, 0) && 2024<COALESCE(?end, 0)) ||
           (1940<COALESCE(?end, 0) && 2024>COALESCE(?end, 0)))
} GROUP BY ?period
```

Figure 7. Requête SPARQL : termes désignant le concept « fibre manufacturée cellulosique obtenue à partir d'une solution cupro-ammoniacale » avant et après 1940.

La recherche fournit quatre termes pour le premier intervalle et trois termes pour le deuxième, les deux ensembles n'ayant en commun que le terme *cupro* :

- [T = 1884 – 1940] : cupro, rayonne cupro-ammoniacale, soie cupro-ammoniacale, soie au cuivre ;
- [T = 1940 – présent] : cupro, CU, CUP.

Ce type de requête peut être construit de manière plus ciblée en demandant des informations précises relatives à la période où les termes repérés sont ou ont été utilisés. Relativement au concept « fibre de cellulose fabriquée par le procédé viscose sous forme de filaments continus », par exemple, la requête SPARQL de la Figure 8 a permis d'identifier trois termes, en leur associant la période dans laquelle ils sont /ont été utilisés pour désigner le concept en question, à savoir, dans l'ordre chronologique :

- terme 1 : *soie artificielle*, période : 1898 – 1935 ;
- terme 2 : *rayonne*, période : 1935 – 1976 ;
- terme 3 : *viscose*, période : 1976 – présent.

```
SELECT ?term ?start (COALESCE(?_end, "today") as ?end)
WHERE {
    ?le ontolex:canonicalForm [ ontolex:writtenRep ?term ] .
    ?le ontolex:sense [ diaterm:temporalReference ?refEvent ] .
    ?refEvent diaterm:temporalReference ?concept ;
          time:hasTime ?t .
    ?concept rdfs:subClassOf [ a owl:Restriction ;
                               owl:onProperty fibreOnto:resultingFrom ;
                               owl:allValuesFrom fibreOnto:PROCÉDÉ_VISCOSE  ] ;
             rdfs:subClassOf [ a owl:Restriction ;
                               owl:onProperty fibreOnto:hasForm ;
                               owl:allValuesFrom fibreOnto:FILAMENTS_CONTINUS  ] .
    OPTIONAL { ?t time:hasBeginning [ time:year ?start ] }
    OPTIONAL { ?t time:hasEnd [ time:year ?_end ] }
}
```

Figure 8. Requête SPARQL : termes désignant le concept « fibre de cellulose fabriquée par le procédé viscose sous forme de filaments continus » avec les périodes de leur emploi.

Les quatre requêtes présentées montrent le potentiel de notre ressource terminologique à fournir des réponses précises à des interrogations complexes, de nature linguistique et/ou conceptuelle, tout en considérant la dimension diachronique des terminologies.

6. Conclusion

Dans cet article, nous avons présenté la modélisation des données terminologiques concernant les fibres textiles, depuis l'émergence des fibres manufacturées à la fin du XIX^e^ siècle jusqu'à aujourd'hui, dans le cadre d'une ressource plus large en cours de construction. Conçue selon le modèle DIATERM, qui permet de représenter l'évolution temporelle de la terminologie aux niveaux linguistique et conceptuel, cette ressource vise à favoriser la préservation et la transmission des savoir-faire dans ce domaine spécialisé, en réponse aux besoins de différents acteurs du secteur textile, allant des chercheurs et développeurs aux professionnels du marketing. Face aux enjeux croissants liés à l'empreinte écologique de l'industrie textile, qui suscitent de plus en plus de débats tant au niveau local qu'international, notre base de données accorde une attention particulière à la description des aspects relatifs

à la durabilité des fibres textiles, fournissant ainsi des informations précieuses pour les consommateurs éco-responsables.

D'un point de vue technique, la construction de la ressource a permis de proposer des solutions innovantes, notamment dans la modélisation des dimensions temporelles. Comme démontré dans cet article, il est essentiel d'attribuer une validité temporelle non seulement aux relations (par exemple, la dénotation d'un concept), mais aussi aux entités elles-mêmes, qui peuvent exister ou non sur une période donnée.

La solution proposée par le modèle DIATERM, introduit ici, repose sur le paradigme de la réification, ce qui permet de modéliser les dimensions temporelles dans le cadre des Données Liées, et ainsi de développer une ressource conforme aux principes FAIR. Cependant, cette approche présente certaines limitations en termes d'inférence automatisée, principalement dues à la représentation des propriétés en tant que classes.

À l'avenir, nous prévoyons d'explorer des solutions alternatives pour améliorer l'équilibre entre expressivité et raisonnement, tout en enrichissant la ressource avec de nouveaux concepts et termes relatifs aux fibres textiles, afin d'élargir sa couverture et son utilité dans divers contextes.

Références

Artale, Alessandro, Franconi, Enrico. 2000.« A survey of temporal extensions of description logics. » *Annals of Mathematics and Artificial Intelligence* 30 (1–4) : 171–210.

Agulhon, Henri. 1962. *Les textiles chimiques*. Paris : Presses universitaires de France, coll.« Que sais-je ? ».

Batsakis, Sotiris, Petrakis, Euripides, Tachmazidis, Ilias and Antoniou, Grigoris. 2017.« Temporal representation and reasoning in OWL 2 » *Semantic Web* 8 (6) : 981–1000.

Baum, Maggy et Boyeldieu, Chantal. 2018. *Dictionnaire encyclopédique des textiles*. Paris : Eyrolles.

Berners-Lee, Tim, Hendler, James, Ora, Lassila. 2001.« The semantic web. » *Scientific American* 284 (5) : 34–43.

Bizer, Christian, Heath, Tom, Berners-Lee, Tim. 2011.« Linked data: The story so far ». Dans *Semantic services, interoperability and web applications: emerging concepts*, edited by Amit Sheth, 205–227. Hershey, Pennsylvania : IGI Global. 1st edition.

Buitelaar, Paul. 2010.« Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions ». Dans *Ontology and the Lexicon: A Natural Language Processing Perspective. Studies in Natural Language Processing*, edited by Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari, Laurent Prévot, 212–223. Cambridge : Cambridge University Press.

Chiarchos, Christian, Moran, Steven, Mendes, Pablo N., Nordhoff, Sebastian, Littauer, Richard. 2013.« Building a Linked Open Data cloud of linguistic resources: Motivations and developments ». Dans *The People's Web Meets NLP: Collaboratively Constructed Language Resource*s, edited by Iryna Gurevych and Jungi Kim, 315–348.

Dankova, Klara. 2023. *Les fibres textiles entre synchronie et diachronie : études terminologiques*, Bern : Peter Lang, coll.« Linguistic Insights ».

DGE/UBIFRANCE. 2006. *Textiles Techniques. Le futur se tisse en France*. Paris : ministère de l'Économie, de l'industrie et de l'emploi.

Fauque, Claude et Bramel, Sophie. 1999. *Une seconde peau : fibres et textiles d'aujourd'hui*. Paris : Éditions Alternatives.

Flouris, Giorgos, Manakanatas Dimitris, Kondylakis Haridimos, Plexousakis Dimitris, Antoniou Grigoris. 2008.« Ontology Change: Classification and Survey. » *The Knowledge Engineering Review* 23, 117–52.

Kanmani, A. Clara, Chockalingam, Thiagarajan, Guruprasad, Nagraj. 2016.« RDF data model and its multi reification approaches: A comprehensive comparative analysis ». Dans *International Conference on Inventive Computation Technologies (ICICT)* 1, 1–5.

Klein, Michel CA., Fensel, Dieter. 2001.« Ontology versioning on the Semantic Web ». Dans *Proceedings of the First International Conference on Semantic Web Working*, 75–91. California : CEUR-WS.org.

Lenci, Alessandro, Bel, Nuria, Busa, Federica, Calzolari, Nicoletta, Gola, Elisabetta, Monachini, Monica, Ogonowski, Antoine, Peters, Ivonne, Peters, Wim, Ruimy, Nilda, Villegas, Marta and Zampolli, Antonio (2000).« SIMPLE: A general framework for the development of multilingual lexicons ». *International Journal of Lexicography* 13(4) : 249–263.

Lutz, Carsten, Wolter, Frank, Zakharyaschev, Michael. 2008.« Temporal Description Logics: A Survey ». In *Proceedings of the Fifteenth International Symposium on Temporal Representation and Reasoning*, 3–14. IEEE Computer Society Press.

Makri, Polyxeni. 2011.« 4D-Fluents Plug-In: A Tool for Handling Temporal Ontologies in Protégé. » Master's Thesis. Department of Electronic and Computer Engineering, Technical University of Crete.

McCrae, John P., Bosque-Gil, Julia, Gracia, Jorge, Buitelaar, Paul, Cimiano, Philipp. 2017.« The OntoLexLemon Model: Development and Applications ». Dans *Electronic lexicography in the 21st century*. Proceedings of eLex 2017 conference, September 19-21, Leiden, Netherlands. Lexical Computing CZ s.R.O. Brno, Czech Republic, 19–21.

Mehta, Sunidhi. 2023.« Biodegradable Textile Polymers: A Review of Current Scenario and Future Opportunities ». *Environmental Technology Reviews* 12 (1) : 441–457. Disponible sur : https://www.tandfonline.com/doi/full/10.1080/21622515.2023.2227391#abstract [consulté le 16 novembre 2024].

Miraftab, Mohsen, Qiao, Q., Kennedy, John F., Anand, Subhash C. et Collyer, Graham. 2001.« Advanced materials for wound dressings: biofunctional mixed carbohydrate polymers ». Dans *Medical Textiles. Proceedings of the 2nd international Conference, 24th and 25th August 1999, Bolton Institute, UK*, édité par Subhash C. Anand, 164–172. Abington : Woodhead Publishing Ltd.

Niu, Jiantao, Xiaoying Zhang, Nana Pei et Qilei Tian. 2012.« Biodegradability of cellulose fibers and the fabrics in activated sludge ». *Applied Mechanics and Materials* (217–219) : 918–922. Disponible sur : https://www.scientific.net/AMM.217-219.918 [consulté le 16 novembre 2024].

Noy, Natasha, Rector, Alan, Hayes, Pat, and Welty, Chris. 2006. "Defining N-ary Relations on the Semantic Web". *W3C Working Group Note* 12(4).

Piccini, Silvia, Bellandi Andrea, Giovannetti Emiliano. 2021.« A Model for Representing Diachronic Terminologies: the Saussure Case Study. » *Digital Humanities Quarterly* 15(3).

Piccini, Silvia, Ruimy, Nilda, Giovannetti, Emiliano. 2016.« Le lexique électronique de la terminologie de Ferdinand de Saussure : une première ». Dans D. Trotter, A. Bozzi, C. Fairon (dir.), *Actes du XXVII^e Congrès international de linguistique et de philologie romanes, Nancy*, 15-20 juillet 2013. Section 16 : Projets en cours ; ressources et outils nouveaux, 255–267. Nancy: ATILF.

Première Vision Yarns. 12-14 février 2019 [catalogue des exposants].

Weidmann, Daniel. 2010. *Aide-mémoire Textiles techniques*. Paris : Dunod.

Welty, Chris, Fikes, Richard, Makarios, Selene. 2006.« A reusable ontology for fluents in OWL ». In *FOIS (Formal Ontology in Information Systems) Proceedings* 150, 226–236.

Zanola, Maria Teresa. 2016.« L'espace du concept, la parole de l'image : pour une typologie des représentations non-verbales dans la terminologie des tissus ». Dans *Verbal and non verbal representation in terminology: Proceedings of the TOTh 2013*, édité par Susanne Lervad *et al.*, 65–80. Copenhague : DNRF's Centre for Textile Research, Institut Porphyre, Savoir et connaissances.

Zanola, Maria Teresa. 2018.« La terminologie des arts et métiers entre production et commercialisation : une approche diachronique ». *Terminàlia* 17 : 16–23.

Zanola, Maria Teresa. 2019.« Néologie de luxe et terminologie de nécessité. Les anglicismes néologiques de la mode et la communication numérique ». *Neologica* 13 : 71–83.

Sitographie

Fulgar. *AMNI SOUL ECO®*. Disponible sur : https://www.fulgar.com/prodotti/60/amni-soul-eco [consulté le 16 novembre 2024].

Orange Fiber. Disponible sur : https://orangefiber.it/it/ [consulté le 16 novembre 2024].

Parlement européen. *Comment parvenir à une économie circulaire d'ici 2050 ?*. Disponible sur : https://www.europarl.europa.eu/topics/fr/article/20210128STO96607/comment-parvenir-a-une-economie-circulaire-d-ici-2050 [consulté le 16 novembre 2024]

Des représentations du monde des technologies de l'information : un travail terminologique en patrimoines scientifiques et techniques

Caroline Djambian*, Micaela Rossi**, Giada D'Ippolito***

*Université Grenoble Alpes, Laboratoire GRESEC (Groupe de recherche sur les enjeux de la communication)
caroline.djambian@univ-grenoble-alpes.fr

**Università di Genova, Dipartimento di Lingue e culture moderne
micaela.rossi@unige.it

***Università di Genova
giadadippolito30@gmail.com

Résumé. La sauvegarde du patrimoine des technologies de l'information (TI) est une démarche d'intérêt général. C'est à quoi s'attache le projet ITinHeritage partant d'une approche patrimoniale basée sur les musées de l'informatique en France et à l'international. L'étude inédite de ce patrimoine nous amène à comprendre de façon plus large, comment pérenniser et médiater les savoirs scientifiques et techniques contemporains. Pour ce faire, il est impératif de comprendre les représentations du monde des TI et les dénominations des objets qui le peuplent. Nous y œuvrons dans cette première étape du projet ITinHeritage, en situant notre étude dans une approche innovante en Humanités Numériques qui mobilise des méthodes et outils issus d'approches linguistiques traditionnelles et d'explorations en sciences de l'informatique. Les TI sont ainsi mises au service de l'étude, de la valorisation et de la pérennisation de leur propre patrimoine.

Abstract. *Safeguarding our information technology (IT) heritage is in the public interest. This is the aim of the ITinHeritage project, which takes a heritage-based approach through the IT museums in France and around Europe. The unprecedented study of this heritage leads us to a broader understanding of how to perpetuate and mediate contemporary scientific and technical knowledge. To do this, it is essential to understand how the world of IT is conceptually represented, starting*

with the study of the names of the objects that populate it. We are working on it in this first stage of the ITinHeritage project, by situating our study within an innovative Digital Humanities approach that mobilizes methods and tools derived from traditional linguistic and computing sciences approaches, and explorations in Artificial Intelligence. In this way, IT is being used to study, enhance, and perpetuate its own heritage.

1. Le patrimoine : entre objet, langage et représentations

Le patrimoine est essentiellement vu dans la matérialité d'objets, outils de mémoire des individus et sociétés. Cela vient du fait qu'il est un objet politique que l'Occident a longtemps imposé sous la forme d'une véritable emphase du matériel, du monumental et de l'esthétique. Également, parce que le patrimoine matériel est plus facilement accessible et gérable, parce qu'aussi, il légitime, valorise, donne autorité…

Mais si nous surmontons cette obsession du matériel pour nous intéresser à l'immatériel, nous ouvrons un espace conceptuel infini dans lequel on peut considérer non seulement la nature du patrimoine, mais aussi les usages qui en ont été faits et son ancrage dans la société. Car il n'est pas seulement un recueil d'objets matériels inertes et figés dans le temps. Il est aussi vivant, dans des moments, actions, pratiques, autour de ces objets, qui sont eux-mêmes des « lieux » de patrimoine. C'est là qu'il devient compréhensible en tant qu'expérience, acte social, culturel, scientifique et technique. Le patrimoine est encore la construction d'une identité qu'il incarne et transmet à travers des représentations du monde détenues par les experts « passeurs du passé ». Les objectivations de cette identité nous en fournissent, par leur matérialité, les bribes intangibles des représentations qui sous-tendent ledit patrimoine. C'est cette tension entre l'action dans la pratique sensorielle, l'objet dont on fait l'expérience et les représentations qui en émergent et s'actualisent sans cesse, à leur tour, dans la pratique, qui fonde en réalité un patrimoine. Cette tension dynamique se concrétise dans l'acte de mise en termes. Le langage fait du patrimoine un processus culturel dans lequel les concepts et leurs noms ne cessent de se créer, évoluer, disparaître, témoins d'une mémoire qu'ils réactivent et transmettent, narrateurs d'objets qu'ils rendent lisibles.

C'est dans ce contexte que nous ouvrons un chapitre jamais exploré, d'une science qui a pourtant un impact sans précédent sur nos sociétés : les

technologies de l'information (TI). Œuvres fondatrices de notre humanité contemporaine, elles sont définies par l'Unesco (2023) comme l'« ensemble d'outils et de ressources technologiques permettant de transmettre, enregistrer, créer, partager ou échanger des informations… ». Ces objets qui peuplent notre vie privée et professionnelle, deviennent innombrables, et rendent ce domaine difficile à circonscrire, bien qu'innervant, forgeant nos sociétés, car leur évolution trop récente, rapide et massive n'a guère laissé le temps à l'étude et à la patrimonialisation. Se pencher sur ce patrimoine est un enjeu crucial de compréhension du monde actuel et de son évolution sociétale depuis la seconde moitié du XX^{e} siècle.

Le projet de recherche ITinHeritage se base sur une approche résolument patrimoniale et interdisciplinaire, partant des espaces muséaux des TI qui ont amorcé le courageux labeur d'établissement d'une mémoire des technologies de l'information. Le Conservatoire ACONIT (Grenoble, France), le London Science Museum (Londres, Royaume-Uni), le NAM-IP (Namur, Belgique), le Museo degli strumenti per il calcolo (Pise, Italie), le Homecomputermuseum (Helmond, Pays-Bas) et le HNF (Paderborn, Allemagne) participent activement à notre réseau en expansion.

Notre projet vise ainsi à mener une réflexion épistémologique sur la question des TI en mettant au cœur de la question du patrimoine, l'histoire d'une mutation sociétale et l'émergence de nouveaux savoirs en termes de représentations du monde. Nous partons pour ce faire de la définition encore inexplorée de ce patrimoine labile, sous la forme de cette tension qui fonde un patrimoine : action-objet-représentation. Les objets du patrimoine des TI ont une dimension technologique forte qui les rend peu lisibles. Pourtant « il faut attribuer à l'objet technique un statut ontologique à côté de celui de l'objet esthétique ou de l'être vivant, en comprenant le sens de sa genèse » (Simondon, 1958), car il témoigne de la relation de l'homme à la technique et plus généralement, au savoir et au monde. Les pratiques de l'homme autour de ces objets, relèvent de la connaissance empirique, l'*emperia* (Kant) ou *métis* (Aristote), qui est « ésotérique », car tacite et révélée seulement dans l'acte sur l'objet par son porteur ; mais elles relèvent aussi de la « connaissance pure » (raisonnement) (Kant, 1997) « exotérique » (Jacob, 2001), concrétisée dans l'acte de formalisation dans l'écrit ou l'objet. Ces trois formes de réalités du monde des TI – connaissances ésotériques, formalisation dans l'écrit, formalisation dans l'objet – en portent les représentations. Le langage de spécialité est l'expression de ce patrimoine. C'est sur lui que nous fondons notre étude.

C'est dans cette logique que le projet ITinHeritage s'axe sur l'interaction entre approche linguistique et ontologique. Ce lien n'est pas un thème nouveau. Dans cette perspective théorique, Sager (1990) adopte une approche onomasiologique, mais basée sur des corpus des collections de données lexicales, adaptée à la création de vocabulaires linguistiques spéciaux. La théorie descriptive/socioterminologique en France, théorisée par Gaudin (1993) et Boulanger (1995), amène à l'étude de l'usage réel du langage, l'importance des influences extérieures et l'évolution diachronique du langage. Cette dernière étape a ensuite été valorisée dans la théorie cognitive, qui étudie les processus cognitifs dans l'utilisation et le traitement du langage. Son principal représentant est Geeraerts, qui s'intéresse principalement à l'évolution des mots dans le langage général. Dans notre cas, nous nous projetons dans les travaux de Temmerman (2000) et sa formulation de l'*unité de compréhension* (UoU), c'est-à-dire des catégories dont les structures prototypiques évoluent sans cesse, ce qui nécessite de prendre en compte la nature conceptuelle des terminologies, ainsi que dans le sillage des recherches de Christophe Roche et de son approche onto-terminologique (Roche, Papadopoulou, 2019).

2. Rassembler la mémoire des TI : le graphe de connaissances

Le mariage heureux du numérique et du patrimoine fait des données les nouvelles collections, mais aussi les nouveaux corpus. Notre travail de sauvegarde du patrimoine des TI a ainsi débuté par le recueil des métadonnées des collections des musées partenaires, sous la forme d'un graphe de connaissances, dans le paradigme du Web Sémantique (Knowledge Graph), qui favorise les liens sur le web avec d'autres données ouvertes et qui est facilement augmentable par l'arrivée de nouvelles données. Creuset du patrimoine des TI, notre Knowledge Graph a été élaboré sur l'environnement Neo4j, après un lourd travail d'harmonisation des métadonnées disparates des musées (formats natifs XML, PDF…, champs libres), puis la mise aux standards CSV (Comma-separated values), l RDF (Resource Description Framework) et FAIR (Findable, Accessible, Interoperable, Reusable), afin de les ouvrir et les lier sur le web (Linked Open Data, LOD).

Le graphe de connaissances est structuré par une ontologie de premier niveau, construite sur la base des deux méta-modèles fondateurs dans le domaine patrimonial : CIDOC-CRM (Comité International pour la DOCumentation du Conseil international des musées, Conceptual Reference Model) et EDM (Europeana Data Model) sur l'environnement Protégé. Cette première ontologie n'a pour vocation que d'organiser les champs des

métadonnées contenues dans le graphe de connaissances et n'est, à ce stade, pas une ontologie de domaine.

L'alignement entre les deux méta-modèles a été réalisé en utilisant SKOS (Simple Knowledge Organisation System ou Système simple d'organisation des connaissances), qui est un langage de représentation d'arborescences conceptuelles recommandé par le W3C (World Wide Web Consortium) notamment, dans l'alignement de thésaurus avec des ontologies. Cet alignement est représenté dans la figure suivante, où les classes EDM indiquées en vert, sont mises en correspondance avec les classes CIDOC-CRM, en orange. En noir sont indiquées les classes de notre propre ontologie. Ici, seules les sept principales classes de l'EDM ont été considérées, les métadonnées de nos musées partenaires n'en nécessitant pas d'avantage.

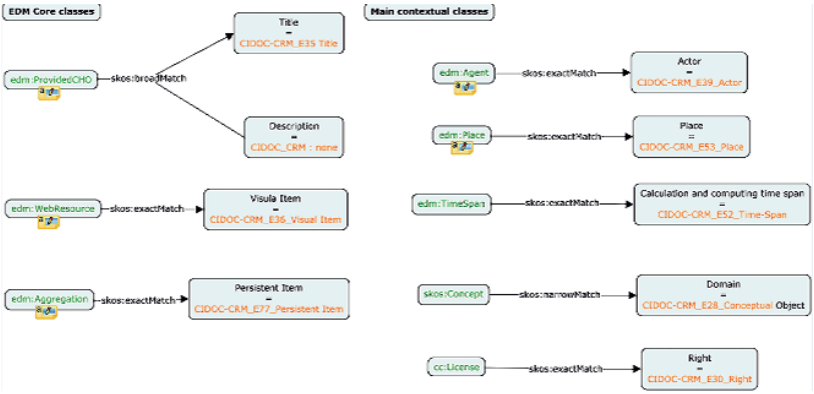


Figure 1. Alignement des méta-ontologies CIDOC-CRM et EDM au sein de l'ontologie ITinHeritage de premier niveau.

Le graphe de connaissances compte à ce jour plus de 25 500 artefacts et est amené à croître. Il pérennise et ouvre le patrimoine des TI, en permet la valorisation, et nous offre surtout un corpus textuel de choix, représentatif des nouveaux corpus auxquels les terminologues sont aujourd'hui confrontés, issus des *big data*, mais avec dans notre cas, des données harmonisées, dont les termes sont déjà relativement normés et travaillés en amont par des experts, pour l'organisation et la représentativité des collections muséales. Ce corpus multilingue (français, anglais, allemand et italien) est pour nous la base d'une étude des discours du domaine.

3. Une première approche sémasiologique basée sur l'analyse de corpus

L'étude de la terminologie des TI, à travers son sous-domaine qu'est l'informatique, est sans doute l'un des domaines qui a le plus intéressé les terminologues, qui se sont penchés sur ses aspects morphologiques et sémantiques (entre autres, voir Claveau & L'Homme, 2005), sur ses usages dans les discours spécialisés et sur sa dimension textuelle, avec une référence particulière à la linguistique de corpus (pour une première étude fondamentale, voir Condamines, 2005), et sur les questions liées à la comparaison interlinguistique (entre autres, Humbley, 2005). L'Homme (2008) qui compare les ressources terminologiques, ontologiques et générales dans la description de la terminologie informatique, est particulièrement intéressant dans notre perspective. Le projet ITinHeritage est basé sur le patrimoine des TI, profondément interdisciplinaire par nature, et le dialogue entre les disciplines présuppose une interrogation sur des notions fondamentales telles que le terme et le concept à partir de différents points de vue – à savoir, à partir d'une approche basée sur la connaissance et à partir d'une approche basée sur la langue. Comme l'affirme L'Homme, «*given the differences between the assumption of lexical-based and knowledge-based approaches and the principle on which they rely, the question is whether they can be used simultaneously in terminology work*» (L'Homme, 2022). Dans cette perspective, le projet ITinHeritage peut être considéré comme une tentative de réponse à cette question, par son orientation vers une vision patrimoniale et interdisciplinaire. La question principale qui nous intéresse à ce stade du projet est l'identification des unités terminologiques et des relations lexicales, nécessaires à la structuration des connaissances du domaine ; ainsi, notre analyse s'inscrit-elle dans la continuité de nombreuses études dans le domaine de la terminologie, qui s'interrogent sur l'interface entre la terminologie et les approches basées sur les connaissances (L'Homme, 2005 ; Malaisé, 2005 ; Meyer, 2022).

Dans un premier temps, nous postulons que l'étude des usages linguistiques (Rossi, 2021) dans le domaine du patrimoine des TI nous permettra d'identifier des noms de concepts et leurs significations, socialement stabilisés au sein des différentes communautés qui fondent ce domaine, en identifiant les relations évoquées plus haut. Cette analyse fine permettra à terme de dynamiser une dialectique entre les experts, la science et le public, et d'observer l'évolution conjointe de la langue et de la technologie (Djambian *et al*., 2023).

Le projet ITinHeritage adopte donc une méthodologie basée sur l'intégration de deux approches parallèles : l'approche sémasiologique, axée sur l'analyse linguistique et l'approche onomasiologique, visant la construction d'une représentation ontologique de domaine formelle. L'objectif est celui de développer une onto-terminologie (dans le sillage des recherches de l'équipe Condillac, dirigée par Christophe Roche) conforme aux normes internationales (ISO 704 et 1087), capable de décrire de manière structurée et interopérable les concepts liés au patrimoine historique des technologies de l'information.

Afin de valoriser ce patrimoine, le projet combine l'approche de l'onto-terminologie avec les développements récents de la terminologie cognitive (Geeraerts, 1993 ; Temmerman, 2000 ; Faber, 2022), dans le but de décrire une terminologie qui reflète les catégories prototypiques et l'évolution des concepts, en tenant compte des variations diachroniques et des nuances conceptuelles au fil du temps.

Ce travail se base sur le graphe de connaissances composé des métadonnées de plus de 25 500 objets patrimoniaux. Ce corpus va nous permettre d'organiser la terminologie du secteur des TI. Comme nous l'avons signalé, il s'agit d'un corpus inédit, singulier dans ses caractéristiques, mais dense en « contextes riches en connaissances » (Meyer, 2001). Pour éviter les distorsions dues à la nature particulière des données muséales, un corpus complémentaire a été créé à partir de sources autorisées et normalisées sur l'histoire de l'informatique à seule fin de consultation et non d'extraction, en respect du droit des auteurs (Mounier-Kuhn & Lazard, 2022 ; Haigh & Ceruzzi, 2021). La sélection du corpus a été guidée par un examen approfondi de l'équilibre entre la pertinence, l'exhaustivité, l'actualité et l'originalité (Cabré, 1999), en maintenant un équilibre entre le langage contrôlé des musées et la terminologie spécialisée.

Les critères de sélection des textes et d'extraction de la terminologie sont conformes aux critères définis dans la norme ISO/DIS 5078:2023. Le processus d'extraction, applicable à toutes les langues incluses dans le projet (anglais, français, allemand et italien), suit une approche semi-automatique, combinant des outils d'extraction et des interventions manuelles de nettoyage lexical. La bibliothèque NLTK du langage de programmation Python a permis un nettoyage préliminaire du lexique et une première extraction de termes candidats (mots-clés, bigrammes et syntagmes nominaux).

Pour une première extraction des termes simples et complexes, le logiciel TermoStat a été utilisé, en sélectionnant des catégories lexicales telles

que les noms, les adjectifs et les verbes. Une fois extraits, ces termes ont été importés dans Microsoft Excel, où ils peuvent être organisés et analysés plus en profondeur. Dans le tableau ci-dessous, nous présentons un exemple de termes extraits de la collection du Conservatoire ACONIT de Grenoble.

Candidat de regroupe	Fréque	Spécificité	Variantes orthographiques	Matrice
clavier	502	27815	clavier___claviers	Nom
connecteur	250	23309	connecteur___connecteurs	Nom
disquette	265	22612	disquette___disquettes	Nom
unité centrale	199	20343	unité centrale___unités centrales	Nom Adjectif
bouton	233	18025	bouton___boutons	Nom
lecteur de disquette	148	17992	lecteur de disquette_lecteurs de disquette_lecteurs de disque	Nom Préposition Nom
micro-ordinateur	160	17692	micro-ordinateur___micro-ordinateurs	Nom
touche	316	17322	touche___touches	Nom
imprimante	129	15559	imprimante___imprimantes	Nom
boîtier	100	14742	boîtier	Nom
machine	427	14681	machine___machines	Nom
bit	104	14364	bit___bits	Nom
boîtier	146	13658	boîtier___boîtiers	Nom
alimentation	221	13642	alimentation___alimentations	Nom

Figure 2. Termes extraits du corpus en langue française, composé ici de la collection du Conservatoire ACONIT de Grenoble.

Cependant, ce sont les expressions régulières et le Corpus Query Language (CQL) de la section Concordance du logiciel d'extraction terminologique Sketch Engine (SkE) qui ont permis d'analyser les mots, les phrases et les documents dans différents contextes possibles, et d'extraire et d'analyser des modèles grammaticaux plus complexes, tels que les acronymes et les séquences de lettres et de chiffres, qui sont fondamentaux pour l'identification de machines et de modèles historiques (par exemple, le « Gamma 30 »).

Les données extraites du corpus ont ensuite, sur la base de l'étude des relations hypéronymiques et méronymiques, permis de dresser une ébauche de réseau lexical.

4. Un travail terminologique basé sur une approche complémentaire

À ce stade du travail terminologique, se sont fait jour les limites d'un travail uniquement basé sur une approche sémasiologique. La densité et la technicité du langage des TI furent un obstacle majeur. Effectivement, l'échelle de complexité d'un langage de spécialité varie énormément d'un domaine à l'autre, d'autant plus pour le profane car, ne perdons pas de vue, que c'est la transmission de ce patrimoine au grand public que nous visons. Ainsi, tous les langages de sciences ne sont pas égaux dans la difficulté d'appréhension

de leurs notions et n'entretiennent pas les mêmes inclusions dans le langage courant. Le langage de spécialité n'est pas un tout. Tant d'obstacles se dressent à la compréhension par le non initié : termes complexes (« ateliers mécanographiques à cartes perforées »), éponymes (« machine de Turing » (qui n'en est, en réalité, pas une au sens du grand public), « loi de Moore »), morphologies lexicales telles qu'acronymes, compliqués de numéros de versions (« IBM 1130 »), caractères spéciaux… Ces termes de spécialité se mêlent à de nombreux termes qui font partie intégrante de nos connaissances, du fait que les TI soient si socialement ancrées dans notre quotidien, tels que « clavier » ou « souris d'ordinateur ».

La complexité terminologique se retrouve donc entre langue commune et langue de spécialité, ce qui peut à la fois apporter des accroches référentielles et un brouillage dénominationnel et notionnel. Mais, ce brouillage peut être accentué par des habitudes langagières différentes au sein de notre même spécialité, sur le plan géographique (diatopie), dans le temps, suivant les évolutions technologiques (diachronie), et entre communautés de pratiques (diastratie). Ainsi, un « calculateur de première génération » pour l'amateur d'informatique, sera dénommé « instrument de calcul mécanique » par l'expert historien de l'informatique ou « calcul antique » par un expert historien généraliste. L'un des objets incarnant ce concept est pour l'initié un « abaque », ou « premier instrument de calcul » ou «*first digital tablet*» en anglais. On comprend immédiatement que le référent sera tout autre pour le grand public anglais du XXIe siècle, pour qui la notion est rattachée à une tablette numérique moderne, alors que l'objet, datant de l'antiquité (on en retrouve jusqu'à 500 av. J.-C.), s'avère en réalité composé de boules et jetons en argile, les «*calculi*» (lat. « petits cailloux ») dont on retrouve l'utilisation jusqu'à 7 000 av. J.-C., servant au calcul arithmétique et ayant évolué en un dispositif à rangs de pièces mobiles, plus connu à compter de 500 av. J.-C., sous le nom de « boulier », le boulier chinois que nous connaissons, datant lui, du XVIe siècle. Les divergences langagières, notamment parmi les experts, tiennent de représentations du monde et de stratégies différentes : « …nous ne manipulons la réalité qu'à travers les représentations que nous en avons » (Roche, 2005). Dans le domaine des technologies de l'information, nous trouvons donc des langues d'usages très diversifiées. D'une part, la langue usuelle, d'autre part les langues de spécialité, qui n'ont pas le même statut et relèvent de réalités extralinguistiques partagées au sein de sous-communautés d'un même domaine.

Ce qui va lier un ensemble d'usages linguistiques variables, correspondant à des pratiques du monde différentes et qui organisent chaque représentation communautaire du monde, est la mise en mots du concept, ce que Robert Lafon (1993) nomme une logosphère : « la logosphère, c'est le maillage du sens enveloppant le réel ». Elle établit un lien direct entre langue et objets du monde réel, elle englobe variétés linguistiques individuelles et normes consensuelles communautaires. Ce constat nous conduit à une inéluctable complétion de l'approche sémasiologique par l'approche onomasiologique. Nous pouvons en effet, déduire de nos précédentes réflexions (Djambian, 2011), et de la tentative d'ébauche d'un réseau sémantique sur la base des extractions lexicales conduites dans le projet ITinHeritage, que le travail terminologique, s'il se borne à l'approche sémasiologique, aboutit à des constructions terminologiques qui reflètent le corpus initial et non une conceptualisation consensuelle. Ces constructions rendent saillantes plus qu'elles n'englobent la variété langagière d'un domaine. La construction d'un réseau sémantique sur la seule base d'extractions lexicales pose alors le problème de ne trouver dans les textes que la variété des dénominations des concepts du domaine et non les concepts eux-mêmes. Ce type de modélisation, bien qu'essentiel pour poser les fondations d'un travail terminologique, s'arrête au sens des mots observés en discours.

Il s'agit donc pour nous, de « mettre en culture les concepts scientifiques » (Gaudin, 1995). Car la matérialité de la langue est tout autant un obstacle que le manque de médiation des objets scientifiques et techniques des TI, pour la communication d'un savoir spécialisé. Les formes linguistiques de la science peuvent s'avérer tout aussi inintelligibles pour le profane, que l'objet qu'elles nomment. Dans la transmission du savoir, c'est donc bien le concept qui est à intégrer, plus que sa dénomination. L'appropriation d'un système lexical ne suffit pas et ne saurait qu'entretenir le manque d'intelligibilité de ce domaine technique, c'est l'acquisition d'un système notionnel que l'on vise.

5. Des représentations du monde des TI : conceptualisation et définition de l'ontologie

Mais comment se fondent les conventions de l'intercompréhension, alors que les dénominations, élaborées socialement, tout comme nos représentations du monde, ne font lien à la notion que par « notre interaction avec les autres locuteurs de la communauté, interaction en vertu de laquelle nous sommes liés au référent lui-même » (Kripke, 1982). Deux éléments sont ici plus stables : l'objet

du monde réel et le concept du monde symbolique. Entre eux deux, évolue le monde de la communication, bien plus labile. Mais la définition exacte d'objets du monde réel, passe pour Kripke, par le fait de substituer « à l'idée de propriétés nécessaires et suffisantes, l'idée d'un faisceau de propriétés, dont seulement quelques-unes doivent être satisfaites dans chaque cas particulier ». Un objet rassemble donc des propriétés essentielles consensuelles, au-delà des propriétés nécessaires que lui attribue de façon variable, chaque communauté de pratique au sein d'un même domaine. Cela revient à un prototype, que Wittgenstein désigne par « air de famille » (1961).

Les objets, en tant que « tout ce qui peut être perçu ou conçu » (ISO 1087), se définissent effectivement, dans la réalité sensible ou intellectuelle, selon les propriétés que l'expérience de la pratique empirique leur forme. Les variétés langagières expriment des représentations du monde différentes, car elles sont basées sur des vérités contingentes liées à leur contexte, à leur communauté. Par une approche méta-linguistique, il est alors possible de définir des propriétés nécessaires et vraies dans tous les mondes possibles aux objets que nous expérimentons, de capter la nature des substances.

La logique antique, si elle s'intéresse aux lois d'un raisonnement bien construit, se penche aussi sur les concepts qui l'animent et sur les mots qui les nomment. En cela, la logique aristotélicienne est fondatrice du travail terminologique moderne. L'objet y est une connaissance singulière, ce sur quoi porte le concept. « On appelle individu Socrate, ou cette chose blanche que voici […]. Les êtres de cette sorte sont appelés individus, parce que chacun d'eux est composé de particularités dont la réunion ne saurait être jamais la même dans un autre être » (Porphyre, 1981). Les concepts reflètent et s'organisent par abstraction, selon ces propriétés qui se traduisent en ensemble de caractéristiques issues du raisonnement logique.

L'organisation des concepts les uns par rapport aux autres au sein du système notionnel, nécessite donc de s'interroger sur l'essence même des choses, leurs caractères eidétiques. À ce propos, Condillac (2023) soulève deux sortes de vérités (en tant que rapport aperçu entre deux idées) qui interviennent dans la formation de nos connaissances. Si « nous nous bornons à observer des qualités qui ne sont pas essentielles aux choses », elles peuvent varier selon la circonstance. Ces vérités sont dites contingentes. Mais si nous raisonnons sur des qualités essentielles aux choses, elles seront vraies hors de toute variation contextuelle et dans tous les mondes possibles. Ces vérités sont dites nécessaires et éternelles. « Les idées distinctes et les vérités nécessaires nous présentent, au contraire, des connaissances exactes et des rapports

appréciés. Elles dévoilent l'essence des choses qu'elles considèrent, elles en développent les propriétés ». Le faisceau de ces caractéristiques essentielles forme un concept de longue durée.

Le concept se définit comme une « unité de connaissance créée par une combinaison unique de caractères » que l'on nomme son intention, et organise les objets en « un groupe en raison de propriétés communes », l'ensemble de ces objets étant son extension (ISO 1087). C'est donc « une connaissance portant sur une pluralité de choses […] répondant à une même loi » (Roche, 2011). « [Les philosophes] ont défini le genre en disant qu'il est l'attribut essentiel applicable à une pluralité de choses ». Il est aussi « ce sous quoi est rangée l'espèce », il est « une sorte de principe pour toutes les espèces qui lui sont subordonnées, et il semble aussi contenir toute la multitude rangée sous lui », ce qui nous ramènera plus bas, à la notion de subsomption (Porphyre, 1981).

Le concept se fonde sur la notion de caractère « propriété abstraite d'un objet ou d'un ensemble d'objets ». On distingue les caractères essentiels, « caractère[s] indispensable[s] pour comprendre un concept », qui en sont l'essence, des caractères distinctifs « caractère[s] essentiel[props] utilisé[s] pour distinguer un concept d'autres concepts associés ». Les types de caractères regroupent les caractères « servant de critère de subdivision lors de l'établissement de systèmes de concepts » (ISO 1087). Roche (2011) signale que le caractère diffère de l'attribut tel qu'il est utilisé en ingénierie des connaissances et bien qu'il y soit associé, en ce qu'il est une description et non une définition du sens du terme, distinction reprise par la logique de Port-Royal : « Il y en a deux sortes [de définitions] : l'une plus exacte, qui retient le nom de définition ; l'autre moins exacte, qu'on appelle description. La plus exacte est celle qui explique la nature d'une chose par ses attributs essentiels, dont ceux qui sont communs s'appellent genre, et ceux qui sont propres, différence. […] La définition moins exacte, qu'on appelle description, est celle qui donne quelque connaissance d'une chose par les accidents qui lui sont propres, et qui la déterminent assez pour en donner quelque idée qui la discerne des autres » (Arnaud et Nicole, 1992).

Il convient à ce stade, d'introduire la notion de principe de différenciation spécifique. « La conceptualisation du domaine repose sur les notions d'objet » (individu ou instance), d'« *attribute* » (caractéristique), de « *concept* » (classe ou prédicat) et de « *relation* » (rôle ou description). Sa construction dépend directement de leur définition – ainsi considérer un concept comme une fonction unaire à valeur de vérité ou comme un ensemble d'attributs valués conduira à des modélisations différentes (Roche, 2011).

Dans son Isagogé, Porphyre (1981) se centre sur la perception que nous avons des objets et comment nous structurons les représentations que nous en faisons. L'objet (individu) est ainsi « composé de particularités dont la réunion ne saurait être jamais la même dans un autre être ». Ces particularités sont héritées par les cinq prédicables (*quinque voces*) ou cinq catégories d'appréhension de notre réalité : le genre, la différence, l'espèce, le propre et l'accident. « Le genre, c'est, par exemple, l'animal ; l'espèce, l'homme ; la différence, le raisonnable ; le propre, la faculté de rire ; l'accident, le blanc, le noir, le "s'asseoir" ».

Porphyre présente dans l'Isagogé, le concept comme regroupant sous le même terme les notions de genre et d'espèce : le genre « est l'attribut essentiel applicable à une pluralité de choses différant entre elles spécifiquement ». Le système notionnel se structure selon des concepts définis en genre et espèces (la définition se compose du genre et de la différence, à l'exclusion des accidents (Porphyre, 1981), dissociés par leurs différences spécifiques, selon une arborescence dite « arbre de Porphyre », où « dans chaque catégorie, il y a certains termes qui sont les genres les plus généraux, d'autres qui sont les espèces les plus spéciales, d'autres enfin qui sont intermédiaires […] qui sont à la fois genres et espèces » et où ce qui est vrai pour un concept l'est également pour ses spécialisations : « Les termes auxquels l'espèce est attribuée recevront aussi nécessairement pour attribut le genre de l'espèce, et le genre du genre, jusqu'au genre le plus général ».

La norme internationale ISO 704 (2001) relative aux principes terminologiques décrit la notion de définition, à la base de toute science, comme devant : « refléter le système de concepts […]. Les caractères retenus dans une définition par intension doivent indiquer les différences qui distinguent les concepts les uns des autres […] ». La base de cette définition trouve de toute évidence son fondement dans les catégories d'Aristote (logique aristotélicienne), que Porphyre s'attache à présenter dans son Isagogé […] : « C'est donc selon les différences qui font la chose autre que se produisent les divisions des genres en espèces et que se forment les définitions, lesquelles se composent du genre et des différences de cette sorte ».

C'est en effet sur la notion de différence que se fondent les catégories d'Aristote, car « la démonstration a pour fonction de démontrer le propre de l'espèce par la difference ». « D'une manière générale, toute différence venant s'ajouter à un être le modifie » (Porphyre, 1981). Les différences peuvent être séparables du sujet ou inséparables : « en effet, se mouvoir, être en repos, se bien porter, être malade, et autres différences similaires, sont

séparables, tandis qu'aquilin ou camus, raisonnable ou irraisonnable, sont des différences inséparables ». Parmi les différences inséparables certaines sont des différences essentielles de l'individu et d'autres le décrivent : « Et parmi les différences inséparables, les unes sont des attributs par soi, et les autres des attributs par accident : le raisonnable appartient par soi à l'homme, ainsi que le mortel et l'apte à recevoir la science, tandis que l'aquilin ou le camus sont des différences accidentelles et non par soi ». « Les différences appartenant par soi au sujet sont comprises dans la définition de la substance et font le sujet autre » et « celles qui le font autres s'appellent spécifiques ». Celles-ci ne peuvent être values : « Les différences par soi n'admettent pas le plus et le moins ». Les différences inséparables par accident, quant à elles, ne sont pas essentielles : « L'accident est ce qui se produit et disparaît sans entraîner la destruction du sujet ». Elles ne visent qu'à décrire le sujet en lui faisant « la qualité autre [et] ne constituent que les diversités et les changements de la façon d'être ». Elles peuvent être values : « les différences par accident, tout inséparables qu'elles puissent être, sont susceptibles d'une intensité plus ou moins grande », comme « être coloré d'une certaine façon, est susceptible d'une intensité plus ou moins grande ».

Le travail terminologique actuel fonde la construction d'une ontologie de domaine sur la différenciation. Le système notionnel est organisé par « un ensemble de concepts structuré selon les relations qui les unissent » (ISO 1087). La relation générique vient compléter heureusement la différenciation spécifique dans la définition d'une ontologie : elle est la « relation entre deux concepts dans laquelle la compréhension de l'un des concepts inclut celle de l'autre concept et au moins un caractère distinctif supplémentaire ». Cette relation, dite aussi « de subsumption » (ou « est une sorte de »), relie un concept spécifique « concept ayant la plus grande compréhension dans une relation générique » à un concept générique « concept ayant la plus petite compréhension dans une relation générique ». L'entière organisation du système notionnel repose alors : sur le plan vertical, la relation générique, et sur le plan horizontal, la différenciation. De la même façon, Porphyre reprend les catégories d'Aristote en divisant les genres en espèces, où l'« espèce se dit de la forme de chaque chose » et « est subordonnée au genre donné », elle est ce « à quoi le genre est attribué essentiellement », et en différences spécifiques, au sein de l'arbre de Porphyre : « Les termes auxquels l'espèce est attribuée recevront aussi nécessairement pour attribut le genre de l'espèce, et le genre du genre, jusqu'au genre le plus general » (Porphyre, 1981). On doit ainsi, « nécessairement dans la définition de l'un se servir de la définition de l'autre ». Car le

genre est aussi « le point de départ de la génération de chaque chose », « tout comme le père lui-même » (on parle aussi de relation parent-enfant).

Sur ces fondements, nous avons construit une ontologie du domaine des technologies de l'information (TI), comme extension de la première ontologie construite sur la base de CIDOC-CRM et qui structure notre graphe de connaissances. L'ontologie n'a pu être élaborée qu'à partir du travail sémasiologique préalablement conçu, de l'étude approfondie du corpus et à l'aide des experts du domaine.

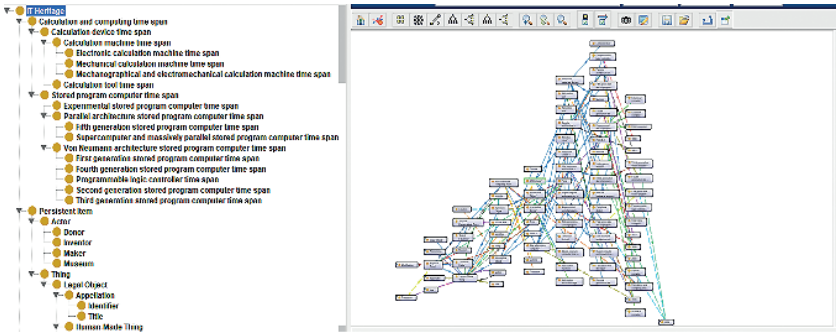


Figure 3. Vue partielle de l'ontologie ITinHeritage du domaine des technologies de l'information (TI).

Cette ontologie du domaine des TI a été alignée avec OWLTIme (Pan & Hobbs, 2005) pour modéliser diachroniquement les connaissances du domaine et représenter l'évolution scientifique et technique. Dans la figure ci-dessous, les classes CIDOC-CRM indiquées en orange sont alignées avec les classes OWLTime en bleu, et celles de notre ontologie ITinHeritage (noir), toujours grâce à SKOS. Les relations entre les classes empruntent préférentiellement au modèle CIDOC-CRM, mais également OWLTime lorsque cela est nécessaire. L'utilisation conjointe des modèles CIDOC-CRM et OWLTime nous permet de répondre à la problématique de représentations des connaissances figées, que connaissent les ontologies. Nous y répondons également, en utilisant l'Intelligence Artificielle (IA) générative pour peupler et étendre automatiquement notre ontologie, grâce aux Large Language Models (LLM) tels que Llama3, malgré les difficultés liées au corpus d'entraînement et aux données annotées. Un Knowledge RAG (Retrieval Augmented Generation), technique d'IA issue du Traitement Automatique des Langues (TAL), basé sur notre

ontologie de domaine, permet de pallier deux problématiques non triviales : d'une part celle du multilinguisme, car ces techniques sont opérationnelles sur toutes les principales langues du monde ; d'autre part, celle de recherche en langage naturel par le grand public, car seules des requêtes en SPARQL (Protocol and RDF Query Language) permettent l'interrogation d'un graphe de connaissances en RDF (Resource Description Framework). L'éditeur Sparnatural (Clavaud & Francart, 2022), que nous avions choisi en premier lieu, permet cette interrogation en utilisant l'ontologie. Mais, en présentant les concepts sous forme d'une liste à plat, il fait perdre tout le sens qu'apportent les relations précédemment citées, à la base du réseau notionnel. *A contrario*, les IA que nous mobilisons ouvrent de nouvelles possibilités d'exploitation des ontologies, en les complétant heureusement, pour un transfert de connaissances plus proche de nos modes de communication naturels.

6. Conclusion

Dans le projet ITinHeritage, il nous tient à cœur de confronter les limites et apports de la technologie pour la modélisation, la pérennisation et la transmission des connaissances expertes, ici, celles qui forment le patrimoine des technologies de l'information. En particulier, les avancées des recherches dans le domaine de l'intelligence artificielle (IA) et de ses applications dans le contexte linguistique, permettent de dépasser les limites existantes concernant le multilinguisme et les variétés langagières en général. Ainsi, repousser les frontières est peut-être le maître mot du projet : ouvrir le patrimoine hors des murs et du temps circonscrits de la visite muséale ; pousser le travail terminologique au-delà de l'approche linguistique, en travaillant conjointement sur les méthodes sémasiologiques et onomasiologiques pour la définition d'ontologies ; ne pas considérer l'ontologie construite à l'instant T comme une fin en soi, mais la rendre évolutive et en faire la base de nouveaux développements d'IA, dans une approche tirant profit de techniques plus anciennes, comme l'IA symbolique et de techniques actuelles, comme l'IA générative. Les réflexions issues du projet permettent de questionner d'une manière innovante, des notions à la frontière entre les disciplines de linguistique, science de l'information et de la communication, et informatique, autour de la notion centrale d'ontologie.

Références

Bourigault, D., Jacquemin, C. 2000. « Construction de ressources terminologiques », dans Jean-Marie Pierrel, *Ingénierie des langues*, Paris : Hermès.

Bourigault, D., Slodzian, M. 1999. « Pour une terminologie textuelle », *Terminologies nouvelles*, nº 19.

Cabré, M.T. 1999. *Terminology. Theory, Methods and Applications*. Amsterdam/Philadelphia : John Benjamins.

Clavaud, F., & Francart, T. (2022). « Sparnatural, un éditeur graphique souple et intuitif pour explorer des graphes de connaissances », *Colloque Humanistica*, Montréal, Canada.

Claveau, V., L'Homme, M.-C. 2005. « Apprentissage par analogie pour la structuration de terminologie-utilisation comparée de ressources endogènes et exogènes », *Actes de la conférence terminologie et intelligence artificielle (TIA-2005).*

Condamines, A. 2005. « Corpus linguistics and terminology », *Langages*, nº 157.

de Boer, V., *et al*. 2013. « Amsterdam museum linked open data ». *Semantic Web*, vol. 4, nº 3.

Djambian, C., Rossi, M., D'Ippolito, G. 2023. « La médiation des objets aux savoirs scientifiques et techniques », *Actes de la conférence TOTh 2023*. Chambéry : Presses Universitaires Savoie Mont Blanc.

Djambian, C. 2011. *Valorisation d'un patrimoine documentaire industriel et évolution vers une base de connaissances orientée métiers*, thèse de Doctorat, Université Lyon 3.

Doerr, M., *et al*. 2010. « The Europeana Data Model (EDM) », *Actes de IFLA 76*, Gottenburgh, Suède.

Faber, P., L'Homme, M.-C. 2022. *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge*, Amsterdam & Philadelphia : John Benjamins.

Gaudin, F. 1995. « Dire les sciences et décrire les sens : entre vulgarisation et lexicographie, le cas des dictionnaires de sciences », *TTR : traduction, terminologie, rédaction*, vol. 8, nº 2.

Geeraerts, D. 1993. « Cognitive linguistics and the history of philosophical epistemology », dans Richard Geiger et Brygida Rudzka-Ostyn, *Conceptualizations and mental processing in language*. Berlin : Mouton de Gruyter.

Haigh, T., Ceruzzi, P. E. 2021. *New History of Modern Computing*, Cambridge : The MIT Press.

Humbley, J., 2005. « La traduction des métaphores dans les langues de spécialité : le cas des virus informatiques », *Linx. Revue des linguistes de l'université Paris X Nanterre*.

ISO. 2019. *ISO 1087-1. Travaux terminologiques. Vocabulaire. Partie 1 : Théorie et application*.

ISO. 2009. *ISO 704 NF ISO 704. Travail terminologique. Principes et méthodes.*

Jacob, C. 2001. « Rassembler la mémoire », *Diogène*, nº 4.

Kant, E. 1997. *Critique de la raison pure*. trad. A. Renaut. Paris : Aubier.

Kripke, S. 1982. *La logique des noms propres*, Paris : Minuit.

Lafont, R. 1993. *Le dire et le faire*, Montpellier : Praxiling.

Lerat, P. 2002. « Qu'est-ce qu'un verbe spécialisé ? Le cas du droit », *Cahiers de lexicologie*, nº 80, 2002.

L'Homme, M.-C. 2008. « Ressources lexicales, terminologiques et ontologiques : une analyse comparative dans le domaine de l'informatique », *Revue française de linguistique appliquée*.

L'Homme, M.-C. 2022. « Terminology and lexical semantics », dans P. Faber, M.-C. L'Homme, *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialised languages*, Amsterdam/Philadelphia : John Benjamins.

L'Homme, M.-C. 2012. « Le verbe terminologique : un portrait de travaux récents », dans F. Neveu *et al*., *Actes du 3e congrès mondial de linguistique française*, Lyon, France.

L'Homme, M.-C. 2022. « Terminology and Lexical Semantics », dans P. Faber, M.-C. L'Homme, *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialised languages*, Amsterdam/Philadelphia : John Benjamins.

Malaisé, V. 2005. *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels*, thèse de doctorat, Université Paris-Diderot-Paris VII.

Meyer, I. 2022. « Concept management for terminology. A knowledge engineering approach », dans P. Faber, M.-C. L'Homme, *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialised languages*, Amsterdam/Philadelphia : John Benjamins.

Pan, F., & Hobbs, J. R. 2005. « Temporal Aggregates in OWL-Time », *FLAIRS*.

Porphyre. 1981. *Isagoge,* Paris : Librairie Philosophique J. Vrin.

Prandi, M. 2023. « De la dimension relationnelle du lexique à la syntaxe : une frontière à retracer », dans A. Roig et A.-G., Toutain, *Congrès mondial de linguistique française. Mélanges offerts à Franck Neveu*, Lyon : ENS éditions.

Putnam, H. 1984. *Raison, vérité et histoire*. Paris : Éditions de Minuit.

Roche, C., Papadopoulou, M. 2019. «Mind the Gap: Ontology Authoring for Humanists», *Proceedings of JOWO: The Joint Ontology Workshops*, Graz, Autriche.

Roche, C. 2011. « L'Isagoge de Porphyre ». *Actes de Terminology & Ontology: Theories and applications, TOTh*, Chambéry : Presses universitaires Savoie Mont Blanc.

Roche, C. 2005. « Terminologie et ontologie ». *Langages*, nº 1.

Rossi, M. 2021. « Termes et métaphores, entre diffusion et orientation des savoirs », *La linguistique*, nº 57.

Simondon, G. 1958. *Du mode d'existence des objets techniques*. Paris : Éditions Aubier-Montaigne.

Smith, L. 2011. « Heritage and its Intangibility », dans Ahmed Skounti y Ouidad Tebbaa, *De l'immatérialité du patrimoine culturel,* Marrakech : Bureau régional de l'Unesco de Rabat.

Temmerman R., 2000. *Towards New Ways of Terminology Description: The Sociocognitive-Approach*, Amsterdam/Philadelphia : John Benjamins.

Unesco. 2023. «*Technologies de l'Information et de la Communication (TIC)* », Glossaire.

Wittgenstein, L. 1961. *Investigations philosophiques*. Paris : Gallimard.

Médiatisation des insectes comestibles dans la presse généraliste : pour une approche communicationnelle et terminologique

Cécile Frérot*, Marcin Trzmielewski**, Yekta Akhbarifar***

*Laboratoire Lidilem, Université Grenoble Alpes, Grenoble
cecile.frerot@univ-grenoble-alpes.fr

**Laboratoire d'Études et de recherches appliquées en sciences sociales
Université Paul Valéry Montpellier 3, Montpellier
marcin.trzmielewski@univ-montp3.fr

***Groupe de Recherche sur les enjeux de la communication (Université Grenoble Alpes), Institut de la communication et des médias (Échirolles)
yekta.akhbarifar@univ-grenoble-alpes.fr

Résumé. À travers une rencontre entre les Sciences de l'Information et de la Communication et la Terminologie, nous nous intéressons aux logiques de communication mises en œuvre dans un corpus de presse généraliste et aux moyens terminologiques mobilisés pour y parvenir. Cette approche interdisciplinaire convoque les méthodes et les outils déployés par chaque discipline pour analyser, dans une approche qualitative, un corpus portant sur l'alimentation à base d'insectes, composé de 190 articles issus de la presse généraliste française publiés entre 1997 et 2023. Les résultats exploratoires montrent que la terminologie spécialisée et institutionnelle, véhiculant les thématiques liées aux questions de la nutrition et de l'environnement, coexiste avec des outils lexicaux et terminologiques de vulgarisation. Une telle « hybridation discursive » conduit à établir un rapport de proximité avec le lecteur, témoignant du succès des aliments à base d'insectes et rendant l'information publiée avérée. Les thématiques relayées dans la presse deviennent ainsi des arguments communicationnels, accompagnés des données spécialisées diffusées par des organismes publics, et soutenant les enjeux des acteurs du développement de la filière d'insectes comestibles en France. Les dangers relatifs à cette filière, tout comme les controverses sont faiblement portés à l'attention du grand public.

Abstract. *In this study, Information & Communication Sciences and Terminology meet with an interdisciplinary view to investigate how a linguistic perspective on terminology can provide insights on the communication logics used in a corpus of French newspaper articles. This interdisciplinary approach draws on the methods and tools used by each discipline to analyze a corpus on insect-based food made up of 190 articles released between 1997 and 2023. The qualitative analysis we carried out leads us to identify a set of linguistic tools involved in constructing the discursive and communication patterns of media coverage on a topic relevant to society, with multiple issues at stake.*

1. Introduction

Ces dernières années, la terminologie textuelle (Bourigault & Slodzian, 1999) s'est ouverte à d'autres disciplines, au-delà des disciplines avec lesquelles elle s'est initialement constituée au début des années 2000 : ingénierie des connaissances, linguistique de corpus et traitement automatique des langues. En témoigne notamment l'émergence d'une linguistique ergonomique (Condamines, 2018) dans le prolongement d'une linguistique située (Condamines & Narcy-Combes, 2005), de la socioterminologie (Gaudin, 2005) et de la pragmaterminologie (De Vecchi, 2004 ; 2016) dans une dynamique discursive. Dans un mouvement plus vaste, l'ouverture de la terminologie à d'autres disciplines est aujourd'hui encouragée (Pecman, 2023) par-delà les disciplines qui ont contribué à sa constitution, notamment la traduction. Dans le contexte de cette ouverture disciplinaire, c'est à travers une rencontre avec les Sciences de l'Information et de la Communication (désormais SIC) que nous proposons de contribuer à cette ouverture.

Le présent article prolonge des travaux initiaux menés en SIC portant sur la médiatisation du sujet des insectes comestibles (Trzmielewski, 2021), réalisés dans le cadre d'un programme CDP intitulé *PUNAISES?, Pour Une Alimentation à base d'insectes : questions scientifiques, économiques et sociétales ?* Dans le cadre de cette recherche, nous avons réalisé des analyses de contenu et lexicométriques d'un corpus de presse généraliste qui ont conduit à identifier les thématiques qui révèlent des formes de médiatisation centrées sur les arguments nutritionnels et écologiques, favorables au développement du marché d'insectes comestibles. Nous avons également montré que les journalistes reprenaient de façon inégale la parole des différents acteurs, privilégiant celles des entrepreneurs du marché par rapport

aux chercheurs et professionnels de santé. Dans le prolongement de ce travail, la présente étude vise à analyser la médiatisation (Lafon, 2019) par le prisme de la terminologie, guidée par une analyse linguistique fondée sur corpus. Nous explorons notamment par la voie terminologique les moyens linguistiques mobilisés pour la construction des tendances discursives reflétant des logiques communicationnelles de la médiatisation. L'exploration porte sur l'alimentation à base d'insectes, au croisement de connaissances spécialisées dans les domaines de l'entomologie, de la biochimie, de la toxicologie, de la nutrition, ou bien encore de l'écologie. L'alimentation à base d'insectes est un sujet porteur d'enjeux sociétaux (Trzmielewski, 2021) et l'approche interdisciplinaire dans laquelle s'inscrit notre travail offre un terrain à explorer. Les travaux ne sont en effet pas légion, et nous citerons à ce titre les recherches d'Hélène Ledouble (2021 ; 2023) sur la médiatisation de la science et la diffusion des connaissances dans le domaine de l'agroécologie.

Dans le présent article, nous nous intéressons aux logiques de communication dans un corpus de presse généraliste et, par la terminologie, nous analysons différents niveaux linguistiques, notamment lexical. Le cadre méthodologique de notre travail repose sur la constitution d'un corpus de presse généraliste ainsi que sur son exploration outillée, au croisement de méthodes et d'outils des deux disciplines convoquées (partie 2). Nous cherchons ainsi à identifier les moyens linguistiques qui participent à la construction des tendances discursives et communicationnelles de la médiatisation (partie 3). La conclusion comporte les apports de l'étude, accompagnés de discussions, puis souligne ses limites et dessine des perspectives de travail qui seront de nature à prolonger cette première étude interdisciplinaire sur les insectes comestibles (partie 4).

2. Méthodologie

2.1. Constitution du corpus

Nous avons effectué l'analyse d'un corpus de presse, obtenu *via* une recherche documentaire et un codage d'articles (avril-juillet 2023) portant sur l'alimentation à base d'insectes. Les articles ont été extraits par une recherche avancée dans le moteur de recherche d'*Europresse* en utilisant les mots-clés « insectes comestibles » et « entomophagie ». Les premiers résultats, filtrés par domaine de recherche (« France »), période (« toutes les archives ») et langue (« fr ») ont comporté 698 articles. La lecture attentive de ces articles, en fonction des critères de sélection prédéfinis a conduit à constituer un corpus de 101 804

mots, comportant 190 articles de presse généraliste française. Ces articles sont parus entre 1997 et 2023 dans une version imprimée et web, majoritairement dans des journaux quotidiens (153 articles), mais également des hebdomadaires (23), des magazines mensuels (12) et bimensuels (2). Un total de 98 articles est issu de la presse locale et 90 de la presse nationale diffusée en France. Ces chiffres, relativement bas par rapport à d'autres sujets peu publicisés (par exemple les exoplanètes ; Lafon, 2022), témoignent de la faible médiatisation du sujet, justifiée par son statut émergent. Soulignons également que les articles de presse ont été rédigés par des journalistes généralistes, certains étant spécialisés dans les rubriques relatives à l'environnement, la santé et l'économie. On note une absence de journalistes spécialisés sur le sujet des insectes comestibles.

L'analyse de contenu du corpus a porté sur trois variables. D'abord, les variables endogènes propres aux caractéristiques descriptives des médias ont été étudiées : le titre, l'auteur, la date, la périodicité, le périmètre couvert et le niveau de spécialisation. Ensuite, nous avons pris en considération les variables liées au contenu, notamment les catégories portant sur les acteurs socioprofessionnels dont la parole a été rapportée par les journalistes : entrepreneurs, chercheurs et professionnels de santé, experts mandatés, consommateurs. Les variables exogènes, enrichissant l'analyse avec des éléments contextuels, ont apporté un cadre politique, économique et socioculturel, notamment.

2.2. Exploration outillée

La nature interdisciplinaire de notre étude se reflète dans les outils utilisés pour analyser le corpus. Tout d'abord, IraMuTeQ (2024 ; Marty *et al.*, 2020), un logiciel libre d'analyse statistique de corpus textuels, a été mobilisé. Nous avons effectué des analyses statistiques et thématiques selon les univers lexicaux : la classification hiérarchique descendante *via* la méthode Reinert (1983), l'analyse factorielle des correspondances, les calculs de chi2 et de corrélation, sur l'intégralité du corpus et sur les sous-corpus créés à partir des catégories d'acteurs émergées précédemment (2.1). L'analyse des articles codés avec les variables relatives à des catégories d'acteurs a conduit notamment à reclasser les segments de texte et séparer les univers lexicaux en fonction des variables données.

Le logiciel d'extraction automatique de terminologie, TermoStat (Drouin, 2003) et la plateforme d'analyse de corpus Sketch Engine (Kilgariff *et al.*,

2014) ont complété le dispositif outillé. Ils ont permis l'exploration de mots et de suites de mots en contexte, donnant accès à leur environnement linguistique et mettant au jour de potentielles unités terminologiques. L'étiquetage morpho-syntaxique, dont ces outils dotent les corpus analysés, a conféré aux analyses menées des possibilités d'exploration lexico-syntaxique complémentaires à IraMuTeQ.

3. Résultats exploratoires

3.1. Le lexique : porteur de thématiques et révélateur d'arguments communicationnels

L'analyse lexicométrique réalisée sur l'intégralité du corpus révèle la coexistence de deux grandes thématiques. D'un côté, les segments du corpus portent sur les produits à base d'insectes, en apportant des descriptions de goût ou encore de type de produits transformés ; de l'autre nous constatons la présence d'arguments nutritionnels et environnementaux en faveur de la commercialisation de ces produits.

Les articles codés par la seule variable « chercheur » font émerger des questions scientifiques, notamment en lien avec la santé et la nutrition alors que dans ceux qui relaient la parole des experts, la majorité des segments renvoie à des questions environnementales. Dans le cas des segments identifiés par la variable « entrepreneur », deux univers lexicaux se distinguent : d'un côté, le lexique révèle des enjeux liés à la commercialisation des insectes comestibles, y compris les stratégies du marketing ; de l'autre, le lexique renvoie également à des questions nutritionnelles et environnementales.

De plus, nos analyses portant sur des sous-corpus montrent que les choix lexicaux varient en fonction des catégories d'acteurs. En effet, les contenus qui rapportent la parole des chercheurs contiennent des marqueurs du discours spécialisé. Nous pouvons observer, par exemple, la présence d'unités lexicales telles qu'*acide, fer, vitamine, minéral* ou encore *protéine.* Le lecteur est ainsi exposé à des questions de nature scientifique associées à des arguments nutritionnels. En revanche, dans les articles où la parole rapportée est celle des entrepreneurs, l'unité lexicale *nutriment* est souvent utilisée. Ces marqueurs spécialisés peuvent être observés dans les contenus reprenant la parole des experts mandatés, de même que chez les chercheurs, lorsqu'il s'agit des questions liées à la santé, bien que ces dernières soient moins souvent abordées par cette catégorie d'acteur. De plus, une recherche manuelle sur les

segments identifiés par la variable « entrepreneur » permet de constater l'utilisation récurrente de l'unité lexicale *besoin*, introduisant les discours sur les besoins nutritionnels qui seraient satisfaits par la consommation d'insectes. Cette unité renvoie également aux discours sur le moindre besoin en espace, eau et nourriture pour l'élevage d'insectes par rapport à l'élevage d'autres animaux, par exemple le bétail. En revanche, dans les articles qui ne relaient pas la parole des acteurs, les journalistes utilisent des unités lexicales telles que *risque*, *danger*, *autorisation* ou *réglementation*, soulignant des enjeux réglementaires en lien avec la commercialisation des insectes comestibles dans l'Union européenne.

3.2. Coexistence terme spécialisé, terme vulgarisé

L'analyse des données de notre corpus montre également la présence d'une combinatoire associant lexique élémentaire et lexique spécialisé. C'est le cas dans l'exemple (a) : *Encore faut-il faire avaler la pilule aux consommateurs, que l'idée d'ingérer ces bestioles protéinées rebute encore* (*Courrier international*, 2022) dans lequel *bestiole protéinée* associe un mot relevant du registre familier (*bestiole*, synonyme de petite bête) au lemme spécialisé *protéine* qui appartient au domaine de la biochimie. Cette combinaison témoigne de la proximité de deux univers lexicaux que nous explorons plus particulièrement en mettant au jour la coexistence de termes spécialisés et de termes vulgarisés. Nous identifions cette coexistence par la présence de marqueurs métalinguistiques qui correspondent à des éléments lexico-syntaxiques pouvant être associés à des éléments typographiques. Dans notre corpus, l'ensemble de ces marqueurs est au service d'une opération de vulgarisation scientifique qui se déploie à travers les articles de presse. Nous l'illustrons à travers des exemples (b) à (d). C'est tout d'abord le cas dans l'exemple (b) : *Il s'est plus spécialement intéressé au ténébrion meunier (tenebrio molitor) plus communément appelé* « *ver de farine* » (*Paris-Normandie*, 2015) où le marqueur verbal *appeler,* précédé de la combinatoire *plus communément*, introduit le terme vulgarisé « *ver de farine* ». L'exemple (c) illustre un phénomène proche avec la combinatoire *aussi appelé* qui apparaît dans une incise en position médiane (que nous soulignons) :

> (c) avait conclu mi-janvier que les larves du ténébrion meunier, aussi appelées « vers de farine », pouvaient être consommées sans danger (*Ouest France*, 2021).

L'exemple (d) est intéressant en cela que le terme vulgarisé est inscrit en premier dans le discours, suivi du terme spécialisé au moyen du marqueur verbal *connaître* dans la combinatoire *aussi connu sous le nom de* (que nous soulignons) :

> (d) Depuis 2014, l'entrepreneur s'est installé dans une zone d'activité anonyme d'un petit village de la région entre Belfort et Montbéliard, où il élève des vers de farine, aussi connu sous le nom de molitors ou ténébrions (*La Croix*, 2019).

Ces marqueurs métalinguistiques mettent en évidence une équivalence conceptuelle et impliquent un changement de degré de spécialisation associant de façon systématique terme spécialisé (voire scientifique) et terme vulgarisé. Cette coexistence terminologique nous semble avoir une double fonction : ancrer le discours dans un environnement scientifique, en diffusant une information avérée au lecteur non spécialiste, et permettre précisément à ce lecteur de se sentir concerné par le sujet, en lui proposant des termes vulgarisés qu'il est davantage susceptible de connaître. En d'autres termes, l'ancrage scientifique et la « captation » du lecteur nous apparaissent ici définitoires des logiques communicationnelles mises en œuvre dans le discours. Cette coexistence pourrait témoigner d'une « hybridation discursive », terme que nous empruntons à Catellani et Errecart (2018), et qui, appliquée à notre étude, combine deux univers au cœur des enjeux de la médiatisation : univers spécialisé et univers vulgarisé.

3.3. Approche analogique

En outre, la mise en correspondance des aliments à base d'insectes avec d'autres aliments connus des non-spécialistes, extérieurs au domaine des insectes, s'opère sur le plan linguistique au moyen d'expressions figées. C'est le cas avec des expressions du registre familier sous forme d'idiotismes alimentaires tels que *partir comme des petits pains* ou *s'arracher comme des petits pains* qui dans les exemples (e) et (f) expriment la vente extrêmement rapide et facile de petits fours à base de vers ou de tartelettes au chocolat et aux vers de farine :

> (e) des petits fours à base de vers sont partis comme des petits pains (*Libération*, 2010) ;
>
> (f) les 140 brownies et tartelettes au chocolat et aux vers de farine, proposés, hier midi, aux demi-pensionnaires du lycée Saint-Jo, se sont presque arrachés comme des petits pains (*Le Télégramme*, 2014).

Cette approche analogique a pour fonction de témoigner du succès des aliments à base d'insectes auprès des lecteurs. Elle permet également de diffuser des arguments aux potentiels consommateurs en recourant notamment à des arguments nutritionnels ; c'est ce qu'illustre l'exemple (g) avec l'apport protéinique et l'analogie établie entre une masse équivalente de criquets et de steak-haché :

> (g) C'est prouvé, il y a cinq fois plus de protéines dans 100 g de criquets que dans 100 g de steak haché (*Le Parisien*, 2014) ;

ou bien encore l'exemple (h) avec l'analogie établie entre la valeur énergétique de 20 g de criquets et un bifteck de 110 g :

> (h) Une dizaine de criquets cuits, soit 20 grammes, correspond à la valeur énergétique d'un bifteck de 110 g ! (*Le Figaro*, 2017).

Des arguments liés à la sécurité alimentaire mondiale sont également mobilisés à travers une mise en correspondance entre la sous-alimentation d'un côté (exemple (i) ci-dessous, *près de 1 milliard d'individus dans le monde ne mangent pas à leur faim*) et le surpoids ou l'obésité de l'autre (*dans les pays développés, une proportion à peu près équivalente souffre de surpoids ou d'obésité*) :

> (i) L'innovation, foisonnante en la matière, pourrait pourtant fournir une alternative alimentaire prometteuse pour l'homme et la planète, alors que près de 1 milliard d'individus dans le monde ne mangent pas à leur faim tandis que, dans les pays développés, une proportion à peu près équivalente souffre de surpoids ou d'obésité (*Le Figaro*, 2017).

Ces analogies nutritionnelles et de sécurité alimentaire nous semblent être utilisées dans le but de représenter la consommation des insectes en tant qu'une pratique légitime. Tout comme nous l'avons vu (3.2), la coexistence terme spécialisé/terme vulgarisé s'inscrit dans le discours médiatique de notre corpus d'étude, les analogies que nous observons participent elles aussi de la logique communicationnelle établie en faveur de la consommation d'aliments à base d'insectes.

Dans l'exemple (g) mentionné précédemment, l'assertion directe non étayée par des données scientifiques (*c'est prouvé*) côtoie des éléments de connaissance vulgarisés faisant appel à un cadre thématique connu du grand public (*protéines – viande – steak-haché*). Afin de moduler l'absence de sources scientifiques, nous observons que la partie scientifique de l'« hybridation discursive » que nous avons évoquée en sollicitant les travaux de Catellani et

Errecart (2018) se déploie plus largement dans notre corpus médiatique par une manifestation de légitimité du discours.

3.4. Sources officielles et terminologie institutionnelle

Dans notre corpus, la légitimité du discours relatif aux insectes comestibles passe par la présence de sources officielles et par la présence d'une terminologie institutionnelle. Les sources sont désignées par des formes développées d'organismes nationaux et internationaux spécialisés dans l'alimentation et la santé : *Organisation des Nations unies* [pour l'agriculture et l'alimentation] (48 occ.), *Autorité européenne de sécurité des aliments* (16 occ.), et leurs sigles respectifs : *FAO* (120 occ.), *EFSA* (34 occ.). La FAOa publié en 2013 un rapport sur les enjeux d'alimentation à base d'insectes pour la sécurité alimentaire et l'alimentation animale (Van Huis *et al.*, 2013), alors que l'EFSA est l'auteur d'un rapport sur la commercialisation du ver de farine jaune séché (EFSA Panel on Nutrition *et al.*, 2021). Dans notre corpus d'étude, ces organismes accompagnent souvent des données précises, des recommandations et des opinions relatives aux atouts et bénéfices nutritionnels et environnementaux, permettant aux lecteurs d'évaluer positivement un sujet aux multiples enjeux. Ainsi, les questions nutritionnelles et environnementales deviennent des arguments favorables pour le développement de la filière d'insectes comestibles, comme en témoignent les exemples (j) à (m) (nous mettons en gras et nous soulignons) :

> j) L'**organisation onusienne** ne tarit pas d'éloges sur leurs **atouts nutritionnels**. Leurs teneurs en protéines de qualité (de 13 % à 77 % de matière sèche) rivalisent avec celles de la viande et du poisson (*Le Monde*, 2019) ;
>
> k) Depuis 2008, la **FAO** (l'**Organisation des Nations unies pour l'agriculture et l'alimentation**) recommande de consommer des insectes plutôt que de la viande pour des considérations […] **écologiques** (*Sud Ouest*, 2012) ;
>
> l) L'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO) et l'Efsa encouragent l'utilisation des insectes qui peuvent contribuer à la sécurité alimentaire et présente des bénéfices pour l'environnement (*Courrier international*, 2022) ;
>
> m) La consommation d'insectes (95 % d'entre eux sont comestibles) apparaît comme une réelle opportunité pour **lutter contre la sous-nutrition**. C'est l'**Organisation des Nations unies (ONU)** elle-même qui le dit. (*Ouest-France*, 2014).

Le caractère élogieux et véridique des passages indiqués permet de considérer la citation de sources officielles en tant que pratique éditoriale, visant à représenter les arguments rapportés par les journalistes comme légitimes. Toutefois, les termes indiquant les organismes publics introduisent également des questions liées aux dangers, ceux-ci étant néanmoins moins présents dans le corpus analysé par rapport aux arguments favorables. Le discours de l'*Agence nationale de sécurité sanitaire* [de l'alimentation, de l'environnement et du travail] (11 occ.), autrement dit *ANSES* (24 occ.), qui a publié un rapport évaluant les risques sanitaires associés à la consommation d'insectes (ANSES, 2015), est notamment repris dans les articles de presse pour indiquer et avertir le grand public sur les dangers potentiels (nous mettons en gras et nous soulignons) :

> n) L'**Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (Anses)** a pointé dans son rapport de 2014 le manque d'études scientifiques concernant les **risques sanitaires** liés à leur élevage, leur transformation et leur consommation (*Libération*, 2016) ;
>
> o) En avril, l'**Agence nationale de sécurité sanitaire** a émis un avertissement sur les **risques** liés à la consommation d'insectes. L'ANSES cible les substances chimiques, les parties dures, les parasites portés par l'animal et les risques allergènes, tout en incitant néanmoins à accentuer les efforts de recherche (*Courrier Picard*, 2015).

La présence de citations de sources officielles comme de la terminologie institutionnelle attire notre attention sur la coexistence des logiques communicationnelles divergentes dans la sphère médiatique. Les phénomènes linguistiques révèlent d'une part des arguments « pour », présentant les bénéfices favorables au développement de la filière d'insectes comestibles, et d'autre part, des arguments « contre », associés à des risques sur la santé humaine et animale.

4. Conclusion

4.1. Apports de l'étude et discussion des résultats

Notre étude exploratoire est une contribution originale, qui a mis en évidence des moyens linguistiques, révélateurs des logiques communicationnelles dans la médiatisation des insectes comestibles. Le lexique présent dans les articles issus de la presse généraliste française reflète notamment la parole des catégories d'acteurs tels que les entrepreneurs et les chercheurs et professionnels

de santé, et véhicule les thématiques spécialisées liées aux questions de la nutrition et de l'environnement. Ces thématiques, à la croisée de différents domaines scientifiques, se sont hybridées avec des outils lexicaux et terminologiques de vulgarisation établissant un rapport de proximité avec le lecteur, témoignant du succès des aliments à base d'insectes et rendant l'information diffusée avérée. Ils deviennent ainsi des arguments communicationnels, accompagnés souvent des données spécialisées diffusées par des organismes publics, et établis en faveur du développement de la filière d'insectes comestibles en France.

Cependant, les arguments environnementaux et nutritionnels de la presse généraliste sont en réalité controversés. Les données issues des articles de revues scientifiques montrent notamment que la vitesse de production d'insectes est plus particulièrement dépendante du maintien de la température très élevée (Ayieko, 2015). Les impacts environnementaux les plus considérables sont associés à la consommation d'électricité utilisée tout au long de la chaîne de production pour le chauffage, le refroidissement et la ventilation des usines (Smetana *et al.*, 2019). De plus, en fonction du lieu de production et des espèces d'insectes utilisées, l'émission de gaz à effet de serre provenant de la production de protéines d'insectes peut même être supérieure à celle des protéines de poulet de chair (Vauterin *et al.*, 2021). Enfin, les apports nutritionnels des insectes séchés sont plus élevés que ceux des autres viandes (Mitsuhashi, 1997), mais en ce qui concerne les insectes crus, leur apport est inférieur à celui du saumon, du porc, du poulet et du bœuf (Liceaga, 2022). Nos analyses ont montré que ces connaissances controversées, tout comme les dangers associés à la production et à la consommation d'insectes sont faiblement portés à l'attention du grand public par la presse généraliste.

4.2. Limites et perspectives

Notre corpus de taille réduite, ayant servi de preuve à la démonstration des tendances discursives et communicationnelles, est un indicateur qui se trouve fragilisé et constitue une limite pour notre analyse. Par ailleurs, au moment de la constitution du corpus (courant 2023), les autorisations de la mise sur le marché des insectes comestibles étant très récentes (2021-2022), nous étions face à un marché peu connu, non seulement par le grand public mais également par les spécialistes. Depuis, nous avons pu augmenter la taille du corpus qui contient désormais 372 articles représentant un total de 178 369 mots. Ce corpus enrichi par de nouveaux articles sera prochainement analysé.

En ce qui concerne l'approche mise en œuvre dans notre étude, la dimension qualitative de notre analyse doit à présent se doubler d'une analyse quantitative. Il s'agira de quantifier les moyens linguistiques mis au jour, par exemple les marqueurs métalinguistiques mobilisés pour faire coexister termes spécialisés et termes vulgarisés (marqueur verbal *appeler* ou *connaître* dans les extraits de notre corpus cités dans la partie 3.2). Cette approche hybride pourra trouver un prolongement dans une perspective contrastive liée au genre textuel. En effet, notre corpus médiatique met en jeu le genre article de presse ; un corpus spécialisé constitué d'articles scientifiques impliquerait un autre genre textuel. La perspective de constitution d'un corpus spécialisé pose d'emblée la question de l'existence de sources spécialisées étant donné la pénurie de connaissances spécialisées sur le sujet (Akhbarifar *et al*., 2024).

Sur un plan épistémologique, la rencontre des SIC et de la Terminologie participe d'un mouvement d'ouverture de la terminologie à d'autres disciplines (Pecman, 2023) dans un contexte de questionnements disciplinaires dont rendront compte, par exemple, les orientations du prochain colloque du réseau Lexicologie Traduction Terminologie consacré aux approches épistémologiques de la terminologie. La fertilisation scientifique passe par l'ouverture disciplinaire (Delavigne et De Vecchi, 2021, 10, citant Stengers et Schlanger, 1989) et chaque discipline, par ses méthodes et ses outils peut nourrir et éclairer l'objectif commun de construction et d'analyse de connaissances spécialisées et de circulation des savoirs.

Références

Akhbarifar, Yekta, Brion, Jules, Villain, Victor. 2024. « Mediatization of Edible Insects in France : poor coverage and limited knowledge ». Dans *Book of Abstracts – INSECTA 2024 International conference –14th-16th Mays 2024 Potsdam, Germany*, édité par Giacomo Rossi *et al.*, 152. Potsdam : Leibniz-Institut für Agrartechnik und Bioökonomie e.V. (ATB) : https://www.insects.plus/

ANSES. 2015. « Avis de l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail relatif à la valorisation des insectes dans l'alimentation et l'état des lieux des connaissances scientifiques sur les risques sanitaires en lien avec la consommation des insectes ». Maisons-Alfort : ANSES. https://www.anses.fr/fr/system/files/BIORISK2014sa0153.pdf

Ayieko, M. A., J. Kinyuru, R. Makhado, M. Potgieter, P. Tshikudo, K. Mawela, H. Maluleke-Nyathela, D. Oonincx, J. Cloutier. 2015. « Locusts and grasshoppers (Orthoptera) ». Dans *Edible insects in Africa: an introduction to finding, using and eating insects*, édité par Josianne Cloutier, 45-56. Wageningen : Agromisa Foundation/CTA Publications Distribution Service. https://www.echocommunity.org/resources/b3350fa1-e1f3-4746-8f8e-72d1370c537f

Bourigault, Didier et Slodzian, Monique. 1999. « Pour une terminologie textuelle ». *Terminologies nouvelles*, nº 19, p. 29-32.

Catellani, Andrea et Amaia Errecart A. 2018. « L'hybridation discursive dans la communication sur la responsabilité sociétale des entreprises : le cas des banques "engagées" ». *Recherches en communication*, nº 47, p. 23-43.

Condamines, Anne. 2018. « Nouvelles perspectives pour la terminologie textuelle ». Dans *Terminology and Discourse*, édité par Jana Altmanova, Maria Centrella, Katherine E. Russo. Berne : Peter Lang.

Condamines, Anne et Narcy-Combes, Jean-Paul. 2015. « La linguistique appliquée comme science située ». Dans *Cultures de recherche en linguistique appliquée*, édité par Francis Carton, Jean-Paul Narcy-Combes, Marie-Françoise Narcy-Combes, Denyse Toffoli. Paris : Riveneuve éditions.

Delavigne, Valérie et de Vecchi, Dardo. 2021. *Termes en discours. Entreprises et organisations*. Presses de la Sorbonne Nouvelle, Paris.

EFSA Panel on Nutrition *et al.* 2021. « Safety of dried yellow mealworm (Tenebrio molitor larva) as a novel food pursuant to Regulation (EU)

2015/2283 ». *EFSA Journal*, vol. ,19, nº 1, p. 6343. https://www.efsa.europa.eu/en/efsajournal/pub/6343

Gaudin, François. 2005. « La socioterminologie ». *Langages*, nº 157, p. 80-92.

IRaMúTeQ. 2024. « Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires. Un logiciel libre construit avec des logiciels libres ». http://www.iramuteq.org

Kilgariff, Adam Baisa, Vít Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel, Suchomel, Vít. 2014. « The Sketch Engine Ten Years On », *Lexicography*, vol. 1, nº 1, p. 7-36.

Lafon, Benoît. 2019. « Des médiatisations au processus de médiatisation ». Dans *Médias et médiatisation. Analyser les médias imprimés, audiovisuels, numériques*, édité par Benoît Lafon, p. 157-189. Grenoble : Presses universitaires de Grenoble.

Lafon, Benoît. 2022. « Médiatisation des exoplanètes ». *Questions de communication*, nº 42, p. 97-131.

Ledouble, Hélène. 2021. « Contextes et connaissances dans les discours de vulgarisation du scientifique dynamique définitoires et problématiques cognitives ». Dans *Des corpus numériques à l'analyse linguistique en langues de spécialité*, édité par Cécile Frérot et Mojca Pecman. Grenoble : UGA Éditions.

Ledouble, Hélène. 2023. *Médiatisation de la science et diffusion des connaissances Approche terminologique et cognitive en agroécologie*. Londres : ISTE Editions.

Liceaga, Andrea M. 2022. « Edible insects, a valuable protein source from ancient to modern times ». *Advances in Food and Nutrition Research*, nº 101, p. 129-152.

Marty, Emmanuel, Marchand, Pascal, Ratinaud, Pierre. 2020. « La méthode ALCESTE et IRAMUTEQ : rappel de quelques principes et fonctionnalités ». Communication présentée à la formation URFIST, URFIST Strasbourg, 2020. https://urfist.unistra.fr/uploads/media/diapo_formation_URFIST.Aprofdmt.2020.pdf

Mitsuhashi, Jun. 1997. « Insects as traditional foods in Japan ». *Ecology of Food and Nutrition*, vol. 36, nº 2-4, p. 187-199.

Pecman, Mojca. 2023. « Les connaissances ont besoin de la terminologie et la terminologie a besoin de corpus ». Communication présentée au Colloque COSEDI (Corpus de genres spécialisés : caractérisation, méthodes et applications didactiques), Université Grenoble Alpes, 6-8 décembre 2023.

Reinert, Max. 1983. « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte ». *Cahiers de l'analyse* des *données*, vol. 8, nº 2, p. 187-198.

Smetana, Sergiy, Schmitt, Éric, Mathys, Alexander. 2019. « Sustainable use of *Hermetia illucens* insect biomass for feed and food: Attributional and consequential life cycle assessment ». *Resources, Conservation and Recycling*, nº 144, p. 285-296.

Stengers, Isabelle et Schlanger , Judith. 1989. *Les Concepts scientifiques : invention et pouvoir*. La Découverte, Conseil de l'Europe, Unesco, Paris.

Trzmielewski, Marcin. 2021. « Alimentation à base d'insectes en France : un "non-problème" public ? ». *Sciences de la Société*, nº 108. doi.org/10.4000/137ie

van Huis, Arnold, van Itterbeeck, Joost, Klunder, Harmke, Mertens, Esther, Halloran, Afton, Muir, Giulia, Vantomme, Paul. 2013. *Edible insects: future prospects for food and feed security*. Rome : Food and Agriculture Organization of the United Nations.

Vauterin, A., B. Steiner, J. Sillman, H. Kahiluoto. 2021. « The potential of insect protein to reduce food-based carbon footprints in Europe: The case of broiler meat production ». *Journal of Cleaner Production*, nº 320, p. 128799.

Vecchi (de), Dardo. 2016. « Approche pragmaterminologique des termes des entreprises et des organisations ». *Synergie Italie*, nº 12, p. 125-139.

Vecchi (de), Dardo. 2004. « La terminologie dans la communication de l'entreprise : approche pragmaterminologique ». Dans *Cahiers du CIEL*, p. 71-83. Paris : Université Paris 7 EILA.

DICOADVENTURE: A Bilingual (EN-ES) Terminological Database of the Language Used in Adventure Tourism

Isabel Durán-Muñoz*, Eva Lucía Jiménez-Navarro**

*Facultad de Filosofía y Letras, Córdoba
iduran@uco.es

**Facultad de Filosofía y Letras, Córdoba
lucia.jimenez@uco.es

Abstract. The purpose of this paper is to present *DicoAdventure*, an online bilingual (English-Spanish) lexico-semantic specialized resource about the language used in adventure tourism (http://olst.ling.umontreal.ca/dicoadventure/). The framework of this work integrates three areas of study: corpus linguistics, lexical semantics, and specialized lexicography. First, all the information to be included is retrieved from a specialized representative corpus (*i.e.*, ADVENCOR); second, the data is based on the Frame Semantics theory (Fillmore, 1982, 1985, 2006); and third, the theories of specialized lexicography (*e.g.*, Bergenholtz & Tarp, 1995, 2003; Tarp, 2009) have allowed the development of a tool that considers its users' needs and contributes to effective communication in specialized situations. So far it has mainly covered motion verbs after noticing that they overwhelmingly predominate (Durán-Muñoz & Jiménez-Navarro 2023; Durán-Muñoz & L'Homme 2020; Jiménez-Navarro & Durán-Muñoz 2024), but other types of concepts are under study.

Résumé. *L'objectif de cet article est de présenter DicoAdventure, une ressource lexico-sémantique spécialisée en ligne et bilingue (anglais-espagnol) sur la langue utilisée dans le tourisme d'aventure. Le cadre de ce travail intègre trois domaines d'étude : la linguistique de corpus, la sémantique lexicale et la lexicographie spécialisée. Premièrement, toutes les informations à inclure sont extraites d'un corpus représentatif spécialisé (ADVENCOR); deuxièmement, les données sont basées sur la théorie de la sémantique du cadre (Fillmore, 1982, 1985, 2006); et troisièmement, les théories de la lexi-*

cographie spécialisée (par exemple, Bergenholtz & Tarp, 1995, 2003 ; Tarp, 2009) ont permis de développer un outil qui prend en compte les besoins de ses utilisateurs et contribue à une communication efficace dans des situations spécialisées. Jusqu'à présent, il a principalement couvert les verbes de mouvement après avoir remarqué qu'ils prédominaient largement (Durán-Muñoz & Jiménez-Navarro, 2023 ; Durán-Muñoz & L'Homme, 2020 ; Jiménez-Navarro & Durán-Muñoz, 2024), mais d'autres types de concepts sont en cours d'étude.

1. Introduction

The *DicoAdventure* dictionary is an online terminological database that intends to house the terminology of adventure tourism discourse, including different grammatical categories (nouns, adjectives, verbs, and adverbs) both in English and Spanish. It is a new resource (currently under construction) which tries to offer a novel way of representing the terminological information in its entries. It is mainly aimed at meeting the consultation needs of professional translators, for this reason, it takes into account the study carried out by Durán-Muñoz (2012). This work states that translators prefer to solve their terminology problems by consulting databases produced by reliable sources rather than developing their own resources. Besides, as far as the microstructure of terminological entries is concerned, they consider the following fields as the most essential: definitions, equivalents, derivatives and compounds, domain specification, real examples, phraseological information, definition in both languages for bilingual resources, and abbreviations and acronyms (Durán-Muñoz, 2012, 82). Furthermore, the same study specifies that, according to translators, terminological resources should: 1) allow for exportability and/or importability in different formats, 2) include more pragmatic information on usage and difficult translations (previous usage, false friends, specific usage in a domain or region, etc.), 3) provide links to other resources to improve or enhance results, 4) improve search options, and 5) provide examples taken from real texts.

The elaboration of this dictionary is based on the consultation needs indicated by the translators in the aforementioned study and is marked by the methodology used to select and analyse the motion verbs of adventure tourism proposed by Durán-Muñoz and L'Homme (2020), which is divided into four different steps: 1) compilation of a specialized corpus, 2) extraction and selection of motion verbs, 3) semantic and syntactic annotation of contexts,

and 4) definition of the argumentative structure of each verb. Furthermore, this methodology is based on Fillmore's Frame Semantics (Fillmore, 1982, 1985, 2006) and on the FrameNet project (Ruppenhofer *et al*., 2016), as well as on previous projects dealing with the terminology of some specialized domains which also apply this theory, such as the *Kicktionary* project (Schmidt 2009), the *JuriDiCo* resource (Pimentel, 2013), *DiCoInfo: A Framed Version* (Ghazzawi, 2016), *A Framed DiCoEnviro* (L'Homme, 2018), and the *EcoLexicon* database (Faber *et al*., 2016).

In the following sections, we describe the *DicoAdventure* resource, both in terms of its macro and microstructure. To provide an overview of this novel terminological database, the paper is structured as follows: first, it focuses on the background of the resource and explains its main purpose; then, the macrostructure is presented, followed by its microstructure. Finally, some concluding remarks are highlighted.

2. Macrostructure and microstructure of *DicoAdventure*

As indicated above, the *DicoAdventure* terminological database is still under construction and currently hosts a reduced number of entries, specifically, 68 entries in total for English and 74 for Spanish. Most of the entries are motion verbs, both representing real motion and fictive motion, but also other types of categories are under study, especially nouns referring to adventure activities, such as *trekking* or *kayaking.*

Regarding the macrostructure of the dictionary, the entries can be accessed through an alphabetical entry list in each language (Figure 1), but also by using a search box (Figure 2) which allows refining the search in terms of which section in the entry it should be conducted (*e.g.*, head terms, definitions), how precise the word typed in is (*e.g.*, exact, starts with), and the language (English, Spanish, or both). Access to information by means of hyperlinks within the entries is also granted.

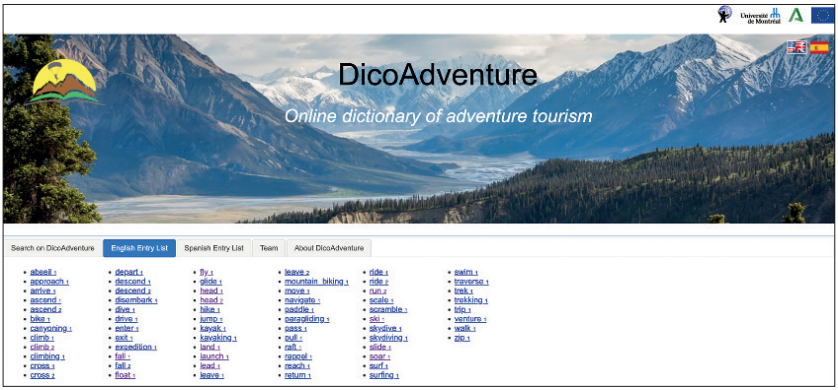


Figure 1. Alphabetical list of entries in DicoAdventure *(English).*

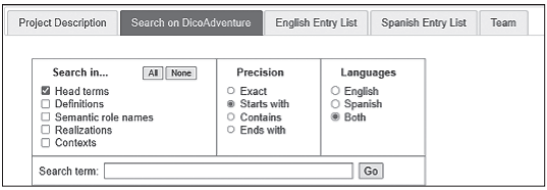


Figure 2. Search tab of DicoAdventure.

Regarding the microstructure of this resource, it intends to meet the needs of potential users, mainly professional translators. For this reason, it offers comprehensive linguistic, pragmatic, and semantic information that is easy to access. In addition, each entry is organized in such a way that, at a glance, users can consult and access a large amount of information, as shown in Figure 3.

In *DicoAdventure*, all terminological entries are structured in the same way (Figure 3): (1) the term, meaning number (since this is a concept-based dictionary; it is particularly useful when a term can convey more than one meaning, *e.g.*, *fall* –fall from a plane vs fall on the ground–), and grammatical category; (2) definition and argument structure of the term; (3) linguistic realizations of the arguments and examples; (4) equivalents in the other language; (5) contexts of use, annotated contexts, and a summary; (6) lexical relations with other terms (*e.g.*, synonyms, different parts of speech, etc.); and (7) administrative information. Furthermore, all entries include a descriptive

image that aims to support the definition of the terminological unit and to facilitate the understanding of the meaning.

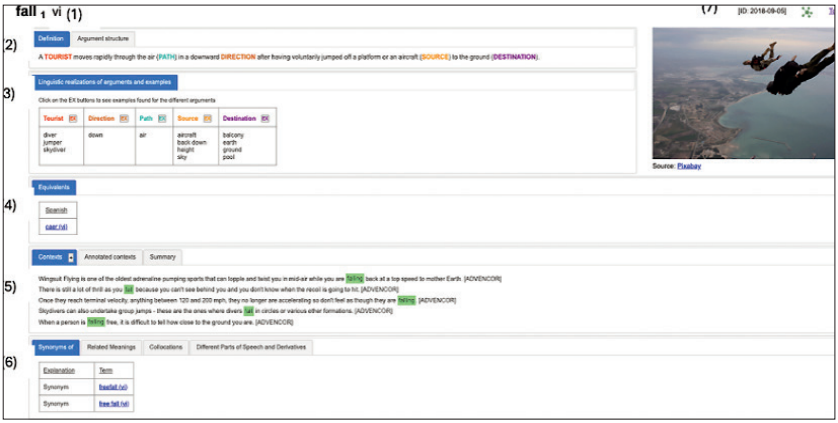


Figure 3. Entry for the verb fall$_1$ *in* DicoAdventure.

Both the definition and the argument structure (section 2) are based on Frame Semantics and provide a description of the lexico-semantic properties of the units. By default, the definition is the field displayed first, although the argument structure is easily accessible by clicking on the tab right next to *Definition*. Both contain the same semantic information based on semantic arguments, although organized differently as they serve different purposes. On the one hand, definitions aim to describe or clarify the meaning of a terminological unit; on the other hand, argument structures show the relationship of the unit with respect to its main semantic arguments in the form of a structure and show how each unit functions in context.

In the case of the verb *fall*$_1$, we find the *Definition* (Figure 4) and the *Argument structure* (Figure 5) below. As can be observed, the arguments of this term are TOURIST, DIRECTION, PATH, SOURCE, and DESTINATION. Additionaly, notes can be added in both sections, that is, linguistic or pragmatic details in the *Definition* and syntactic or semantic details in the *Argument structure*.

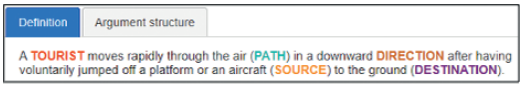


Figure 4. Definition of the verb fall$_1$ *in* DicoAdventure.

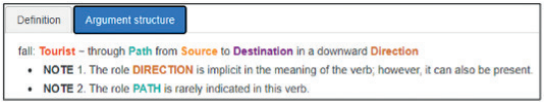


Figure 5. Argument structure of the verb fall$_1$ in DicoAdventure.

The next section displays the linguistic realizations and examples of each argument of the terminological unit. The number of columns provided will vary according to the number of arguments of the verb. In the case of *fall*$_1$, five different arguments are represented (Figure 6), which are the same arguments that we found previously in its definition and in its argument structure: Tourist, Direction, Path, Source, and Destination.

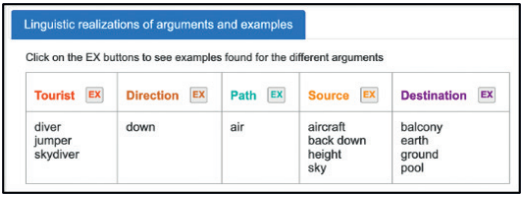
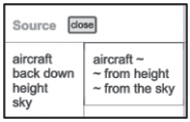
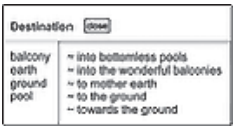


Figure 6. Linguistic realizations and examples of the arguments of the verb fall$_1$ in DicoAdventure.

In each of these columns, we gather the realizations found in the corpus used for this study. Thus, the argument Tourist is linguistically instantiated as *diver*, *jumper*, and *skydiver*; the argument Direction, with *down*, and so on. In addition, by clicking on the *Ex* [Examples] button, the resource allows us to access the different options that have been annotated in the corpus. For example, the cases found for Source and for Destination are shown in Figure 7 and Figure 8, respectively.

7.


8.


9.


Figure 7. Examples of the argument Source of the verb fall$_1$ in DicoAdventure. *Figure 8. Examples of the argument Destination of the verb fall$_1$ in* DicoAdventure. *Figure 9. Equivalents of the verb fall$_1$ in* DicoAdventure.

The following section refers to the equivalents in Spanish (in this case) or in English. It includes a hyperlink to easily forward users to the corresponding entry and it also indicates the grammatical category of the unit.

Next, *DicoAdventure* displays contexts in two different forms: raw contexts, that is, contexts without any annotations except for the highlighted term at stake, and semantically and syntactically annotated contexts (Figure 10 and Figure 11, respectively).

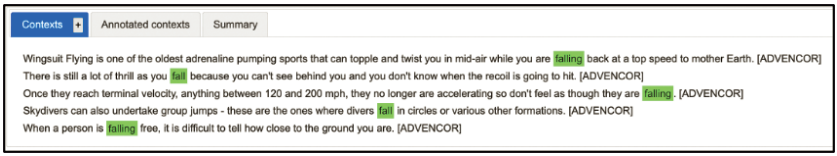


Figure 10. A sample of raw contexts extracted from the corpus for the verb fall$_1$ *in* DicoAdventure.

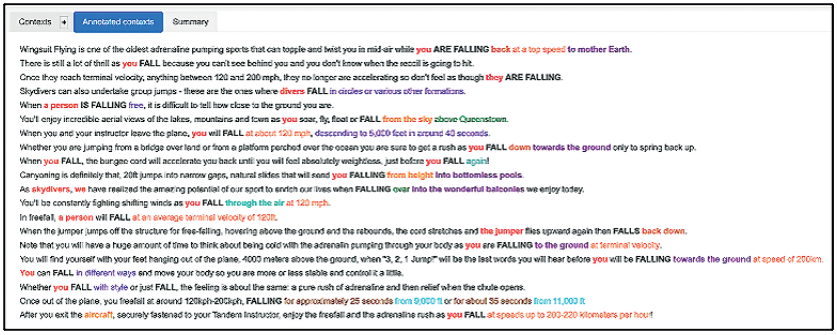


Figure 11. Annotated contexts of the verb fall$_1$ *in* DicoAdventure.

Finally, there is a third tab providing the summary of the annotated elements (arguments and circumstantials), including the annotated examples from the corpus and the syntactic specifications used (Figure 12).

Contexts + | Annotated contexts | Summary

Arguments		
Destination	Complement (PP-to) Complement (PP-towards) Complement (PP-into)	into bottomless pools (1) into the wonderful balconies (1) to mother earth (1) to the ground (1) towards the ground (2)
Direction	Modifier (AdvP)	back (1) back down (1) down (1)
Path	Complement (PP-through)	through the air (1)
Source	Complement (PP-from) Indirect link (NP)	aircraft (1) from height (1) from the sky (1)
Tourist	Indirect link (NP) Subject (NP)	a person (2) divers (1) skydivers (1) the jumper (1) they (1) we (1) you (14)
Others		
Distance	Complement (PP-from)	from 11,000 ft (1) from 9,000 ft (1)
Frequency	Complement (AdvP)	again (1)
Manner	Complement (PP-in) Complement (VP) Complement (PP-with) Modifier (AdvP)	descending to 5,000 feet in around 40 seconds (1) free (1) in circles or various other formations (1) in different ways (1) with style (1)
Place	Complement (PP-above) Complement (AdvP)	above queenstown (1) over (1)
Sequence	Complement (AdvP)	then (1)
Speed	Complement (PP-at)	at 120 mph (1) at a top speed (1) at about 120 mph (1) at an average terminal velocityof 120ft (1) at speed of 200km (1) at speeds up to 200-220 kilometers per hour (1) at terminal velocity (1)
Time	Complement (PP-for)	for about 35 seconds (1) for approximately 25 seconds (1)

Figure 12. Summary of annotations of the verb fall$_1$ *in* DicoAdventure.

This part of the input is of great interest to potential users, since, on the one hand, it allows them to access real contexts extracted from a specialized textual corpus and, on the other hand, it makes it possible to observe

the different arguments (and circumstancials) of the units. Thus, both the annotated contexts and the summary table of arguments and circumstancials provide quality information on how the unit functions in context, how it relates to its arguments, and which are the most frequent arguments and circumstancials of the unit.

The final section offers lexical relations. The list of relations is richly varied, but in each entry only those corresponding to each unit are shown. In the case of *fall*$_1$, we find the following: 1) units that show synonymy with the entry, 2) units that have a related meaning, 3) collocations, and 4) units that belong to other grammatical categories (Figure 13). Each of the units included in these relations has (or will have) a specific entry in the resource, and they are (or will be) linked by hyperlinks. In this way, to access information about, for example, units belonging to different grammatical categories, the user will only have to click on the *Different Parts of Speech and Derivatives* tab, which will greatly speed up the consultation, giving direct access to a large number of specialized units from a single entry.

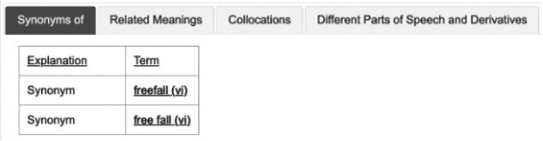


Figure 13. Lexical relations of the verb fall$_1$ *in* DicoAdventure.

These sections of the entry complement each other and provide an overview of the terminology related to the unit consulted, as well as an adequate understanding of the meaning of the unit in question and how it functions linguistically and syntactically. In short, they provide relevant information to enable users to use the adventure tourism terminology appropriately.

3. Conclusions

The terminology of the adventure tourism sector is incredibly rich and eclectic, displaying its own characteristics and making it necessary to develop terminological resources such as *DicoAdventure*.

This resource is aimed at professional translators and is intended to cover their main needs, such as efficient consultation of linguistic, pragmatic, and semantic information that is provided in different sections, namely:

1) definition, argument structure (along with notes, if necessary), and descriptive image, which allows information on the use and meaning of terminological entries to be acquired; 2) linguistic realizations of arguments and examples, which allow quick access to terms semantically related to the consulted entry; 3) lexical relations, which include the relations that the terminological entry holds with other terms (*e.g.*, synonymy, antonymy, derivatives, related meanings, among others); and 4) contexts (annotated and non-annotated contexts and summary of participants), which offer users the possibility of observing real contexts extracted from a textual corpus, together with the semantic and syntactic annotations that allow them to observe the actual functioning of these units. All this information is presented in a user-friendly and intuitive way to facilitate understanding and to provide a complete picture of each term.

In order to complete this resource, we intend to increase the number of entries in English and Spanish, by adding both verbs and other grammatical categories, such as nouns, adjectives, and adverbs. The overall aim is to provide a comprehensive database that includes the most relevant terminology of the adventure tourism segment and that, therefore, offers high-quality and reliable information that satisfies the main needs of professional translators.

This research was partially carried out within the framework of the R&D projects RECOVER (ProyExcel_00540) and VIP II (PID2020-112818GB-I00).

References

Bergenholtz, Henning & Tarp, Sven. 1995. *Manual of specialized lexicography. The preparation of specialized dictionaries*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bergenholtz, Henning & Tarp, Sven. 2003. "Two opposing theories: On H. E. Wiegand's recent discovery of lexicographic functions." *Hermes, Journal of Linguistics* 31: 171–196.

DicoAdventure. An online dictionary of adventure tourism. 2024. OLST. Accessed October 9. https://olst.ling.umontreal.ca/dicoadventure/

Durán-Muñoz, Isabel. 2012. *La ontoterminografía aplicada a la traducción. Propuesta metodológica para la elaboración de recursos terminológicos dirigidos a traductores*. Berlin: Peter Lang.

Durán-Muñoz, Isabel & Jiménez-Navarro, Eva Lucía. 2023. "Motion verbs in adventure tourism: A lexico-semantic approach to fictive meaning." *International Journal of English Studies* 23(1): 27–48.

Durán-Muñoz, Isabel & L'Homme, Marie-Claude. 2020. "Diving into English motion verbs from a lexico-semantic approach. A corpus-based analysis of adventure tourism." *Terminology* 26(1): 33–59.

Faber, Pamela, León-Araúz, Pilar & Reimerink, Anne. 2016. "EcoLexicon: New features and challenges." In I*GLOBALEX 2016: Lexicographic Resources for Human Language Technology in conjunction with the 10th edition of the Language Resources and Evaluation Conference*, edited by Ilan Kernerman, Iztok Kosem, Simon Krek, & Lars Trap-Jensen, 73–80. Portorož.

Fillmore, Charles. J. 1982. "Frame semantics". In *Linguistics in the morning calm*, edited by Linguistic Society of Korea, 111–137. Seoul: Hanshin Publishing Company.

Fillmore, Charles J. 1985. "Frames and the semantics of understanding." *Quaderni di Semantica* 6(2): 222–254.

Fillmore, Charles J. 2006. "Frame semantics". In *Cognitive linguistics: Basic readings*, edited by Dirk Geeraerts, 373–400. New York: De Gruyter Mouton.

Ghazzawi, Nizar. 2016. "Du terme prédicatif au cadre sémantique: méthodologie de compilation d'une ressource terminologique pour les termes arabes de l'informatique." PhD diss., Université de Montréal.

Jiménez-Navarro, Eva Lucía & Durán-Muñoz, Isabel. 2024. "Collocations of fictive motion verbs in adventure tourism: A corpus-based study of the

English language." *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 37(2): 371–395.

L'Homme, Marie-Claude. 2018. "Maintaining the balance between knowledge and the lexicon in terminology: A methodology based on Frame Semantics." *Lexicography* 4(1): 4–21.

Pimentel, Janine. 2013. "Methodological bases for assigning terminological equivalents. A contribution." *Terminology* 19(2): 237–257.

Ruppenhofer, Josef, Ellsworth, Michael, Petruck, Miriam R., Johnson, Christopher R., Baker, Collin F. & Scheffczyk, Jan. 2016. *FrameNet II: Extended Theory and Practice*. Accessed October 9. https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the_book

Schmidt, Thomas. 2009. "The Kicktionary. A multilingual Lexical Resource of Football Language." In *Multilingual FrameNets in computational lexicography*, edited by Hans C. Boas, 101–134. New York: De Gruyter Mouton.

Tarp, Sven. 2009. "Reflections on lexicographical user research." *Lexicos* 19: 275–296.

Pratique de la capitalisation des connaissances en environnement industriel : l'importance d'une démarche de cartographie associant les experts, un exemple en ingénierie nucléaire

Marie Calberg Challot*, Julien Roche**

Onomia

* marie.calberg-challot@onomia.com, ** julienroche@onomia.com

Résumé. Cet article se propose de mener une analyse approfondie sur l'importance de la capitalisation des connaissances en s'appuyant sur notre expérience dans des environnements industriels à haute technicité. L'objectif est de faciliter la préservation et le partage des savoirs parmi les ingénieurs et les experts en développant des systèmes de cartographie des connaissances adaptés et signifiants pour les utilisateurs actuels et futurs dans des secteurs où les défis techniques sont considérables.

Abstract. *This article aims to conduct an in-depth analysis of the importance of knowledge capitalization based on our experience in highly technical industrial environments. The goal is to facilitate the preservation and sharing of knowledge among engineers and experts by developing knowledge mapping systems that are relevant and meaningful to current and future users in sectors where technical challenges are considerable.*

Face aux multiples défis tels que les départs à la retraite, la gestion des carrières et le transfert de connaissances, il convient de souligner la nécessité cruciale de s'interroger sur la mise en œuvre de stratégies de capitalisation des connaissances au sein des organisations. L'enjeu est de garantir la pérennité des savoir-faire et des décisions stratégiques pour les décennies à venir ou plus trivialement de se mettre en capacité de répondre positivement à la question si l'on saura faire demain ce que l'on sait faire aujourd'hui.

Pour répondre concrètement à ces enjeux, l'expérience nous a conduits à insister sur des approches fondées sur la création d'une base de connaissances explicite, structurée et accessible, servant de fondement à diverses

applications telles que la formation, la maintenance, l'ingénierie et l'innovation. La méthodologie repose sur une compréhension profonde de l'origine et de la localisation des connaissances, principalement dans l'expertise humaine accumulée au fil des années d'expérience et de réflexion.

L'article réitère l'importance des experts humains dans la gestion des connaissances, soulignant leur capacité à naviguer et enrichir les données mieux que toute solution automatisée. Il met en lumière les limitations intrinsèques d'approches purement textuelles et la valeur irremplaçable de l'expertise humaine dans l'interprétation et l'application des connaissances.

Le parti pris de nos travaux est donc, dans une démarche onomasiologique, de partir de la connaissance là où elle se trouve. Cela suppose alors évidemment d'être en mesure de cibler les connaissances d'intérêt. Notre positionnement et notre expérience nous conduisent à identifier que les connaissances qualifiées pour répondre aux besoins se trouvent essentiellement dans « la tête des sachants » qui, après l'étude et la pratique dans leurs domaines des années durant, connaissent la logique, l'articulation ou l'architecture d'ensemble des éléments de connaissances les uns avec les autres car ce sont eux soit qui l'ont construite, soit qui en ont hérité d'un compagnonnage avec leurs pairs. Ils maîtrisent donc l'historique, les réussites, les écueils, les risques et sont donc en mesure de résoudre ou d'expliquer facilement les incomplétudes, imprécisions, ambiguïtés, contradictions et savent ce qui est mais aussi et surtout ce qui n'est pas dans la documentation.

Après ces constats, nous revenons donc dans un premier temps sur les bases théoriques de notre approche (Roche, 2008, Calberg Challot *et al.*, 2009, Roche, 2015) que nous positionnerons également vis-à-vis de certains fondamentaux de la terminologie (Wüster, 1968 ; Felber, 1987 ; Norme ISO 1087 2021 et Medhat-Lecocq, 2022). Ces bases nous conduisent alors à l'établissement et l'explicitation d'une méthode pour mise en œuvre sur le terrain que nous présenterons.

Puis, dans un second temps, une illustration des bénéfices de cette démarche est présentée sur un exemple en ingénierie nucléaire par comparaison entre des résultats accessibles par une analyse textuelle et le produit d'un travail de cartographie directement mené avec des experts sur le sujet des réacteurs nucléaires (CEA, 1975, 2008).

Finalement, il nous semble donc indispensable de placer le sachant au cœur de la démarche car il est le dépositaire de l'architecture des connaissances. L'expérience nous montre également que l'analyse de la documentation

n'est certainement pas suffisante, ni la meilleure source d'information pour commencer un travail de capitalisation des connaissances car celle-ci ne contient pas nécessairement les connaissances essentielles pour concevoir, fabriquer ou maintenir des équipements techniques à longue durée de vie. Cependant, une fois des cartographies établies, l'intégration d'un échantillon de documentation dûment sélectionné dans cette architecture prendra tout son sens pour apporter du contenu au sein d'une démarche structurée facilitant ainsi la diffusion et le partage des connaissances au sein de l'entreprise.

1. Introduction

L'article défend la place des experts au cœur d'une démarche de **cartographie des connaissances** explicite, structurée et accessible pour la capitalisation des connaissances dans les industries à haute technicité. Face aux départs à la retraite et au besoin de transfert des connaissances techniques, il s'agit d'assurer la pérennité des savoir-faire au-delà de la simple accumulation de documents. La démarche, **onomasiologique**, part de la connaissance telle qu'elle est portée par les sachants — capables d'interpréter, d'analyser et de comprendre la documentation technique — et montre les limites des approches purement textuelles ou automatisées. Après un rappel théorique sur la terminologie, et l'ontoterminologie et une présentation de la méthode, un **cas en ingénierie nucléaire** illustre les bénéfices de la cartographie par rapport à l'analyse textuelle. Nous montrerons finalement comment la cartographie avec les experts prime sur l'analyse textuelle d'un échantillon de documents pertinents à cette architecture.

Nous proposons de présenter un témoignage qui s'appuie sur notre expérience dans des environnements industriels à haute technicité.

L'objectif est de faciliter la préservation et le partage des savoirs parmi les ingénieurs et les experts en développant des systèmes de cartographie des connaissances adaptés et signifiants pour les utilisateurs actuels et futurs dans des secteurs où les exigences techniques sont considérables.

2. Les raisons d'une telle approche

Les contraintes incitatives peuvent être différentes (départ en retraite et pyramide des âges, roulement dans le poste plus court etc…) Le point de départ est toujours le même, à savoir comment m'assurer que ce que nous

savons faire aujourd'hui dans l'entité, nous sachions encore le faire, comprendre les choix ou en utiliser les savoirs dans 10 ans, 20 ans ou plus ?

La capitalisation des connaissances pour en faire une base explicite, organisée et accessible pouvant servir de support à de nombreuses utilisations ou décisions telles que la formation, la maintenance, l'ingénierie ou l'innovation est la vocation de cette approche.

Face aux multiples défis tels que les départs à la retraite, la gestion des carrières et le transfert de connaissances, il convient de souligner la nécessité cruciale de s'interroger sur la mise en œuvre de stratégies de capitalisation des connaissances au sein des organisations.

L'enjeu est de garantir la pérennité des savoir-faire et des décisions stratégiques pour les décennies à venir ou, plus trivialement, de se mettre en capacité de répondre positivement à la question si l'on saura faire demain ce que l'on sait faire aujourd'hui.

Une approche est centrée sur la connaissance et les experts qui la maîtrisent et cible les connaissances à forte technicité et à longue durée de vie pour établir des cartes, des architectures de connaissances, des graphes et des chemins de connaissances.

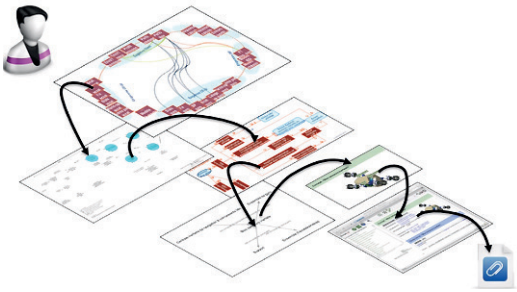


Figure 1.

Les experts ou sachants ont en effet des années d'études, de pratiques et de réflexions qui leur permettent de comprendre la logique d'ensemble car ce sont eux qui l'ont construite.

Ils maîtrisent également l'historique, les réussites, les écueils, les risques techniques et résolvent facilement les incomplétudes, imprécisions, ambiguïtés, contradictions.

Ils savent enfin ce qui est et ce qui n'est pas dans la documentation et que la compréhension des documents techniques requiert des connaissances métiers extralinguistiques.

Ces sachants savent identifier la documentation pertinente, mieux encore ils savent l'interpréter et l'enrichir soulignant la valeur irremplaçable de leur intervention qui ne peut être accessible par des outils numériques d'analyse de la documentation qui par essence contient peut-être beaucoup d'information sans garantie de contenir l'essentiel.

Ces faits s'imposent à nous au quotidien. En effet, entre, d'une part, consulter l'expert qui saura nous orienter, nous répondre ou, d'autre part, analyser soi-même la documentation volumineuse qu'il faudrait des années voire une vie pour assimiler, synthétiser et articuler, quel est le chemin à privilégier car « on ne voit jamais personne devenir médecin par la simple étude de recueils d'ordonnances » (Aristote, 1997), le texte et la documentation étant rarement autosuffisants.

3. L'approche par étapes

Après ces constats, nous revenons donc dans un premier temps sur les bases théoriques de notre approche (Roche, 2008 ; Calberg Challot *et al.*, 2009 ; Roche, 2015) que nous positionnons également vis-à-vis de certains fondamentaux de la terminologie (Wüster, 1968 ; Felber, 1987 ; Norme ISO1087, 2021 et Medhat-Lecocq, 2022). Ces bases nous conduisent alors à l'établissement et l'explicitation d'une méthode pour mise en œuvre sur le terrain que nous présenterons.

Concrètement la méthode outillée avec les experts et ingénieurs se déroule de la manière suivante.

Tout débute par une co-construction avec des experts pour formaliser les connaissances et établir les domaines cibles où les connaissances implicites au départ deviennent explicites à des fins de capitalisation et d'exploitations ultérieures.

Figure 2.

Le résultat est donc des cartes, des architectures ou arborescences ou des graphes et des chemins et peuvent être établis potentiellement selon les besoins et selon plusieurs dimensions — par exemple matérielle : de quoi c'est constitué, fonctionnelle : à quoi ça sert ou historique : évolution des savoirs dans le temps — sans avoir besoin de repenser et reconstruire la structure.

Figure 3.

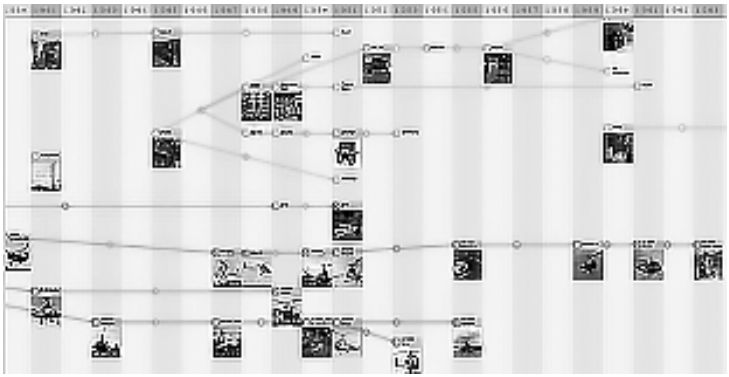


Figure 4.

Ensuite, ces cartes sont à relier et à parcourir, par les sachants, avec des éléments d'intérêt pour la formation, pour la transmission des connaissances comme des documents clés, standards, documents de capitalisation ou méthodologique, schéma, fiche, publications majeures, enrichissement thématique.

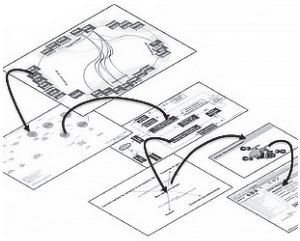
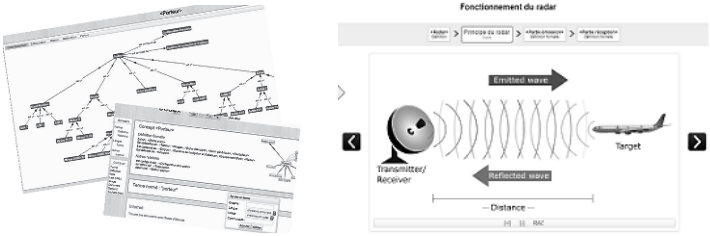


Figure 5.

In fine, cette structure enrichie devient une base de connaissance de l'entité à utiliser soit en navigation libre, soit en navigation dirigée comme dans le cadre de parcours de formation ou par thème centré sur une problématique choisie.



Figures 6 et 7.

4. Un exemple en ingénierie nucléaire

Nous allons maintenant présenter les bénéfices de cette démarche à travers un exemple en ingénierie nucléaire par comparaison entre des résultats accessibles par une analyse textuelle et le produit d'un travail de cartographie directement mené avec des experts sur le sujet des réacteurs nucléaires (CEA, 1975 ; 2008).

D'après deux éditions du *Dictionnaire des sciences et techniques nucléaire* du CEA dont un extrait est présenté ci-dessous, nous avons, au travers de l'étude des définitions à propos des différents types de réacteurs nucléaires, contruit un réseau sémantique à partir des données textuelles.

RÉACTEUR À CYCLE INDIRECT – Indirect cycle reactor.
Réacteur nucléaire dans lequel le *fluide primaire de refroidissement* transfère sa chaleur à un *fluide secondaire de refroidissement* pour produire la puissance utile.

RÉACTEUR À DÉRIVE SPECTRALE – Spectral shift reactor.
Synonyme de : *Réacteur à spectre réglable*.

RÉACTEUR À DOUBLE CYCLE – Dual-cycle reactor.
Réacteur nucléaire utilisant simultanément le mode d'extraction de la chaleur du r*éacteur à cycle direct* et du *réacteur à cycle indirect*.

RÉACTEUR À EAU BOUILLANTE – Boiling WATER reactor ; B.W.R.
Réacteur nucléaire dont le *fluide primaire de refroidissement* est de l'eau qui est portée à l'ébullition à l'intérieur du *cœur* du réacteur.

RÉACTEUR À EAU (OU À FLUIDE) SOUS PRESSION – Pressurized reactor.
Réacteur nucléaire dont le *fluide primaire de refroidissement* est maintenu sous une pression telle qu'une ébullition en masse ne peut pas se produire.
Nota : Le terme « réacteur à eau pressurisée » utilisé pour désigne ce type de réacteur est impropre, le mot pressurisé n'étant pas admis par l'Académie des Sciences.

RÉACTEUR À ÉCHANGEUR INTÉGRÉ – Integral reactor.
Réacteur nucléaire dans lequel l'échangeur de chaleur entre les *circuits primaire* et *secondaire de refroidissemen*t est contenu dans le caisson.

RÉACTEUR À FAISCEAUX SORTIS – Beam reactor.
Réacteur de recherche spécialement conçu pour produire des faisceaux de *neutrons* à l'extérieur du réacteur.

RÉACTEUR À HAUTE TEMPÉRATURE – High temperature reactor.
Réacteur nucléaire dont la conception et le fonctionnement mettent en jeu des problèmes technologiques spécifiques des hautes températures.

RÉACTEUR À LIT DE BOULETS – Pebble bed reactor.
Réacteur nucléaire dans lequel tout ou partie des matériaux essentiellement du réacteur (*combustible*, matière *fertile*, *modérateur*) se trouvent sous forme de boulets en contact les uns avec les autres et disposés en ligne stationnaire. L'avantage de cette disposition est de permettre d'ajouter ou de retirer du combustible sans arrêter le réacteur. Il est parfois désigné sous le nom de réacteur à boulets.

RÉACTEUR À LIT FLUIDISÉ – Fluidized bed reactor.
Réacteur nucléaire dans lequel le *combustible* se trouve sous la forme d'un lit de fines particules qui, en cours de fonctionnement, sont maintenues à l'état de suspension par le courant du *fluide de refroidissement*, mais non entraînées par celui-ci.

RÉACTEUR À NEUTRONS ÉPITHERMIQUES – Epithermal reactoc.
Réacteur nucléaire dans lequel la *fission* est produite principalement par des *neutrons épithermiques*.

RÉACTEUR À NEUTRONS INTERMÉDIAIRES – Intermediate neutron reactor.
Réacteur nucléaire dans lequel la *fission* est produite principalement par des *neutrons intermédiaires*. Synonyme : Réacteur à spectre intermédiaire.

RÉACTEUR À NEUTRONS RAPIDES – Fast neutron reactor.
Réacteur nucléaire dans lequel la *fission* est produite principalement par des *neutrons rapides*.

RÉACTEUR À NEUTRONS THERMIQUES – Thermal neutron reactor.
Réacteur nucléaire dans lequel la *fission* est produite principalement par des *neutrons thermiques*. Synonyme : Réacteur thermique.

RÉACTEUR AQUEUX – Aquaeous reactor.
Réacteur homogène dans lequel le *combustible* est en solution aqueuse.

RÉACTEUR À SELS FONDUS — Molten salt reactor.
Réacteur à combustible circulant dans lequel le *combustible* (souvent additionné de produits *fertiles*) se présente sous la forme de sels fondus.

RÉACTEUR À SPECTRE INTERMÉDIAIRE – Intermediate spectrum reactor.
Synonyme de : *Réacteur à neutrons intermédiaires*.

RÉACTEUR À SPECTRE MIXTE – Mixed-spectum reactor.
Réacteur nucléaire ayant des *spectres de neutrons* fortement différents en diverses parties du cœur.

RÉACTEUR À SPECTRE RÉGLABLE – Spectral shift reactor.
Réacteur nucléaire dans lequel le *spectre de neutrons* peut être ajusté pour assurer la *commande* du réacteur ou dans un autre but en modifiant les propriétés du *modérateur* ou la quantité de celui-ci. Synonyme : Réacteur à dérive spectrale.

RÉACTEUR À TUBES DE FORCE – Pressure tube reactor.
Réacteur nucléaire dans lequel *les assemblages combustibles* et le *fluide caloporteur* sont enfermés dans des tubes appelés tubes de force qui supportent la pression du fluide caloporteur. L'ensemble des tubes de force est entouré d'un réservoir, appelé calandre, contenant le *modérateur* sous une faible pression.

RÉACTEUR AU PLUTONIUM – Plutonium reactor.
Réacteur nucléaire alimenté en *combustible* dont le plutonium est la principale *matière *fissile*.

Source : Dictionnaire des sciences et techniques nucléaires.

L'ensemble des définitions ne permet de construire que des réseaux partiels et incomplets car toutes les définitions ne comportent pas le même niveau d'informations comme l'illustrent les deux schémas ci-après.

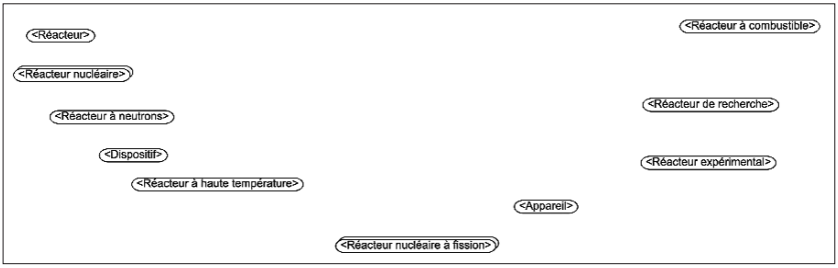


Figure 8.

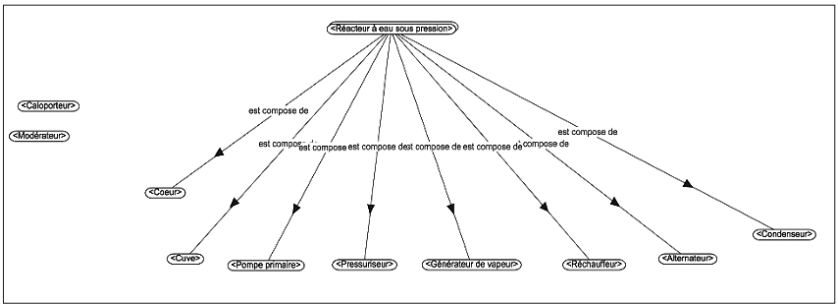


Figure 9.

Le travail avec un expert permet de construire très rapidement une représentation conceptuelle plus exhaustive comme illustré ci-après.

Finalement, il nous semble donc indispensable de placer le sachant au cœur de la démarche car il est le dépositaire de l'architecture des connaissances.

L'expérience nous montre également que l'analyse de la documentation n'est certainement pas suffisante, ni la meilleure source d'information pour commencer un travail de capitalisation des connaissances car celle-ci ne contient pas nécessairement les connaissances essentielles pour concevoir, fabriquer ou maintenir des équipements techniques à longue durée de vie.

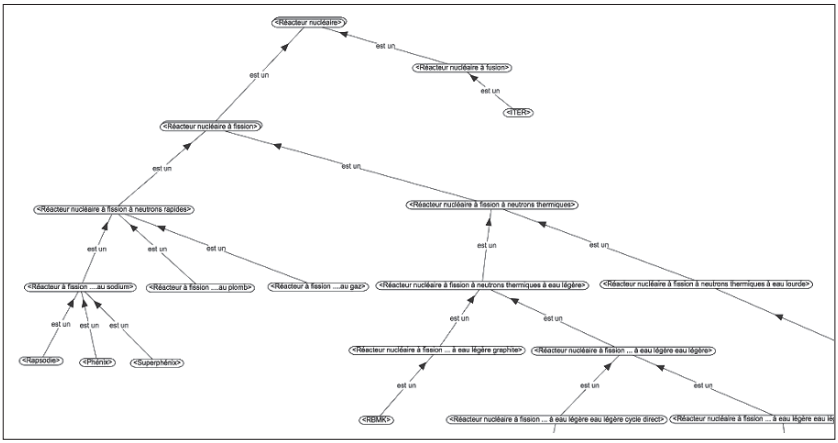


Figure 10.

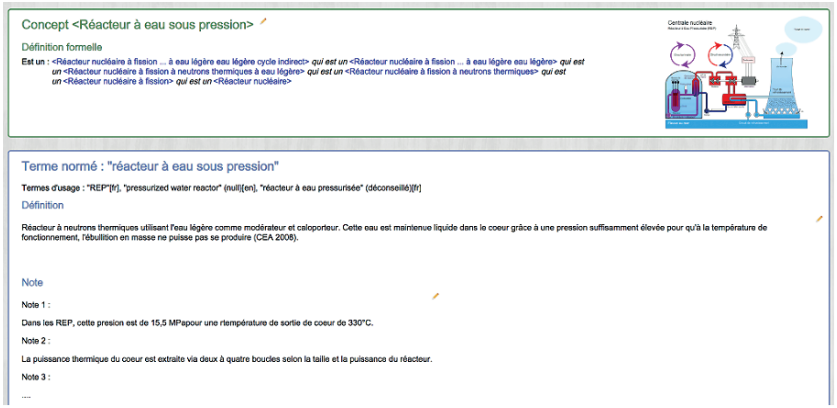


Figure 11.

Cependant, une fois des cartographies établies, l'intégration d'un échantillon de documentation dûment sélectionné dans cette architecture prendra tout son sens pour apporter du contenu au sein d'une démarche structurée facilitant ainsi la diffusion et le partage des connaissances au sein de l'entreprise.

5. En guise de conclusion

L'**illusion d'exhaustivité documentaire** constitue le premier écueil : tout n'est pas dans les textes et l'on ne capitalise pas en commençant par la documentation. L'**architecture des connaissances** est d'abord portée par les experts ; c'est à partir d'eux que l'on élabore la carte, à laquelle on **relie ensuite**, de manière sélective et critique, un **échantillon de documents** venant étayer et compléter le contenu.

Cette approche requiert une **disponibilité réelle, quoique mesurée**, des sachants. Sans leur participation, la démarche n'avance pas. Obtenir l'**adhésion** suppose d'expliquer clairement la méthode, son déroulé et le rôle de chacun à l'ensemble des intervenants.

Enfin, il ne s'agit ni d'un **système clé en main** ni d'un logiciel promettant un résultat automatique. La qualité de la cartographie tient à un **travail d'explicitation**, de mise à l'épreuve et d'arbitrage conceptuel - les outils n'en sont que les supports. C'est à ce prix que la capitalisation produit une **structure fiable et transmissible**, utile à l'ingénierie, à la maintenance et à la formation.

Et les outils de type LLM ne sont pas encore au niveau pour proposer des conceptions fiables comme le montrent ces deux illustrations.

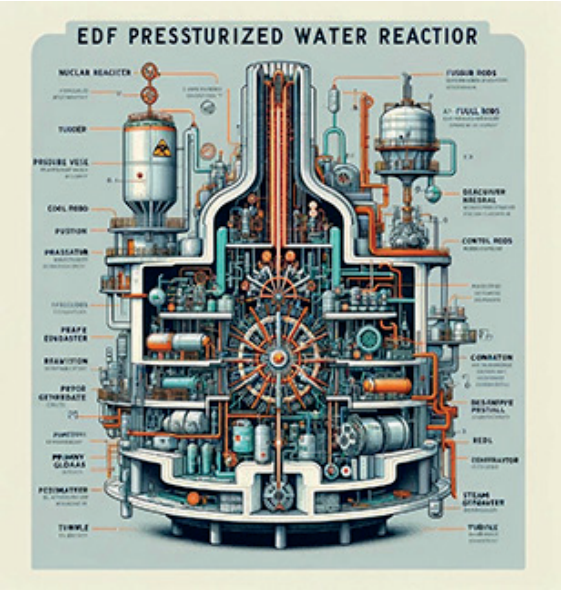


Figure 12.

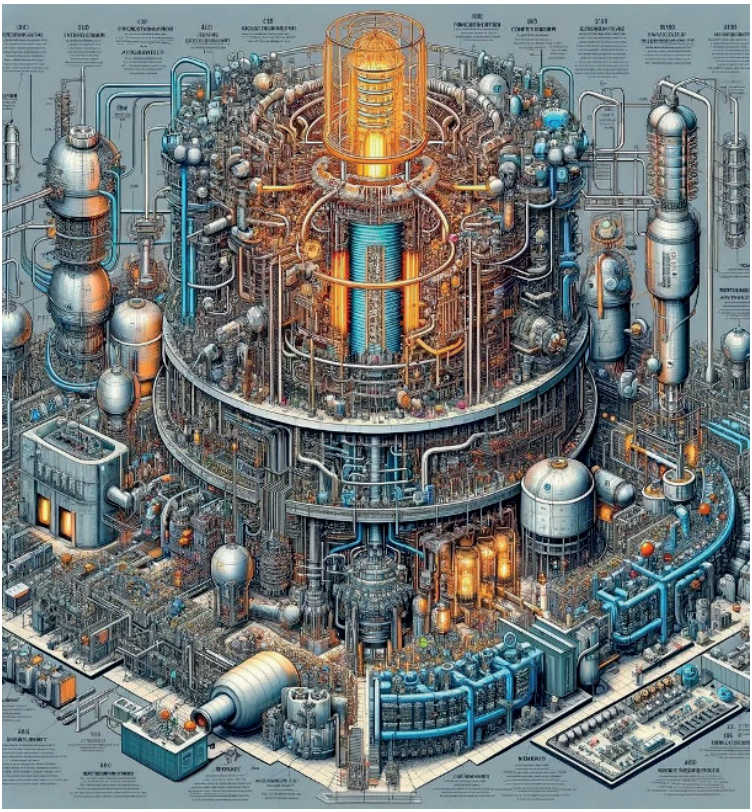


Figure 13.

Images générées artificiellement.

Références

Akrich, M. (1987). Comment décrire les objets techniques ?. *Techniques et culture*, n° 9, p. 49-64.

Aristote. (1997). *Éthique à Nicomaque, Traductions et notes par J. Tricot*. Vrin, « Bibliothèque des Textes Philosophiques, Poche ».

Calberg Challot, M., Lerat, P. & Roche, C. (2009). « Quelle place accorder aux corpus dans la construction d'une terminologie ». *Actes de la 3e conférence TOTh « Terminologie et Ontologie : Théories et Applications »*, Annecy, éd. Institut Porphyre, *Savoir et Connaissance*, p. 33-52.

Dictionnaire des sciences et techniques Nucléaires. (1975). Commissariat à l'énergie atomique (CEA), [3e éd.], Paris : Eyrolles.

Dictionnaire des sciences et techniques Nucléaires. (2008). Commissariat à l'énergie atomique (CEA) [4e éd.], Paris : Onisciences.

Felber, H. (1987). *Manuel de terminologie*. Unesco/INFOTERM.

Lerat, P. (2016). *Langue et technique*, Paris : Hermann, coll. « Vertige de la langue ».

Leroi-Gourhan, A. (1945). *Milieu et technique*s, Paris, Albin Michel.

Medhat-Lecocq, H. (2021). *Terminologie comparée et traduction*. *Approche interdisciplinaire*, Éditions des archives contemporaines, coll. « Inter-Culturel ».

Norme française ISO 1087. (2020). *Travail terminologique et science de la terminologie. Vocabulaire*, AFNOR/Normalisation.

Roche, C. (2008). « Le terme et le concept : fondements d'une ontoterminologie », *Actes de la 1ère conférence TOTh 2007, Terminologie & Ontologie : Théories et applications*, Annecy, éd. Institut Porphyre, *Savoir et Connaissance*, p. 1-22.

Roche, C. (2005). « Terminologie et ontologie », *Langages*, n° 157, « La terminologie : nature et enjeux », p. 48-62.

Roche, J. (2015). *Le tournant ontologique de la terminologie*, thèse en informatique et langage, Université Grenoble Alpes.

Simondon, G. (2008). *Du mode d'existence des objets techniques* [1958]. Paris : Aubier.

Wüster, E. (1968). *Dictionnaire multilingue de la Machine-Outil, Notions fondamentales, définies et illustrées, présentées dans l'ordre systématique et dans l'ordre alphabétique*, Londres, Technical Press, [1ère édition : Organisation des Nations unies].

Annotation automatique des classes sémantiques dans des comptes rendus de bilans orthophoniques : intérêt de la coopération entre terrain clinique et de recherche

Frédérique Brin-Henry*, Tiphaine Le Clercq de Lannoy**, Romaric Besançon**, Olivier Ferret**, Julien Tourille**, Bianca Vieru**

Résumé. L'exploitation automatique des comptes rendus médicaux offre des opportunités dans le champ de la recherche clinique ou de la pratique médicale. En particulier, caractériser automatiquement les concepts médicaux dans les dossiers patients est une brique de base permettant de mettre en œuvre des tâches plus complexes comme l'étude de la faisabilité d'études cliniques ou l'exploration des pratiques médicales spécifiques. Or, pour un domaine de spécialité comme l'orthophonie, les ressources nécessaires pour construire des modèles automatiques pour cette caractérisation sont rares. Nous présentons dans cet article une approche pour la reconnaissance automatique de concepts médicaux dans des comptes rendus adaptés spécifiquement au domaine de l'orthophonie. Nous mettons en évidence certaines caractéristiques terminologiques propres au domaine de l'orthophonie et proposons une approche pour construire un modèle spécifique en exploitant à la fois des ressources propres au domaine et l'affinage de modèles neuronaux sur des données pseudo-annotées. Nous montrons que cette approche permet d'améliorer les performances du modèle sur un corpus de comptes rendus de bilans orthophoniques.

Abstract. *The automatic exploitation of medical health reports (EHRs) offers opportunities in the field of clinical research or medical practice. Specifically, the automatic identification of medical concepts in EHRs is a fundamental building block for implementing more complex tasks, such as assessing the feasibility of clinical studies or exploring specific medical practices. However, for specialized domains such as speech and language therapy (SLT), the resources required to build automatic models for this identification are scarce. In this article, we present an approach for the automatic recognition of medical concepts in EHRs specifically adapted to the domain of SLT. We highlight certain terminological characteristics unique to the domain and propose an approach to build a specific model by leveraging both*

domain-specific resources and the fine-tuning of neural models on pseudo-labeled data. We demonstrate that this approach enhances the model's performance on a corpus of SLT EHRs.

1. Introduction

Le projet POPEX a pour objectif d'évaluer, pour les établissements de santé, l'intérêt des outils de recherche d'information permettant la sélection d'une population ciblée. En effet, pour bien répondre aux études de faisabilité d'études cliniques, l'investigateur principal d'un centre hospitalier doit s'assurer, préalablement à la mise en place de l'étude, qu'il dispose théoriquement de suffisamment de cas de patients répondant aux critères de sélection. Il s'agit souvent d'identifier des patients ayant présenté une combinaison spécifique de pathologies ou de symptômes : par exemple des personnes présentant un diabète de type 2, traitées par pompe à insuline depuis plus d'un an.

Par ailleurs, lorsque le professionnel de santé souhaite évaluer ses propres pratiques, il lui est possible, sauf non-accord du patient, de réutiliser des données de santé (recueillies dans le cadre du soin) à des fins de recherche. Pour les praticiens, l'intérêt est de questionner des pratiques locales, en les confrontant à la littérature scientifique pour élaborer des procédures cliniques, comme le préconise l'*evidence-based-practice* (Sackett *et al.*, 1996). Nous nous intéressons dans cet article au cas de l'orthophonie. Dans cette perspective de pratiques probantes actuellement enseignées (Witko *et al.*, 2021), l'orthophoniste pourra ainsi souhaiter interroger ses pratiques au sein d'un parcours de soins, comme les décisions relatives à la pose d'une sonde naso-gastrique dans le cadre d'une dysphagie neurologique post-AVC. L'orthophonie pourra également souhaiter explorer des variables telles que les différents symptômes constatés ou les comorbidités mises en évidence.

2. Objectifs

Pour répondre à ces besoins de terrain, il est souhaitable de mettre à disposition des moyens dédiés facilitant cette démarche. En effet, les systèmes actuels de codification (PMSI), constituant des données structurées, ne permettent pas d'accéder aux informations diverses présentes dans les notes ou les comptes rendus de bilan orthophonique (CRBO). L'automatisation des processus permet un gain de temps quant à l'identification des patients concernés et l'exploration du contenu des dossiers médicaux ou paramédicaux permet la

réflexion sur des pratiques de terrain et contribue à la qualité des soins (HAS, 2023[1]). Le projet POPEX vise la mise au point d'une méthode d'identification automatique de données non structurées dans les textes médicaux, permettant d'extraire des unités lexicales pertinentes du discours des professionnels de santé. Nous supposons que la plupart de ces unités lexicales sont des termes désignant des concepts au sens terminologique défini par la norme ISO 1087-1[2] (Desprès *et al.*, 2020).

Le défi que posent ces objectifs est la qualité de cette extraction terminologique automatique, puisqu'elle dépend des méthodes utilisées. En effet, comme pour tout usage terminologique, des variantes et des équivalents peuvent être en usage dans la pratique orthophonique (exemple : « sonde nasogastrique », « alimentation entérale », « SNG »), limitant ainsi la qualité de la réponse aux requêtes du soignant (Walsh, 2005). De plus, on note l'utilisation fréquente en orthophonie d'unités polylexicales (Brin-Henry & Knittel, 2016), qui favorisent l'ambiguïté et compliquent l'extraction.

Cet article abordera ces enjeux ainsi que les méthodes utilisées, en coopération pluridisciplinaire, pour répondre progressivement aux problématiques suivantes :

- Comment procéder à une validation de l'extraction des candidats termes pour permettre la généralisation ?
- Quelle est la place de l'expert dans la validation de la précision de l'extraction ?
- Comment identifier les axes d'amélioration de l'annotation en s'appuyant sur des analyses linguistiques ?
- Comment ces travaux contribuent à distinguer la langue orthophonique de la langue médicale ?

3. Méthode

Dans le but d'annoter des CRBO et d'identifier les termes spécifiques du domaine de l'orthophonie, nous avons utilisé des méthodes connues en visant une adaptation au domaine par l'utilisation de ressources du domaine médical général complétées par d'autres ressources du domaine de l'orthophonie, en particulier un corpus non annoté et une terminologie de l'orthophonie. L'idée

1 https://www.has-sante.fr/upload/docs/application/pdf/2023-09/fiche_pedagogique_dpa.pdf.

2 Le terme est une « désignation verbale d'un concept général dans un domaine spécifique » (ISO 1087-1).

est de partir de modèles déjà capables de reconnaître un certain nombre de termes, puis d'exploiter leur capacité de généralisation pour acquérir de nouveaux termes à partir d'un corpus représentatif du domaine. L'utilisation d'un corpus cible de CRBO permet de valider la méthodologie.

3.1. Modèles d'extraction d'entités médicales

Afin d'extraire les candidats termes, nous avons combiné, avec un vote majoritaire, les résultats de trois méthodes : une approche à base de connaissances (qui permet d'identifier des termes connus dans les textes) et deux approches par apprentissage automatique exploitant des modèles neuronaux (qui permettent de généraliser la reconnaissance des termes en exploitant le contexte de la phrase).

Pour l'approche à base de connaissances, nous avons défini et implémenté un outil, QuickMatching, fondé sur l'algorithme SimString (Okazaki & Tsujii, 2010) et comparable à QuickUMLS (Soldaini & Goharian, 2016). Il permet de calculer la similarité entre des termes de référence et les mots des textes sur la base d'un découpage en n-grammes. Ces termes de référence proviennent de diverses sources biomédicales, telles que le méta-thésaurus UMLS (*Unified Medical Language System* [Lindberg *et al.*, 1993]) et la base de données publique des médicaments (ANSM, 2021). La méthodologie adoptée vise notamment l'extension de la terminologie utilisée par Quickmatching au domaine de l'orthophonie.

Les approches à base d'apprentissage automatique considèrent quant à elles la tâche d'identification des candidats termes comme une tâche d'annotation en entités nommées médicales.

Le premier modèle à base d'apprentissage que nous utilisons s'appuie, de façon standard, sur une architecture neuronale de type *transformers* (Vaswani *et al.*, 2017), en suivant les travaux de (Devlin *et al.*, 2019). Dans notre cas, nous avons utilisé le modèle entraîné pour le français CamemBERT (Martin *et al.*, 2020) comme modèle de langue initial. Ce modèle a été adapté au domaine médical en effectuant une poursuite du pré-entraînement sur des textes du domaine (*cf.* tableau 1 pour les corpus utilisés) et finalement affiné pour la tâche d'annotation en entités grâce au corpus QUAERO (Névéol *et al.*, 2014).

Corpus	Description	Taille
OrthoCorpus (2019)	Articles de la revue spécialisée *Rééducation orthophonique* (Brin-Henry, 2018)	6,7 M
ISTEX	Articles de revues médicales indexées par ISTEX (Inist)	42,6 M
EQueR	Articles scientifiques et de recommandations de bonne pratique médicale (CIS-MeF) (Ayache *et al.*, 2006)	16,8 M
PMC OA	Articles de revues médicales (PubMed Central Open Access)	3,8 M
Cochrane	Résumés d'articles de l'organisation Cochrane	5,0 M
EMA	Notices de l'Agence européenne des médicaments	21,2 M
CRTT	Articles de revues, extraits de *Science Direct*.	21,7 M
E3C	Résumés d'articles, articles de revues, cas cliniques (Magnini *et al.*, 2021)	12,1 M
Wikipédia	Articles Wikipédia dans le domaine medical (Bawden *et. al.*, 2020).	6,6 M

Tableau 1. Collections de textes médicaux utilisées. Tailles : en millions de mots.

La seconde approche neuronale considérée est le modèle Pyramid (Wang *et al.*, 2020), dédié plus spécifiquement au traitement des termes imbriqués. Ces derniers sont en effet fréquents dans le domaine médical et en particulier dans le corpus QUAERO, sur lequel le modèle a également été entraîné. Les deux approches neuronales n'ayant pas la même architecture, les résultats diffèrent mais se complètent.

Ces trois approches ont été combinées à l'aide d'un vote majoritaire sur les résultats : pour qu'une entité (mot ou ensemble de mots associés à une étiquette) soit sélectionnée dans le résultat final, elle doit être prédite par au moins deux des modèles.

Cette approche est motivée par le fait que, les modèles ayant des fonctionnements et des entraînements différents, les entités se retrouvant dans l'intersection de leurs résultats sont les meilleures candidates pour l'annotation. Cette combinaison donne de fait les meilleurs résultats expérimentaux sur le corpus QUAERO.

3.2. Application de l'extraction des entités médicales à l'orthophonie

3.2.1. Extraction des termes de la collection OrthoCorpus

L'application combinée des trois approches d'identification d'entités médicales a été appliquée sur un corpus représentatif du domaine de l'orthophonie, OrthoCorpus[3], et a permis de collecter un ensemble de plus de 11 000 candidats termes, dont plus de la moitié sont des expressions polylexicales (Candito *et al.*, 2020), organisées selon dix des classes sémantiques de l'UMLS. Certains de ces types ont une pertinence réelle du point de vue de la pratique de l'orthophonie (*Anatomy*, *Disorders*, *Physiology*, *Procedures*, *Devices*, *Living Beings*), notamment pour ce qui concerne les unités polylexicales. De plus, les candidats termes très fréquents semblent bien correspondre à la spécificité de la pratique orthophonique. Par exemple, on trouve « langue », « cérébral », « cordes vocales » et « système nerveux » parmi les termes d'anatomie les plus fréquents, « trouble du langage » et « trouble de la déglutition » parmi les pathologies les plus citées, même si cette catégorie est très générale et mêle étiologie (« TC », « lésion cérébrale »), symptômes (modifications posturales, refus alimentaire) et résultats (« échec scolaire »).

3.2.2. Extraction de patrons syntaxico-sémantiques

Nous avons filtré les résultats obtenus par les votes sur OrthoCorpus par l'utilisation d'un ensemble de patrons syntaxico-sémantiques de termes médicaux. Pour ce faire, nous utilisons les classes sémantiques du méta-thésaurus UMLS. Plus précisément, les noms et adjectifs de l'UMLS sont d'abord extraits et associés automatiquement à leurs classes sémantiques. D'autres classes d'adjectifs sont également ajoutées, comme les adjectifs de quantification. Les termes complexes sont alors représentés comme des patrons syntaxico-sémantiques associant les catégories grammaticales et les types sémantiques de l'UMLS : par exemple, on identifie le patron *NOUN_diso ADJ_anat* indiquant un nom de pathologie suivi d'un adjectif anatomique. Ces patrons sont alors associés à des types sémantiques spécifiques en les projetant dans une annotation de référence (QUAERO). Certains patrons sont peu ambigus alors que d'autres, comme *NOUN_proc ADJ_anat* (nom de procédure suivi d'un adjectif anatomique), se retrouvent en pratique dans des concepts

3 Analyse et traitement informatique de la langue française - UMR 7118 (ATILF) (2023). *OrthoCorpus* [Corpus]. ORTOLANG (*Open Resources and TOols for LANGuage*) en ligne : https://hdl.handle.net/11403/orthocorpus/v3.

de types différents. Dans la référence, nous obtenons ainsi ce patron pour le terme « endartériectomie carotidienne », annoté comme *Procedures*, et pour « coagulation sanguine », annoté en tant que *Physiology*. Pour la seconde expression, la classification de chaque terme étant ambiguë, cela explique l'obtention d'un tel patron.

Globalement, ces premiers résultats obtenus à partir du corpus OrthoCorpus n'étaient pas totalement satisfaisants du point de vue de l'expert et montraient beaucoup d'erreurs d'attribution des concepts UMLS dans les textes. Par exemple, « système phonologique », « lait », « mémoire » ont été classés comme *Anatomy*, et « intervention chirurgicale » sous le concept *Disorders*.

3.3. Développement d'un modèle d'extraction de concepts dédié à l'orthophonie : OrthophoNER

3.3.1. Exploitation de ressources supplémentaires

Notre approche d'extraction d'entités s'appuie sur plusieurs modèles, dont un modèle à base de connaissances. Pour une meilleure prise en compte des entités spécifiques de l'orthophonie, nous proposons donc d'étendre la source des connaissances exploitées par ce modèle en y intégrant des termes provenant de ressources supplémentaires. En l'occurrence, nous avons décidé d'exploiter le Dictionnaire d'orthophonie (Brin-Henry *et al*., 2020).

Une première étape de cette intégration a consisté à catégoriser les termes s'identifiant à des entrées du dictionnaire dans le schéma des types d'entités existants, qui correspondent à des types sémantiques de l'UMLS. Cette étape a été réalisée en combinant une pré-classification automatique des termes et une vérification manuelle effectuée par une orthophoniste experte. Cette classification a ainsi offert une perspective de comparaison entre le domaine médical général et l'orthophonie en même temps qu'elle a servi de filtre pour identifier de nouveaux termes non connus. La démarche nous permet d'attribuer des classes aux entrées du dictionnaire, qui devient intégrable dans QuickMatching malgré certaines ambiguïtés dans l'attribution des classes sémantiques. Par exemple, le terme « langue » peut être classé en tant que « *Anatomy* » dans un contexte relatif à la perte sensitive de la langue du patient, ou bien en tant que *Concepts_&_Idea*s dans une discussion sur la langue maternelle du patient. De même, « articulation » peut faire référence à une jointure entre deux os, comme celle reliant la mandibule au crâne, mais apparaître également en tant que *Concepts_&_Ideas* en référence à la deuxième ou troisième articulation du langage.

Il est donc nécessaire d'effectuer une désambiguïsation des annotations de QuickMatching. Pour cela, nous proposons un vote avec ambiguïtés : la désambiguïsation n'est pas réalisée pendant la prédiction de QuickMatching mais pendant l'étape de vote afin de profiter des connaissances des autres modèles sur les entités à désambiguïser. Si CamemBERT ou Pyramid prédit l'une des étiquettes considérées comme ambiguës par QuickMatching, alors cette étiquette est choisie pour annoter l'expression.

3.3.2. Exploitation de modules supplémentaires : extension grammaticale des entités

L'extraction de patrons syntaxico-sémantiques à partir des termes extraits de la collection Orthocorpus nous a permis de comparer la taille des patrons issus de textes biomédicaux et des textes orthophoniques et nous a amené à constater la présence d'entités souvent plus longues dans le domaine de l'orthophonie par rapport au domaine médical général. En effet, les orthophonistes peuvent avoir besoin de combiner des termes afin d'obtenir des expressions satisfaisantes pour traduire la réalité des pathologies auxquelles ils sont confrontés.

Comme les modèles biomédicaux ne sont pas entraînés à repérer des entités d'aussi grande taille, la plupart des entités trouvées dans des textes orthophoniques ne sont pas complètes. Nous avons donc développé un module prenant en entrée des entités déjà repérées dans un texte et cherchant à compléter ces entités à l'aide de compléments du nom et d'adjectifs. Par exemple, dans le cas où un modèle biomédical détecte le mot « trouble » issu de l'expression « trouble de la compréhension orale » dans un texte orthophonique, le module d'extension ajoute à l'entité la partie « de la compréhension orale » au mot « trouble ».

Cette méthode permet d'obtenir davantage d'entités orthophoniques correctes et d'adapter les modèles biomédicaux à l'orthophonie.

3.3.3. Entraînement des modèles à base d'apprentissage automatique

Les modèles neuronaux de reconnaissance d'entités médicales s'appuient sur des données annotées pour leur entraînement. Or, des données de ce type ne sont pas directement disponibles pour un corpus de spécialité comme l'orthophonie.

Pour entraîner les modèles, nous avons donc choisi de constituer une collection de textes spécifiques à l'orthophonie, d'effectuer une annotation

automatique de ces textes avec le modèle existant dédié à l'orthophonie (enrichi par les ressources et modules supplémentaires présentés dans les sections précédentes) et d'exploiter cette collection pseudo-annotée pour l'entraînement des modèles.

Plus précisément, nous avons constitué un corpus de 413 CRBO, qui ont été anonymisés par une correction manuelle de l'anonymisation automatique réalisée par l'outil MEDINA (Grouin & Zweigenbaum, 2013), conformément aux procédures de traitement et de protection des données conventionnellement fixées entre les différentes parties prenantes. Ces CRBO ont été annotés avec le modèle dédié à l'orthophonie, et 16 CRBO ont été extraits de ce corpus et corrigés manuellement en étroite collaboration avec un expert (ces documents correspondent à environ 7 000 mots et 2 867 entités annotées). Ces textes constituent le corpus de référence CRBO-test utilisé pour notre évaluation.

Les 397 CRBO restants annotés automatiquement ont été utilisés, en combinaison avec le corpus QUAERO, pour entraîner un modèle CamemBERT, pré-entraîné sur l'ensemble des corpus biomédicaux et orthophoniques présentés dans le tableau 1 *supra*, et affiné pour la tâche de détection d'entités nommées. Ce modèle CamemBERT spécialisé pour l'orthophonie est appelé OrthophoNER.

4. Expériences

4.1. Cadre d'évaluation

Afin d'évaluer nos modèles biomédicaux et orthophoniques, nous utilisons la collection CRBO-test constituée des 16 CRBO annotés manuellement.

Les métriques utilisées pour les expériences sont celles proposées pour l'évaluation du corpus QUAERO, soit la précision (qui indique la proportion d'entités correctes parmi les entités prédites), le rappel (qui indique la proportion d'entités prédites sur l'ensemble des entités attendues) et le F1-score, qui est une moyenne harmonique de la précision et du rappel. Les mesures sont calculées en mode strict (les frontières des entités doivent être strictement identiques) et agrégées en mode micro. Elles ont été calculées à l'aide de l'outil BRATEval ehealth, développé par Verspoor *et al.* (2013) et modifié par Névéol *et al.* (2014).

4.2. Résultats

Les résultats des modèles biomédicaux et orthophoniques sont présentés sur le corpus CRBO-test afin de mieux visualiser l'impact des adaptations à l'orthophonie réalisées sur les modèles d'annotation.

Tout d'abord, nous pouvons noter l'effet de l'ajout de termes issus du dictionnaire orthophonique sur les résultats de QuickMatching dans les deux premières lignes du tableau 2 *infra* : l'ajout de termes permet d'améliorer le rappel de 22,93 % à 35,84 % grâce à l'amélioration de la couverture du modèle mais conduit également à une diminution de la précision, ce qui est probablement causé par l'apparition d'ambiguïtés.

L'absence d'adaptation à l'orthophonie du modèle Pyramid permet d'expliquer le faible F1-score (16,95 %). De la même façon, le modèle CamemBERT biomédical est pré-entraîné sur un corpus de textes biomédicaux (*cf.* tableau 1 *supra*) contenant le corpus OrthoCorpus. Toutefois, son entraînement a été réalisé sur un corpus biomédical (QUAERO) et non orthophonique et ses résultats ne dépassent pas les résultats du modèle à base de connaissances.

En revanche, la combinaison des résultats de ces trois modèles (QuickMatching, Pyramid et CamemBERT) et du vote avec ambiguïtés (ligne « vote avec ambiguïtés ») permet d'améliorer le F1-score d'environ 5 points par rapport au meilleur modèle. La précision est améliorée par rapport à l'ensemble des modèles : en effet, une entité est plus probable si elle est prédite par au moins deux modèles.

Modèles	**Mesures d'évaluation sur CRBO-test**		
	Précision	**Rappel**	**F1-score**
QuickMatching non adapté à l'orthophonie	50,56 %	22,93 %	31,55 %
QuickMatching adapté à l'orthophonie	42,52 %	35,84 %	38,90 %
Pyramid	40,42 %	10,73 %	16,95 %
CamemBERT biomédical	53,38 %	20,06 %	29,12 %
Vote avec ambiguïtés	58,76 %	35,13 %	43,97 %
Vote avec post-traitements	38,73 %	43,57 %	41,00 %
OrthophoNER	50,52 %	41,73 %	**45,70 %**

Tableau 2. Évaluation des différents modèles proposés pour le domaine de l'orthophonie (corpus de test des CRBO).

Nous pouvons également constater que l'ajout de post-traitements spécifiques (extension de compléments du nom et d'adjectifs) au vote avec ambiguïtés permet certes d'améliorer la couverture des annotations (le rappel est amélioré d'un peu plus de 8 points) mais ajoute du bruit dans les résultats. Cela est dû à des extensions excessives d'entités : le module ajoute tous les compléments du nom et adjectifs possibles, sans tenir compte du sens de l'entité.

Enfin, nous pouvons noter que le modèle OrthophoNER, entraîné conjointement sur QUAERO et le corpus de CRBO annotés automatiquement, apporte une amélioration significative d'environ 2 points de F1-score par rapport au vote avec ambiguïtés. Cette amélioration nous montre que l'entraînement sur le corpus CRBO permet d'adapter le modèle au domaine de l'orthophonie, même en exploitant des pseudo-annotations automatiques (*silver standard*). Les erreurs potentielles de ces pseudo-annotations dans le corpus d'apprentissage sont en partie contrebalancées par l'entraînement sur le corpus QUAERO, ce qui permet au modèle de bénéficier également d'entités de qualité, car annotées manuellement (*gold standard*).

Un avantage du modèle OrthophoNER réside également dans le fait qu'il n'est composé que d'un seul modèle, là où le vote avec post-traitements nécessite une combinaison spécifique de plusieurs modèles et modules, ce qui est plus coûteux en temps et ressources.

5. Conclusion et discussion

Nous avons présenté une approche hybride pour l'extraction terminologique à partir de comptes rendus orthophoniques, approche fondée sur des modèles de reconnaissance des entités nommées. Pour compenser l'absence de corpus annotés dans le domaine cible de l'orthophonie, cette approche combine des méthodes à base de règles et des modèles d'apprentissage profond pour la reconnaissance d'entités nommées biomédicales générale et leur associe l'utilisation d'une terminologie du domaine de l'orthophonie pour annoter automatiquement des comptes rendus d'orthophonie. Les rapports ainsi annotés sont ensuite utilisés pour entraîner un modèle CamemBERT pour le domaine de l'orthophonie, appelé OrthophoNER. Les évaluations menées sur un ensemble de comptes rendus anonymisés et annotés manuellement ont montré que l'entraînement d'un modèle utilisant une combinaison d'annotations en entités générées automatiquement pour notre domaine cible, l'orthophonie, et d'entités annotées manuellement pour notre domaine source, le domaine médical, donne de meilleurs résultats qu'un modèle s'appuyant uniquement sur les

annotations pour le domaine source. En outre, le modèle ainsi entraîné permet une certaine généralisation conduisant à identifier de nouvelles entités pour le domaine cible, ce qui peut contribuer à enrichir les terminologies spécifiques de ce domaine.

Globalement, la méthode présentée montre la nécessaire progression par étapes, en interdisciplinarité, de l'annotation automatique de concepts en orthophonie. Sur un plan clinique, cette extraction permet de rendre compte de l'étendue et de la spécificité du discours médical et orthophonique. Sur un plan fondamental, le projet permet d'enrichir les modèles et de réfléchir sur les plans conceptuels et linguistiques.

Nous avons rencontré des difficultés liées à la structure complexe des unités polylexicales. Il a fallu évaluer dans quelle mesure il était pertinent d'étendre l'extraction, au risque de multiplier des erreurs d'identification de structures faussement équivalentes. Le problème des entités disjointes représente également un défi pour la juste sélection des structures syntaxico-sémantiques. Par exemple, le terme « trouble de l'expression et de la compréhension » peut être identifié comme un terme complet (c'est un trouble qui affecte à la fois l'expression et la compréhension), qui n'est cependant pas totalement équivalent à un « trouble de l'expression » associé à un « trouble de la compréhension ». Dans ce dernier cas, il pourrait s'agir d'une entité disjointe qui nécessite alors une annotation manuelle. Cela est également le cas pour l'exemple « compréhension orale et écrite ».

Par ailleurs, le problème de l'homonymie des unités lexicales est particulièrement présent, pour des raisons sémantiques mais également parce qu'aucune annotation syntaxique n'a été ajoutée. Ainsi le terme « marqueurs », identifié comme Nom est classé dans les produits chimiques dans l'UMLS (dans le cas de marqueurs sanguins ou biologiques), alors qu'il fait référence à des notions linguistiques (« des marqueurs du nom ») ou revêt une fonction de support de considérations cliniques (« la présence de dysarthrie est un marqueur de mauvais pronostic »).

Plus largement, il reste difficile de distinguer ce qui fait partie d'un lexique transdisciplinaire en santé de ce qui est spécifique à l'orthophonie. Une des réponses possibles serait de renforcer l'analyse linguistique en identifiant des couples coexistants tels que « troubles de la déglutition »/« troubles de déglutition », ainsi que l'utilisation du singulier et du pluriel comme dans « trouble de l'oralité » *vs* « troubles de l'oralité ». La distinction entre des groupes nominaux ne peut se faire qu'au sein d'une discussion experte relative aux concepts que désignent ces termes.

6. Perspectives

La poursuite des travaux permettra de terminer l'annotation manuelle sur un échantillon de comptes rendus de bilan pour disposer d'un corpus annoté de référence permettant d'évaluer l'ensemble de la démarche. Nous pourrons également valider un lexique permettant de corriger plusieurs erreurs à partir du recueil des équivalents tel que celui effectué concernant la sonde nasogastrique (SNG = sonde gastrique = sonde nasogastrique).

Une autre piste envisagée est d'utiliser un modèle conceptuel tel que SNOMED-CT, ou une termino-ontologie spécifique au domaine de l'orthophonie, comme celle résultant du projet MOCOLANG-O (Brin-Henry, Costa & Desprès, 2019). Cela permettrait de garantir une interopérabilité grâce à l'utilisation des concepts et des relations entre les concepts. Par exemple, les « implants cochléaires » peuvent être considérés comme une sous-partie des prothèses auditives en tant que dispositif médical mais restent spécifiques en tant que procédure.

Ces travaux ont bénéficié d'un financement dans le cadre du programme e-Meuse Santé, porté par le Département de la Meuse et soutenu par les Départements de la Haute-Marne et de la Meurthe et Moselle, les GIP Objectif Meuse et Haute-Marne, la Région Grand Est, l'Agence Régionale de Santé Grand Est, et la Banque des Territoires au titre du programme France 2030. Ils ont été réalisés grâce au supercalculateur Factory-IA financé par le Conseil régional d'Île-de-France.

Références

ANSM 2021. Base de données publique des médicaments.

Brin-Henry, F. & Knittel, ML. 2016. « Étude lexicosémantique du nom difficulté(s) dans les comptes rendus de bilan orthophonique : apports structuraux et conceptuels ». *LIDIL*, nº 53.

Brin-Henry, F., Courrier, C., Lederle, E., Masy, V. 2020. *Dictionnaire d'orthophonie. 4e édition révisée*. Isbergues : Ortho-Edition.

Brin-Henry, F., Costa, R., Desprès, S. 2019. « TemPO : towards a conceptualisation of pathology in speech and language therapy ». Terminologie et intelligence artificielle. Atelier TALN-RECITAL et IC (PFIA 2019), Toulouse, France. hal-02193859

Candito, M., Constant, M., Ramisch, C., Savary, A., Guillaume, B., Parmentier, Y. & Cordeiro, S. R. 2020. « A French corpus annotated for multiword expressions and named entities ». *Journal of Language Modelling*, vol 8, nº 2, p. 415-479. doi.org/10.15398/jlm.v8i2.265 , hal-03016721

Desprès, S., Roche, C., Papadopoulou, M. 2020. « Étude comparative de deux méthodes outillées pour la construction de terminologies et d'ontologies », *Terminology & Ontology: Theories and applications*. hal-02904475

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 2019. « BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding ». Dans *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2019), p. 4171-4186, Minneapolis, Minnesota.

Grouin, C., Zweigenbaum, P. 2013.« Automatic De-Identification of French Clinical Records: Comparison of Rule-Based and Machine-Learning Approaches », *MEDINFO*, 2013.

Lindberg, D. A., Humphreys, B. L. & Mccray, A. T. 1993. *The unified medical language system*. *Methods of information in medicine*, vol. 32, nº 4, p. 281-291.

Névéol, A., Grouin, C., Leixa, J., Rosset, S. & Zweigenbaum, P. 2014. « The QUAERO French medical corpus: A resource for medical entity recognition and normalization ». Dans *LREC Workshop BioTxtM2014*, p. 24–30.

Okazaki, N. & Tsujii, J. 2010. « Simple and Efficient Algorithm for Approximate Dictionary Matching ». Dans *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING 2010), p. 851-859, Beijing, China.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B. & Richardson, W.S. 1996.« Evidence based medicine: what it is and what it isn't », *BMJ,* vol. 312, nº 7023, p. 71-72. doi-org.bases-doc.univ-lorraine.fr/10.1136/bmj.312.7023.71

Soldaini, L. & Goharian, N. 2016. « QuickUMLS: a fast, unsupervised approach for medical concept extraction ». Dans *SIGIR MedIR workshop*, p. 1-4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. 2017. « Attention is all you need. Advances in Neural Information ». *Processing Systems*, p. 5998-6008.

Verspoor, K., Jimeno Yepes, A., Cavedon, L., Mcintosh, T., Herten-Crabb, A., Thomas, Z. & Plazzer, J.-P. 2013. « Annotating the biomedical literature for the human variome ». *Database*.

Walsh, R., 2005. « A response to eight views on terminology: is it possible to tame the wild beast of inconsistency? » *Advances in Speech-Language Pathology*, vol. 7, nº 2, p. 105-111.

Wang, J., Shou, L., Chen, K. & Chen, G. 2020. « Pyramid: A layered model for nested named entity recognition ». *ACL 2020. Association for Computational Linguistics*.

Witko, A., Tourmente, B., Dessez, B., & Decullier, E. 2021. « French speech-language therapy students' interest in evidence-based practice: A survey ». *International Journal of Language & Communication Disorders*, vol. 56, nº 5, p. 989-1008.

Standardizing Legal Feminine Terms in the Web Dictionary of Ukrainian Feminine Personal Nouns

Olena Synchak

Department of Philology, Faculty of Humanities, Ukrainian Catholic University
o_synchak@ucu.edu.ua

Abstract. This article examines nearly 200 Ukrainian feminine personal nouns related to the legal domain, sourced from dictionaries and corpus queries in the General Regionally Annotated Corpus of Ukrainian (GRAC). A multifaceted approach is applied, combining corpus dynamics, cross-corpus studies, and Frame-Based Terminology Analysis. The paper first analyzes the dynamics of these terms within the media subcorpus, comparing their frequency and variety over the past two decades with political terms. Next, it contrasts their presence in the GRAC media subcorpus with their use on a court decisions website, offering a detailed cross-corpus analysis. Finally, it employs Frame-Based Terminology Analysis to create Ukrainian-language ontology of legal feminine terms. This approach provides a balanced analysis and effective lexicographical strategies, contributing to updates in the Web Dictionary and advancing the standardization of feminine terms in Ukrainian legal language.

Résumé. *Cet article examine près de 200 noms personnels féminins ukrainiens liés au domaine juridique, issus de dictionnaires et de requêtes dans le* General Regionally Annotated Corpus of Ukrainian *(GRAC). Une approche multifactorielle est appliquée, combinant la dynamique du corpus, des études comparatives inter-corpus et l'analyse terminologique fondée sur les cadres. L'article analyse d'abord la dynamique de ces termes dans le sous-corpus médiatique, en comparant leur fréquence et leur variété au cours des deux dernières décennies avec celle des termes politiques. Il compare ensuite leur présence dans le sous-corpus médiatique du GRAC avec leur usage dans un site web de décisions judiciaires, offrant ainsi une analyse inter-corpus approfondie. Enfin, il utilise l'analyse terminologique basée sur les cadres pour créer la première ontologie ukrainienne des termes juridiques féminins. Cette approche permet une analyse équilibrée et propose des stratégies lexicographiques efficaces, contribuant aux mises à jour du dictionnaire en ligne et à la normalisation des termes féminins dans la langue juridique ukrainienne.*

1. Introduction

Feminine personal nouns form one of the most dynamic and rapidly evolving layers of the Ukrainian lexicon. Media monitoring reveals a striking trend: the use of feminine terms in Ukrainian media has doubled since 2019. This surge in usage has drawn increasing attention from linguists, who study these terms from a variety of sources—dictionaries (Synchak & Starko, 2022), the press (Kravets, 2021), fiction (Zayets, 2020), and official documents (Sheremet, 2021). Research has largely focused on the derivational models of these nouns (Neliuba, 2011), their diachronic development (Brus, 2019), and the socio-cultural forces shaping their modern use (Arkhanhelska, 2019). Despite this interest, the standardization of feminine personal nouns has been slow and inconsistent. Of the approximately 9,000 feminine terms registered by the National Commission for the Standards of the State Language, only around 3,500 have been codified in dictionaries, highlighting the gap between usage and official recognition.

Feminine personal nouns encompass various professions, positions, and activities, the latter commonly referred to as "nomina agentis." In this article, legal feminine terms are defined as those used specifically within the fields of notary, judiciary, and law, emphasizing the rights, duties, and authority of women in these areas. These nouns refer to both permanent professions and primary roles in the legal sector and titles related to specific legal actions, such as property transactions, rentals, and inheritances etc. Notably, this article offers the first conceptualization of women's legal roles in the Ukrainian language, marking a significant step forward in understanding and codifying the gender-fair language in legal contexts.

Although the derivation of feminine personal nouns through suffixation is a structural and typological feature of Ukrainian, opinions on their use remain sharply divided in both society and academic circles. Some view these terms as an "artificial intervention" in the natural evolution of the language, driven, in their opinion, by gender equality movements (Arkhanhelska, 2019). Others see them as a "revival of Ukrainian identity and originality," promoting the broader decolonization of Ukrainian from Russian language standard (Styshov, 2020). The central debate, therefore, is whether the use of feminine terms can promote gender equality and support decolonization in Ukrainian society.

In official Ukrainian language, debates about feminine terms are especially contentious. While most researchers favor masculine forms when referring to women in formal communication, arguing they are more appropriate

(Lahdan, 2019, 169), others contend that feminine derivatives can be stylistically neutral and suitable for different professional and legal contexts (Klymenko, 2019). However, both perspectives often rely on hand-selected examples and subjective judgments. This study aims to address the gap by comparing legal feminine terms in both media resources and official documents using consistent corpus data and an ontological framework, while avoiding preconceived conclusions.

2. Standardization Efforts in Legal Feminine Terms

2.1. Regulation of Use in Official Documents

The State Classifier of Professions has prohibited the use of feminine terms in official documents since Soviet times up to the present. On August 18, 2020, the Ministry of Economy approved Amendment No. 9 to Ukraine's State Classifier of Professions[1]. The amendment mandates that job titles be listed in the masculine form, except for those exclusively used in the feminine gender, of which only a few examples are provided: *сестра медична* "nurse", *швачка* "seamstress", *манікюрниця* "(female) manicurist", *нянька* "baby-sitter", *покоївка* "chambermaid", etc. This norm seems gender discriminating today, since it prescribes women the less prestige and less paid jobs. To temper this discriminating rule, a small remark is added. At the request of the user, when entering a record of the job title in the personnel documentation, the professional job titles may be adapted to indicate the female gender. As this issue is left to the discretion of users, detailed guidelines for forming and using feminine personal nouns in official documents need to be established.

Some scholars argue that the norm established by the Classifier of Professions to avoid feminine professional titles in official documents is a vestige of Soviet-era practices. Despite the rich variety of feminine terms in Ukrainian that signify women's roles in legal contexts, their usage in official documents was long deemed a stylistic error. This convention was not based on the intrinsic logic of Ukrainian but was instead driven by a language assimilation policy intended to align Russian and Ukrainian more closely (Overchuk & Maiboroda, 2020, 283-284). In Russian, such feminine forms are typically confined to colloquial speech and are excluded from the literary norm[2].

1 https://www.me.gov.ua.

2 In January 2024, the Russian Supreme Court banned feminine terms, associating them with the LGBT movement and branding the last one as extremist.

M. Brus argues that even Old Ukrainian, from the 11th to 15th centuries, included feminine terms denoting property rights, such as *господариня* "(female) landowner," *отчичька* "(female) heiress," and *дѣдичьна* "(female) inheritor" (Brus, 2019, 95). While these feminine terms were primarily used in colloquial speech, other feminine terms appeared in medium- to low-style texts, including chronicles, legal documents, and business records (*ibid*., 107). This historical evidence underscores a longstanding tradition in Ukrainian of using feminine personal nouns in official documents, especially when women held power or owned property.

The gender equality issues are now on the agenda of the newly established National Commission for the Standards of the State Language. Aimed to regulate the use of feminine terms in official documents and in the Classifier of Professions, this Commission in 2021 published two sets of recommendations regarding the formation of feminine derivatives in diplomacy and medicine, respectively. These authoritative statements represent a significant change in the treatment of feminine personal nouns in the formal Ukrainian language. However, further investigation is needed to fully understand how these terms are used in official documents and how they ought to be codified in dictionaries.

2.2. Codification in Dictionaries

Ukrainian lexicography has a long tradition of codifying legal feminine terms, as highlighted by L. Tomilenko (2019, 209). For instance, Russian-Ukrainian dictionaries from the early 20th century feature nearly a hundred feminine derivatives specific to jurisprudence, representing 2.9% of all feminine forms in these works. Similarly, legal terms constitute 3% of the feminine derivatives found in the 11-volume Dictionary of Ukrainian published in Soviet Ukraine (SUM-11, 1970–1980). These feminine personal nouns primarily denote women as participants in civil and legal relations or as subjects of criminal law. Examples include: *позивальниця* "(female) barrator," *довірителька* "(female) principal," *злочинниця* "(female) criminal," etc.

The Dictionary of Business Ukrainian (1930), a product of the Ukrainianization era's dictionary boom, lists 533 feminine terms, such as *посвідчиця* "(female) attestor", etc. Notably, 25% of the terms recorded in this dictionary did not make it into the main Soviet-era dictionary, SUM-11. This omission includes legal terms related to ownership and finance—*e.g.*, *спадкодавиця* "(female) testator," *роботодавиця* "(female) employer,"

and з*емлекористувачк*а “(female) land user”—as well as certain professional titles, such as *діловодка* “(female) clerk,” *податковиця* “(female) tax official,” and agent nouns, such as *сплатниця* “(female) payer” and *промовниця* “spokeswoman”, etc. Furthermore, some Ukrainian-origin terms were replaced with internationalized or Russified alternatives, such as *співпозиванка* becoming *співвідповідачка* “(female) co-defendant” (compare with Russian *соответчица*).

These examples prompt the question of whether the masculinization of official language was a deliberate element of Soviet language policy aimed at Russifying Ukraine. In Ukrainian, as in Bulgarian (Bocale, 2010), Russian influence has restricted the use of feminine terms to colloquial contexts. Ukrainian research has identified suppressed terms from Soviet times (Masenko *et al.*, 2005), though whether feminization was explicitly banned is debated (Arkhanhelska, 2019). Given this context, it appears that there is a significant need to enhance the conceptualization of legal feminine terms. To address this, a multifaceted approach is applied, combining corpus dynamics, cross-corpus studies, and Frame-Based Terminology Analysis to better support their comprehensive representation in dictionaries.

3. Corpus Investigation

3.1. Method and data

This section analyzes the dynamics of Ukrainian feminine legal terms compared to political terms in newspaper texts. Political titles were chosen for comparison as they closely relate to the legal field, emphasizing women’s roles as agents of social change. The list of terms was compiled both manually from lexicographic sources and automatically through the General Regionally Annotated Corpus of Ukrainian (GRAC) (Shvedova *et al.*, 2016-2024).

Two key lexicographic resources were used to collect legal feminine terms. First, the Dictionary of Business Ukrainian (1930), previously mentioned, contains terms excluded from the SUM-11, which are of particular interest. Second, additional terms were sourced from the Web Dictionary of Ukrainian Feminine Personal Nouns (WDUF), published in January 2022 on the r2u.org.ua dictionary portal. This dictionary fills a significant gap in Ukrainian linguistic resources by providing comprehensive definitions, corpus-based examples, and expert surveys. Notably, legal feminine terms make up 5% of WDUF’s entries, comprising 135 lexical items.

New words were also identified through searches in the GRAC corpus. For this study, Version 16 was used, as it includes many media texts, with over 140,000 texts across various styles and nearly 1.8 billion tokens. GRAC, equipped with an NLP package and the Large Electronic Dictionary of Ukrainian (VESUM), enables the search for feminine personal nouns through CQL queries with morphological and semantic tagging. Since most of legal feminine terms are registered in VESUM, the following CQL lemma query was constructed for them:

[lemma="адвокатка|…|юристка"].

New lexical items were classified as out-of-vocabulary (OOV), and regular expressions were used to identify them, for example:

[word=".*спадкодавиц(я|і|ю|ею|е|ь|ям|ями|ях)"].

For some feminine derivatives that appear in multiple compound nouns and vary in spelling, regular expressions of the following type were added to the query:

[word=".*відповідачк.*|.*відповідачок.*"].

From both dictionaries and the corpus, I extracted 193 legal feminine derivatives and 97 political feminine derivatives. Detailed methods of word extraction and corpus analysis are available in previous studies (Synchak, 2024; Starko & Synchak, 2023).

GRAC's rich metadata and the ability to create custom subcorpora through its web interface make it an invaluable tool for diachronic analysis. This study focuses on newspaper texts from 2000 to 2022, using 23 journalism subcorpora labeled JOU_2000 to JOU_2022. These subcorpora, created by Vasyl Starko for our broader joint research (Starko & Synchak, 2023), range from 2.3 million tokens in 2000 to 255 million tokens in 2022.

Three CQL queries were consolidated into one for both groups of vocabulary, and the query was run across each journalism subcorpus (JOU_2000 to JOU_2022) within GRAC (see Fig. 1.). To account for variations in subcorpus sizes, relative frequencies were used for comparison. These were calculated by dividing absolute frequencies by the corpus size in tokens.

3.2. Results

The dynamics of the frequency of appearances for legal and political feminine titles in journalistic texts in GRAC are juxtaposed in Figure 1 below.

The findings clearly show a rising frequency of feminine terms in legal and political domains within media texts. However, the growth trends differ by thematic group. Legal feminine terms exhibit a steady, gradual increase, while political feminine terms experience a remarkable fivefold surge after 2015. Notably, before 2010, legal titles were more frequent than political ones, with only a minor exception in 2006.

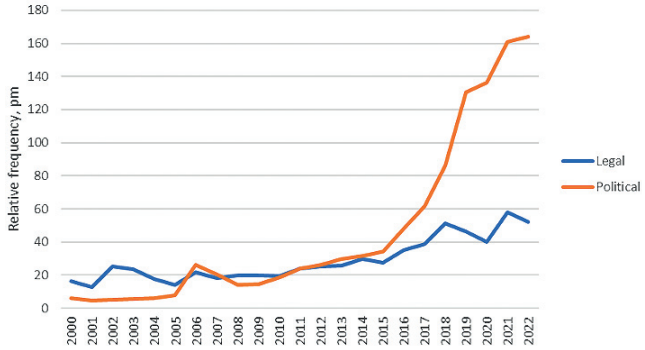


Fig. 1. Dynamics of feminine terms in legal and political reference in GRAC.

To better understand the usage of these feminine terms, it's important to consider not only their relative frequencies but also their variety, represented by the number of distinct lemmas. Lemma counts for each subcorpus were determined by sorting concordance results in GRAC, followed by manual verification. Figure 2 compares the use of feminine terms in political and legal contexts, considering the number of different lemmas used.

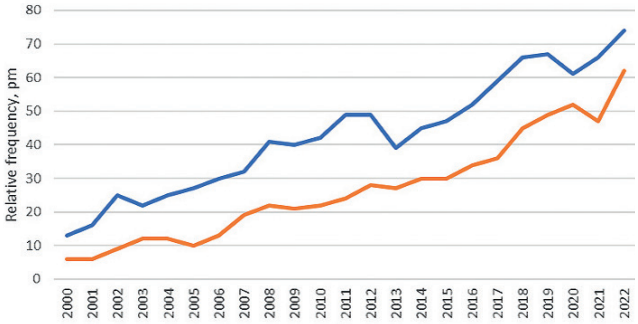


Fig. 2. Lemma variety among feminine terms in legal (blue) and political (orange) reference.

The data reveals greater lemma diversity among legal feminine terms than political ones. Although political titles are used more frequently (Fig. 1), legal terms show more variety (Fig. 2). In both groups, the increase in diversity follows a relatively smooth trajectory until 2013, doubling after the Euromaidan revolution.

Interpreting Figures 1 and 2, we observe two key points. First, each thematic group of feminine terms shows distinct growth patterns in both frequency and variety of usage. The growth dynamics, nevertheless, vary depending on the thematic group of words. Second, the increase in these terms appears to be triggered by major political events, particularly the Euromaidan in late 2013 and the subsequent Revolution of Dignity in 2014-2015. This is consistent with findings from studies on other thematic groups of feminine terms (Starko & Synchak, 2023).

4. Cross-corpora analysis

4.1. Method and data

This section aims to compare the use of legal feminine terms in Ukrainian official texts with their usage in media sources. Among the various subcorpora, GRAC offers a subcorpus of official Ukrainian texts (GRAC_style_OFF_Oficijni_teksty), which includes laws, acts, and other official documents, along with some legal commentary. However, this subcorpus is smaller than the journalism subcorpora used in section 3, and many official documents have yet to be added.

To address this, I created a subcorpus that combines official documents (doc.style=OFF), scientific texts on jurisprudence (doc.style=ACA), and public oral texts (doc.style=SPU), including parliamentary transcripts. Recognizing that this subcorpus still doesn't fully represent official Ukrainian style, I supplemented it by searching the website of court decisions[3]. It's important to note that while corpus data are presented in absolute and relative frequency, court decision data reflects the number of documents, not word occurrences. Additionally, court decisions sometimes contain spelling errors, requiring manual verification, and new documents are added daily, meaning search results change over time.

3 https://reyestr.court.gov.ua/.

To determine if the use of feminine personal nouns varies by text style, a cross-corpus comparison is essential. The 193 extracted legal feminine terms were analyzed across three sources: (1) the website of court decisions, also referred to as the Legal Subcorpus (LS), (2) the Journalism Subcorpus in GRAC-16 (JS), and (3) the Formal Subcorpus in GRAC-16 (FS that includes OFF+ACA+SPU). By comparing these subcorpora, conclusions about the actual usage and stylistic distribution of legal feminine terms will be drawn.

4.2. Results

The coverage of legal feminine terms is different within each of the subcorpus analyzed (Tab. 1). The biggest ammount of words in the list (80%) is coverred by Journalism Subcorpus. Formal Subcorpus includes almost two thirds of the lexical items (64%). Although the website of court decisions identyfies only half of the word list (55%), but it demonstrates the expanded use of feminine terms with extremely high frequency. For example, the noun *позивачка* "(female) prosecutor" appears in over five million documents, while *відповідачка* "(female) defendant" is used more than 1.68 million times. This data clearly shows that feminine terms are used not only in journalistic texts but also in official documents, particularly in court decisions, as indicators of legal status.

Semantic groups	Court Decisions		Journalism Subcorpus			Formal Subcorpus		
	Items	Documents	Items	Abs.	Relat.	Items	Abs.	Relat.
1. Professional and occupational titles in legal sphere	8	3098	21	1568	1,17	18	88,5	0,05
2. Legal nomina agentis	78	98262	99	585	0,54	79	131	0,07
3. Property ownership	6	3	5	30,4	0,03	3	8,7	0,01

Tab. 1. The Cross-corpora Use of Legal Feminine Terms.

The quantitative data in Table 1 confirms that, across all three subcorpora, the most diverse group is legal nomina agentis (LS – 78, JS – 99, FS – 79 items). While their frequency is higher in the court decisions subcorpus (98,

262 documents), the news texts from the GRAC corpus show greater lexical variety (99 items). Notably, court decisions include fewer female professional and occupational titles from the legal sphere compared to the news and formal subcorpora. This suggests that legal feminine nomina agentis are stylistically neutral across different text styles, while legal professional feminine titles, less common in official documents, are more typical of the media sphere and have yet to gain widespread use in official documentation.

For the codification of feminine terms, it is crucial to systematize their various word-formation variants. To achieve this, I compared data from the Formal Subcorpus in GRAC-16 with information from the court decisions website. The results of this comparison are detailed in the following table (Tab. 2).

As shown in Table 2, the preferred derivational variant for dictionary recommendations can be identified by comparing usage across different corpora. The first variant in each pair generally has higher frequency values and is more concise, reflecting greater adherence to linguistic economy. Therefore, using frequency criteria based on corpus data is a valuable strategy for lexicographers in selecting the most appropriate derivational variant for inclusion in dictionaries.

Feminine term	**Translation**	**Formal Subcorpus**		**Court decisions**
		absolute	***relative***	***absolute***
позивачка	"(female) claimant"	1 115	0,59	5 034 580
позивальниця		17	0,01	1 582
відповідачка	"(female) defendant"	215	0,11	1 682 556
відповідальниця		4	< 0,11	53
спадкоємиця	"heritress"	2 334	1,24	187 212
спадкоємниця		505	0,27	16 111

Tab. 2. Frequency of Use of Legal Feminine Derivatives Across Corpora.

Comparing different subcorpora is valuable for uncovering new meanings or nuances of terms, revealing how context influences their interpretation. For instance, in the journalism subcorpus of GRAC, *виборниця* refers to a "(female) elector," while in court documents, it denotes a female worker who selects items, such as coal. Similarly, *виконавиця* can mean a leading lady or performer in artistic contexts but refers to an official executing orders (*e.g.*, *судова виконавиця*) in legal contexts. Therefore, balancing data from various

subcorpora is crucial when compiling a dictionary, especially for terms with multiple meanings.

5. Frame-based Terminology Analysis

5.1. Method and data

In recent years, the theory of Frame Semantics has been increasingly applied across various disciplines, including the study of lexical and terminological fields. One significant development in Terminology is the Frame-Based Terminology (FBT) approach (Faber *et al*., 2009). FBT uses semantic frames to structure specialized knowledge and create conceptual representations, or "event templates," within specific domains. These frames aim to capture the dynamic nature of real-life events or processes by offering prototypical conceptual representations (Faber *et al*., 2006).

The effectiveness of frame-based analysis has been widely validated by Faber and her research group, along with other scholars who have applied this method to specialized terminology (*e.g*., Durán Muñoz, 2016). This study employs FBT to analyze Ukrainian legal feminine nouns. By organizing legal terminology within conceptual frames, this method refines and updates WDUF's definitions of legal feminine terms. Drawing on Durán Muñoz's (2016) ontoterminography, this approach provides a framework for analyzing the conceptual structure of these terms, integrating both WDUF entries and data from various legal subcorpora.

5.2. Results

5.2.1. Conceptual representation of legal feminine terms

This study begins with a bottom-up word extraction process, following Faber's text-based approach, which asserts that "conceptual structure is specified on the basis of linguistic information" (Faber, 2011, 50). By grounding term extraction in corpus data, we ensure that the conceptual organization of legal feminine terms is both theoretically sound and reflective of actual usage. The analysis led to the development of a preliminary conceptual framework (see Fig. 3), which later was refined through a top-down examination of WDUF definitions, envolving corpus data and consultation with Ukrainian legal experts.

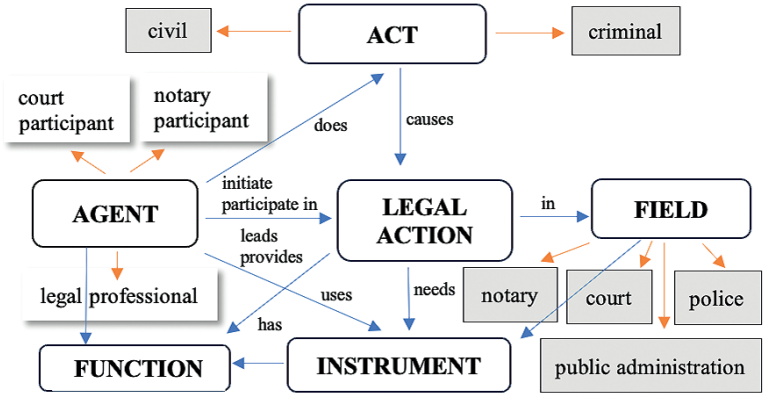


Fig. 3. Preliminary Framework for Categorizing Legal Feminine Terms.

Proposed categorization (Fig. 3) outlines six broad categories within the legal sphere: **<agent>** refers to female legal professionals or women involved in legal relationships; **<act>** denotes a real-world action, distinguishing between civil and criminal activities; **<legal action>** refers to any action that has legal consequences or is performed within the framework of law or any other formal regulation; **<field>** identifies the primary domains in which these agents operate, such as court activities or notary services, along with less prominent areas like state authorities and police; **<instrument>** encompasses legal norms, approved regulations and official documents that regulate specific types of legal relationships, serving as essential tools in legal practice. Lastly, **<function>** refers to the right, duty or authority of the agent concerning a specific legal action. Although this conceptual representation somewhat simplifies the prototypical situation of women in legal activities, it nevertheless reveals the complex relationships and dynamics between female agents, their actions, and the broader legal context.

A key consideration in developing this conceptual framework was whether to classify law and legal provisions as a <field> or an <instrument> of agent activity. While the expression *у правовому полі* "in the legal field" is common in Ukrainian, analysis of use in the GRAC corpus provided clarity. The collocation data revealed a significantly higher frequency of *закон* "law" and *право* "right" used in the instrumental case, suggesting their role as instruments, rather than as a field (Tab. 3). This insight suggests support for the classification of law as an instrumental tool in legal processes.

Defining women as agents in the legal field inherently activates the other categories within the proposed framework (Fig. 3). Thus, female legal <agents>, by performing real-world <acts>, become engaged in various legal actions within specific <field> domains to regulate legal relationships. They rely on <instruments>, such as laws and official documents, to fulfill their <function> within legal contexts.

CONCEPT		Absolute frequency	Relative frequency
LAW AS AN INSTRUMENT – *Instrumental case*		**172 235**	**96,68**
за "by law" згідно із "according to law" захищені "protected by law" керуватись "folow the law" заборонено "prohibited by law"	ЗАКОНОМ	135 944	76,31
користуватися правом "exercise the right"		36 291	20,37
LAW AS A FIELD – *Locative case*		**31 331**	**17,58**
в (у) законі "in the law"		29 238	16,41
в (у) правовому полі "in the legal field"		2093	1,17

Tab. 3. Categorization of Law in the Framework: Frequency Data from the GRAC corpus.

In general, the **<agent>** gain authority in the legal field either through their professional role or by actively engaging in activities to protect their rights. An essential feature of agents in the legal field is the duality of nomina agentis, which highlights the presence of paired or opposing roles. For example, there are two distinct terms to describe the inheritance process in the notary field: *спадкодавиця* "testatrix"—the one who transfers the inheritance (**donor**), and *спадкоємиця* "heiress"—the one who receives it (**recipient**). In court cases, beyond the roles of submitter and receiver, there are additional roles such as victim, witness, suspect, and offender.

The categorization in Fig. 3 presents unidirectional relationships, where the agent influences other categories without reciprocal interaction. Some categories, however, exhibit complex connections. These are evident in dictionary definitions, where agents in the legal field activate another agent's function through involvement in official documents, such as lawsuits or complaints. For instance, *скаржниця* "(female) complainant" is defined as a "woman who

complains about someone in court" (**active agent**), while *свідчиця* "(female) witness" is a "woman summoned to testify in court" (**re-active** or **summoned agent**). In both cases, agency stems from performing a legal action with specific function: *та хто* "woman who" (AGENT) → *скаржиться* "complains" (LEGAL ACTION) → *в суді* "in court" (FIELD) → by filing a complaint (INSTRUMENT) → to formally initiate legal proceedings (FUNCTION).

Semantic relations among the concepts and their activation within contextual situations are traced using two techniques, similar to Durán Muñoz (2016) and Faber (2011). First, through cross-corpora searches analyzing knowledge-rich contexts, and second, by examining definitions from WDUF. Key linguistic patterns are highlighted in Annex 1. Among them the hyperonymy relations (is_a), which indicate the relations between hyperonyms and hyponyms, are the most common. Other types of semantic relations can be categorized under different types of role relations. This approach allows not only identify new and validated the legal feminine terms but also enhances the consistency and coherence of the definitions within the WDUF.

A deeper examination of the specialized domain through corpus analysis, dictionary definitions, and expert consultations enabled a more precise definition of functions and tools in the initial diagram. As a result, the categorization was refined into a more comprehensive frame-based ontology, incorporating additional categories and relationships (Fig. 4). Figure 4 presents the main categories of the ontology, with a more detailed breakdown provided here in the text.

Most of the new subcategories are integrated into the initial categorization through hyperonymy relations (IS_A):

- In the **<agent>** category different subtypes of agents are identified within each group (Tab. 5): <Legal professional> is divided into <permanent legal professional> and <temporary legal agents>. <Court participant> is divided into two subgroups: 1. <Initiator>, which includes active and summoned agents. 2. <Involved party>, which includes roles such as victim, witness, suspect, and offender. <Notary participant> is divided into two subgroups: <donor> and <recipient>.
- In the **<legal action>** category, four subcategories are defined based on their functions: 1. Rights Protection, 2. Rights Acquisition, 3. Managing Finance and Property. 4. Legal Documentation and Validation (see Annex 2 for more details).

- In the **<field>** category, the subcategory of <human rights> is added alongside the subcategories of <notary>, <court>, <police>, and <public administration>.
- In the **<instrument>** category three types are distinguished: 1. <Legislative> which includes legal norms based on wider specification of law as <international> and <national>. The last one is divided into public and private. Public includes human rights, civil, constitutional, corporate, administrative and criminal. Private includes family and labour relationships, as well as finance obligations. 2. <Regulative> includes approaved regulations, such as a. parlaments' law; b. resolutions of the Cabinet of Ministers; c. presidential decrees; d. orders of ministers; e. administrative orders; f. corporative regulations, etc. 3. <Judicial> (Juridical) includes a. Procedural documents such as applications, complaints, inquiries, and petitions. b. Court documents, including proceedings, rulings, and decisions. c. Notarized documents, such as trade and exchange agreements, contracts of sale and purchase, notarial deeds, and civil acts like contracts and agreements.

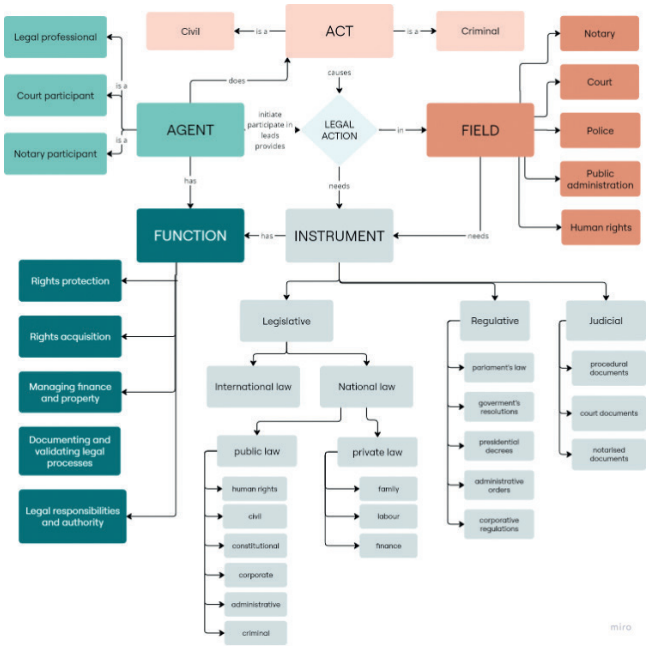


Fig. 4. Frame-based Ontology for Legal Feminine Terms (Miro).

- In the <**function**> category five subcategories are included: 1. Rights Protection. 2. Rights Acquisition. 3. Managing Finance and Property. 4. Legal Documentation and Validation. These first four subcategories were used to classify the legal actions mentioned above. 5. Legal Responsibilities and Authority encompass duties such as obligations, consultation, investigation, monitoring, and prosecution, as well as authorities related to defense, verdicts, legislation, verification, permission, and prohibition.

5.2.2. Producing Definitional Templates to Update the WDUF

Definitional templates provide a framework for creating clear and coherent definitions. They are definied as schematic representations of prototypical relations within a shared semantic frame. According to ontoterminography methodology (Durán Muñoz, 2016), consistent templates should be applied to concepts within the same category, while distinct templates are necessary for different categories. This approach ensures systematic and coherent definitions by reflecting the organization of the domain and clarifying semantic relations between concepts.

The analysis of the Web Dictionary reveals that most definitions of feminine legal terms align with the template AGENT – LEGAL ACTION – FUNCTION – FIELD. For example, *адвокатка* "(female) advocate" is defined as a lawyer (AGENT) who provides LEGAL ACTION by defending the accused or prosecuting a case (FUNCTION) in court (FIELD) and advising (FUNCTION) on legal issues (FIELD). The category INSTRUMENT is only implicitly present here, while the advocate acts within legal norms.

A less common template, AGENT – LEGAL ACTION – INSTRUMENT – FIELD, introduces <instrument> for terms like *оскаржувачка* "(female) complainant", defined as "a woman who files a written complaint (INSTRUMENT) with an authority or court (FIELD)." In this case, a real-world action (ACT) triggers legal consequences (LEGAL ACTION) through the use of a formal INSTRUMENT, emphasizing the crucial link between action and legal procedure.

Based on the frame-based ontology for legal feminine terms (Fig. 4), definitional templates for the <agent> category were created for the current WDUF's entries (Tab. 4).

TYPE OF AGENT		DEFINITIONAL TEMPLATE
1. Permanent legal professional		1) TYPE OF AGENT who [LEGAL ACTION] [FUNCTION] [FIELD] – *адвокатка* 2) TYPE OF AGENT who [INSTRUMENT] – *криміналістка* (knowledge – instrument)
2. Temporary legal agents		1) TYPE OF AGENT who [LEGAL ACTION] [FUNCTION] [FIELD] – *законодавиця* 2) TYPE OF AGENT who [INSTRUMENT] [FIELD] – *землеупорядниця* 3) TYPE OF AGENT who [LEGAL ACTION] [FUNCTION] [INSTRUMENT] [FIELD] – *повірена*
3. Active agent 4. Summoned agent 5. Involved party	*юр.*	TYPE OF AGENT who [LEGAL ACTION [FUNCTION] [INSTRUMENT] [FIELD] *позивальниця, відповідачка, підозрювана*
6. Donor 7. Recipient		TYPE OF AGENT who [LEGAL ACTION] [FUNCTION] [FIELD] – *позикодавиця vs. позичальниця*

Tab. 4. Definitional templates for different types of legal agents.

The definitional templates for the <agent> category (see Tab. 4) suggest that all agent types in the legal field should be defined according to the <legal action> in which they are involved. Notably, the second templates for <Permanent Legal Professional> and <Temporary Legal Agents> omit the <legal action> category. Therefore, this element should be incorporated to enhance the accuracy and completeness of the WDUF's entries. Additionally, the category of <instrument> appears somewhat arbitrary, as only some templates include it.

Going forward, establishing a clearer distinction between permanent professionals and temporary agents in the legal sector, along with a more detailed their classification, is crucial. It seems advantageous to categorize professional titles based on whether they relate to holding public office, such as *суддя* "(female) judge", or the independent provision of legal services, such as *юристка* "(female) lawyer". Based on dictionary definitions, Figure 5 classifies legal professionals into four subcategories: a. *юристка* "(female) lawyer"; b. *службовиця* "(female) clerk"; c. *нотаріуска* "(female) notary";

d. *маклерка* "(female) broker". These terms will serve as key references for updating the WDUF's definitions of corresponding professional legal titles.

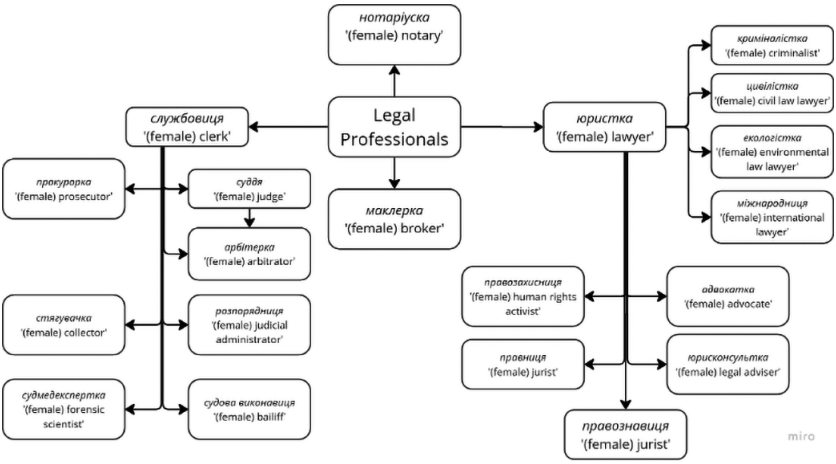


Fig. 5. Classification of Permanent Legal Professionals (Miro).

It is worth noting that the Web Dictionary assigns legal labels (*e.g.*, *юр.* in Ukrainian) to certain feminine terms used in court settings. For consistency, it would be beneficial to apply this labeling to all female participants in court proceedings, including non-professional roles, while removing it from professional titles, such as *прокурорка* "(female) prosecutor".

5.3. Decolonizing Ukrainian Lexicography

It is compelling to investigate whether, and to what extent, the lexicography of legal feminine terms was shaped by political interference in Soviet Ukraine. Quantitative analysis (Fig. 6) reveals that 32% of these terms are newly coined and absent from dictionaries, 27% appear in the Soviet-era SUM-11, while the largest portion—41%—originates from Ukrainianization-era lexicons (1920–1930). This suggests that Soviet-era dictionaries not only omitted several legal feminine nouns recorded in earlier lexicons but may have deliberately excluded certain terms, reflecting intentional language planning policies.

Fig. 6. Codification of Legal Feminine Terms in Dictionaries.

Notably, the SUM-11 excluded terms like *маклерка* "(female) broker", *заповітниця* "(female) bequestand", *споживачка* "(female) consumer," *миротворицяя* "(female) peacemaker" and instead of listing the feminine form of *прокурорка* "(female) prosecutor," it offered a Russified version with suffix *-ш(а) —прокурорша*. Further highlighting the selective and ideologically influenced approach to lexicography, only two out of seven terms with the suffixoid *-давиця* commonly used in legal contexts—*законодавиця* "(female) legislator" and *позикодавиця* "(female) lender"—were included in the SUM-11. Interestingly, other terms like *роботодавиця* "(female) employer", *дародавиця* "(female) donor", and *спадкодавиця* "(female) inheritor", though absent from the SUM-11, are now seeing a revival in modern usage. Additionally, newly coined words following this model, such as *орендодавиця* "(female) lessor" and *рекламодавиця* "(female) advertiser" etc., are emerging.

The Soviet-era Ukrainian dictionary, SUM-11, contains the fewest legal feminine terms compared to dictionaries from the Ukrainianization period. Two strategies of Soviet linguistic planning are evident: (1) excluding 31% of terms from earlier dictionaries and (2) promoting Russian word-formation models for certain terms. Today, however, there's a revival of word-formation models from the Ukrainianization era, with the suffixoid *-давиця* being the most productive, having generated 12 feminine terms. This shift signals a return to more authentic Ukrainian linguistic patterns. Thus, the decolonization of Ukrainian legal terminology is occurring partly through the revival of lexicon and word-formation models from Ukraine's early 1920s-30s.

6. Conclusion and future work

The three approaches used in this article—corpus dynamics, cross-corpus studies, and Frame-Based Terminology Analysis—are time-intensive but well-justified. This multifaceted method offers a more balanced analysis of legal feminine terms in Ukrainian and provides effective strategies for their lexicographical documentation.

The proposed model of frame-based analysis for legal feminine terms, integrating both definitions and real corpus data, offers a valuable approach to enhancing WDUF. Building a linguistically based ontology refines the registry and definitions, supporting the standardization of legal feminine terms. Comparing these terms with those from pre-Soviet and Soviet dictionaries enables this process to unfold in a decolonized manner. This multifaceted method can be extended to other thematic groups of feminine personal nouns. However, to create a comprehensive frame-based ontology, each group should undergo detailed analysis within its specific domain corpus.

The research presented in Sections 5-7 of this paper was conducted as part of the DAAD funding program "Future Ukraine: Research Grants for Ukrainian Master's Students and Researchers" (2024) at Leipzig University, under the supervision of Professor Olav Mueller-Reichau.

References

Arkhanhelska, A. 2019. *Femina cognita*. Ukrainska zhinka u slovi y slovnyku. Kyiv.

Bocale, P. 2010. "The formation of feminine nouns in Bulgarian. The results of a field investigation". Ezikov Svjat, VIII: 2, 102–119.

Brus, M. 2019. *Feminityvy v ukrainskii movi: heneza, evoliutsiia, funktsionuvannia*. Ivano-Frankivsk.

Doroshenko, M. 1930. *Slovnyk dilovoyi movy: terminolohiya ta frazeolohiya*. Kharkiv.

Durán Muñoz, I. 2016. "Producing Frame-Based Definitions: A case study". *Terminology* 22:2. 224–250.

Faber, P. & Martin, San A. 2011. "Linking Specialized Knowledge and General Knowledge in EcoLexicon". In TOTh International Conference, 47–61. Chambéry : Presses universitaires Savoie Mont Blanc.

Faber, P., León, P., Prieto, J. 2009. "Semantic Relations, Dynamicity, and Terminological Knowledge Bases". *Current Issues in Language Studies* 1: 1–23.

Faber, P., Montero Martínez, S., Castro Prieto, M.R. *et al.* 2006. "Process-Oriented Terminology Management in the Domain of Coastal Engineering". *Terminology* 12(2): 189–213. doi:10.1075/term.12.2.03fab

GRAC-16 (2016–2024). Shvedova, M., Waldenfels von R., Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T. *et al. GRAC: General Regionally Annotated Corpus of Ukrainian*. Electronic resource: Kyiv, Lviv, Jena. URL: http://uacorpus.org.

Klymenko, N. 2019. "Tendentsii rodovoi katehoryzatsii imennykiv u suchasnykh ukrainskii ta novohretskii movakh". In *Ukrainska mova v konteksti suchasnoi slavistyky*: Monohrafiia, Kyiv.

Kravets, T. 2021. *Henderni stereotypy v suchasnomu ukrainskomu masmediinomu dyskursi: Dysertatsiia*. Kyivskyi natsion. univ. im. Tarasa Shevchenka, Kyiv.

Lahdan, S. 2019. "Stylove vykorystannia feminityviv v ukrainskiy movi". In *Scientific Research in XXI century: Proceedings of the 1st International Scientific and Practical Conference*, 163–170. Ottawa.

Masenko, L., Kubaichuk, V., Demska-Kulchytska, O. (eds.). 2005. *Ukrayinska mova u XX storichchi: istoriya linhvotsydu*. Dokumenty i materialy. Kyiv.

Neliuba, A. 2011. "Innovatsiini zrushennia v ukrainskomu zhinochomu slovotvori". *Linhvistyka*, 2(23): 49–59.

Overchuk, O. B. & Maiboroda, I. R. 2020. "Leksyko-semantychni osoblyvosti vykorystannia feminityviv u profesiinomu movlenni". In Suchasni problemy pravovoho, ekonomichnoho ta sotsialnoho rozvytku derzhavy: tezy dop. mizhnar. nauk.-prakt. konf., 283-285. Kharkiv.

Sheremet, A. V. 2021. "Aktualni tendentsii funktsionuvannia feminnoi leksyky u tekstakh ofitsiino-dilovoho styliu ukrainskoi movy (20-ti roky XX st.)". In *Naukovyi visnyk KhDU*, 2: 72–76.

Starko, V., Synchak, O. 2023. "Feminine Personal Nouns in Ukrainian: Dynamics in a Corpus". In *COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems*, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, 407–425, URL: https://ceur-ws.org/Vol-3396/paper33.pdf.

Styshov, O. 2020. "Osoblyvosti sufiksalnoho slovotvorennia neolohizmiv na poznachennia osib u suchasnii ukrainskii movi". In *Linhvistychni doslidzhennia: zb. nauk. prats KhNPU imeni H. S. Skovorody*, 53: 127–140.

SUM-11: Slovnyk ukrainskoi movy. In *11 vols*. Naukova Dumka, Kyiv, 2018–2023. URL: http://sum.in.ua.

Synchak, O. 2024. "Language Change in Times of Turbulence: A Corpus-Based Investigation of the Rise of Feminine Personal Nouns in Ukrainian". In *Language, Gender and Politics in Central and Eastern Europe*, Palgrave Macmillan (forthcoming).

Synchak, O., Starko, V. 2022. "Ukrainian Feminine Personal Nouns in Online Dictionaries and Corpora". In *COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems*, Gliwice, Poland, URL: http://ceur-ws.org/Vol-3171/paper58.pdf.

Tomilenko, L. M. 2019. "Feminityvy yurydychnoyi tsaryny v ukrayinskiy zahalnomovniy leskykohrafiyi XX-XXI stolit". In *Ukrainska mova v yurysprudentsii: stan, problemy, perspektyvy*. P. I: 209–212.

WDUF (Web Dictionary of Ukrainian Feminine Personal Nouns). 2022. *Slovnyk zhinochykh nazv ukrayinskoyi movy*. Lviv. https://uacorpus.org/

Website of court decisions, URL: https://reyestr.court.gov.ua/.

Zayets, V. 2020. "Slovotvirni typy feminityviv suchasnoi prozy ta literaturna norma. Colloquium-journal", *Warszawa*, 19(71): 17–22.

Annex 1: Sample of linguistic patterns

Semantic relations	Linguistic patterns	Translation
IS_A	уповноважена представниця потерпіла, постраждала стала виявилась вважала себе була присутня мінімальної пенсії за віком як учасниці війни	authorised representative victim, sufferer became turned out to be considered herself was present minimum retirement pension as a war veteran
DOES_ ACT	скоїла, вчинила украла, привласнила купила будинок підписала скаржилася вимагала звинуватила визнала провину успадкувала	committed stole, appropriated bought a house signed complained demanded accused pleaded guilty inherited
INITIATES	звернулася до суду подала заяву оформила спадщину заявила в суді надала свідчення свідчення потерпілої	went to court filed an application issued an inheritance declared in court testified victim's testimon
PARTICI-PATES	учасниця місії причетна особи, що беруть участь у справі суд вважає за необхідне залучити до участі у справі суд у складі: головуючого – судді за участю секретаря	participant in a mission implicated persons involved in the case the court considers it necessary to involve a court composed of: a presiding judge with the participation of a secretary
LEADS	організувала веде справу в суді керівниця нотаріальної контори виносить вирок у судовій справі	Organised conducts the case in court head of a notary office passes judgement in a court case

PROVIDES	забезпечує примусове виконання судових рішень надає юридичні послуги	enforces court decisions provides legal services
USES_INSTRUMENT	Скористалася діяти за дорученням звернулася із заявою	took advantage of act by power of attorney applied with a statement
HAS_FUNCTION	виконувала такі функції захищати інтереси особи захистити від порушення прав надано право захищати права, свободи та інтереси інших осіб самостійно захистити у суді права набула права користування набула право власності на майно набула права на отримананя коштів / на виплату набула право голосу на виборах набула право на успадкування оштрафувати засудити арештувати призначити запобіжний засіб зобов'язати захистити у суді права	performed functions to protect the person's interests to protect against violation of rights the right to protect the rights, freedoms and interests of others independently defend their rights in court acquired the right to use acquired ownership of the property acquired the right to receive funds / to payment acquired the right to vote in elections acquired the right to inherit to fine to convict to arrest to impose a preventive measure to oblige to defend the rights in court

Annex 2: Classification of Civil Legal Action by Function

1. **Rights Protection**
   - **Exemption**: Protecting the rights of vulnerable individuals, such as refugees or repatriates, during resettlement or integration.
   - **Complain**: Initiating a legal claim or grievance.
   - **Accuse**: Bringing charges or allegations against someone.
   - **Object**: Raising formal objections during legal proceedings.
   - **Appeal**: Requesting a review of a decision by a higher court.
   - **Litigate**: Engaging in legal proceedings to protect rights.
   - **Consumer**: Seeking protection or legal recourse in cases of fraud, false advertising, or defective products (protection of consumer rights).

2. **Rights Acquisition**
   - **Submit**: Filing documents or petitions to establish rights.
   - **Respond**: Answering a complaint or petition to protect one's rights.
   - **Possess**: Taking legal ownership or control of property.
   - **Rent**: Establishing lease agreements to acquire usage rights.
   - **Adopt**: Legally establishing a parent-child relationship.
   - **Bequeath**: Transferring property or rights upon death.
   - **Voter**: Exercising the right to vote in elections (acquisition and exercise of political rights).
   - **Client**: Acquiring legal, financial, or professional services (rights in service contracts or agreements).

3. **Managing Finance and Property**
   - **Debt**: Formalizing obligations regarding financial loans.
   - **Deposit**: Placing items or funds in safekeeping.
   - **Contribute**: Providing support or resources to trusts or shared ownership.
   - **Invest**: Allocating resources for financial returns.
   - **Pay tax**: Fulfilling legal obligations regarding taxation.
   - **Lend**: Providing temporary use of property or funds.
   - **Guard**: Safeguarding property or rights.
   - **Transfer**: Referring to successful transfers of assets or rights.

4. **Documenting and Validating Legal Processes**
   - **Attest**: Providing testimony or evidence in legal contexts.
   - **Declare**: Making formal statements about legal rights or status.
   - **Refuse**: Declining to comply with requests or demands.

5. **Legal Responsibilities and Authority**
   - **Duties**: obligations, consultation, investigation, monitoring, and prosecution.
   - **Authorities**: defense, verdicts, legislation, verification, permission, and prohibition.

Is Domain Important for Automatic Term Extraction in the Era of Large Language Models?

Hanh Thi-Hong Tran*, **, ***, Carlos-Emiliano González-Gallardo*, José G. Moreno****, Antoine Doucet*, Senja Pollak**; ***

*L3i Laboratory, University of La Rochelle, La Rochelle
{firstname.lastname}@univ-lr.fr

** Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

***E8 Department, Jožef Stefan Institute, Ljubljana, Slovenia
senja.pollak@ijs.si

**** IRIT, University of Toulouse, IRIT, 31000 Toulouse
jgmorenofr@gmail.com

Abstract. Automatic term extraction (ATE) is a natural language processing (NLP) task that reduces the effort to manually identify terms from domain-specific corpora by providing a list of candidate terms. This paper explores using open-sourced large language models (LLMs), namely Llama-2 on the ATE task using in-context learning prompts. We evaluate how well LLMs perform with different levels of domain-related information, *i.e.*, in-domain and cross-domain demonstrations with and without domain enunciation. Our findings suggest the model's predictive performance when designing prompts with few demonstrations without specifying the domain to help LLMs capture the candidate terms in the correct formats. Furthermore, the results demonstrate that LLMs offer a valuable compromise by reducing the need for extensive data annotation, making them suitable for real-world applications where labeled data are scarce.

Résumé. *L'extraction automatique de terminologie (EAT) est une tâche de traitement du langage naturel (TLN) qui vise à réduire l'effort d'identification manuelle des termes à partir de corpus spécifiques à un domaine en fournissant une liste de termes candidats. Dans cet article, nous explorons l'utilisation de grands modèles de langage open-source (LML), notamment Llama-2, sur la tâche d'EAT*

à l'aide d'invitations d'apprentissage contextuelles. Nous évaluons les performances des LML avec différents niveaux d'informations liées au domaine, c'est-à-dire des démonstrations dans le domaine et entre domaines avec et sans énonciation du domaine. Nos résultats suggèrent que le domaine a un impact sur les performances prédictives du modèle lors de la conception d'invitations avec peu de démonstrations pour aider les LML à capturer les termes candidats dans les formats corrects. De plus, les résultats montrent que les LML offrent un compromis précieux en réduisant le besoin d'annotation extensive des données, ce qui les rend adaptés aux applications du monde réel où les données étiquetées sont rares.

1. Introduction

Terms can be defined as "*the designation of a defined concept in a special language by a linguistic expression*" (ISO 1087) or a terminological unit extracted from specialized texts (Castellví, *et al.*, 2001). They are beneficial not only for several terminographical tasks performed by linguists (*e.g.*, specialized dictionary construction [Le Serrec *et al.*, 2010]) but also for several complex downstream tasks (*e.g.*, topic detection [El-Kishky *et al.*, 2014] and information retrieval [Lingpeng *et al.*, 2005]). To minimize the time and effort to extract terms from domain-specific corpora, several automatic term extraction (ATE) approaches have been proposed.

The main contribution of our work is two-fold:

1. We investigate the potential of prompting LLMs for our specific task using two distinct prompting approaches: (1) in-domain k-shot demonstration and (2) cross-domain k-shot demonstration.
2. We empirically evaluate how the domain impacts the performance of term extraction using the same k-shot examples.

This paper is organized as follows: Section 2 presents the related work with recent advances in term extraction. Next, we describe the methods and experimental setup, including the datasets and the evaluation metrics in Section 4. The results and the error analysis are discussed in Section 5 before we conclude with future work in Section 6.

2. Related Work

2.1. Automatic Term Extraction

Traced back to the 1990s, automatic term extraction was first proposed under the research of Damerau (1990) and Justeson and Katz (1995) with the two-step procedure: (1) extracting a list of candidate terms and (2) determining their correctness. Traditional methods primarily relied on either linguistic or statistical aspects (Frantzi *et al.*, 1998; Vintar, 2010) or combined both (Kessler *et al.*, 2019; Repar *et al.*, 2019). The advancement of representation learning and neural networks has led to the exploration of various text embedding techniques for term extraction. These include local-global (Amjadian *et al.*, 2016), non-contextual (Zhang *et al.*, 2018), and contextual (Kucza *et al.*, 2018) word embeddings, as well as their combinations (Gao and Yuan, 2019). Language models have also been used for this task, as shown in the TermEval 2020: Shared Task on Automatic Term Extraction (Rigouts Terryn *et al.*, 2020). For example, GloVe embeddings have been fed into a Bi-LSTM (Rigouts Terryn *et al.*, 2020), and all potential extracted *n*-gram combinations have been input into a BERT binary classifier (Hazem *et al.*, 2020), among other approaches. In recent years, the evolution of ATE has seen a shift towards treating the task as a sequence-labeling problem, extending beyond monolingual learning (Kucza *et al.*, 2018) to include cross-lingual and multilingual learning (Conneau *et al.*, 2020; Lang *et al.*, 2021; Liu *et al.*, 2020; Tran *et al.*, 2022b). A systematic review of the tasks can be found in Tran *et al.* (2023).

2.2. In-context Learning with LLMs

After the evolution of transformer-based token classifiers toward term extraction (*e.g.*, BERT [Lang *et al.*, 2021], XLMR [Tran *et al.*, 2022b, c, a]), recent years witnessed the blossoming of large-scale generative models with the introduction of prompting (Radford *et al.*, 2019) in both zero-shot and few-shot demonstration (Brown *et al.*, 2020; Chowdhery *et al.*, 2022). Despite several works with state-of-the-art (SOTA) performance on text generation (Song *et al.*, 2023) or question-answering (Guo *et al.*, 2023), no application has been found on ATE yet. This is in part due to the sequence-labeling nature of the task, which is challenging to generative models to which LLMs belong.

The concept of prompts with few-shot demonstrations was first introduced by Radford *et al.* (2019), followed by an empirical analysis of the in-context learning paradigm on several downstream tasks with GPT-3 (Brown

et al., 2020) and PaLM (Chowdhery *et al*., 2022) LLMs. With the release of ChatGPT[1] by OpenAI in 2022, recent research focuses on evaluating its performance in various NLP tasks (Kocoń *et al*., 2023), such as text generation (Guo *et al*., 2023; Song *et al*., 2023; Wu *et al*., 2023) and question-answering (Bian *et al*., 2023; Tan *et al*., 2023; Omar *et al*., 2023).

To our knowledge, none of these directions had been previously explored in ATE, nor the impact that domain-specificity has on guiding term extraction.

3. Domain Relevance in Term Extraction

Domain greatly influences the selection of terms in a collection of texts. A linguistic unit in a text can be considered (or not) as a term depending on the domain of interest. Nevertheless, how specialized or domain-specific this lexical unit needs to be before it is considered a term is far from being a consensus. In addition, only keeping those terms specific to the domain can be restrictive and limit the vocabulary diversity, leading to certain drawbacks in further tasks like cross-domain alignment. The concept of termhood was introduced by Kageura and Umino (1996) to indicate the relation degree between a lexical unit and specific concepts inside a domain. Rigouts Terryn, Ayla (2021) address this concern by defining termhood as a function of lexicon- and domain-specificity in their annotation guidelines. On the one hand, lexicon-specificity denotes if a lexical unit is only known by specialists or if it is part of a common language. On the other hand, domain-specificity denotes whether a lexical unit is relevant or unrelated to the researched domain. The combination of these two indicators generates four different types of classes: specific terms, common terms, out-of-domain terms, and no terms. This classification, far from perfect and with a considerable degree of subjectivity, showed to be very useful leading to more intuitive annotations.

If term selection depends on the annotator's lexical capacity and expertise in a specific domain, how dependent would these two aspects be on an ATE system based on an LLM and in-context learning? LLMs are trained on billions or trillions of tokens (*i.e*., chunks of words) collected mainly from publicly available sources. This amount of training text gives the LLMs the ability to generate plausible-sounding text that will follow the syntax and semantics of a given input prompt. Further, LLMs are domain-agnostic as a result of this amount of diverse text. The versatility of LLMs is enhanced by their domain-agnostic nature, allowing them to be used for numerous applications.

1 https://openai.com/chatgpt

However, it is crucial to consider the importance of domain-specific knowledge in certain cases, such as term extraction.

The quality of the input prompt when querying an LLM is essential to the quality of the response. The popularization of LLMs was followed by the development of prompt engineering, whose goal is to design prompts that best fit the language model to maximize the quality of the generated text. As a rule of thumb, the more explicit the prompt, the better. Some examples of the task can be included to show the language model the sort of answers that are expected; this is known as a few-shot demonstration. Moreover, specifying certain criteria like the field or discipline that is relevant to the task, narrows down the language model into a fixed text generation style.

We find two analogies between termhood (Kageura, 1996; Vezzani, 2020) as used by Rigouts Terryn, Ayla (2021) in their annotation guidelines of the ACTER dataset and LLMs in-context learning for ATE. First, lexicon-specificity can be represented with few-shot demonstrations that illustrate, independently of the domain, the lexicon of input sentences and the desired output list of terms. Second, domain-specificity can be induced into the prompt by explicitly declaring the domain of interest or letting the language model deduce it from the few-shot demonstrations. These two analogies result in four different prompt scenarios: (1) in-domain demonstrations with explicit domain, (2) cross-domain demonstrations with explicit domain, (3) in-domain demonstrations with implicit domain, and (4) cross-domain demonstrations with implicit domain. In the next section, we describe how prompts are structured and detail our methodology to test each scenario.

4. Methodology

4.1. Prompt designs

We conduct a set of experiments on the ACTER v1.5 dataset[2] (Rigouts Terryn *et al.*, 2020). ACTER is a manually annotated collection of 12 corpora covering four domains, *i.e.*, Corruption (Corp), Equitation (Equi), Wind energy (Wind), and Heart failure (Htfl) in English, French, and Dutch. The dataset has two versions of gold standards: one containing only terms (ANN) and the other containing terms and named entities (NES). To test cross-domain configurations we apply the same configuration as in the TermEval 2020 shared task where the Heart failure domain of each language is considered the test set.

2 https://github.com/AylaRT/ACTER

In the scope of this study, we focus on English corpora to verify our hypothesis and follow the general paradigm of in-context (few-shot) learning with a three-step prompt structure as in Figure 1. Given an input sentence *x*, we construct a *Prompt(x)* to give a descriptive overview of the task with the following steps:

(1) Task Description. We define the role that the LLM needs to follow as depicted by the orange box of Figure 1. First, we specify the task the LLM has to perform. In this case, ATE: "As an excellent automatic term extraction (ATE) system, extract the terms in the Heart Failure domain given the following text delimited by triple backquotes". This sentence is the same in all the data and domain versions we experimented with.

The ACTER dataset has two types of gold standards: one that excludes named entities (ANN) and another that includes them (NES). Thus, we provide an additional sentence to clarify the scope of the candidate terms. More specifically, in the ANN version, we mention "Named entities are not considered as terms" and in the NES version, we mention "Named entities are considered as terms". Finally, we clarify the output format that we expect the LLM to return: "Output Format: [list of terms present]". We also detail the output scenario when the input sentence does not contain any term by "If no terms are presented, keep it empty list: []".

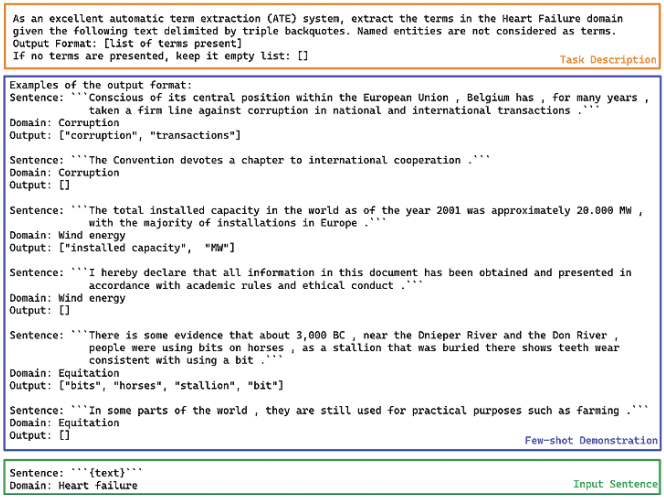


Fig. 1. An example of a complete prompt designed for the ANN version in an explicit cross-domain setting.

(1) Task Description instructs the LLM to detect terms using terminological knowledge with the desired format. (2) Few-shot Demonstration provides the model with few-shot examples. (3) Input Sentence is a placeholder for the input sentence to be fed in the LLM to get the predictions.

(2) Few-shot Demonstration. In this step, we focus on the analogies between term hood and in-context learning. We provide configurations of few-shot demonstrations that are added to the task description to control the output format for each test input. This ensures that the outputs generated by the LLM resemble the format of the demonstrations. Based on our previous studies, we chose the demonstration output that sequentially packs a list of examples, each consisting of both the input and output sequences as shown in the purple box of Figure 1.

```
Examples of the output format:
Sentence: ```Treatment of anemia in patients with heart disease : a clinical practice guideline
           from the American College of Physicians .```
Domain: Heart failure
Output: ["anemia", "patients", "heart disease", "clinical practice guideline", "Physicians"]

Sentence: ```Recommendation 2 : ACP recommends against the use of erythropoiesis-stimulating agents
           in patients with mild to moderate anemia and congestive heart failure or coronary heart disease .```
Domain: Heart failure
Output: ["erythropoiesis-stimulating agents", "patients", "anemia", "congestive heart failure", "coronary heart disease"]

Sentence: ```Moreover , there is yet to be established a common consensus being used in current assays .```
Domain: Heart failure
Output: []
                                                                                    Few-shot Demonstration
```

Fig. 2. In-domain demonstration designed for ANN version in explicit cross-domain setting.

We consider two scenarios regarding the lexicon-specificity domain-related information: cross-domain and in-domain demonstrations. On the one hand, cross-domain demonstration excludes the examples from the test domain (Heart failure) and includes two examples (1 positive and 1 negative) from the other three domains (Corruption, Equitation, and Wind energy), which sums up to 3 positive and 3 negative examples as shown in the purple box of Figure 1. On the other hand, in-domain demonstration includes three examples only from the test domain without incorporating any additional information from the other three domains from the dataset as shown in Figure 2. While the first two examples contain terms (positive examples), the last one does not contain any term inside the input sequence (negative example).

```
Sentence: ```Treatment of anemia in patients with heart disease : a clinical practice guideline
          from the American College of Physicians .```
Domain: Heart failure
Output: ["anemia", "patients", "heart disease", "clinical practice guideline", "Physicians"]

                                                        Explicit Domain Demonstration
-------------------------------------------------------------------------------------
Sentence: ```Treatment of anemia in patients with heart disease : a clinical practice guideline
          from the American College of Physicians .```
Output: ["anemia", "patients", "heart disease", "clinical practice guideline", "Physicians"]

                                                        Implicit Domain Demonstration
```

Fig. 3. Explicit vs. *implicit in-domain demonstration designed for the ANN version.*

Similar to lexicon-specificity, we consider two scenarios for domain-specificity: explicit and implicit domain demonstrations. Explicit domain prompts include the following information as viewed in the upper section of Figure 3:

- *Sentence* feeds the current input sentence into the LLM as found inside the dataset.
- *Domain* specifies the domain of the input sentence to delimit the candidate terms the LLM may extract.
- *Output* specifies that the output format is of vital importance in guiding the LLM to generate the prediction in the desired format.

In contrast, implicit domain prompts only keep the information of Sentence and Output, removing the information of Domain as shown in the lower section of Figure 3.

(3) Input sentence. In this step, we feed the sentence for which we expect the LLM to generate predictions of the list of candidate terms. Depending on the domain-specificity, we provide the LLM only the sentence (implicit domain) or the sentence with the domain of interest (explicit domain) as shown in the green box of Figure 1.

4.2. Postprocessing steps

An additional postprocessing step has been added to clean the prediction output of LLMs. Despite not asking for an explanation of the prediction or any further information and that more than 95% on average of the predictions LLMs returns are only the list of candidate terms, there exists the case LLMs output returns the output that contains the list of the candidate terms with the additional sentence not appearing in the few-shot demonstration and not the sentence we input or so-called "*hallucination*" information (See example 1

in Table 1). Besides, a small proportion of the predictions (less than 0.1% on average) are not well wrapped up inside quotation symbols (See example 2 in Table 1).

Thus, we provide the postprocessing steps to solve the above issues by splitting the predictions into two parts (the list of candidate terms and the rest) and keeping only the part of the candidate list. After that, we correct all the format of the candidate terms that are in the wrong format by encapsulating them with quotation symbols.

Example 1: Output with hallucination sentence
Sentence: "The analysis included a large study sample with more than 60,000 patients across 4372 hospitals." **Domain:** Heart failure **Output:** ["patients", "hospitals"] [INST] Sentence: "The study from that the new drug was effective in reducing symptoms in patients with mild to moderate depression." Domain: Heart failure [/INST] **Expected output:** ["patients", "hospitals"]
Example 2: Output in wrong formats due to lack of quotation marks
Sentence: "The PAF, its key biosynthetic enzymes (lyso-PAF acetyltransferase [lyso-PAF-AT] and dithiothreitol [DTT]-insensitive CDP choline: 1-alkyl-2-acetyl-sn-glycerol cholinephosphotransferase [PAF-CPT]), and its catabolic isoenzymes (PAF-acetylhydrolase [PAF-AH] and lipoprotein-associated phospholipase A2 [Lp-PLA2]) were mesured in serum and leukocytes of participants." **Domain:** Heart failure **Output:** [PAF, lyso-PAF-AT, DTT, CDP choline, PAF-CPT, PAF-AH, Lp-PLA2]] **Expected output:** ["PAF", "biosynthetic enzymes", "lyso-PAF acetyltransferase", "lyso-PAF-AT", "dithiothreitol", "CDP choline", "1-alkyl-2-acetyl-sn-glycerol choliephosphotransferase", "AF-CPT", "catabolic isoenzymes", "PAF-acetylhydrolase", "PAF-AH", "ipoprotein-associated phospholipase A2", "Lp-PLA2", "leukocytes"]

Tab. 1. Examples for error scenarios on the ACTER dataset, where the error or wrong-formatted information is colored red and the expected correct output is colored blue.

5. Results and Discussion

5.1. Results

The LLM we use in this study to address the ATE task is *Llama-2-13b-chat-hf*[3], the medium-size chat version of the Llama-2 family. We find in *Llama-2-13b-chat-hf* a compromise between model size and performance. We assess the performance of each term extractor by strictly comparing the aggregated list of candidate terms identified across the entire test set against the manually designated gold standard list of terms. For this purpose, we use three evaluation metrics: precision, recall, and F1 score. These metrics have also been employed in prior research endeavors (Hazem *et al.*, 2020; Lang *et al.*, 2021; Tran *et al.*, 2022b, c).

In Table 1, we report the results of *Llama-2-13b-chat-hf* applied to the ANN and NES versions of the English subset of the ACTER dataset. To establish a strong baseline reference point, we also report the performance of the *gpt-3.5-turbo* LLM for in-domain demonstrations with the explicit domain. The *gpt-3.5- turbo* together with *gpt-4* are the most popular private LLMs by OpenAI.

Domain-specificity	Lexicon-specificity	English		
		precision	recall	F1 score
ANN version				
Explicit	***In-domain*** *gpt–3.5–turbo*	0.396	0.483	0.435
	In-domain *Llama2–Chat–13B*	0.350	0.634	0.451
	Cross-domain *Llama2–Chat–13B*	0.419	0.629	0.503
Implicit	***In-domain*** *Llama2–Chat–13B*	0.354	**0.640**	0.456
	Cross-domain *Llama2–Chat–13B*	**0.429**	0.623	**0.508**

3 https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

NES version				
Explicit	***In-domain*** *gpt–3.5–turbo*	0.398	0.531	0.455
	In-domain *Llama2–Chat–13B*	0.384	0.661	0.486
	Cross-domain *Llama2–Chat–13B*	0.408	0.657	0.503
Implicit	***In-domain*** *Llama2–Chat–13B*	0.386	0.659	0.487
	Cross-domain *Llama2–Chat–13B*	**0.410**	**0.665**	**0.507**

Tab. 2. Comparative results in term extraction applied to ACTER corpora.

5.2. Discussion

Explicit domain vs. implicit domain (domain-specificity). Across both gold-standard versions, the implicit prompt scenarios outperform the explicit scenarios in all three metrics. This suggests that implicit prompts, which guide the model in identifying relevant terms without explicitly mentioning them, are more effective for term extraction.

In-domain vs. cross-domain demonstrations (lexicon-specificity). We observe a similar behavior related to lexicon-specificity. The cross-domain demonstrations achieve F1 scores that surpass the in-domain demonstrations. This again highlights the potential for transfer learning and the model's ability to adapt to new tasks even without extensive in-domain training. In other words, this suggests that the cross-domain model may be better at generalizing to new unseen data.

ANN vs. NES gold standards. Across all prompt scenarios, F1 scores are consistently higher in the NES version compared to the ANN version in explicit domain-specificity. This suggests that including named entities in the gold standard with a specified domain leads to better term extraction performance, which is typically easier to identify than other terms.

5.3. Conclusion

In this study, we explored the applicability of the open-sourced large language models (LLMs), namely *Llama2-Chat-13B* on the term extraction task using in-context learning prompts. We evaluate how well LLMs perform with different levels of domain-related information, *i.e*., in- and cross-domain demonstrations with and without domain enunciation. Our findings suggest the model's predictive performance when designing prompts with few demonstrations without specifying the domain to help LLMs capture the candidate terms in the correct formats. Furthermore, the results demonstrate that LLMs offer a valuable compromise by reducing the need for extensive data annotation, making them suitable for real-world applications where labeled data are scarce. However, their effectiveness depends heavily on the prompting style, language, and type of response generated.

The work was partially supported by the Slovenian Research and Innovation Agency (ARIS) core research program Knowledge Technologies (P2-0103) and projects Linguistic Accessibility of Social Assistance Rights in Slovenia (J5-50169) and Embeddings-based techniques for Media Monitoring Applications (L2-50070). The work has also been supported by the ANNA (2019-1R40226) and TERMITRAD (2020-2019-8510010) projects funded by the Nouvelle-Aquitaine Region, France. Besides, the work was supported by the project Cross-lingual and Cross-domain Methods for Terminology Extraction and Alignment, a bilateral project funded by the program PROTEUS under the grant number BI-FR/23-24-PROTEUS006.

References

Amjadian, E., Inkpen, D., Paribakht, T., Faez, F., 2016. "Local-Global Vectors to Improve Unigram Terminology Extraction". In *Proceedings of the 5th International Workshop on Computational Terminology (Computerm2016)*, pp. 2–11.

Bian, N., Han, X., Sun, L., Lin, H., Lu, Y., He, B., 2023. "Chatgpt is a knowledgeable but inexperienced solver: An investigation of the commonsense problem in large language models". arXiv:2303.16421

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*, 2020. "Language models are few-shot learners". *Advances in neural information processing systems* 33, 1877–1901.

Castellví, M.T.C., Bagot, R.E., Palatresi, J.V., 2001. "Automatic term detection: A review of current systems. Recent advances in computational terminology" 2, 53–88.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., *et al.*, 2022. "Palm: Scaling language modeling with pathways". arXiv:2204.02311

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. "Unsupervised cross-lingual representation learning at scale". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* Association for Computational Linguistics, 8440–8451.

Damerau, F.J., 1990. "Evaluating computer-generated domain-oriented vocabularies". *Information processing & management* 26, 791–801.

El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J., 2014. "Scalable topical phrase mining from text corpora". *Proc. VLDB Endow.* 8, 305–316. doi.org/10.14778/2735508.2735519

Frantzi, K.T., Ananiadou, S., Tsujii, J., 1998. "The c-value/nc-value method of automatic recognition for multi-word terms". In *International conference on theory and practice of digital libraries, Springer.* pp. 585–604.

Gao, Y., Yuan, Y., 2019. "Feature-less End-to-end Nested Term extraction". In *CCF International Conference on Natural Language Processing and Chinese Computing, Springer*, pp. 607–616.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y., 2023. "How close is ChatGPT to human experts? comparison corpus, evaluation, and detection". arXiv:2301.07597

Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M., 2020. "Retrieval augmented language model pre-training". In *International conference on machine learning*, PMLR, pp. 3929–3938.

Hazem, A., Bouhandi, M., Boudin, F., Daille, B., 2020. "TermEval 2020: TALN-LS2N System for Automatic Term Extraction". In *Proceedings of the 6th International Workshop on Computational Terminology*, pp. 95–100.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., Sontag, D., 2023. "Tabllm: Few-shot classification of tabular data with large language models". In *International Conference on Artificial Intelligence and Statistics, PMLR*. pp. 5549–5581.

Justeson, J.S., Katz, S.M., 1995. "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text". *Natural Language Engineering* 1, 9–27.

Kageura, K., Umino, B. 1996. "Methods of automatic term recognition: A review." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2), 259-289.

Kessler, R., Béchet, N., Berio, G., 2019. "Extraction of terminology in the field of construction". In *2019 First International Conference on Digital Data Processing (DDP), IEEE*. pp. 22–26.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J.,Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radlinski, L., Wojtasik, K., Wozniak, S., Kazienko, P., 2023. "ChatGPT: Jack of all trades, master of none". arXiv:2302.10724

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., Stuker, S., 2018. "Term Extraction via Neural Sequence Labeling a Comparative Evaluation of Strategies Using Recurrent Neural Networks". In *INTERSPEECH*, pp. 2072–2076.

Lang, C., Wachowiak, L., Heinisch, B., Gromann, D., 2021. "Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains". In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3607–3620.

Le Serrec, A., L'Homme, M.-C., Drouin, P., Kraif, O., 2010. "Automating the compilation of specialized dictionaries: Use and analysis of term extraction and lexical alignment." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 16, 77–106.

Lingpeng, Y., Donghong, J., Guodong, Z., Yu, N., 2005. "Improving retrieval effectiveness by using key terms in top retrieved documents". In *European Conference on Information Retrieval, Springer*. pp. 169–184.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L., 2020. "Multilingual denoising pre-training for neural machine translation". *Transactions of the Association for Computational Linguistics* 8, 726–742.

Omar, R., Mangukiya, O., Kalnis, P., Mansour, E., 2023. "ChatGPT *versus* traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots." arXiv:2302.06466

Perez, E., Kiela, D., Cho, K., 2021. "True few-shot learning with language models". *Advances in neural information processing systems* 34, 11054–11070.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*, 2019. "Language models are unsupervised multitask learners". *OpenAI blog* 1, 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer". *The Journal of Machine Learning Research* 21, 5485–5551.

Repar, A., Podpecan, V., Vavpetic, A., Lavrac, N., Pollak, S., 2019. "TermEnsembler: An Ensemble Learning Approach to Bilingual Term Extraction and Alignment". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 25, 93–120.

Rigouts Terryn, A., Hoste, V., Drouin, P., Lefever, E., 2020. "TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset". In *6th International Workshop on Computational Terminology (COMPUTERM 2020), European Language Resources Association (ELRA)*. pp. 85–94.

Roberts, A., Raffel, C., Shazeer, N., 2020. "How much knowledge can you pack into the parameters of a language model?". arXiv:2002.08910

Song, M., Jiang, H., Shi, S., Yao, S., Lu, S., Feng, Y., Liu, H., Jing, L., 2023. "Is ChatGPT a good keyphrase generator? A preliminary study". arXiv:2303.13001

Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., Qi, G., 2023. "Can ChatGPT Replace Traditional KBQA Models? An In-depth Analysis of the Question Answering Performance of the GPT LLM Family". arXiv:2303.07992

Tran, H. T. H., Martinc, M., Repar, A., Ljubešić, N., Doucet, A., & Pollak, S. (2024). "Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling?". *Machine Learning*, 1-30.

Tran, H.T.H., Martinc, M., Caporusso, J., Doucet, A., Pollak, S., 2023. "The recent advances in automatic term extraction: A survey". arXiv:2301.06767

Tran, H., Martinc, M., Doucet, A., Pollak, S., 2022a. "A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction". In *Slovenian Conference on Language Technologies and Digital Humanities*.

Tran, H.T.H., Martinc, M., Doucet, A., Pollak, S., 2022b. "Can cross-domain term extraction benefit from cross-lingual transfer?". In *Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings, Springer*. pp. 363–378.

Tran, H.T.H., Martinc, M., Pelicon, A., Doucet, A., Pollak, S., 2022c. "Ensembling transformers for cross-domain automatic term extraction". In *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries: 24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, November 30–December 2, 2022, Proceedings,* Springer. pp. 90–100.

Vezzani, F. 2019. "La technicité des termes : le v-tech comme paramètre d'évaluation". In *Terminologie & Ontologie : Théories et Applications, Actes de la conférence TOTh 2019*, Presses Universitaires Savoie Mont Blanc, « Terminologica », pp. 215-227.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., Foster, G., 2022. "Prompting palm *for* translation: Assessing strategies and performance". arXiv:2211.09102

Vintar, S., 2010. "Bilingual Term Recognition Revisited: The Bag-of-equivalents Term Alignment Approach and its Evaluation. Terminology". *International Journal of Theoretical and Applied Issues in Specialized Communication* 16, 141–158.

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V., 2021. "Finetuned language models are zero-shot learners". arXiv:2109.01652

Wu, H., Wang, W., Wan, Y., Jiao, W., Lyu, M., 2023. "ChatGPT or grammarly? Evaluating ChatGPT on grammatical error correction benchmark". arXiv:2303.13648

Zhang, Z., Gao, J., Ciravegna, F., 2018. "Semre-rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank". *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 1–41.

Extracting domain-specific terms using contextual word embeddings

Andraž Repar*, Nada Lavrač**, Senja Pollak***

*Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
repar.andraz@gmail.com

**Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
nada.lavrac@ijs.si

***Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Senja.pollak@ijs.si

Abstract. Automated terminology extraction refers to the task of extracting meaningful terms from domain-specific texts. This paper proposes a novel machine learning approach to terminology extraction, which combines features from traditional term extraction systems with novel contextual features derived from contextual word embeddings. Instead of using a predefined list of part-of-speech patterns, we first analyse a new term-annotated corpus RSDO5 for the Slovenian language and devise a set of rules for term candidate selection and then generate statistical, linguistic and context-based features. We use a support-vector machine algorithm to train a classification model, evaluate it on the four domains (biomechanics, linguistics, chemistry, veterinary) of the RSDO5 corpus and compare the results with relevant extraction approaches for the Slovenian language. While the approach does not improve on the state-of-the-art in terms of F1 scores, it proves that contextual word embeddings can provide a valuable contribution to improving term extraction.

Résumé. *L'extraction automatique de terminologie désigne le fait d'extraire des termes clés au sein de textes spécifiques à un domaine. Cet article propose une nouvelle approche d'apprentissage automatique pour l'extraction de terminologie, qui combine certaines caractéristiques des approches traditionnelles avec celles, novatrices, dérivées des* embeddings *de mots contextuels. Au lieu d'utiliser une liste prédéfinie de modèles de parties du discours, nous analysons*

d'abord RSDO5, un nouveau corpus annoté pour la tâche d'extraction de termes pour la langue slovène, et élaborons un ensemble de règles pour la sélection des termes candidats, puis générons des caractéristiques statistiques, linguistiques et contextuelles. Nous utilisons un algorithme de machine à vecteurs de support pour entraîner un modèle de classification, l'évaluons sur les quatre domaines (biomécanique, linguistique, chimie, vétérinaire) du corpus RSDO5, et comparons les résultats avec des approches d'extraction pertinentes pour la langue slovène. Bien que l'approche n'améliore pas l'état de l'art en termes de scores F1, elle prouve que les embeddings *de mots contextuels peuvent apporter une contribution précieuse pour améliorer l'extraction de termes.*

1. Introduction

In this paper, we describe a terminology extraction methodology that combines two traditional aspects of automated terminology extraction with a novel contextual-embedding approach in a machine learning setting. Focusing on the Slovenian language, which is an under-resourced Slavic language with a rich morphology, we conducted the first experiments on a new RSDO5 (see Section 3) corpus of term-annotated texts created as part of the RSDO national project. The proposed approach starts with the corpus — we first analyze the annotated terms in the corpus and study their part-of-speech tags. While traditional systems use a set of pre-defined part-of-speech patterns to identify the initial candidate terms (CTs), we take a different approach and instead define a shallow filter for CTs, which considers only very basic part-of-speech based information. In our approach, we generate three types of features (linguistic, statistical and contextual) and use them to train a linear support vector machine (SVM) classifier. Since the RSDO5 corpus contains four different domains, we train the algorithm on three domains and test its performance on the fourth domain using the standard measures of precision, recall and the F1 score (as in the related work by Rigouts Terryn, Hoste, Drouin *et al.* [2020]).

2. Related work

Terminology extraction refers to structuring terminological knowledge from unstructured text. According to ISO standard 087-1:2000 Terminology work — Vocabulary, terminology is a set of designations belonging to one special language. ATE systems were traditionally classified as either statistical,

linguistic or a combination of these two approaches. The linguistic approach utilizes the distinctive linguistic aspects of terms, most often their syntactic (*i.e.* part-of-speech) patterns. On the other hand, the statistical approach takes advantage of term frequencies in a corpus. However, most traditional systems are hybrid, using a combination of the two approaches. For example, Justeson and Katz (1995) first define part-of-speech (POS) patterns of terms and then use simple frequencies to filter the term candidates.

Many terminology extraction algorithms are based on the concepts of termhood and unithood defined by Kageura and Umino (1996): termhood is "the degree to which a stable lexical unit is related to some domain-specific concepts" and unithood is "the degree of strength or stability of syntagmatic combinations and collocations". Termhood-based statistical measures (Vintar, 2010) function on a presumption that a term's relative frequency will be higher in domain-specific corpora than in the general language, while common statistical measures, such as mutual information (Daille *et al.*, 1994), are used to measure unithood. These two approaches have been used as a basis of several hybrid systems, such as Termolator (Meyers *et al.*, 2018) and TermEnsembler (Repar *et al.*, 2019).

Another distinct approach is to utilize machine learning with feature engineering. It involves first extracting a set of term candidates, followed by the calculation of various features and training of a machine learning model, where term extraction is treated as a bilingual classification task. Various types of machine learning algorithms have been used, such as decision trees (Karan *et al.*, 2012), rule induction (Foo & Merkel, 2010), k-nearest neighbours (Zadeh & Handschuh, 2014), support vector machines (Ljubešić *et al.*, 2019) and random forest (Rigouts Terryn *et al.*, 2021). The last two approaches are particularly relevant for our work.

The first approach, developed by Ljubešić *et al.* (2019) is the state-of-the-art system for Slovene. It first extracts CTs with the CollTerm tool (Pinnis *et al.*, 2012), which uses a complex language-specific set of term patterns originally developed for the SketchEngine terminology extraction module (Fišer *et al.*, 2016). A total of 31 patterns were defined from unigrams up to four grams and CTs (*i.e.* lemmatized versions of terms) with a frequency of 3 or more were considered. The resulting term lists were annotated by four annotators as either in-domain, out-of-domain, academic or irrelevant terms (which is a similar setup as used in the ACTER and RSDO datasets). The annotations were then used as training data for a machine learning approach with the following features: term frequency, 5 statistical measures from the

SketchEngine terminology module (chi-square, dice, pointwise mutual information, t-score, tf-idf), C-value (Frantzi *et al.*, 2000), oversampling (based on instances where the annotators were in agreement), candidate length, average token length, term pattern and context. Since some of the statistical measures can be calculated only for multi-word units, they trained separate classifiers for single-word (SWU) and multi-word (MWU) units. They evaluated the models in a cross-validation setting on the Slovene KAS dataset (Erjavec *et al.*, 2018), obtaining F1 scores of around 0.5 (*i.e.* when only the irrelevant annotation is considered negative and the remaining annotations are treated as valid terms, which is similar to the setting used in our experiments).

The second approach, which also uses the machine learning paradigm with feature engineering (as we do in our work) was developed by Rigouts Terryn *et al.* (2021). It first identifies the linguistic patterns of the annotated terms in the ACTER corpus and then uses these patterns to identify term candidates. They generate 177 features in 6 subgroups: shape (*e.g.*, number of tokens in a CT), linguistic (*e.g.*, POS tag of the first token of the CT), frequency (*e.g.*, relative frequency of the CT in a specialized corpus), statistical (various termhood/unithood measures), contextual (*e.g.*, whether the CT occurs between parentheses or right before/after parentheses), variational (number of different variations of the CT). They experimented with several different classification algorithms in the sklearn Python library and obtained the best F1 scores with the random forest classifier. In the setup with a held-out test set (3 domains are used for training, one for testing and the experiments are run 4 times, each time with a different test domain) they achieved F1 scores between 0.338 and 0.436 for English, between 0.288 and 0.520 for French and between 0.361 and 0.616 for Dutch.

3. Data

The RSDO5 corpus (Jemec Tomažin *et al.*, 2021) was compiled for developing and evaluating terminology extraction methods in Slovene and was inspired by the ACTER corpus (Rigouts Terryn, Hoste, & Lefever, 2020). It consists of 12 texts with 250,000 words and 38,000 manually annotated terms1. The corpus texts, published between 2000 and 2019, belong to the fields of biomechanics, linguistics, chemistry, or veterinary science. For each domain, they include: a PhD thesis, a graduate level textbook, and a journal article. The entire texts were annotated for terminology and allow for evaluation of methods in terms of precision and recall. Apart from the manually annotated terms, the corpus was automatically annotated, *i.e.* performing tokenization,

sentence segmentation, lemmatization, assigning morphological features and dependency syntax using the Classla pipeline (Ljubešić & Dobrovoljc, 2019). It is available in the Clarin.si repository.

The RSDO5 corpus contains a total of 9,657 unique lemma term forms and a large number of these occur only once or twice in an individual domain of the corpus. We have analyzed the annotated terms in the corpus with respect to token and character lengths, POS tags of unigram terms, POS tags of the first token in the terms, POS tags of the last token in the terms, and frequency of different POS tags in the terms.

Most annotated terms have 4 or less tokens. The longest term per domain has 6 tokens in the chemistry domain, 11 tokens in the biomechanics domain, 10 tokens in the veterinary domain and 8 tokens in the linguistics domain. The vast majority of terms also have more than 4 characters. We can also observe that:

- almost all unigram terms are either nouns (NOUN) or proper nouns (PROPN);
- most terms start with either an adjective (ADJ), noun (NOUN) or proper noun (PROPN);
- most multi-word unit terms end with a noun (NOUN) or a proper noun (PROPN);
- nouns (NOUN) and adjectives (ADJ) are by far the most frequent POS tags that appear in the terms, but adverbs (ADV), adpositions (ADP) and proper nouns (PROPN) can also be found; other POS tags occasionally appear in some terms, but the number of occurrences is low and may in some cases be attributed to errors during the syntactic parsing process.

4. System overview

This section describes the architecture of the system. We use a machine learning approach to automated terminology extraction (ATE), which means that for each term candidate, we generate a set of features and then use the corresponding labels for model training. Our approach is similar to the ones developed by Ljubešić *et al.* (2019) and Rigouts Terryn *et al.* (2021) in some respects, but it differs from them in two major aspects: 1) instead of using a pre-defined set of linguistic patterns to identify CTs, we apply 6 filters to all possible *n*-grams in the input corpus up to a user-defined length *n*, and 2) in addition to linguistic and statistical features, we also employ a set of novel contextual features based on ELMo (Peters *et al.*, 2018) contextual word-embeddings.

4.1. Dataset pre-processing

In traditional ATE systems, candidate terms (CTs) are usually selected based on a pre-defined list of POS patterns. For example, all adjective-noun (ADJ+NOUN) sequences, such as "supervised learning" or "basic rule", are considered CTs. Since most terms, in particular high frequent ones, correspond to one of the standard patterns, this allows us to quickly filter out a large number of low-quality CTs. However, defining a POS pattern list that would effectively cover terms across various types of domains is difficult. While some common patterns (*e.g.*, NOUN+NOUN or ADJ+NOUN) can be considered universal, other patterns may be domain-specific and maintaining different pattern lists for many different domains can be cumbersome. Using a traditional pattern-based approach, we also discard potentially valid terms that do not correspond to one of the POS patterns either because they follow a non-standard POS pattern or because they are unusually long. Since most patterns are rarely longer than 4 or 5 tokens, longer terms would automatically be discarded.

Instead of relying on a list of POS patterns to identify CTs we apply a shallow filter to all *n*-grams up to a pre-defined maximum length value. In our case, we set the maximum length to 11, since this is the longest annotated term in the RSDO5 corpus. The shallow filter is based on the analysis of the terms in the RSDO5 corpus and describes the general linguistic characteristics of the terms. The rules are described below:

- Terms have to be longer than 3 characters.
- Only nouns can be single word terms.
- Patterns longer than 1 have to end with a noun (NOUN) or proper noun (PROPN) to be terms.
- Patterns not starting with adjectives (ADJ), adverbs (ADV) or nouns (NOUN, PROPN) are not terms.
- If a pattern contains a verb (VERB), a symbol (SYM), a subordinating conjunction (SCONJ), punctuation (PUNCT), a pronoun (PRON), a particle (PART), an interjection (INTJ), a determiner (DET), a coordinating conjuction (CCONJ), an auxiliary verb (AUX) or other (X), it is not a term.
- If term contains a comma or an underscore, it is not a term.

Using these 6 filters, we are able to significantly reduce the number of CTs while maintaining adequate gold standard coverage. The maximum recall per domain is 0.91 for the biomechanics and linguistics domain, 0.86 for the

veterinary domain and 0.93 for the chemistry domain, while the number of CTs (including valid gold standard terms) is 12,847 for the biomechanics domain, 22,610 for the linguistics domain, 17,996 for the veterinary domain and 15,417 for the chemistry domain.

4.2. Feature construction

Similar to traditional ATE systems, we generate linguistic and statistical features and then we also add a third category of features that are based on contextual word embeddings.

4.2.1. Linguistic features

Contrary to some traditional systems, such as the one developed by Justeson and Katz (1995), we do not use pre-defined linguistic patterns, but instead generate features based on a set of loosely defined pattern rules utilizing the Universal Dependency (UD) POS tags (de Marneffe *et al.*, 2021), to avoid overlooking terms due to missing patterns.

We generate several vectors of UD tags for each CT depending on the UD part-of-speech value as described below. Each vector has a length of 17 corresponding to the number of all possible UD tags:

- StartUD: a one-hot vector where the UD tag of the first token in the CT has a value of 1, while the rest have a value of 0,
- EndUD: a one-hot vector where the UD tag of the last token in the CT has a value of 1, while the rest have a value of 0; in the case of unigram CTs, the StartUD and EndUD vectors are the same,
- AnywhereUD: a vector where the UD tags that appear anywhere in the CT other than the first and last position have a value of 1, while the rest have a value of 0; in the case of unigram and bigram CTs, this vector has only zero values,
- CountOfUD: a vector indicating the number of occurrences of each UD tag anywhere in the CT,
- NoUniquePos: an additional numeric feature that counts the number of unique POS tags in the CT.

The vectors are concatenated, resulting in a representation of 69 features: 51 features with binary values, and 18 (17 from CountOfUD and 1 NoUniquePos) with numeric values.

4.2.2. Statistical features

For statistical features, we use the termhood measure from Vintar (2010), which is based on the premise that domain-specific terms are used more frequently (in relative terms) in domain texts than in the general language. But instead of calculating a single termhood value, we generate the three core variables (general corpus frequency, domain corpus frequency and term length) from the termhood formula as separate features:

- TermGenFreq: the sum of the relative frequencies in a general corpus of individual tokens constituing a CT,
- TermDomFreq: the sum of the relative frequencies in a domain-specific corpus of individual tokens constituing a CT,
- TermLength: which corresponds to the length of the CT (*i.e.* number of tokens).

4.2.3. Contextual features

To generate contextual features, which is also the main novelty of our approach, we utilize a premise that is similar to statistical termhood measures. Just as termhood suggests that domain-specific terms are used more frequently in domain-specific corpora than in general corpora, so we hypothesize that domain-specific terms are used in different contexts in domain-specific corpora compared to general corpora.

For the calculation of contextual embeddings, we used the eLMO model for Slovenian created by Ulčar (2019), which was trained on the Gigafida 2.0 corpus for 10 epochs. General corpus contextual embeddings were calculated for the top 200k most frequent tokens from the publicly available ccGigafida corpus (Logar *et al.*, 2013). To produce the values, the first LSTM layer was used (based on Reimers and Gurevych (2019a) who report that the middle layer is the best single layer and comparable to concatenating all three layers) to produce the vector values. Each instance of a word has its own vector, based on the context it appears in. These vectors have been averaged, so that each word has only one corresponding vector representing its average context in the general corpus. We then calculate average word embeddings for every word in the domain corpus. To do this, we first tokenize the corpus into sentences and then generate word embeddings for every sentence using the AllenNLP Python library (Gardner *et al.*, 2018) using the same ELMo settings as for the generation of the general domain corpus embeddings. We iterate over every word in the sentence and calculate

the average embedding of its lemma in the domain corpus by adding up all vectors and dividing them by the number of occurrences of the lemma in the corpus. While for single word terms the average lemma embedding is the final representation, for every multi-word term, we calculate the average embedding by summing up the average lemma embeddings of all tokens in the term and dividing the sum with the number of tokens in the term. All 1,024 dimensions of the resulting vector are then added to the dataset as features. In a similar manner, we then also generate the average general domain term embedding by summing up average lemma embeddings in the general corpus and dividing the sum with the number of tokens in the term. All 1,024 dimensions of the resulting vector (elmo feature vector) are again added to the representation feature vector.

Finally, we generate three additional features based on contextual embeddings:

- **elmoSim**, which is the cosine similarity of the domain-specific and generalterm embeddings calculated as described above, the motivation being that since true terms are used in different contexts in domain-specific and general language texts, the similarity between the two vectors would be smaller for true terms compared to expressions that are not valid terms.
- **elmoTermSim**, which is the cosine similarity of the domain-specific term embedding and the embedding of a seed term defined by the user 7, the motivation being that true terms would be used in similar contexts to a term representative of the domain.
- **elmoStDev**, which is calculated as follows: for every lemma in the domain-specific corpus, we calculate the standard deviation of all its contextual embeddings and then for each term, we sum up the standard deviations of all lemmas and divide the sum with the number of tokens in the term, the motivation being that true terms in most cases appear in similar contexts within domain-specific texts which would result in smaller standard deviation compared to non-valid terms.

5. Experiments and results

5.1. Experimental setup

For model training, we experimented with five algoritms from the sklearn Python library. The best F1 score was achieved by a SVM binary classifier with a linear kernel using the default settings (c = 1). Following the setup proposed by Hazem *et al.* (2020), we use three domains for training and one for testing, which means that we run the experiments four times with a different domain used for testing each time. Evaluation is performed using the standard measures of precision, recall and F1 scores. Precision is calculated as the number of true positive terms divided by the number of all predicted terms, recall is calculated as the number of true positive terms divided by the number of GS terms and F1 score is calculated as the harmonic mean of precision and recall.

We compare the results to the state-of-the-art for Slovenian (Ljubešić *et al*., 2019) (the code for that approach is freely available), to the LUIZ approach by Vintar (2010) as implemented in Repar *et al*. (2019), where a joint list of single and multi-word terms is produced, and to an approach where we use corpus-based patterns similar to Rigouts Terryn *et al.* (2021) instead of our filtering rules. For Ljubešić *et al.* (2019), we trained the MWU and SWU models for all four combinations of training and testing domains and evaluated their performance against the gold standard terms. Minimum frequency was set to 1, which is the same as in our approach. For the LUIZ approach, which does not classify the terms but produces a ranked list of term candidates, we used a cut-off which corresponded to the number of terms predicted by our approach for a specific domain (*i.e.* if our approach predicted n terms for a domain, we considered the top n terms according to the LUIZ score for this domain).

5.2. Results

Model	Test	Precision	Recall	F1 score
Our approach	bim	0.650	**0.448**	**0.530**
Pattern approach	bim	**0.694**	0.342	0.458
LUIZ	bim	0.359	0.393	0.363
Ljubešić *et al.* (2019)	bim	0.538	0.248	0.339
Our approach	ling	0.672	**0.494**	**0.569**
Pattern approach	ling	**0.678**	0.446	0.538
LUIZ	ling	0.338	0.393	0.363
Ljubešić *et al.* (2019)	ling	0.522	0.254	0.341
Our approach	chem	0.691	**0.472**	**0.561**
Pattern approach	chem	**0.694**	0.374	0.486
LUIZ	chem	0.239	0.444	0.311
Ljubešić *et al.* (2019)	chem	0.478	0.314	0.378
Our approach	vet	**0.688**	**0.523**	**0.594**
Pattern approach	vet	0.670	0.487	0.564
LUIZ	vet	0.400	0.349	0.373
Ljubešić *et al.* (2019)	vet	0.669	0.193	0.299

Table 1. Precision, recall and F1 score compared to state-of-the-art results for Slovenian. Only the test domain is listed, it can be assumed that the other three domains were used for training. Our approach uses the shallow filter described in 4.2.3, whereas Pattern approach uses corpus-based patterns similar to Rigouts Terryn et al. *(2021).*

With our approach, we achieve F1 scores of 0.530 for the biomechanics domain, 0.569 for the linguistics domain, 0.561 for the chemistry domain and 0.594 for the veterinary domain (see Table 1). These results are comparable

with several strong baselines, including results by Rigouts Terryn *et al.* (2021) and Lang *et al.* (2021), but have since been exceeded by a sequence labelling approach developed by Tran *et al.* (2024).

We also see that our approach exceeds the approach by Ljubešić *et al.* (2019) in both precision and recall, which is not surprising given that their method relies heavily on frequency-based features, which work best with high frequency CTs, as well as the more traditional LUIZ approach.

5.3. Ablation study

We wanted to analyze the impact of different feature types on the final results. One approach, particularly often used in the evaluation of deep learning algorithms, is ablation study (Meyes *et al.*, 2019). Analogous to ablation in biology, ablation in machine learning denotes the removal of individual components and studying the effect on the results. In line with this, we have removed the individual feature types from the dataset one-by-one and analyzed the results. We can observe that removing each feature type does reduce the F1 scores in all domains and removing both statistical and linguistic features results in an even bigger drop in F1 scores. When we removed the statistical features but kept linguistic and contextual features, we observed a drop in F1 score performance by 11.70% in the biomechanics domain, 7.91% in the linguistics domain, 13.37% in the chemistry domain and 5.72% in the veterinary domain. When we removed the pattern features but kept statistical and contextual features, we observed a drop in F1 score performance by 18.30% in the biomechanics domain, 13.18% in the linguistics domain, 11.59% in the chemistry domain and 9.43% in the veterinary domain. When we removed both statistical and linguistic features but kept contextual features, we observed a drop in F1 score performance by 34.34% in the biomechanics domain, 27.59% in the linguistics domain, 25.49% in the chemistry domain and 17.68% in the veterinary domain. Using only statistical and pattern features, either together or independently, produces almost no correct predictions.

All three different sets of features contribute to the final result. The drop in F1 score performance when removing linguistic features is somewhat larger than when removing statistical features (with the exception of the chemistry domain) and when we remove both statistical and linguistic features, the results are even worse. However, the results when using just contextual features are still respectable, in particular in terms of precision, which is above 0.630 for all four domains.

5.4. Error analysis

We performed error analysis of the results obtained with the best performing model (*i.e*. the model based on all three feature types described in Section 4.2). When looking at the false positive predictions in all four domains, we were immediately reminded of the issue we mentioned in the introduction, namely that there is no clear definition of the nature of domain-specific terms. On first look and with the caveat that we are not experts in any of these domains (with the possible exception of linguistics), it would seem that many of the false positives could be valid terms. For example, *atletska steza* (*running track*), *živčni končič* (*nerve ending*) and *upogibalka* (*flexor*) in the biomechanics domain, *jezikoslovni model* (*linguistic model*), *kodna tabela* (*code table*) and *nacionalni korpus* (*national corpus*) in the linguistic domain, *prekurzor* (*precursor*), *spektroskopija*, (*spectroscopy*) and *chronbachov koeficient* (*cronbach coefficient*) in the chemistry domain and *žvekalka* (*masseter muscle*), *stomatitis* (*stomatitis*) and *nekrotično vnetje* (*necrotic inflammation*) in the veterinary domain could to an untrained eye look like valid domain-specific terms. We also believe that some of the false positive predicted terms (and many others) would also be useful at least in a "semi-automatic" terminology extraction setting, where CTs are first extracted automatically and then evaluated by a domain expert.

In addition to the terms described above, we noted two additional issues among false positives. The first one is general terms/words being predicted as terms, such as *leto* (*year*), *mesto* (*city*), *sistem* (*system*), *stopnja* (*rate*), *proces* (*process*), *sprememba* (*change*), *skupina* (*group*), *delo* (*work*), *zbirka* (*collection*), *primer* (*example*), *pogoj* (*condition*), etc.

The reason for these wrongly predicted terms could again be related to the unclear nature of domain-specific terms, because the gold standard contains some terms that, on first look, do not look much different than the ones mentioned above, such as *sila* (*force*), *enota* (*unit*), *sejem* (*fair*), *komora* (*chamber*) etc. The second identified issue is related to the lemmatization algorithm in the Classla pipeline. For example, false positives contain wrongly lemmatized CTs, such as "regrgrposs" (lemma of "REGR_r_OsSU") and "doogpodlaht" (lemma of "D_O_podlahti") or "mehanovsprejemnik" (lemma of "mehano-sprejemniki") and "skorajstrokovnjak" (lemma of "skoraj-strokovnjakov").

6. Conclusions

This paper presents a machine learning terminology extraction system, which combines elements of traditional termhood- and unithood-based systems with a novel contextual-word-embedding-based approach that takes advantage of the differences in the domain-specific and general language contexts which terms appear in. In addition, we also introduce a novel method of candidate term selection — instead of being limited to a pre-defined list of part-of-speech patterns, we employ a shallow filter that offers greater flexibility in candidate term selection, in particular when it comes to unseen data and when implementing term extraction in user-facing applications.

We evaluated the system on a new corpus of term-annotated texts RSDO5 1.0 for Slovenian, created as part of the work in the Slovenian national language technology project RSDO. We compared the results to the existing state-of-the-art approach for Slovenian and were able to improve F1 scores by a considerable margin in all four domains. The novel contextual features based on the eLMO embeddings appear to work well even for low-frequency terms (a well-known issue of traditional statistical methods, many of which are based on frequency counts). In addition, the results also exhibit little variation between test domains.

In terms of time, the most time-consuming part is the calculation of contextual embeddings on the large reference corpus. But since this can be computed only once and reused for further calculations, the system is fairly quick for a corpus of modest size (*i.e.* 50 to 100 thousand words). *E.g.*, calculating domain-specific embeddings and applying the model is performed in approximately half an hour on a standard laptop without specialized machine learning hardware. This is an acceptable setting for practical applications, *e.g.* in translation industry or terminology dictionary construction setting. While the current version works only for Slovenian, it would be relatively easy to adapt it to other languages, provided that a suitable general language corpus is available.

References

Amjadian, E., Inkpen, D., Paribakht, T., Faez, F. 2016. "Local-global vectors to improve unigram terminology extraction". In: *Proceedings of the 5th International Workshop on Computational Terminology*, pp. 2–11.

Daille, B., Gaussier, É., Lange, J.-M. 1994. "Towards automatic extraction of monolingual and bilingual terminology". In: *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, Kyoto, Japan.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4171–4186.

Erjavec, T., Fišer, D., Ljubešić, N., Arhar Holdt, Š., Bren, U., Robnik-Šikonja, M., Udovič, B. 2018. "Terminology identification dataset KAS-term 1.0". Language resource repository CLARIN.SI. URL: https://www.clarin.si/repository/xmlui/handle/11356/1198?locale-attribute=sl

Fišer, D., Suchomel, V., Jakubiček, M. 2016. "Terminology extraction for academic Slovene using Sketch Engine". In: *Tenth Workshop on Recent Advances in Slavonic Natural Language Processing*, RASLAN 2016, pp. 135–141.

Foo, J., Merkel, M. 2010. "Using machine learning to perform automatic term recognition". In: *LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods*, European Language Resources Association, pp. 49–54.

Frantzi, K., Ananiadou, S., Mima, H. 2000. "Automatic recognition of multi-word terms: The C-value/NC-value method". *International Journal on Digital Libraries* 3, 115–130.

Gao, Y., Yuan, Y. 2019. "Feature-less end-to-end nested term extraction". In: *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 607–616.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.F., Peters, M., Schmitz, M., Zettlemoyer, L. 2018. "AllenNLP: A deep semantic natural language processing platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pp. 1–6.

Hazem, A., Bouhandi, M., Boudin, F., Daille, B. 2020. "TermEval 2020: TALN-LS2N system for automatic term extraction". In: *Proceedings of the 6th International Workshop on Computational Terminology*, pp. 95–100.

Jemec Tomažin, M., Trojar, M., Žagar, M., Atelšek, S., Fajfar, T., Erjavec, T. 2021. "Corpus of term-annotated texts RSDO5 1.0". Slovenian language resource repository CLARIN.SI. URL: https://www.clarin.si/repository/xmlui/handle/11356/1400

Justeson, J.S., Katz, S.M. 1995. "Technical terminology: Some linguistic properties and an algorithm for identification in text". *Natural Language Engineering* 1, 9–27.

Kageura, K., Umino, B. 1996. "Methods of automatic term recognition. A review. Terminology". *International Journal of Theoretical and Applied Issues in Specialized Communication* 3, 259–289.

Karan, M., Šnajder, J., Bašic, B.D., 2012. "Evaluation of classification algorithms and features for collocation extraction in Croatian". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 657–662.

Khan, M.T., Ma, Y., Kim, J.-J. 2016. "Term Ranker: A graph-based re-ranking approach". In: *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, pp. 310–315.

Krek, S., Holdt, Š.A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., Dobrovoljc, K. 2020. "Gigafida 2.0: The reference corpus of written standard Slovene". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3340–3345.

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., Stüker, S. 2018. "Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks". In: *Proceedings of INTERSPEECH*, pp. 2072–2076.

Lang, C., Wachowiak, L., Heinisch, B., Gromann, D. 2021. "Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3607–3620.

Ljubešič, N., Dobrovoljc, K., 2019. "What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian". In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pp. 29–34.

Ljubešić, N., Fišer, D., Erjavec, T., 2019. "Kas-term: Extracting Slovene terms from doctoral theses via supervised machine learning". In: *International Conference on Text, Speech, and Dialogue*, Springer. pp. 115–126.

Logar, N., Erjavec, T., Krek, S., Grčar, M., Holozan, P. 2013. "Written corpus ccGigafida 1.0". CLARIN.SI. URL: https://clarin.si/repository/xmlui/handle/11356/1035

de Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D. 2021. "Universal Dependencies". *Computational Linguistics* 47, 255–308.

Meyers, A.L., He, Y., Glass, Z., Ortega, J., Liao, S., Grieve-Smith, A., Grishman, R., Babko-Malaya, O. 2018. "The Termolator: Terminology recognition based on chunking, statistical and search-based scores". *Frontiers in Research Metrics and Analytics* 3, 19.

Meyes, R., Lu, M., de Puiseau, C.W., Meisen, T. 2019. "Ablation studies in artificial neural networks". arXiv:1901.08644

Pais, V., Ion, R. 2020. "TermEval 2020: RACAI's automatic term extraction system". In: *Proceedings of the 6th International Workshop on Computational Terminology*, pp. 101–105.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. 2018. "Deep contextualized word representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237.

Pinnis, M., Ljubešić, N., Stefanescu, D., Skadina, I., Tadic, M., Gornostay, T. 2012. "Term extraction, tagging, and mapping tools for under-resourced languages". In: *Proceedings of the 10th Conference on Terminology and Knowledge Engineering*, pp. 20–21.

Reimers, N., Gurevych, I., 2019. "Alternative weighting schemes for ELMo embeddings". arXiv:1904.02954

Repar, A., Podpečan, V., Vavpetič, A., Lavrač, N., Pollak, S. 2019. "TermEnsembler: An ensemble learning approach to bilingual term extraction and alignment". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 25, 93–120.

Rigouts Terryn, A., Hoste, V., Drouin, P., Lefever, E. 2020. "TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset". In: *Proceedings of the 6th International Workshop on Computational Terminology*, pp. 85–94.

Rigouts Terryn, A., Hoste, V., Lefever, E 2021. "HAMLET: Hybrid adaptable machine learning approach to extract terminology". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 27, 254–293.

Terryn, A.R., Hoste, V., Lefever, E. 2020. "In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora". *Language Resources and Evaluation* 54, 385–418.

Tran, H. T. H., Martinc, M., Doucet, A., & Pollak, S. 2022. "Can cross-domain term extraction benefit from cross-lingual transfer?". In: *Discovery Science: 25th International Conference*, *DS 2022, Montpellier, France, October 10–12, 2022, Proceedings,* 363–378. doi.org/10.1007/978-3-031-18840-4_26

Ulčar, M. 2019. "ELMo embeddings models for seven languages". Slovenian language resource repository, CLARIN.SI. URL: http://hdl.handle.net/11356/1277

Vintar, S. 2010. "Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation". *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 16, 141–158.

Wang, R., Liu, W., McDonald, C. 2016. "Featureless domain-specific term extraction with minimal labelled data". In: *Proceedings of the Australasian Language Technology Association Workshop*, pp. 103–112.

Zadeh, B., Handschuh, S. 2014. "Evaluation of technology term recognition with random indexing". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*.

Zhang, Z., Gao, J., Ciravegna, F., 2018. "SemRe-Rank: Improving automatic term extraction by incorporating semantic relatedness with personalised pagerank". *ACM Transactions on Knowledge Discovery from Data* 12, 1–41.

An Exploration into the Knowledge Network of *The Journey to the West*

A Digital Humanities Approach

Hui Liu

College of Foreign Languages, Nanjing University of Aeronautics and Astronautics
luisaliu0339@gmail.com

Abstract. *The Journey to the West* is a masterpiece in Chinese literature with great research value in both literary features and cultural connotation. Among the numerous existing English versions, Anthony Christopher Yu's translation has been widely praised. However, existing research on this translation has been limited to qualitative analyses which usually focus on the selection of translation methods employed. As a result, this study takes a digital humanities approach to delve into the intricate networks of knowledge found within Yu's translation. By employing ConText for LDA topic modeling and AntConc for further analysis, the study uncovers three main channels of knowledge in the footnotes of Yu's translation, namely, Confucianism, Taoism (or Daoism) and Buddhism. This study also highlights the book's philosophical roots in Confucianism, its development from official historical sources, and its integration into philosophical and religious studies within both eastern and central Asian contexts.

Résumé. La Pérégrination vers l'Ouest *est un chef-d'œuvre de la littérature chinoise, dont la valeur de recherche est considérable, tant sur le plan littéraire que culturel. Parmi les nombreuses versions anglaises existantes, la traduction d'Anthony Christopher Yu a été largement saluée. Cependant, les recherches existantes sur cette traduction se sont limitées à des analyses qualitatives qui se concentrent généralement sur la sélection des méthodes de traduction employées. Par conséquent, cette étude adopte une approche des humanités numériques pour explorer les réseaux complexes de connaissances trouvés dans la traduction de Yu. En utilisant ConText pour la modélisation thématique LDA et AntConc pour une analyse plus*

approfondie, l'étude révèle trois principaux canaux de connaissances dans les notes de bas de page de la traduction de Yu, à savoir le confucianisme, le taoïsme (ou taoïsme) et le bouddhisme. Cette étude met également en évidence les racines philosophiques du livre dans le confucianisme, son développement à partir de sources historiques officielles et son intégration dans les études philosophiques et religieuses dans les contextes d'Asie orientale et centrale.

1. Introduction

The late-Ming novel *The Journey to the West* is a monumental work of prose fiction built upon meandering pilgrimage to India by the seventh-century monk Tripitaka. As the protagonist, Sun Wukong (or Monkey King), embarks on his epic journey to retrieve Buddhist scriptures, the narrative unfolds against a backdrop where Confucian values of duty, loyalty, and moral conduct are tested. Additionally, the Taoist principles of harmony with nature, spontaneity, and the pursuit of the Way are embodied in the characters and events. Furthermore, Buddhism plays a central role as the ultimate goal of the pilgrimage is to attain enlightenment. The novel beautifully weaves together the philosophical threads of Confucianism, Taoism, and Buddhism, showcasing their coexistence and influence on the characters' journeys and the broader narrative.

Since its first publication as the one-hundred-chapter novel in 1592 by Wu Cheng'en, studies on this Chinese novel have been fairly continuous for over 400 years. One can glimpse the richness of production through the biography and admit that the English translations have played a pivotal role in introducing this seminal Chinese literary work to a global audience. Arthur Waley's early 20th century translation, while praised for its accessibility, has been critiqued for taking certain liberties in interpreting the original text. Anthony Christopher Yu's four-volume translation, completed in the late 20th century, is widely regarded as a comprehensive and faithful rendition, capturing the nuances of the narrative and the philosophical underpinnings of Confucianism, Taoism, and Buddhism. Yu's translation, however, has faced scrutiny for its scholarly density, making it challenging for casual readers. More recent translations, such as those by Julia Lovell and David Kherdian, have sought to strike a balance between accessibility and fidelity to the source material.

Although scholars from different nationalities have conducted various studies on English translations of *The Journey to the West*, most of these studies are theoretical or documentary in nature, and few scholars have examined the footnotes or their interconnections. As a result, this study takes a digital humanities approach to delve into the intricate network of knowledge found within Yu's translation. By employing ConText for LDA topic modeling and AntConc for further analysis, the study uncovers three main channels of knowledge in the notes of Yu's translation, namely, Taoism, Buddhism and Confucianism. The study also highlights the book's philosophical roots in Confucianism, its development from official historical sources, and its integration into philosophical and religious studies within both eastern and central Asian contexts.

2. Previous research and state-of-the-art

Anthony Christopher Yu, born on Oct. 6, 1938, in Hong Kong, was an eminent scholar of both religious studies and literature. He was the Carl Darling Buck Distinguished Service Professor Emeritus in the Humanities and Professor Emeritus of Religion and Literature in the Chicago Divinity School as well as a member of the Committee on Social Thought at the University of Chicago. In his complete translation of *The Journey to the West*, in addition to the ninety-six-page "Introduction" in the front of the book and the "Index" at the back, there are detailed notes on background knowledge, proper names, special expressions, metaphors, similes and analogies in each of the translated chapters.

Existing studies about Yu's translation can be roughly categorized into three groups. The first group is made up of book reviews and essays aiming at giving a full assessment of his translation (see, *e.g.* Wang and Xu, 2016). This kind of remarks puts emphasis on discussion of translation quality, techniques and strategies as well as praises on the wide-ranging and detailed footnotes attached in the translation. The second group includes essays written in retrospect and prospect of sinology development outside China, especially those made by Yu. For instance, many sinologists' achievements would be listed with just a few words summarizing their thinking and logic. He (2011) points out that Yu's research of sinology was based on psychology, and the major research approach He adopted is cross-cultural analysis. The last group centers on Yu's research on Chinese classics and religion. While considerable attention has been paid in the past to the translation of *The Journey to the West*, interest of Yu's study on Chinese classics has emerged only very slowly and in a more scattered way. These lone voices include Thomas DuBois and Blaine

Gaustad. DuBois (2007) provides a framework to address the indivisibility of statecraft and religion, therefore "warrants the effort of multiple readings" and Gaustad (2007) gives high praise on Yu's *State and Religion in China: Historical and Textual Perspectives*.

3. Obtaining and handling of data

Firstly, 1,240 footnotes in Yu's translation of *The Journey to the West* were collected and saved in txt format to form a corpus. Then the corpus was cleaned with the documents segmented, manually proofread, and stop-words removed. Topic modeling[1] is performed using ConText developed by Jana Diesner and her team in University of Illinois. Finally, AntConc is used for further analysis of selected themes.

4. Preliminary findings of the footnotes

Topic modeling can give a string of words that best represent a topic. And the collocate[2] function of AntConc is a powerful feature used for identifying and analyzing collocates of a given search term within a corpus. Based on table 1, it is clear that the words "Chinese," "daoist," "buddhism," and "liji" suggest an interplay of cultural, religious, and literary themes of *The Journey to the West*.

Topic	Weight	Topic Members
1	7.8655	chinese - chapter - term - text - daoist - poem - line - literally - body - buddhism
2	1.3866	appeared - grove - phenomena - works - smith - pool - allude - shaped - evil - issues
3	1.3769	tiantai - xue - quatrain - chi - leading - rhetorical - stick - energetics - empire - powerful -
4	1.3696	stock - ways - purple - sticks - black - lotus - dame - guo - venerated - national
5	1.3619	passage - scheme - provinces - flying - liji - designation - stages - commandments - boar - myth

1 This is a method of natural language processing that centralizes and lists the words in a corpus that express similar topics as topic strings, which allows for detecting the semantic content of the corpus.

2 A collocate is a word that frequently appears near another specific word, and analyzing collocates can reveal patterns of language usage and associations between words.

6	1.3617	power - flower - star - common - alternately - warring - flowing - ghosts - belief - rabbit
7	1.3556	schafer - magic - death - qian - production - memorial - astronomical - rivers - cave - pan

Tab. 1. Topic modeling of footnotes in Yu's translation of The Journey to the West.

In addition, "Chinese" as a topic member indicates the cultural and linguistic origin of *The Journey to the West*. As a classic of Chinese literature, the novel is deeply embedded in the historical, cultural, and philosophical milieu of China. Daoism ("daoist") is a religious and philosophical tradition deeply rooted in Chinese culture. Its presence in the context highlights the influence of Daoist beliefs, practices, and characters in the narrative. For example, the Monkey King encounters various Daoist figures and concepts throughout the story. Moreover, "buddhism" plays a significant role in the novel because the main plot of the book revolves around the pilgrimage of the monk Tripitaka, who traveled to India to obtain Buddhist scriptures. This journey symbolizes the quest for enlightenment, a central theme in Buddhism. Last but not least, "liji", the pinyin transliteration[3] for "礼记," which is translated as the *Book of Rites,* is the most intriguing. As one of the Confucian classics which deals with Chinese rituals, social forms, and court etiquette, liji in this context implies an analysis or reference to Confucian concepts of propriety and ritual.

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	way	1	5	3	2	1	26.976	5.271
2	entanglement	2	2	0	2	1	22.355	9.341
3	cong	3	2	2	0	1	19.366	8.341
4	bandits	3	2	2	0	1	19.366	8.341
5	livia	3	2	2	0	1	19.366	8.341
6	kohn	6	2	2	0	1	17.679	7.756
7	lao	6	2	0	2	1	17.679	7.756
8	shiji	8	2	1	1	1	13.230	6.171
9	chan	9	3	1	2	1	12.531	4.372
10	lord	10	2	0	2	1	12.431	5.882

Fig. 1. Collocates of dao.

3 Pinyin transliteration is a system for transcribing Chinese characters into the Latin alphabet. It is used to represent the pronunciation of Mandarin Chinese.

Figure 1 shows the collocates of *dao* generated by AntConc[4], a software designed by Laurence Anthony from Waseda University in Japan. Several interesting findings include: the word ranked first is *way* which could be associated with the Daoist concept of "The Way" or "Dao", a central philosophical idea in Chinese thought. The word ranked second is *entanglement* which might refer to complex narrative plots or philosophical discussions about the nature of reality, which are common in classical Chinese literature. The collocate *lao* is likely a reference to Laozi, the ancient Chinese philosopher and founder of Daoism. The collocate *shiji* refers to "史记", or the *Records of the Grand Historian*, a foundational historical text in Chinese literature. And the collocate *chan* refers to Chan Buddhism, a branch of Buddhism in China that could be discussed in the context of the religious elements of the novel.

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	chan	1	7	6	1	1	22.266	3.596
2	stanley	2	3	2	1	1	18.311	5.743
3	weinstein	2	3	2	1	1	18.311	5.743
4	china	4	7	1	6	1	18.084	3.124
5	daoism	5	5	1	4	1	15.948	3.605
6	powerless	6	2	1	1	1	15.162	6.743
7	mahayana	6	2	1	1	1	15.162	6.743

Fig. 2. Collocates of Buddhism.

Figure 2 shows the collocates of *Buddhism* which reflect a complex interplay of religious traditions, scholarly influences, and cultural context. The prominence of "Chan[5]" and "Daoism" suggests a particular emphasis on Zen Buddhism and its relationship with Daoism, while the inclusion of "Mahayana" points to broader Buddhist themes. The presence of scholarly names such as "Stanley" and "Weinstein" indicates that academic perspectives are integrated into the translation, potentially influencing its interpretation. The collocate "powerless" could indicate themes of humility, suffering, the relinquishment of worldly power, or the challenges faced by characters in this narrative.

4 AntConc 4.3.1 is used for this study.

5 Chan is associated with a school of Mahayana Buddhism known as Zen in Japan. It is the major branch of Buddhism practiced in East Asia

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	neo	1	6	6	0	1	53.525	7.782
2	traditions	2	3	3	0	1	24.934	7.367
3	analects	3	3	0	3	1	19.350	6.045
4	doctrine	4	3	0	3	1	16.932	5.460
5	transmitted	5	2	1	1	1	14.953	6.782
6	song	6	4	2	2	1	14.897	4.027
7	zeng	7	2	1	1	1	14.049	6.460

Fig. 3. Collocates of Confucius.

Figure 3 shows the collocates of *Confucius* which include: 1) Neo: This word could refer to "Neo-Confucianism", which is a later renaissance of Confucian thought that occurred after the classical period, integrating Taoist and Buddhist concepts; 2) Traditions: The term suggests discussions about Confucius within the context of cultural or philosophical traditions, which is consistent with the role of Confucius as a transmitter of traditional values; 3) Analects: This refers to the *Analects of Confucius*, a collection of sayings and ideas attributed to the Chinese philosopher and his disciples; 4) Transmitted: This implies that the text may discuss how Confucian teachings were passed down, which is central to the way Confucianism has been disseminated through generations: 5) Song: This word refers to the Song dynasty, which was a period of significant development for Neo-Confucianism; 6) Zeng: This refers to Zengzi, one of the disciples of Confucius, who was traditionally credited with writing one of the *Four Books* of Confucianism.

5. Categorizations of the footnotes

Discussions above highlight the blend of religious and philosophical traditions, namely, Confucianism, Daoism and Buddhism within the Chinese cultural framework in *The Journey to the West*. This blend of themes is a testament to the rich tapestry of ideas and beliefs of the fiction in classical Chinese literature. The following part will divide the 1,240 footnotes into four categories[6] based on thick translation theory to further examine the network of knowledge presented.

6 The four groups include (1) notes that trace the source; (2) notes that explain the background information; (3) notes that integrate the relevant information; and

5.1. Thick translation

Kwame Anthony Appiah (1993, 817) defines thick translation as a translation that not only seeks to render a text from one language into another but also aims to provide a rich cultural and contextual understanding of the text for the reader[7]. This approach goes beyond mere linguistic translation and includes detailed annotations, commentary, and explanations that help the reader grasp the cultural, historical, and social contexts of the original text. Meanwhile, Genette's theory of paratexts is a significant contribution to literary theory, particularly in understanding how texts are framed and contextualized. Genette Gérard (1997, 5) divides paratexts into two main categories: 1) peritexts-elements that are physically attached to the main text within the same volume which usually are titles, prefaces, forewords, footnotes, chapter titles, epilogues, and cover designs and 2) epitexts-elements related to the text but are not physically attached to it. Epitexts usually include author interviews, letters, diaries, promotional material, reviews, and critical essays.

Based on thick translation theory and Genette's categorization of paratexts, the following four categories of notes are defined: notes category Nº 1 are those that trace the source (these kinds of notes can include an account of the source and the source of information); notes category Nº 2 are those that explain the background information; notes category Nº 3 are those that integrate the relevant information (these kinds of notes can also summarize extra-textual knowledge, other people's comments, related research, etc.); and notes category Nº 4 are those that serve as commentary or criticism. The categorization statistics are shown in Figure 4-5.

Figure 4 illustrates that there are significant fluctuations in the number of footnotes across chapters, indicating varying levels of commentary or explanation required depending on the chapter's content. Notable peaks are observed around chapters 10, 20, 40, 50, 60, 80, and 100, suggesting that these chapters may contain more complex or culturally significant content requiring additional footnotes.

(4) notes that serve as commentary or criticism.

7 This concept of thick translation is influenced by the anthropological idea of thick description, which involves an in-depth explanation of cultural phenomena.

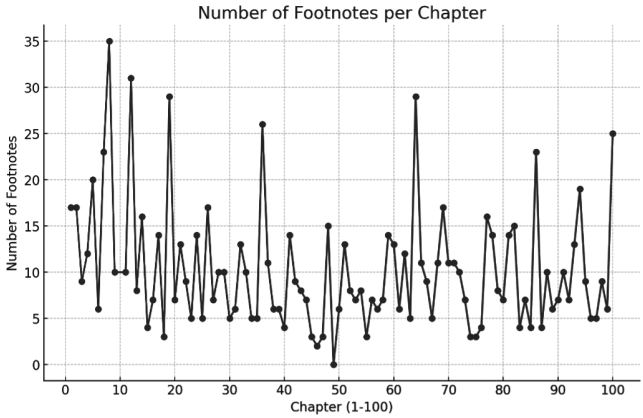


Fig. 4. Number of Footnotes per Chapter in Yu's Translation of The Journey to the West.

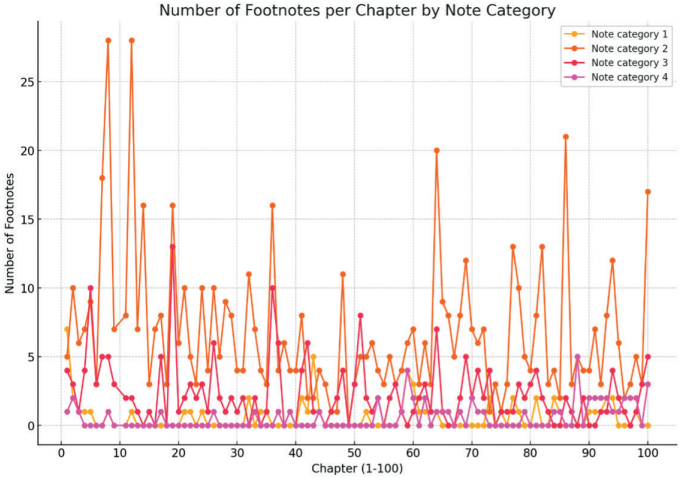


Fig. 5. Footnote categorization in Yu's translation of The Journey to the West.

Figure 5 shows that footnotes category N° 2, *i.e.* notes that explain the background information accounts for over 50 percent of all the footnotes in

Yu's translation, followed by category 3, 1 and 4. Several spikes indicate a higher density of notes category no.2 in certain chapters, particularly in the early chapters (around chapters 1-10) and some later chapters (around chapters 50 and 100). This indicates a significant emphasis on providing context and background information to aid reader understanding. Footnotes category Nº 3, which integrate relevant information, are the second most common, suggesting a strong focus on relating the text with external knowledge and research. Footnotes category Nº 1, *i.e.* notes tracing the source, are the third most common, showing a moderate emphasis on sourcing and referencing original material. And footnotes category Nº 4, *i.e.* commentary or criticism, are the least common, indicating that while there is some evaluative or critical content, it is not the primary focus. Although footnotes category 1 and 3 show much lower frequency compared to Note Category 1, they have consistent appearances across various chapters.

5.2. Findings and discussions

I further use topic modeling for each type of notes in Yu's translation to see if there are differences among these four types. Figures 6-8 show the results.

To start with, figure 6 reveals the topic modeling result of footnotes category Nº 1 which predominantly explore the religious, philosophical, and historical origins of the fiction. Among the seven topics, Topic 1 appears to cover a broad exploration of Daoist, Buddhist, and alchemical influences, which are central to understanding the source material for *The Journey to the West*. The other topics provide insights into more specific areas such as historical references (*e.g.* officials or abbots), discussions of temporal and regional elements, and the tracing of allegorical or symbolic elements (*e.g.* "baopuzi" is a significant Daoist text and "daxue" refers to *The Great Learning*, a classical Confucian text).

Topic	Weight	TopicMembers
Topic1	8.9601	chinese - chapter - daoist - text - buddhism - xiyouji - ele...
Topic2	1.7514	real - tianzun - wood - baopuzi - city - shidetang - world ...
Topic3	1.7368	officials - autumn - abbot - scholarship - good - underw...
Topic4	1.7362	hour - days - argued - region - spirit - huanden - andrew...
Topic5	1.722	discourse - assembly - preserved - imagined - stages - d...
Topic6	1.7015	gold - pilgrimage - antiquity - disciple - attained - years - ...
Topic7	1.7007	wisdom - symbolic - daxue - pearl - guardians - allegory ...

Fig. 6. Topic modeling of footnotes category Nº 1 in Yu's translation of The Journey to the West.

Secondly, figure 7 reveals the topic modeling result for footnotes that explain background information (category Nº 2). The outcome indicates that the most prominent background information notes relate to textual, cultural, and religious contexts, particularly with a heavy focus on Chinese literary and Daoist elements (as shown by "daoist" and "term"). Other significant themes include scholarly debate ("conversation"), mythological places ("zhejiang"), religious concepts, poetry, divine law ("cyclical" and "universe"), and literary analysis. The distribution of weights suggests that while all these topics are relevant, the notes explaining cultural and religious background are particularly important for understanding the translation.

Topic	Weight	TopicMembers
Topic1	6.2966	chinese - chapter - daoist - term - poem - line - text - bu...
Topic2	0.7887	student - daxue - argued - tianshi - conversation - fair - k...
Topic3	0.787	place illusion triple pool worthy emperors saliva...
Topic4	0.7632	selection - temple - study - cyclical - universe - asked - lu...
Topic5	0.733	poetic - fish basket - zhouyi - dark - relevant - zhejiang - ...
Topic6	0.7109	divine - law - submission - festival - realize - traditions - ...
Topic7	0.6957	fsz - emphasis - rhyme - reprint - theories - bamboo - pa...

Fig. 7. Topic modeling of footnotes category Nº 2 in Yu's translation of The Journey to the West.

Moreover, figure 8 shows the topic modeling result of footnotes category Nº 3. As these notes are those that integrate relevant information, this result suggests a detailed and nuanced analysis of the text's themes and elements. The dominant topic (Topic 1) deals with key literary and cultural elements, while the other topics provide insights into various aspects like character development ("man" and "transformation"), historical context ("Han" and "Indian"), imperial themes ("imperial" and "rule"), as well as philosophical influences ("Zhuangzi" and "premodern").

Topic	Weight	TopicMembers
Topic1	10.5891	chapter - chinese - poem - body - term - palace - text - cl...
Topic2	2.3289	man - transformation - past - form - earliest - mensus - lijl...
Topic3	2.3273	vols - bamboo - han - novel's - zhao - indian - sui - peaco...
Topic4	2.3025	found king level phases discussion section water...
Topic5	2.2923	chu - pun - structural - spirit - wrote - time - one's - trans...
Topic6	2.2654	follow - imperial - story - rod - zhu - club - rule - apricot -...
Topic7	2.2607	lines - zhuangzi - premodern - front - texts - east - plan - ...

Fig. 8. Topic modeling of footnotes category Nº 3 in Yu's translation of The Journey to the West.

Last but not least, figure 9 shows the topic modeling result of footnotes category Nº 4, *i.e.* notes that serve as commentary or criticism. These notes encompass a variety of critical perspectives and analyses on different aspects, including the novel's philosophical meaning, narrative techniques, reception, imagery, and historical context. For instance, keywords in Topic 1 ("chapter", "chinese", "term", "line", "literally", "moon", "text" and "poem") suggest that the fiction might involve a detailed commentary on the language, structure, and literary devices used in specific chapters, perhaps focusing on interpretation and literal translation issues. Topic 2 suggests a focus on early influences, didactic elements, and possibly the tonal qualities in the text, criticizing or commenting on how these aspects are represented or interpreted. While Topic 5 implies a commentary on the depiction of nature, the fusion of various elements (such as "action" and "antiquity"), and possibly visual imagery within the narrative.

Topic	Weight	TopicMembers
Topic1	10.7389	chapter - chinese - term - line - literally - moon - text - p...
Topic2	2.4802	zwx - earliest - leaves - teaching - dyadic - tonal - emper...
Topic3	2.4088	famous - correlated - cycle - a.m - received - smith - iii - ...
Topic4	2.3101	meaning - souls - chen - xyj - sun - quotation - taking - b...
Topic5	2.2509	nature - action - antiquity - fusion - images - narrated - s...
Topic6	2.2295	flower - episode - west - narration - textual - opening - q...
Topic7	2.2135	chapters - estrade - dwar - process - quoted - patriarch's ...

Fig. 9. Topic modeling of footnotes category Nº 4 in Yu's translation of The Journey to the West.

6. Conclusion

This study tries to explore the intricate knowledge networks embedded within Anthony Christopher Yu's translation of *The Journey to the West* through a digital humanities approach. By employing topic modeling and collocation analysis, the research has uncovered the profound interconnections between Confucianism, Taoism, and Buddhism within the footnotes of Yu's translation. These findings highlight not only the translation's dedication to preserving the philosophical and religious dimensions of the original text but also the depth of cultural and historical knowledge interwoven into the narrative.

The categorization of footnotes reveals a significant emphasis on providing contextual background and integrating external knowledge, demonstrating Yu's commitment to offering a rich, multi-layered reading experience. This approach aligns with the principles of thick translation, ensuring that

readers can fully appreciate the cultural and intellectual heritage that *The Journey to the West* represents.

Moreover, the detailed analysis of the footnotes shows how Yu's translation serves as a bridge between Eastern and Western literary traditions, making the text accessible to a global audience while maintaining its original philosophical and cultural essence.

I hope the concepts and methods presented in this paper can lay the groundwork for demonstrating the potential of computational tools in literary analysis and exploring various classical texts to deepen our understanding of translation as a vehicle for cross-cultural exchange and knowledge preservation.

This study is supported by Fundamental Research Funds for Central Universities-Research on the English Translation and Dissemination of Chinese Historical Books from the Perspective of Digital Humanities (Grant nº ND2024013).

References

Appiah, Kwame Anthony. 1993. "Thick Translation." *Callao* 64: 801–819.

Du Bois, Thomas. 2007. "Reviewed work: State and Religion in China: Historical and Textual Perspectives by Anthony C. Yu." *China Review International* 1(Spring): 292–296.

Gaustad, Blaine. 2007. "Review on State and Religion in China: Historical and Textual Perspectives by Anthony C. Yu." *The Journal of Asian Studies* 264: 139.

Genette, Gérard. 1997. *Paratexts: Thresholds of Interpretation*. Translated by Lewin, Jane E., Cambridge: Cambridge University Press.

Giulianelli, Mario, Tredici, Marco Del, & Fernández, Raquel. 2020. "Analysing lexical semantic change with contextualised word representations". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Association for Computational Linguistics. doi.org/10.18653/v1/2020.acl-main.365

He, Min. 2011. "A survey on the English translations of Qing fictions and their characteristics." *Journal of Xi'an International Studies University* 19(4): 109–112.

Kozlowski, Austin C., Taddy, Matt, & Evans, James A. 2019. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 443: 905–949.

Lansdall-Welfare, Thomas *et al.* 2017. "Content analysis of 150 years of British periodicals." In *Proceedings of the National Academy of Sciences of the United States of America*, E457-E465. doi.org/10.1073/pnas.160638011

Michel, Jean-Baptiste *et al.* 2011. "Quantitative analysis of culture using millions of digitized books." *Science* 601: 176–182.

Newberry, Mitchell G. *et al.* 2017. "Detecting evolutionary forces in language change." *Nature* 551: 223–226.

Newberry, Mitchell G., & Plotkin, Joshua B. 2022. "Measuring frequency-dependent selection in culture." *Nature Human Behaviour* 68: 1048–1055.

Wang, Richard G., & Xu, Dongfeng. 2016. "Three decades' reworking on the Monk: The Monkey, and the fiction of allegory." *The Journal of Religion* 381: 102,108–109.

Zypf ranks of modifying collocations as a criterion for mining bilingual terminological equivalents

Serhii Fokin

Taras Shevchenko National University of Kyiv, Ukraine
sergiyboryvovych@ukr.net

Abstract. Mining bilingual terminological equivalents is a challenging task due to numerous lacunas in specialized domains, given constantly emerging neologisms. While corpora prove to be powerful sources with large amounts of data, most of them are monolingual. Therefore, to extract potential equivalents, it is crucial to resort to large comparable corpora and to rely on formal properties of terms, among which the frequency rank correlation of their modifying collocations seems a promising criterion to rely on. At the same time, besides the meaning of the term and its frequency rank in different corpora, the degree of semantic correlation among them may depend on other factors, such as their semantic class or their culturally marked character. To verify the rank/semantic correlation, we calculate Pearson's and Spearman's correlation for 16 equivalent bigram sets in English and Ukrainian. Pearson's and Spearman's correlation of equivalent bigram ranks proves strong for terms denoting artifacts, location, event and cognition.

Résumé. *L'extraction d'équivalents terminologiques bilingues est une tâche difficile en raison de nombreuses lacunes dans des domaines spécialisés, compte tenu de l'émergence constante de néologismes. Bien que les corpus s'avèrent être des sources puissantes avec de grandes quantités de données, la plupart d'entre eux sont monolingues. Par conséquent, pour extraire des équivalents potentiels, il est crucial de recourir à de grands corpus comparables et de s'appuyer sur les propriétés formelles des termes, parmi lesquelles la corrélation fréquence-rang de leurs collocations modificatrices semble être un critère prometteur. Parallèlement, outre le sens du terme et son rang de fréquence dans les différents corpus, le degré de corrélation sémantique entre eux peut dépendre d'autres facteurs, tels que leur classe sémantique ou leur caractère culturellement marqué. Pour vérifier la corrélation rang/sémantique, nous calculons les corréla-*

tions de Pearson et de Spearman pour 16 ensembles de bigrammes équivalents en anglais et en ukrainien. Les corrélations de Pearson et de Spearman des rangs de bigrammes équivalents s'avèrent fortes pour les termes désignant des artefacts, locations, événements et cognition.

1. Introduction

Many terminological equivalents are not explicitly paired with their correlative terms in other languages since they do not appear as headwords in terminological databases, bilingual dictionaries, multilingual encyclopedias, parallel concordances or other sources. By reading comparable texts on the same subject in different languages, translators or lexicographers can find equivalence correlations between some domain-specific terms. Although there are various *ad hoc* intuitive solutions for this task, it is important to objectivize and generalize them from the methodological point of view.

Semantic analysis, based upon comparing definitions of two terms in different languages, may seem promising, but is prone to difficulties since definitions of a concept provided by various institutions may feature different relevant characteristics. Thus, semantic analysis does not appear to be the best choice for establishing equivalents. That is why an alternative methodology should be preferred. As E. Lozanova-Bogomilova remarks, "the cases of complete equivalence between the concepts in two languages, that is, the identity of content between two compared terms, result from the fact that they occupy the same position in both systems" (2006, 681). This statement, crucial for translational terminology, is not as easy to implement in practice as it may seem.

Although the expression *position in the system* is rather metaphorical, its intuitive interpretation as the "place" a concept occupies in the network of its semantic relations with other concepts of the domain in question looks convincing. We are fully aware that there are cases of multiple candidates for a given relation, *i.e*., *the ribosome* is part of a *cell*, but it is also part of *mitochondria.* The degree of its hierarchical relation granularity to a parent term may vary and be theoretically unlimited.

Comparing a schema of a biological cell in two different languages, we are likely to find dozens of equivalents -in this case, meronyms- by simply matching their disposition in the respective images (*e.g.* for the English-French

language pair: *nucleus – noyau*; *cell membrane – membrane cellular*; *cytoplasm – cytoplasme*; *mitochondria – mitochondrie*s*; endoplasmic reticulum – réticulum endoplasmique*, etc.).

Many concepts are not provided with illustrated schemas, images, tables or graphs. In such cases, lexical databases like *WordNet*, developed by G. Miller at Princeton University (USA) and built upon semantic relationships among lexemes in English and some other languages, look like the primarily recommended source for extracting the relations network of a particular concept. However, WordNet-annotated relations rely mainly on common vocabulary rather than domain terms and are still unavailable for many lexicographically low-resource languages. Hence, another problem may arise with narrow-domain neologisms, even for the rich-resource languages not listed in the lexicographic sources. Subsequently, alternative methods for extracting hyponyms, meronyms or other relations of a given concept, such as large-volume corpora and other specific data sources, become of priority in the lack of more suitable data.

Different researchers have repeatedly experimented with correlative multilingual texts in order to extract equivalent fragments from them. For example, Shao (2004) developed the principle of finding equivalents in Chinese and English correlated texts from different sources based on the analysis of transcription, as well as semantic and partial language information in the context. Stefanescu proposed an algorithm for extracting potential equivalent fragments in correlated texts given the presence of cognates and the relative length in Romanian and English (2012). Wołk, Reimund, and Marasek achieved successful results of equivalent selection by analyzing correlated articles in Polish and English from *Wikipedia* and *Euronews* news texts on the same topic (Wołk 2015).

In this study, we explore the possibility of finding translational equivalents among modifiers of equivalent terms in accordance with their frequency rank in large-volume corpora. Our objective is to evaluate the frequency distribution correlation between modifiers of terms of the military domain in English and Ukrainian and to show its application for mining terminological equivalents. This objective stems from the idea that terms correlated by semantic meaning are expected to possess similar syntagmatic links and have similar frequency ranks.

2. Theoretical background

2.1. Hierarchical relation as data mining tool

Hierarchical relations (generic-specific, part-whole, cause-consequence, etc.), expressed through thesaurus functions, feature the "term surroundings" in the system outlining its unique "portrait" and providing reliable ground for its comparisons with other terms, even in other languages. This comparison makes it possible to judge whether they are equivalent. Whereas some narrow-domain or neological terms may be absent in bilingual sources, their hierarchically parental terms (holonyms or hyperonyms), as much more frequently used lexemes, have a much higher likelihood of being strongly correlated in bilingual termbases or dictionaries. For example, the article *magnetic core* as a separate entry in *Wikipedia* is currently linked to 17 languages, whereas the parent terms *inductor* and *electric device* in which a *core* is a fundamental component, appear connected to articles in 99 languages. Thus, the chances of mining holonym equivalent in the observed case are more than five times higher than for the specific term in question (similar approaches to searching equivalents through hyperonymic or holonymic relations are described in Fokin (Фокін 2014, 29) and Anwarus Salam (2018, 81).

Texts concerning the parent term in question are likely to contain mentions of meronyms or hyponyms, as the articles about *inductors* should include mentions of its components. Thus, we could schematize the method of searching for a particular equivalent in the following way:

> Term in SL -> find its parent term -> translate the parent term into the TL -> find child terms for the translated term in the TL -> find among them candidates for the original term translation

This schema may be suitable for the translation of a specific term, although in the case of collecting multiple equivalent pairs for a domain glossary, a somewhat simpler algorithm is required:

> Term in SL -> find its translation in the TL -> find child terms for both terms in SL and TL

Logically, once established the equivalence between two parent terms, the chances of mining translation pairs among their child terms are high. This observation posits several methodological questions: 1) finding formal criteria and tools for extracting lists of potential child terms; 2) finding the most suitable and effective data source; 3) finding a methodology for establishing

equivalent relations among the members of the extracted child terms sets; 4) automating steps 2 and 3.

2.2. Exploring formal markers of semantic relations

Gromann *et al.*, in line with the idea of equivalency, develop a fine-tuned approach for modeling the terminological system by creating an algorithm focused on extracting terminological equivalents relying on their relations to other terms in a *tbx* dataset with the aid of additional neural processing (2023, 220). In the absence of such sets, textual data managed by corpus tools come into play as an alternative and valuable data source.

Semantic relations are implicit data, and most modern corpora are not semantically annotated. However, these relations can find projections in some formal characteristics, particularly, syntagmatic properties, mineable through corpus queries. Hassani mentions the possibility of extracting semantic data based on distribution (2011, 359). It is worth noticing other methods, such as the one explored in the paper by Vintar, Špela & Grcic, Larisa & Martinc, Matej & Pollak, Senja & Stepišnik, Uroš (2020) where equivalent adjectives in comparable texts (Croatian, English) are used to establish semantic relations.

To build successful corpus queries aimed at extracting semantic relations through syntagmatic ones, it is essential to rely on formal syntactic markers. At the exploratory stage, we expected some significant results for mining particular semantic relations (*e.g.*, meronyms) from such lexical markers as the verbs *include*/*comprise*/*consist*/*involve* and other synonyms, but in most cases, the search outcomes were scarce, probably due to relatively low usage of these verbs in comparison to other markers, thus, we explored other candidate markers.

One of the promising syntagmatic features, as our preliminary observations for Ukrainian suggest, that consistently yields much richer results and can be used as a marker of semantic relations of a term is the genitive case (in its formal and functional meaning). Most commonly, the following functions are characteristic of the genitive case: subjective, objective, properly genitive and partitive (Jespersen, 1924, 180). Since the genitive case does not formally exist in modern English, it is necessary to resort to other syntactic patterns. Among them, one of the richest syntactic relations from the standpoint of semantic connections, closely associated with the genitive case and serving as a source for searching child terms, are modifiers. Modifiers may hold diverse

types of relation with their head words: kind, quality, location and quantity (Rijkhoff, 2014, 132). Nikonova illustrates by means of the transformational analysis of compound lexemes, are subject-predicate, place, means of movement, material, and purpose of usage (Ніконова, 2020, 44). Therefore, possible relations held by modifiers with their heads include a much broader range than hierarchical (child-parent) connexions, but nonetheless useful in the search for equivalents.

While hierarchical relations seem to be comprehensively described in literature and universally accepted as such in ISO standards, such as ISO 1087, ISO 704, the rest type of relations are never presented in an exhaustive form. As per ISO 704, the associative relations include: *raw material – product, action – equipment/tool, quantity – unit, material – property, material – state, matter/substance – property, concrete item – material, concrete item – shape, action – target, action – place/location, action – actor, time sequence, cause – effect, container – contained, activity – tool, steps of a cycle, producer – product, duration – measurement device, profession – typical tool, object – associated tool, organization* – associated building, although this list is not exhaustive (ISO 704, 2002). Due to the broad spectrum of approaches to semantic classifications and space limitation, at this stage, it is impossible to attempt to elaborate a single relation grid for all cases. At the same time, it is clear that, regardless of the syntactic dependency, underlying modifier semantic relation to the head is potentially helpful to match equivalents stemming from modifiers.

The most common syntactic patterns for modifiers of a term are:

Adjective + head term (in English and Ukrainian);

Noun + head term (in English);

Head term + genitive case (in Ukrainian);

Head term + preposition “of” + Noun.

Aware that other modifying structures involving entire nominal syntagms (not only nouns) are possible, for the sake of simplification of this first stage of the survey, we limit to one token the extension of modifiers being searched. At the same time, the head term can be a modifier as well, which opens the possibility of mining additional equivalent bigrams. Thus, we intend to run corpus queries as per the following patterns:

Adjective + head term (in English and Ukrainian, e.g.: atomic weapons – ядерна зброя);

Noun + head term (in English, e.g.: cannon artillery).

Since the head term may also be a modifier, we also include the following patterns:

Head term + Noun (in English, e.g.: mine detection);

Adjective derived from the head term + Noun (in Ukrainian: e.g.: мінне поле);

Noun + head term in genitive case (in Ukrainian: командир роти).

2.3. Hypothesis of rank and semantic correlation

The most commonly known rank properties in the context of language are reflected in the Zypf law. The Zypf law, although in many cases raises debates, conveys an intuitively undeniable harmony of big data properties (Corominas-Murtra & Sole, 2010). Semple, Ferrer-I-Cancho & Gustison described a range of facts proving that the Zipf's rank-frequency law and its sister laws apply across molecular, organismal, and ecological levels, describing patterns in DNA sequences, RNA structures, gene expression, animal communication, species distributions, and even epidemiological data (2022, 4).

Many attempts were made to explore Zypf's law and its possible practical applications in linguistics and translation. Mohammadi proposes a methodology for identifying the degree of parallel relationship between corpora in two languages and finds that it is possible to use Zipf's index to identify the parallel documents (2016, 25). In the domain of copyright and marketing, it was discovered that web content with the law Zipf index is more likely to appear on the top of relevant results. These and other observations allow formulating recommendations on how to improve the text for a better website promotion (Korsun 2024). Thus, the consequences of Zypf law find their applications in practical tasks and can extend far beyond the linguistic field.

Among many potentially significant aspects of this law, the relevant questions for the present study are:

1. Does Zypf's law hold for all the languages?
2. Does Zypf's rank correlate to the semantic meaning?
3. Does Zypf's law hold for word combinations?

Bentz & Ferrer-i-Cancho hold an experiment to put in evidence that shorter words tend to be more frequent (2016, 1-2). The exploration of the mentioned direction can shed light upon important properties of languages, particularly, the vocabulary. Bolea *et al.* found that Zipf´s law holds better for words with certain etymologies (2024, 146). Provided that many terms are of Greko-Latin

origin, the terminological internationalization may have a strong effect on Zypf's properties of terms, exploitable in the mining of equivalences. Piantadosi emphasizes the relationship between Zipf's law and extralinguistic factors (psycholinguistics, cognition and human memory), as well as stresses the strong correlation between the frequency and semantics (2014, 1128). Claude and Pagel examine the semantic and rank correlation of Swadesh list[1] in 17 languages of 6 different families and found a strong correlation between equivalent Swadesh word frequencies and their meanings (2011, 1107).

While most researches focus on the rank of isolated words, it is also important to explore the Zipf properties of word combinations since from the beginning of the corpus linguistics, the concept of meaningful units was extended from words to *n*-grams. An additional advantage of resorting to *n*-grams rather than isolated tokens is that the language speakers tend to build their texts from "semi-preconstructed" phrases rather than words (Sinclair, 1991, 110). Le, Sicilia-Garcia, Ming, and Smith confirm the Zipf's law validity for predicting the likelihood of the occurrence of a phrase according to the phrase frequency distribution; it turned out that the validity of the Zipf law proved even more reliable for bigrams and trigrams in Mandarin (2004, p. 5). It seems appropriate, then, to employ this property in the data mining.

Assuming Zipf's law validity for phrases, it is logical to take a step forward to infer its validity for mining terminological bigram equivalents from comparable corpora. In other words, the question we furtherly address is whether it is possible to find bilingual equivalents in both lists based upon the frequency rank of compound terms, even though they do not necessarily denote language universals like in the Swadesh list.

3. Methodology

3.1. Observation

To subject to prove the hypothesis of semantic and frequency correlation, let us generate frequency lists of bigrams consisting of adjectival collocates of the noun *interview*/*інтерв'ю* preceded by adjectives in English and Ukrainian, summarized in Table 1 from the corpora enWebTenTen21 (SketchEngine, 2021) and ukTenTen22 (SketchEngine, 2022).

1 The Swadesh list is a core set of hypothetically universally used words proposed by Morris Swadesh in 1952 to compare and study the relationships between different languages. The list has been adjusted several times since then.

	Adjectival phrases with *interview* in English	Relative frequency	Adjectival phrases with *інтерв'ю* in Ukrainian	Relative frequency	Adjectival phrases with *entrevista* in Spanish	Relative frequency
1	recent interview	1.07	ексклюзивне інтерв'ю	1.62	entrevista personal	0.84
2	exclusive interview	0.94	телефонне інтерв'ю	0.79	entrevista exclusivo	0.73
3	in-depth interview	0.54	особисте інтерв'ю	0.42	entrevista telefónico	0.47
4	video interview	0.52	велике інтерв'ю	0.4	entrevista completo	0.41
5	semi-structured interview	0.42	нещодавнє інтерв'ю	0.39	entrevista reciente	0.29
6	personal interview	0.39	повне інтерв'ю	0.38	entrevista radial	0.24
7	first interview	0.33	перше інтерв'ю	0.37	entrevista televisivo	0.22
8	new interview	0.31	недавнє інтерв'ю	0.34	entrevista laboral	0.22
9	full interview	0.3	формалізоване інтерв'ю	0.3	entrevista individual	0.19
10	great interview	0.24	відвертий інтерв'ю	0.29	entrevista previo	0.16

Tab. 1. The top frequent adjectival collocations of the noun interview, entrevista and інтерв'ю in English, Spanish and Ukrainian.

Table 1 demonstrates that a significant correlation holds not only for the Swadish list, but also for the beginning of the raw word list generated in corpora tools. Although this correlation is not 100 percent precise, it provides grounds to expect that the semantically related terminological bigrams are expected to appear in near positions in comparable frequency lists.

Besides illustrating evident semantic correlations between different frequency lists for adjectival modifiers, it is clear that local data also have their traits in the list and interfere with the high correlation of the list, which is

better visible through Table 2 with the adjectival modifiers of *deporte*/*sport* in Spanish and French. The corpora used for data mining are:

- spWebTenTen23, "Spanish Web 2023", over 28,500,000,000 tokens (SketchEngine, 2023);
- frWebTenTen23, "French Web 2023", over 7,500,000,000 tokens (SketchEngine, 2023).

It is fully understandable that in the respective corpus of Spanish the *French sport* does not have a high usage rate and, conversely, *Cuban sport* does not appear among the top phrases of the French corpus. Thus, the possibilities of mining equivalents are expected to be ontologically specific.

	Adjectival phrases with *deporte* in Spanish	**Relative frequency**	**Adjectival phrases with *sport* in French**	**Relative frequency**
1	deporte acuático	1.28	sport nautique	1.39
2	deporte extremo	0.59	sport automobile	1.01
3	deporte nacional	0.5	sport collectif	0.67
4	deporte favorito	0.49	sport extrême	0.51
5	deporte náutico	0.46	sport mécanique	0.44
6	deporte electrónico	0.42	sport favori	0.34
7	deporte profesional	0.33	sport national	0.32
8	deporte femenino	0.31	sport professionnel	0.27
9	deporte español	0.25	sport féminin	0.24
10	deporte olímpico	0.25	sport aquatique	0.23
11	deporte individual	0.23	sport équestre	0.22
12	deporte colectivo	0.18	sport français	0.19
13	deporte escolar	0.16	sport individuel	0.18
14	deporte tradicional	0.14	sport scolaire	0.15
15	deporte local	0.12	sport canin	0.15
16	deporte mundial	0.12	sport olympique	0.14
17	deporte cubano	0.11	sport amateur	0.13
18	deporte paralímpico	0.11	sport électronique	0.13
19	deporte universitario	0.11	sport populaire	0.12

Tab. 2. Top frequent adjectival collocations of the noun deporte and sport in Spanish and French.

The results from Table 3 show that the ontologically dependent words are to be expected to appear in different positions in the frequency lists. Conversely, highly conventionalized domains are expected to show significantly correlated frequency lists from the semantic point of view.

To illustrate the large comparable corpora potentiality for data mining of term modifiers, we generate the frequency lists of nominal modifiers of the term *DNA* (*ДНК*/*AND*) in Ukrainian and Spanish that represent a highly conventionalized domain. For the sake of search simplification, we used Ukrainian and Spanish combinations since for both languages the prominent types of modifiers are formed according to similar syntactic patterns: modifier postponed in the genitive case (Ukrainian) or after the preposition *de* in Spanish:

The frequency ranks of semantically correlated modifiers of *ДНК*/*AND* are illustrated in Table 3.

	Bigram in Spanish	**Bigram in Ukrainian**	**Spanish bigram frequency**	**Ukrainian bigram rank**
1	аналіз ДНК	análisis de ADN	1	5
2	ланцюг ДНК	cadena de ADN	11	5
3	тест ДНК	test de ADN	12	20
4	фрагмент ДНК	fragmento de ADN	8	6
5	спіраль ДНК	hélice de ADN	14	27
6	нитка ДНК	hebra ADN	16	12
7	код ДНК	código de ADN	31	37
8	зразок ДНК	Muestra de ADN	9	3
9	частина ДНК	parte de ADN	18	18
10	основа ДНК	base de ADN	22	23
11	кількість ДНК	cantidad de ADN	24	13
12	рівень ДНК	nivel de ADN	25	36
13	дані ДНК	dato de ADN	37	22
14	слід ДНК	rastro de ADN	38	32

Tab. 3. Frequency lists correlation of the modifiers of the term DNA in Spanish and Ukrainian.

Statistical analysis confirmed that indeed there is a significant correlation between the term rank and its meaning: Pearson's correlation for the given rank correlation is strong reaching 0.701268, and Spearman's correlation is 0.782348.

In the following sections, we subject to prove empirical data to judge whether is sort of correlation is accidental or regular.

3.2. Data selection

3.2.1. *Ad hoc* comparable corpora

Since the objective of this study is to explore the potentiality of comparable texts of a given subject domain, initially, we attempted to create domain-specific two comparable corpora in English and Ukrainian using the automated tools of corpus generation developed by *SketchEngine*.

Once observed the first screens of frequency lists in both corpora, we found scarce mutual equivalents. Thus, within the total volume used up to 1 million tokens, the results of data mining did not yield any significant results. At the same time, similar queries run on large-volume corpora of the -*TenTen* family on the *SketchEngine* platform summing billions or dozens of billions of tokens output noticeably richer results.

Since the results obtained from *ad hoc* created comparable corpora turned out to be worse than in previous observations from large general corpora, we decided to change the strategy and stake on the volume of the corpus rather than on their subject specificity.

3.2.2. Large volume comparable corpora

Since the -*TenTen* corpora are of the same order of volume and taken from the same source (largely auto-downloaded from open sources including press and social media), similar volume order and precedence of the data makes it possible to qualify them as comparable. The largest available corpora of English and Ukrainian are the following:

- enTenTen21 (English Web 2021, over 52,000,000,000 tokens) (Sketch Engine, 2021);
- ukTenTen22 (Ukrainian Web 2022, over 7,500,000,000 tokens) (SketchEngine, 2022).

3.2.3. Building corpus queries

To run the queries, we use the CQL (*Corpus Query Language*) (Jakubíček, 2010) in two corpora: *English Web TenTen 2021* and *Ukrainian Web 2022*. The queries and the patterns they target are specified below:

Adjective + head term in English:

[tag="JJ"][lemma="<term base form>"];

Adjective + head term in Ukrainian:

[tag="A.*"][lemma="<term base form>"];

Noun + head term (in English, e.g.: *cannon artillery*):

[tag="NN"][lemma="<term base form>"];

Head term + Noun (in English, e.g.: *mine detection*):

[lemma="<term base form>"][tag="NN"];

Adjective derived from the head term + Noun (in Ukrainian):

[lemma="< adjective derived from the term >"] [tag="NN"];

Noun + head term in genitive case (in Ukrainian):

[tag="N.*"][lemma="< term in genitive case >"];

Head term + noun in genitive case (in Ukrainian):

[tag="N.*"][lemma="< term in genitive case >"].

3.3. Selecting terms for experiment

As shown in Section 3.1, the degree of equivalency of comparable corpora is expected to be dependent on the ontological similarities between the domains. Thus, highly internationalized areas, such as mathematics, and health care, are likely to yield more reliable results of corpus queries. But these domains usually pose fewer challenges either of translation or creation of bilingual dictionaries, while other, ontologically less similar domains often conform to a gap in bilingual lexicography. One of these challenging areas is the military domain due to the objective practical needs of compiling bilingual terminological glossaries in view of the full-scale war in Ukraine. Furthermore, the military domain consists of both highly internationalized (such as tactical medicine) and, on the contrary, significantly local subfields (strategy, military degrees, some kinds of weapons). To cover different semantic groups, we resorted to

WordNet Based Categorization Dictionary for nouns: *Act, Animal, Artifact, Attribute, Body, Cognition, Communication, Event, Feeling, Food, Group, Location, Motive, Object, Person, Phenomenon, Plant, Possession, Process, Quantity, Relation, Shape, State, Substance, Time* (Provalis Research, 2025). At the current stage, we focused the study exclusively on nominal terms.

Among the mentioned above, we selected the following domain-relevant categories along with two individual representatives of each one: *Act (offensive, demining), Artifact (howitzer, missile), Cognition (reconnaissance, strategy), Event (assault, siege), Group (army, regiment), Location (garrison, front), Person (colonel, lieutenant), Substance (gunpowder, TNT).*

Taking into account possible debates concerning the differences between *Phenomenon, Process, Act* and *Event*, let us consider them as a common category only one of them for the concepts of "offensive" and attack and similar. Areas like *Feeling*, *Communication*, *Plant*, *Animal*, *Food* and others are not domain-specific categories. Semantic diversity should allow us to evaluate (*i.e*., to confirm or reject) whether statistical significance applies to all of them or to some specific items.

4. Results and Discussion

In the current section, we delineate the outcomes obtained: frequency lists of bigrams containing an equivalent term with a modifier. The null hypothesis is that there is no correlation between the frequency ranks of the equivalent bigrams in both languages, whereas, under the alternative hypothesis, we assume that there is a significant correlation between the ranks of an equivalent bigrams containing modifiers of equivalent term in English and Ukrainian, a property that can be furtherly used to mine equivalents (*i.e*., search for translations) and to compile bilingual dictionaries. Extremely different ranks may reveal significant ontological differences or partial equivalence.

For the sake of the experiment, we limited the measurement of correlations to bigram level and ruled out equivalent *n*-grams of any order higher than two. For example, some concepts turn out to be very polysemic nouns in one language and are often used with additional collocates, *i.e*., are expressed by means of two or more words in one of the languages: *штаб – general staff.* We also excluded unigrams, such as isolated cases of hyphenated appositions: *howitzer-cannon – гаубиця-гармата.* In some equivalence pairs, the terms were expressed through word blending comprising both the head and a modifier (*артполк*/*artillery regiment*), which were also excluded.

We also discarded one-to-many (or partial) equivalents (*e.g.*, *trench – окоп, траншея*), as they would bias the result and terms denoting *realia* as prone to show considerably different frequency usage in different language communities, for example *ракетний полк – missile regiment, Gunpowder plot – Порохова змова*.

To hold the experiment, we listed all bigram-to-bigram equivalents with one-word modifiers from the respective term entries of the *English-Ukrainian military dictionary.* In the next stage, we searched for the ranks for each bigram in the corpus-based frequency lists. Most of the bigrams contained in the dictionaries were present in the frequency lists.

To assess the correlation between the bigram ranks and their semantics, we computed Pearson's and Spearman's correlation coefficients between the ranks of semantically correlated bigram terms in English and Ukrainian. Since the hypothesis being proved stems from the correlation between the rank and the semantics, the rank (Spearman's) correlation should be periodized to either confirm or rule out the hypothesis.

We also manually extracted additional equivalents visible within the window of the first 100 lines of the frequency list to illustrate the possibilities of enriching the entry of a specialized bilingual dictionary.

The most significant results were achieved for seven out of sixteen terms as shown in Table 6. Among them, the highest correlations were observed for the terms denoting artifacts: *howitzer* and *missile*, whose data is specified and commented on below. Table 4 integrates the frequency ranks of equivalent bigrams with the term howitzer-гаубиця.

	Bigram in English	**Bigram in Ukrainian**	**Frequency rank of the bigram in English**	**Frequency rank of the bigram in Ukrainian**
1	self-propelled howitzer	самохідна гаубиця	1	1
2	howitzer artillery	гаубична артилерія	3	3
3	howitzer battalion	гаубичний дивізіон	2	2
4	howitzer fire	гаубичний вогонь	7	12
5	howitzer section	гаубична секція	13	31
6	field howitzer	польова гаубиця	4	3

7	artillery howitzer	артилерійська гаубиця	20	12
8	heavy howitzer	важка гаубиця	8	2
9	wheeled howitzer	колісна гаубиця	64	33
10	long-range howitzer	далекобійна гаубиця	39	10
		Pearson's correlation	**0.694055**	
		Spearman's correlation	**0.774879**	

Tab 4. Frequency rank of bigrams with modifier collocates of "howitzer".

Within the first 100 rows of the frequency lists, we can find additional equivalents that are currently not included in the dictionary, which are listed hereafter along with their ranks in parenthesis:

- *field howitzer* (3) – *польова гаубиця* (4) – *artillery howitzer* (12) – *артилерійська гаубиця* (20)
- *heavy howitzer* (2) – *важка гаубиця* (8)
- *wheeled howitzer* (39) – *колісна гаубиця* (33)
- *long-range howitzer* (39) – *далекобійна гаубиця* (10)
- *howitzer tube* (27) – *ствол гаубиці* (8)
- *howitzer carriage* (16) – *лафет гаубиці* (11)
- *howitzer gun* (8) – *навідник гаубиці* (12)
- *howitzer ammo* (33) – *боєкомплект гаубиці* (14)
- *howitzer turret* (48) – *башта гаубиці* (62)

Pearson's correlation between the additional equivalents is also strong 0.668012, whereas Spearman's correlation between additional equivalents is 0.613278. In the found bigrams, the modifiers hold both hierarchical and associative relation of different types with the head term.

The resulting list of equivalents with their ranks for the modifiers of *missile – ракета* are shown in Table 5.

	Bigram in English	Bigram in Ukrainian	Frequency rank of the bigram in English	Frequency rank of the bigram in Ukrainian
1	cruise missile	крилата ракета	1	2
2	ballistic missile	балістична ракета	1	1
3	anti-ship missile	протикорабельна ракета	2	6
4	anti-aircraft missile	зенітна ракета	3	9
5	anti-tank missile	протитанкова ракета	4	7
6	hypersonic	гіперзвукова ракета	8	15
7	enemy	ворожа раекта	12	5
8	nuclear missile	ядерна ракета	2	9
9	air missile	зенітна ракета	15	9
10	antiradar missile	протирадіолокаційна ракета	27	62
11	tactical missile	тактична ракета	24	24
12	guided missile	керована ракета	2	4
13	intercontinental missile	міжконтинентальна ракета	29	21
14	strategic missile	стратегічна ракета	17	25
15	single-stage missile	одноступенева ракета	290	276
16	missile battery	ракетна батарея	29	35
17	missile division	ракетна дивізія	92	26
18	missile site	ракетна позиція	25	100
19	missile system	ракетний копмлекс	2	2
20	missile battalion	ракетний дивізіон	92	22
21	missile corvette	ракетний корабель	69	18
22	missile cruiser	ракетний крейсер	14	9
23	missile transporter	ракетний транспортер	242	349
24	missile attack	ракетний удар	4	1
25	missile boat	ракетний катер	32	11
26	missile canister	ракетний контейнер	255	81
		Pearson's correlation	**0.910632**	
		Spearman's correlation	**0.82112**	

Table 5. Frequency rank of bigrams with modifier collocates of "missile".

The lists of equivalent bigrams in Table 4 and Table 5 illustrate diverse types of relations (hierarchical, associative) between the head terms and their modifiers. There were no additional equivalents since potential correlates were one-to-many equivalents *ракета – missile*/*rocket* or exceeded two-word combinations: *air deffence missile – зенітна ракета*.

	Head term in English	Head term in Ukrainan	Pearson's correlation	Spearman's correlation
1	army	армія	0.79213	0.567196
2	assault	штурм	0.621482	0.483333
3	colonel	полковник	-	-
4	demining	розмінування	-	-
5	front	фронт	0.639987	0.866025
6	garrison	гарнізон	0.682182	1
7	howitzer	гаубиця	0.694055	0.774879
8	gunpowder	порох	-	-
9	lieutenant	лейтенант	0.39736	0.5
10	missile	ракета	0.910632	0.82112
11	offensive	наступ	-0.20448	-0.18587
12	reconnaissance	розвідка	-0.12568	0.466667
13	regiment	полк	0.827017	0.2
14	siege	облога	0.773046	-0.1
15	strategy	стратегія	0.529469	0.512668
16	TNT	тринітротолуол	-	-

Tab 6. Integrated Pearson's and Spearman's correlation for all the term sets.

Table 6 shows that a significant correlation among bigram equivalents is possible, which holds among the following groups (in descendant correlation order): *Artifact* (*howitzer*, *missile*), *Location* (*garrison, front*), *Event* (*assault, siege*), *Cognition* (*reconnaissance, strategy*), *Group* (*army, regiment*). For the groups *Person (colonel, lieutenant), and Substance* (*gunpowder, TNT*), it was impossible to establish a correlation due to the lack of sufficient modifiers. The aforementioned results testify that there might be a powerful factor determining the possibility of correlation of equivalence and term frequency ranks, which is to be subject to the ANOVA analysis on a wider range of data. Even though, for the current dataset obtained, the p-value as per the ANOVA single-factor test is 0.005295, which is far lower than the conventionally accepted threshold of 0.05.

In six cases, Pearson's correlation was greater than Spearman's. Conversely, in four cases, Spearman's correlation proved stronger than Pearson's. A possible explanation of this fact is that the frequency factor might have a stronger impact on the semantic meaning of term modifiers than their frequency ranks, whereas subtle oscillations of ranks, particularly, of terms with close frequency rates, may drastically influence the results.

5. Conclusion

The idea of large quantitative corpora data providing the possibility of mining paradigmatic relations on the basis of the syntagmatic ones is widely explored, while the correlation between the frequency rank and semantics remains mainly unexamined in the domain of bilingual lexicography or the practice of translation.

The initial hypothesis relied on the existence of a semantic correlation between modifying collocates of a head term and their frequency ranks computed on large comparable corpora.

Equivalence based on frequency range proved valid for collocates of terms denoting *Artifacts, Location, Event* and *Cognition*. Among these groups, either Pearson's or Spearman's correlation was strong or very strong, or both values proved to be strong or very strong. We could not establish any correlation for the groups denoting *Person* and *Substance* due to the lack of sufficient modifiers. The frequency factor might have a stronger impact on the semantic meaning of term modifiers than their frequency ranks.

The possibility of establishing rank correlation among the bigram positions in the frequency lists and their meaning does not appear to be universal nor cover most terms due to several reasons:

- correlative lexemes may be ambivalent in one language, being monosemantic in another language;
- some equivalences are partial, *i.e*., characterized by one-to-many interlinguistic relations, which is why the rank of the equivalents may appear scattered among two or three synonyms;
- ontological differences of different language communities significantly influence the frequency rank of a particular term (some terms are rarely used in one community and show high frequency in another, which may be the case of *realia*);
- one-word prepositional or postpositional modifiers hold both hierarchical or associative relations with the head term.

Thus, the frequency rank cannot be used exclusively to automate the extraction of terminological equivalents as a reliable marker, although as an additional feature in a complex approach, it can substantially contribute to a complex automation algorithm.

Despite the lack of significant results for most groups, generating frequency lists of term modifiers allows enriching the existing bilingual specialized dictionaries with new equivalent pairs, which can be used to search for compound terms translation and in partial automation of bilingual glossary generation.

References

Anwarus Salam, Khan Md & Yamada, Setsuo & Tetsuro, Nishio. 2018. "Improve Example-Based Machine Translation Quality for Low-Resource Language Using Ontology." In *International Journal of Networked and Distributed Computing*, Vol. 5, No. 3 (July 2017), 176–191. 10.1007/978-3-319-64051-8_5

Bentz, Christian & Ferrer-i-Cancho, Ramon. 2016. "Zipf's law of abbreviation as a language universal." In Bentz, Christian, Jäger, Gerhard & Yanovich, Igor (eds.). *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tübingen, online publication system (pp. 1–4). 10.15496/publikation-10057

Bolea, Speranta & Pirnau, Mironela & Bejinariu, Silviu-Ioan & Apopei, Vasile & Gîfu, Daniela & Teodorescu, Horia-Nicolai. 2024. "Some Properties of Zipf's Law and Applications." In *Axioms*, 13, 146. 10.3390/axioms13030146

Claude, Andreea S. & Pagel, Mark. 2011. "How do we use language? Shared patterns in the frequency of word use across 17 world languages." *Phil. Trans*. R. Soc. B366. 1101–1107 10.1098/rstb.2010.0315

Corominas-Murtra, Bernat & Sole, Ricard. 2010. "Universality of Zipf's Law." *Physical review. E, Statistical, nonlinear, and soft matter physics*, 82, 011101-1–011102-9. 10.1103/PhysRevE.82.011102

English-Ukrainian Military Dictionary. Ukrop Austria. 2025. Accessed March 12. https://english-military-dictionary.org.ua

Gromann, Dagmar, Heinisch, Barbara, Wachowiak, Lennart and Lang, Christian. 2023. "A novel way of extracting full-fledged terminologies from multilingual texts." *Lexicographica*, vol. 39, no. 1, 209–223. 10.1515/lex-2023-0011

Hassani, Ghodrat. 2011. "Corpus-Based Evaluation Approach to Translation Improvement." *Meta. Revue des Traducteurs*, Vol. 56, No. 2, 351–373 Retrieved June 29. 10.7202/1006181ar

Hernández, Chantal. 2002. "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento." *Estudios de Lingüística del Español*. Universidad de Málaga. Vol. 18. http://elies.rediris.es/elies18/

ISO 12616. 2002. *Translation-oriented terminography*. Accessed March 12. https://ssu.elearning.unipd.it/pluginfile.php/425691/mod_resource/content/1/BS%20ISO%2012616-2002--%5b2019-01-23--10-13-39%20AM%5d_TranslationorientedTerminography.pdf

ISO-1087. 2019. *Terminology work*. *Vocabulary*. Accessed March 12. https://ediscipli nas.usp.br/pluginfile.php/%20312608/mod_resource/content/1/ISO_1087-1_2000_PDF_version_%28en_fr%29_CPDF.pdf

ISO 704. 2022. Terminology work — Principles and methods. Accessed March 12. https://www.academia.edu/35054077/Terminology_work_Principles_and_methods. 2000.

Jespersen, Otto. 1924 [1992]. *The Philosophy of Grammar*. The University of Chicago Press. Accessed March 12. http://ctlf.ens-lyon.fr/volumes/5317_eng_Jespersen_01_1924.pdf

Korsun, Oksana. 2024. "What is Zipf's law in Copirighting?" Accessed March 12. https://marketing.link/what-is-zipfs-law-in-copyrighting/

Le, Quan Ha & Sicilia-Garcia, E., Ming, Ji & Smith, F. 2004. "Extension of Zipf's Law to Words and Phrases." COLING. 10.3115/1072228.1072345

Lozanova Bogomilova, Elena. 2006. "La diversidad lingüística y cultural. Experiencia en la elaboración del banco de datos terminológico multilingüe BT-FRASJURE." In M. Cabré, R. Estopà y C. Tebé (eds.) *La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad: actas del IX Simposio Iberoamericano de Terminología*. RITMER04, 675–686.

Mohammadi, Mehdi. 2016. "Parallel Document Identification using Zipf's Law." Accessed March 12. https://comparable.limsi.fr/bucc2016/pdf/BUCC04.pdf

Piantadosi, Steven Thomas. 2014. "Zipf's word frequency law in natural language: A critical review and future directions." *Psychon Bull Rev* 21, 1112–1130. 10.3758/s13423-014-0585-6

Provalis Research. 2025. *WordNet-based categorization dictionary*. Accessed March 12. https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/wordnet-based-categorization-dictionary/

Rijkhoff, Jan. 2014. "Modification as a propositional act". In María de los Ángeles Gómez González, Francisco José Ruiz de Mendoza Ibáñez, Francisco Gonzálvez-García and Angela Downing (eds.). *Theory and Practice in Functional-Cognitive Space*, John Benjamins Publishing Company, 2014, (pp. 129–150). 10.1075/sfsl.68.06rij

Semple, Stuart, Ferrer-i-Cancho, Ramon, Gustison, Morgan. 2022. "Linguistic laws in biology." *Trends Ecol Evol*. 37(1), 53–66. 10.1016/j.tree.2021.08.012

Shao, Lei, & Ng, Hwee Tou. 2004. "Mining new word translations from comparable corpora." In Barry Smith & Christiane Fellbaum (eds.), *COLING 2004: Proceedings of the 20th International Conference on Computational*

Linguistics (pp. 618–624). COLING. Accessed March 12. https://aclanthology.org/C04-1089

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sketch Engine. *UkrainianWebTenTen 2022*. Accessed March 12. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fuktenten22_rft2

Sketch Engine. 2021. *English Web TenTen Corpus*. 2021. Accessed March 12. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen21_tt31

Sketch Engine. 2023. *French Web TenTen Corpus*. 2023. Accessed March 12. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Ffrtenten22_fl4

Jakubíček, Miloš, Kilgarriff, Adam, McCarthy, Diana, Rychlý, Pavel. 2010. "Fast Syntactic Searching in Very Large Corpora for Many Languages". *PACLIC*, 741–47.

Sketch Engine. 2023. *Spanish Web TenTen Corpus 2023*. Accessed March 12. https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Festenten23_fl6

Stefanescu, Dan. 2012. "Mining for Term Translations in Comparable Corpora." *Heuristics. 5th Workshop on Building and Using Comparable Corpora*. Accessed March 12. https://www.researchgate.net/publication/233808024_Mining_for_Term_Translations_in_Comparable_Corpora

Tan, Ah-Hwee, *et. al.* 2000. "Text Mining: The state of the art and the challenges. Text Mining: The state of the art and the challenges." In Ning Zhong, Lizhu Zhou (eds.) *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*. Berlin/Heidelberg/New York/Barcelona/Budapest/Hong Kong/London/Milan/Paris/Singapore/Tokyo: Springer (pp. 65–70). Accessed March 12. https://www.researchgate.net/publication/2471634_Text_Mining_The_state_of_the_art_and_the_challenges

Vintar, Štefan, Grčić Simeunović, Ljiljana, Martinc, Miha, Pollak, Štefan, & Stepišnik, Urban. 2020. "Mining Semantic Relations from Comparable Corpora through Intersections of Word Embeddings." In Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff (eds.). *Proceedings of the 13th Workshop on Building and Using Comparable Corpora* (pp. 29–34). European Language Resources Association.

Wołk, Krzysztof, Rejmund, Elżbieta, & Marasek, Krzysztof. 2015. "Harvesting Comparable Corpora and Mining Them for Equivalent Bilingual

Sentences Using Statistical Classification and Analogy-Based." In Floriana Esposito, Olivier Pivert, Stefano Ferilli, Zbigniew W. Raś, Mohand-Saïd Hacid (eds.). *22nd International Symposium on Methodologies for Intelligent Systems*. (pp. 443–441). 10.1007/978-3-319-25252-0_46

ГРАК, Генеральний регіонально анотований корпус української мови. 2023. Accessed February 12. https://parasol.vmguest.uni-jena.de/grac_crystal/#dashboard?corpname=grac14_full

Ніконова, Валентина, Никитченко, Катерина. 2020. *Курс зіставної лексикології англійської та української мов [A Course in Contrastive Lexicology of the English and Ukrainian Languages]*. Київ.: Вид. центр КНЛУ. Accessed March 12. http://rep.knlu.edu.ua/xmlui/bitstream/handle/787878787/1192/Nikonova7.pdf

Фокін, Сергій. 2014. *Комп'ютерна лексикографія і переклад: іспанська та українська мови [Computational lexicography and translation: Spanish and Ukrainian]*. Київ: КНУ імені Тараса Шевченка.

Microbrasserie, *brasserie agricole*, *écobrasserie* : les dimensions durable et néolocale de la terminologie brassicole en français et en italien

Lorenzo Devilla, Nicla Mercurio

Université de Sassari
ldevilla@uniss.it, nmercurio@uniss.it

Résumé. Par une approche socioterminologique (Gambier, 1987 ; Gaudin, 1993, 2003, 2005) et comparative français-italien, cette étude se propose d'explorer la terminologie des lieux de consommation et de production de bière, un nano-domaine qui manifeste la vocation territoriale et les pratiques durables de la filière brassicole artisanale. De fait, l'essor du secteur coïncide avec le succès du néolocalisme (Shortridge, 1996), un mouvement qui s'appuie sur le sens d'appartenance au lieu, l'identité régionale et l'attention envers l'environnement. Nous avons collecté des termes candidats à partir de sites Web de (micro) brasseries installées en France et en Italie, et examiné leur composition (substantifs, adjectifs, préfixes, suffixes) en tenant compte du contexte discursif et des définitions éventuelles fournies par les entreprises, des dictionnaires ou d'autres documents. Les résultats de l'analyse confirment, d'un côté, la portée globale des enjeux environnementaux et, de l'autre, la relation entre les terminologies et les sociétés.

Abstract. *This study uses a socioterminological (Gambier, 1987; Gaudin, 2003, 2005) and comparative French-Italian approach to explore the terminology of the places where beer is consumed and produced, a nano-domain that suggests the territorial vocation and sustainable practices of the craft brewing sector. Indeed, the rise of the sector coincides with the success of Neolocalism (Shortridge, 1996), a movement that draws on a sense of place, regional identity and care for the environment. Therefore, we have collected candidate terms from the Websites of (micro)breweries based in France and Italy, and we have examined their composition (nouns, adjectives, prefixes, suffixes), considering the discursive context and any definitions provided by the companies, the dictionaries or other documents. The results of the analysis confirm, on the one hand, the global relevance of environmental issues and, on the other, the relation between terminologies and societies.*

Lorenzo Devilla a rédigé les paragraphes 1, 3 et 5, alors que Nicla Mercurio a rédigé les paragraphes 2 et 4. Cette contribution s'inscrit dans le cadre du projet de recherche Plurilinguismo, patrimonio culturale e sviluppo sostenibile (« Plurilinguisme, patrimoine culturel et développement durable ») dirigé par Lorenzo Devilla (Université de Sassari) et financé par la Fondazione di Sardegna-Progetti di ricerca di base dipartimentali 2022 e 2023.

1. Introduction

Le changement climatique, la dégradation de l'environnement et toute question écologique impactent chaque aspect de la vie quotidienne, entrant inévitablement en conflit avec l'industrie de production. Ainsi, les entreprises de différents secteurs tentent de concilier leurs activités avec les valeurs sociétales et environnementales, dans le but également de susciter l'intérêt d'un public de plus en plus exigeant et confiant à l'égard des marques qui montrent une responsabilité sociétale (*Corporate Social Responsibility*, désormais CSR, Carroll, 1991).

L'importance d'adopter une approche soucieuse du développement durable est l'un des noyaux du mouvement néolocal (Cipollaro *et al.*, 2021) : le néolocalisme (*Neolocalism*) est défini comme la quête délibérée des traditions locales en réaction à la perte des liens ancestraux avec la communauté et la famille (Shortridge, 1996). Ce mouvement se base sur le sens d'appartenance au lieu (*Sense of Place*, Flack, 1997 ; Hede, Watne, 2013), sur l'identité régionale et sur des narrations plus larges impliquant l'environnement et les économies locales durables (Gatrell *et al.*, 2018). Puisque les entreprises embrassant la cause écologiste souhaitent montrer leur CSR, le néolocalisme peut aussi se manifester dans des stratégies publicitaires qui promeuvent cet engagement, afin de fidéliser les consommateur·trice·s et, le cas échéant, de les sensibiliser[1].

Comme nous l'avons mis en évidence dans une étude précédente (Devilla, Mercurio, 2024), le succès du néolocalisme se reflète dans l'essor de la bière artisanale (BoE 2022, 2023 ; *cf.* Bowen, Miller, 2023). Dans cette contribution, nous aborderons le domaine brassicole en nous penchant sur la terminologie

1 En revanche, si l'objectif est celui de dissimuler des actions qui vont dans le sens inverse, on parle de *greenwashing* ou écoblanchiment, surtout de la part des multinationales (*cf.* Biros, 2014, 55).

de lieux de production et de consommation de cette boisson, le terme *brasserie* étant enrichi d'adjectifs, de préfixes et de suffixes qui évoquent le caractère local et artisanal et l'attention envers la communauté et l'environnement. Par une approche socioterminologique (Gambier, 1987 ; Gaudin, 1993, 2003, 2005 ; Delavigne, Gaudin, 2022 ; Humbley, 2022) et comparative français-italien, nous retiendrons des termes candidats et les définitions éventuelles des lieux de production et de consommation de bière à partir des sites web de (micro)brasseries artisanales en activité en France et en Italie. Cela nous permettra de mettre en évidence la relation entre la filière brassicole et la durabilité, un sujet très débattu lors du *39th EBC Congress and 6th Brewers Forum* qui a eu lieu à Lille en mai 2024[2], et de confirmer le rôle de la terminologie dans la société, à travers l'analyse de termes candidats d'un nano-domaine qui fait émerger l'orientation des consommateur·trice·s contemporain·e·s face à la durabilité et à l'authenticité.

2. Néolocalisme, durabilité et bière artisanale

Le mouvement néolocal prône l'urgence d'adopter une approche durable et date déjà de la fin des années 1990, quand Shortridge (1996, 38) parle d'une «*deliberate seeking out of regional lore and local attachment by residents (new and old) as a delayed reaction to the destruction in modern America of traditional bonds to community and family*». Plus tard, Schnell et Reese (2003, 56) y ajoutent l'aspect conscient, puisque le néolocalisme est le résultat de l'engagement de personnes qui cultivent des liens locaux par choix et non par besoin, comme c'était le cas dans le passé. Il s'agit également d'une manière de fuir l'uniformisation générale et la production en série du consumérisme ; étant en mesure de séduire une certaine catégorie de consommateur·trice·s, le néolocalisme se révèle comme une véritable stratégie de promotion d'entreprise. Les activités qui adoptent des pratiques durables, telles que la réduction de l'empreinte carbone ou le recours aux fournisseurs locaux et aux circuits courts (Erhardt *et al.*, 2022), veulent évidemment mettre en évidence leur CSR.

2 À titre d'exemple, on mentionne les communications *Reducing Water Use and Waste Water in General. Challenges and Solutions* (Garnavault, Carmona Miura), *Reducing Scope 3 Emissions across the Beer Value-Chain* (Adda-Dailly, Valet) et *The Future of Packaging and Packaging Waste in Europe* (Radmilovic, Dubourg, Leferman) (Lille, 2024).

Comme nous l'avons dit, la réussite du néolocalisme, tant au niveau idéologique qu'au niveau publicitaire, se manifeste dans le succès de la bière artisanale, dont le marché s'est accru de manière considérable (BoE, 2023, 9-11). Les buveur·euse·s contemporain·e·s, à l'instar des consommateur·trice·s d'autres produits et des expert·e·s du secteur brassicole, sont plus attentif·ve·s à la qualité des bières et de leurs ingrédients (Ellis, Bosworth, 2015 ; Melewar, Skinner, 2020), ainsi qu'à la vocation néolocale des brasseries. D'ailleurs, l'orientation territoriale (Garavaglia, 2020 ; Mercurio, 2023) et durable des (micro)brasseries artisanales et, par imitation, de quelques grands groupes brassicoles (songeons au phénomène *crafty*, BA 2012 ; Rice, 2016), est évidente non seulement dans le *storytelling* et dans l'image diffusée par les entreprises à travers le nom de marque et de produit, le logo, l'emballage et l'étiquette[3], mais aussi, de façon plus ample, dans la terminologie qui désigne les lieux de production et de consommation de bière : cela correspond à une « effervescence sociétale » qui « se manifeste linguistiquement par l'émergence et la diffusion de nouvelles lexies » (Altmanova *et al.*, 2022, 85).

L'intérêt de la filière brassicole pour l'environnement est également témoigné par les initiatives durables de Brewers of Europe (BoE), l'organisation qui représente les associations nationales de brasseur·euse·s auprès de l'Union européenne. De plus, nous constaterons que des unités lexicales et des termes candidats du domaine brassicole font ressortir les dimensions durable et néolocale, ce qui élargit, d'une part, la terminologie brassicole, d'autre part, la terminologie de l'écologie, de l'environnement et de la durabilité. En raison de son importance et de son actualité, ce domaine a fait l'objet de plusieurs travaux (Légifrance, 2009 ; Zanola, 2010 ; Foulquier, 2012 ; Sayhi, 2012 ; Dury, 2013 ; Romagnoli, Vecchio, 2015 ; Cagninelli 2023), sur lesquels nous nous appuierons pour notre analyse.

3. Démarche méthodologique

L'analyse que nous présentons ici repose sur les principes de la socioterminologie (Gambier, 1987 ; Gaudin, 2003, 2005 ; Delavigne, Gaudin, 2022 ; Humbley, 2022) : sous le signe de la sociolinguistique, cette branche soutient la circulation sociale des termes et la nécessité de tenir compte des pratiques

3 Il est très intéressant d'observer des jeux de mots utilisés dans la promotion des « piliers de responsabilité sociétale », tels que : « Pour la nature, on se met la pression ! » ou « Le CO_2, on le préfère dans nos bouteilles ! » (Brasserie Licorne, https://www.brasserielicorne.com/nos-piliers-rse).

langagières des locuteur·trice·s et du discours dans lequel les termes sont utilisés (Cabré, 2003). Nous avons donc examiné les termes candidats répertoriés qui désignent les lieux de production et de consommation de bière en fonction du contexte discursif et des procédés de formation (*cf.* Altmanova *et al.*, 2022 ; Dal, Namer, 2022 ; Kiegel-Keicher, 2022). Nous avons aussi envisagé les définitions présentes dans certaines sections des sites Web (« Qui sommes-nous ? », « Histoire de la brasserie », « Nos bières »), en les comparant à celles figurant dans des dictionnaires et des documents législatifs, caractérisés par un langage moins accessible au grand public que celui des textes rédigés par les brasseries. Il est à préciser que ces sources ne sont pas nombreuses, le domaine étant assez récent : en effet, certains termes candidats apparaissent seulement dans le Grand dictionnaire terminologique (désormais GTD)[4].

Le corpus a été collecté à partir de 100 sites Web de (micro)brasseries françaises et italiennes, classées dans deux bases de données : *Liste des brasseries françaises en activité* du *Projet Amertume* et *Microbirrifici.org. The Italian Beer Database*[5]. Les brasseries sélectionnées sont artisanales (les critères seront explicités dans le sous-paragraphe suivant) et en activité, et disposent d'un site Web officiel actualisé et fonctionnel. Nous avons suivi l'ordre chronologique fourni par les bases de données afin de recueillir des données à jour.

Nous avons considéré notamment les termes candidats qui renvoient à l'écologie, à l'environnement et à la durabilité, ou qui véhiculent l'authenticité recherchée de nos jours par la clientèle (*cf.* Mac Cannel, 2013 [1976] ; Cipollaro *et al.,* 2021), ce qui nous a fait écarter des éléments moins connotés dans ce sens, comme *salle de brassage* et *cave de fermentation*. Le corpus se compose au total de 22 termes candidats en français (Tab. 1) et 39 en italien (Tab. 2), affichés dans les tableaux par ordre de fréquence. Nous indiquons en gras les termes principaux, dont certains dérivés se caractérisent notamment par l'accumulation d'adjectifs, comme *brasserie artisanale indépendante et biologique* ou *microbirrificio artigianale indipendente*.

4 https://vitrinelinguistique.oqlf.gouv.qc.ca.

5 http://projet.amertume.free.fr ; https://www.microbirrifici.org.

FR		
1. **Brasserie artisanale** [co. Brassarìa artisanarìa]	9. **Taproom**	17. **Ferme brassicole**
2. Brasserie artisanale bio	10. **Brewpub**	18. **Brewing company**
3. Brasserie artisanale indépendante et familiale	11. **Bar à bières**	19. **Cuve de garde**
4. Brasserie artisanale indépendante et biologique	12. **Fabrique**	20. **Laboratoire à bières**
5. Brasserie artisanale paysanne	13. Fabrique de bières artisanales	21. **Atelier de fabrication**
6. Brasserie de bières artisanales	14. Fabrique de bières bio	22. **Écobrasserie**
7. **Microbrasserie**	15. **Craft brewery**	
8. Microbrasserie associative	16. **Cave à bières**	

Tab. 1. Termes candidats en français.

IT		
1. **Birrificio artigianale**	14. **Brewpub**	27. **Beer garden**
2. Birrificio artigianale indipendente	15. **Agribirrificio**	28. **Beer room**
3. Birrificio artigianale a gestione indipendente	16. Agribirrificio artigianale	29. **Birrificio biologico** [de. Bio Bierbrauerei]
4. **Taproom**	17. **Beerfirm**	30. **Hofbrauerai**
5. **Brewery**	18. **Brewing company**	31. **Cantina brassicola**
6. **Birrificio agricolo**	19. **Nano-birrificio artigianale**	32. **Brettificio artigianale**
7. Birrificio agricolo artigianale	20. **Azienda agricola**	33. **Birrificio sociale**
8. Birrificio agricolo indipendente	21. **Birrificio rurale**	34. **Cooperativa con birrificio**
9. **Microbirrificio**	22. **Laboratorio**	35. **Spaccio**
10. Microbirrificio artigianale	23. Laboratorio birrario	36. **Opificio brassicolo**
11. Microbirrificio artigianale indipendente	24. **Micropub**	37. **Public house**
12. Microbirrificio agricolo	25. **Birrificio micro**	38. **British pub**
13. Microbirrificio solidale	26. **Birreria**	39. **Home restaurant**

Tab. 2. Termes candidats en italien.

3.1. Terminologie brassicole : *petite brasserie indépendante*

Plusieurs travaux en sciences du langage se penchent sur le discours de dégustation (*brutoglossia*, Konnelly, 2020), sur le marketing (Temmerman, 2018 ; Parizot, 2022) et sur les terminologies (Malin, 2019 ; Mercurio, 2023 ; Parizot *et al.*, 2023) de la bière, en signalant parfois une sorte d'imprécision (*cf.* Devilla, Mercurio, 2023), qui est aussi due au fait qu'il s'agit d'un domaine de recherche récent. En outre, la variété des cultures brassicoles des différents pays, certaines plus anciennes que d'autres, justifie en partie l'absence d'une définition de bière artisanale officielle et normalisée.

Nous nous concentrerons sur le lieu où la bière artisanale est produite. Il est désigné par un terme qui marque la rencontre entre la terminologie brassicole et le domaine juridique. De fait, la loi italienne (la seule à notre connaissance) présente une définition du terme *birra artigianale* (*cf.* Baiano, 2020, 1830-1831), alors que dans d'autres pays des explications ont été formulées par des associations de catégorie. La définition italienne figure dans la modification de 2016 à un article d'une loi de 1962 (art. 2 comma 4-bis, loi n° 1354/1962). Ce qui nous intéresse, c'est la référence au lieu de production, le *piccolo birrificio indipendente*, à savoir une brasserie juridiquement et économiquement indépendante par rapport à d'autres entreprises ou de grandes enseignes, qui utilise des installations physiquement séparées de celles de toute autre brasserie et dont la production annuelle n'excède pas 200 000 hectolitres (Normattiva, 2016) :

> birrificio […] legalmente ed economicamente indipendente da qualsiasi altro birrificio, che utilizzi impianti fisicamente distinti da quelli di qualsiasi altro birrificio, che non operi sotto licenza di utilizzo dei diritti di proprietà immateriale altrui e la cui produzione annua non superi 200.000 ettolitri.

L'équivalent français «*petite brasserie indépendante*» apparaît dans le Code général des impôts français (art. 178-0-bis, décret n° 2006/1026, modifié en 2007, Légifrance, 2024), qui décrit cette activité par les mêmes paramètres que la définition italienne :

> […] une petite brasserie indépendante s'entend d'une brasserie établie dans un État membre de la Communauté européenne qui respecte cumulativement les critères suivants :
> 1° elle produit annuellement moins de 200 000 hectolitres de bière ;
> 2° elle est juridiquement et économiquement indépendante de toute autre brasserie ;

> 3° elle utilise des installations physiquement distinctes de celles de toute autre brasserie ;
> 4° elle ne produit pas sous licence.

Autour de ces termes se trouvent les autres termes candidats énumérés dans les tableaux 1 et 2.

4. Analyse

4.1. Anglicismes

La terminologie brassicole est très riche en anglicismes et notre corpus ne fait pas exception, surtout en ce qui concerne les brasseries italiennes. En général, l'anglais est utilisé pour dénommer la plupart des styles de bière (sauf ceux qui ne sont pas originaires de territoires anglophones et qui sont nés dans des pays où la tradition brassicole est plus enracinée, comme la Belgique et l'Allemagne), certains métiers du secteur (par exemple, *cellarman*, en français *caviste*, GDT), mais aussi des lieux emblématiques de production (*brewery*) et de consommation (*pub*). *Pub*, abréviation de *public house*, est l'un des termes les plus anciens du domaine et désigne un « établissement typiquement anglo-saxon où l'on sert des boissons alcoolisées, particulièrement de la bière, et des repas simples » (GDT), associé parfois à *inn*, encore présent au Royaume-Uni sur les enseignes de quelques établissements. En effet, parmi les termes candidats italiens, nous retrouvons *british pub*.

L'un des termes anglais pour lesquels des brasseries, tant françaises qu'italiennes, fournissent des descriptions est *brewpub*, en français *bistrot-brasserie*, « un bistrot où l'on vend pour consommation immédiate des bières brassées sur place » (GDT) ou encore :

> [1] le rendez-vous incontournable de tous les amateurs de bières made in Pays Bourbons. À deux pas des cuves de Mousse Bourbonnaise, assis autour d'un de nos barils de bières, profitez d'un moment convivial entre amis[6].
>
> [2] *il posto ideale in cui trascorrere serate in compagnia, festeggiare occasioni importanti. o, semplicemente, rendere particolarmente piacevole e unica anche una fugace pausa pranzo*[7].

6 Les brasseries citées sont énumérées à la fin des références bibliographiques.

7 « L'endroit idéal pour passer des soirées en compagnie, pour célébrer des occasions importantes… ou simplement pour rendre la pause déjeuner fugace plus agréable et unique » (nous traduisons).

> [3] [*brewpub*] *con vista impianto dove bere la nostra birra artigianale e gustare un menù sfizioso*[8].

Outre l'idée de la qualité garantie par le brassage sur place et la possibilité d'associer la bière au repas, ces exemples suggèrent également la convivialité du moment et du contact qui peut s'établir entre les client·e·s du *brewpub*. Cela revient aussi dans un autre concept, qui est souvent désigné par le terme anglais *taproom*, en français *bar à bières*, la « salle où se trouvaient les barriques de bière, et donc leurs robinets » (GDT) :

> [4] pour que vous puissiez contempler nos cuves inox en buvant une bière ;
>
> [5] bar à bières de la Boussole ;
>
> [6] *un luogo d'incontro* […] *di qualità*[9] ;
>
> [7] [*è qua che*] *serviamo direttamente le nostre birre e dove potrai trovarle al massimo della loro freschezza*[10] *;*
>
> [8] *un posto ideale per degustare e bere le birre alla spina, oltre che per acquistare direttamente tutte le bottiglie prodotte*[11] *;*
>
> [9] per bere alla spina e in bottiglia a pochi passi dalla bottaia[12].

Dans ce cas, le contact s'instaure aussi entre les consommateur·trice·s et les producteur·trice·s ou les vendeur·euse·s. Le plus célèbre *beer garden* est la version en plein air.

4.2. Artisanale, indépendante et petite

Le terme générique *brasserie* (*birrificio*), suivi d'un ou des deux adjectifs *artisanale* (*artigianale*) et *indépendante* (*indipendente*), est le plus récurrent dans les deux langues. Nous avons déjà mentionné des définitions officielles, néanmoins les termes [*petite*] *brasserie artisanale* [*indépendante*] et [*piccolo*] *birrificio artigianale* [*indipendente*] n'apparaissent pas dans les dictionnaires.

---

8 « [Brewpub] avec vue sur l'installation de brassage où vous pouvez boire notre bière artisanale et déguster un menu savoureux » (nous traduisons).

9 « Un lieu de rencontre […] de qualité » (nous traduisons).

10 « [C'est là que] nous servons directement nos bières et où vous pouvez les trouver dans toute leur fraîcheur » (nous traduisons).

11 « Un lieu idéal pour déguster et boire des bières à la pression, ainsi que pour acheter toutes les bouteilles produites sur place » (nous traduisons).

12 « Pour boire des bières à la pression et en bouteille à quelques pas de la cave » (nous traduisons).

Comme le confirment les explications fournies par certaines brasseries, la nature de ces établissements est déterminée par les procédés de production et les matières premières utilisées :

> [10] toutes nos bières sont locales, de fermentation haute et produites en petite quantité.
>
> [11] bières artisanales, ni filtrées, ni pasteurisées, brassées à l'eau des Écrins !
>
> [12] elles ne subissent ni stérilisation, ni pasteurisation.
>
> [13] toutes les bières sont brassées artisanalement avec des matières premières naturelles.
>
> [14] *nasce dalla volontà di creare una birra pulita con ingredienti naturali, un prodotto fresco che soddisfi i palati più esigenti e non solo, di far conoscere il vero sapore della birra*[13].

Les adjectifs *artisanale* et *indépendante* qualifient aussi le terme *microbrasserie* (*microbirrificio*), « dont la production est réalisée de manière artisanale et en quantité limitée » (GDT). Le préfixe *micro-* abrège l'adjectif *petit* de *petite brasserie* (*piccolo birrificio*), se configurant comme un préfixe fonctionnel à la condensation. Dans les termes candidats italiens, on le retrouve comme adjectif postposé (*birrificio micro*), en tant qu'abréviation de *microscopico* (en français *microscopique*). Bien que moins courant que *micro-*, le préfixe *nano-* (*nano-birrificio*), qui est également utilisé dans le langage technique et scientifique, a la même function et évoque une production qui se distingue de l'industrielle à grande échelle :

> [15] les quantités par brassin sont limitées.
>
> [16] nous brassons des bières artisanales, familiales, à échelle humaine.
>
> [17] notre local de brassage est installé à la maison.
>
> [18] une toute petite brasserie, d'où le *micro-*, concentrée sur 35 m².
>
> [19] nos bières […] ne contiennent aucun additif, sont non filtrées, non pasteurisées.

13 « […] naît du désir de créer une bière pure avec des ingrédients naturels, un produit frais qui satisfasse les palais les plus exigeants et, en plus, de faire découvrir le vrai goût de la bière » (nous traduisons).

4.3. Paysanne et locale

Un autre terme très fréquent toujours issu de la loi italienne est *birrificio agricolo*, à savoir une brasserie qui, afin de rester dans le régime agricole, utilise au moins 51 % de sa propre orge pour le maltage des bières qu'elle produit. À cet égard, nous avons relevé des termes candidats tels qu'*agribirrificio* (sur le modèle d'*agriturismo*[14]) ou *birrificio rurale*, alors qu'en français un aspect comparable est mis en évidence par des éléments tels que l'adjectif [*brasserie*] *paysanne* ou le substantif *ferme* [*brassicole*] : ces termes renvoient à l'authenticité du territoire et de la nature, sans avoir pourtant la même implication juridique que *birrificio agricolo*.

> [20] *insieme di una azienda agricola, una produzione, un agriturismo, un negozio e una sola famiglia*[15].
>
> [21] *azienda agricola coltiva le proprie materie prime [...] e produce birre speciali in linea con la stagionalità delle spezie coltivate*[16].
>
> [22] *birre agricole del nostro territorio*[17].

Dans cette lignée, certaines unités lexicales communiquent le concept d'artisanat. En français : *laboratoire à bières*, *fabrique de bières artisanales*, *atelier de fabrication* ; en italien : *laboratorio birrario*, *opificio brassicolo*, *brettificio artigianale*. *Spaccio*, au lieu de *negozio* (en français, magasin), en tant que point de vente au détail dans un milieu plutôt rural («*negozio, o anche esercizio di rivendita al dettaglio, soprattutto presso comunità, collettività e sim*», Treccani)[18], semble faire allusion à la dimension locale ou au temps passé.

14 Barbanti et Venturi (2007) observent l'émergence d'*agri-entreprises* dénommées justement avec le préfix *agri-*.

15 « Combinaison d'une entreprise agricole, d'une production, d'un agritourisme, d'un magasin et d'une famille » (nous traduisons).

16 « Entreprise agricole cultive ses propres matières premières […] et produit des bières spéciales en fonction de la saisonnalité des épices cultivées. » (nous traduisons).

17 « Bières agricole de notre territoire » (nous traduisons).

18 https://www.treccani.it/vocabolario/.

4.4. Environnement, communauté et territoire

En ce qui concerne l'attention plus explicite portée aux questions environnementales et sociétales, nous avons observé que certaines brasseries françaises se définissent comme *brasserie biologique*, *microbrasserie associative*, *écobrasserie*, ou des brasseries italiennes se désignent comme *birrificio biologico*, *birrificio sociale*, *cooperativa con birrificio*, *micro-birrificio solidale*. Les préfixes *bio-* et *éco-*, abréviations de *biologique* et *écologique* ou *durable*, suivent les mêmes règles que *micro-* et *nano-*, bien qu'ils relèvent d'un autre domaine.

Ces termes candidats évoquent de manière évidente l'engagement sociétal et environnemental des entreprises concernées en reflétant les préoccupations de l'époque actuelle. Il faudrait donc les normaliser ou les accompagner par des explications afin que les consommateur·trice·s puissent comprendre de quoi il s'agit. Cependant, dans notre corpus, nous n'avons trouvé que deux explications :

> [23] [FR. *écobrasserie*] construit en bois et coiffé de panneaux photovoltaïques, notre établissement se distingue aisément. reflétant notre engagement indéfectible envers la création de produits d'exception, authentiques et naturels. Le coffrage du bâtiment est issu de bois recyclé, nous disposons également d'une innovation permettant de réduire drastiquement notre consommation d'eau et d'électricité. Nos bières artisanales, reconnues pour leur saveur distinctive, sont élaborées à partir d'ingrédients naturels et biologiques.

> [24] *[IT. birrificio sociale] impresa sociale genera occasioni di inclusione e spazi di aggregazione per la collettività* […] *trovano spazio anche dei lavoratori svantaggiati che sono coinvolti nelle attività di produzione meno complesse.* […] *consente di uscire dalle logiche assistenziali e di riattivare legami di solidarietà nella comunità*[19].

La sauvegarde de l'environnement et le soutien à la communauté locale font partie des principes du néolocalisme et confirment l'approche de certaines (micro)brasseries artisanales en faveur du développement durable tous azimuts.

19 « Entreprise sociale crée des opportunités d'inclusion et des espaces d'agrégation pour la communauté […] il y a aussi de la place pour les travailleurs défavorisés qui sont impliqués dans des activités de production moins complexes. […] permet de sortir de la logique de l'assistanat et de réactiver les liens de solidarité au sein de la communauté » (nous traduisons).

Parmi les stratégies linguistiques visant à évoquer le territoire, nous avons relevé l'emploi, dans le corpus en italien, de dénominations territoriales (*birrificio artigianale sardo*, *birrificio agricolo toscano*), alors qu'en français le recours aux langues locales ou régionales (*brassarìa artisanarìa* en corse) est plus fréquent : leur valeur symbolique et exotique (Devilla, 2022 ; Devilla, Mercurio, 2024) contribue à la promotion du territoire et de l'authenticité de l'entreprise.

5. En guise de conclusion

Dans cette étude, nous avons analysé des termes candidats désignant les lieux de production et de consommation de bière, un nano-domaine qui représente bien la tendance des consommateur·trice·s de notre époque, plus attentif·ve·s à la qualité et à l'origine des produits ainsi qu'aux pratiques durables des entreprises. Le corpus examiné est constitué de termes candidats en français et en italien, tels qu'ils sont utilisés par les brasseries dans le web et sur les emballages des bières. Des dictionnaires et des documents officiels ont été également consultés.

Nous avons constaté que la terminologie brassicole est encore peu normalisée, étant liée à une filière récente et en évolution, surtout par rapport au secteur du vin, par exemple. Toutefois, en dépit des différences culturelles et brassicoles qui caractérisent forcément les pays considérés, les terminologies et les stratégies promotionnelles adoptées par les brasseries artisanales évoluent de la même façon, soumises qu'elles sont aux besoins de la clientèle contemporaine. Dans le sillage du néolocalisme, les consommateur·trice·s sont en effet à la recherche de valeurs authentiques et de traditions locales, et plus intéressé·e·s aux défis sociétaux et environnementaux, ce qui témoigne de la « médiatisation globale du problème climatique et de la préservation de l'environnement » (Altmanova *et al.*, 2022, 97, 100).

Bien que parfois il s'agisse de marketing d'entreprise, les brasseries se proposent également de valoriser le territoire dans une perspective durable et néolocale. Se reflétant dans la terminologie, les phénomènes observés attestent encore une fois le lien profond entre les langages et les sociétés. À l'avenir on continuera à surveiller les terminologies du domaine brassicole pour aborder l'analyse du discours environnemental des (micro)brasseries et des brasseries industrielles.

Références

Altmanova, Jana, Cartier, Emmanuel, Luzzi, Jimmy, Pinto, Sarah et Piscopo, Sergio. 2022. « Innovations lexicales dans le domaine de l'environnement et de la biodiversité : le cas de bio en français et en italien ». *Neologica*, nº 16, p. 85-110.

Baiano, Antonietta. 2021. « Craft Beer: An Overview ». *Comprehensive Reviews in Food Science and Food Safety*, nº 20, p. 1829-1856.

Barbanti, Lorenzo et Venturi, Gianpietro. 2007. *Produzioni vegetali*. Parma : Monte Università Parma Editore.

Biros, Camille. 2014. « Les couleurs du discours environnemental ». *Mots. Les langages du politique* 105. Consulté le 20 juillet 2024. http://journals.openedition.org/mots/21688

Bowen, Robert et Miller, Maggie. 2023. « Provenance Representations in Craft Beer ». *Regional Studies* 57, nº 10, p. 1995-2005.

Brewers Association (BA). 2012. *Craft* vs. *Crafty: A Statement from the Brewers Association*. Consulté le 20 juillet 2024. https://www.brewersassociation.org/press-releases/craft-vs-crafty-a-statement-from-the-brewers-association

Brewers of Europe (BoE). *Brewing 4 Eu*. Consulté le 20 juillet 2024. https://brewing4.eu/energy-and-climate-change

Brewers of Europe (BoE). *Beer – Europe's Choice. Our 2024-2029 Manifesto for a Sustainable Brewing Future*. Consulté le 20 juillet 2024. https://brewersofeurope.eu/wp-content/uploads/2024/03/manifesto-for-a-sustainable-beer-future.pdf

Brewers of Europe (BoE). 2022. *European Beer Trends. Statistics Report 2022 Edition*. Consulté le 20 juillet 2024. https://brewersofeurope.org/uploads/mycms-files/documents/publications/2022/european-beer-trends-2022.pdf

Brewers of Europe (BoE). 2023. *European Beer Trends. Statistics Report 2023 Edition*. Consulté le 20 juillet 2024. https://brewersofeurope.eu/wp-content/uploads/2023/11/european-beer-trends-2023-web.pdf

Cabré Castellví, Maria Teresa. 2003. « Theories of Terminology. Their Description, Prescription and Explanation». *Terminology*, vol. 9, nº 2, p. 163-200.

Cagninelli, Claudia. 2023. « Du développement durable à la transition écologique dans les tweets. Des phraséotermes concurrents du domaine écologique ? ». Dans *Phraséologie et terminologie*, édité par Paolo Frassi, p. 305-330. Berlin/Boston : De Gruyter.

Carroll, Archie B. 1991. « The Pyramid of Corporate Social Responsibility: Toward the Moral Management of Organizational Stakeholders ». *Business Horizons*, vol. 34, nº 4, p. 39-48.

Cipollaro, Maria, Fabbrizzi, Sara, Alampi Sottini, Veronica, Fabbri, Bruno et Menghini, Silvio. 2021. « Linking Sustainability, Embeddedness and Marketing Strategies: A Study on the Craft Beer Sector in Italy ». *Sustainability*, vol. 13, nº 19, p. 10903.

Dal, Georgette et Namer, Fiammetta. 2022. « Éco- lave plus vert, et il lave toute la famille ». *Neologica*, nº 16, p. 111-128.

Delavigne, Valérie et Gaudin, François. 2022. « Founding Principles of Socioterminology ». Dans *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, édité par Pamela Faber et Marie-Claude L'Homme, p. 177-196. Amsterdam : John Benjamins.

Devilla, Lorenzo. 2022. « La communication touristique numérique des organismes institutionnels de Sardaigne et de Corse. Les langues régionales entre authenticité et marchandisation. Une étude de cas ». *CMLF 2022, SHS Web of Conferences*, nº 138, p. 12004.

Devilla, Lorenzo et Mercurio, Nicla. 2023. « Imprécisions terminologiques et marketing : le cas de la bière *doppio malto* ». *XIXe Journée scientifique Realiter. Terminologie et interdisciplinarité : défis et perspectives de recherche futures. Livret des résumés*, p. 31-33.

Devilla, Lorenzo et Mercurio, Nicla. 2024. « Communiquer la bière artisanale : valorisation du territoire et promotion des langues locales ». Colloque international : *Local, authentique, durable : la gastronomie face à la globalisation. Perspectives multidisciplinaires*, Université de Sassari, 7-8 octobre 2024.

Dury, Pascaline. 2013. « Quelle(s) traduction(s) pour le terme anglais *greenwashing* ? Quelques observations croisées en terminologie ». *Traduire*, nº 229, p. 26-35.

Ellis, Victoria et Bosworth, Gary. 2015. « Supporting Rural Entrepreneurship in the UK Microbrewery Sector ». *British Food Journal*, vol. 117, nº 11, p. 2724–2738.

Erhardt, Niclas, Martin-Rios, Carlos, Bolton, Jason et Luth, Matthew. 2022. « Doing Well by Creating Economic Value through Social Values among Craft Beer Breweries: A Case Study in Responsible Innovation and Growth ». *Sustainability*, vol. 14, nº 5, p. 2826.

Flack, Wes. 1997. « American Microbreweries and Neolocalism: "*Ale-ing*" for a Sense of Place ». *Journal of Cultural Geography*, vol. 16, nº 2, p. 37-53.

Foulquier, Luc. 2012. « Le parcours des mots : le cas du préfixe “éco” et l'écologie ». *Environnement, Risques & Santé*, vol. 11, nº 3, p. 230-239.

Gambier, Yves. 1987. « Problèmes terminologiques des pluies acides : Pour une socio-terminologie ». *Meta*, vol. 32, nº 3, p. 314-320.

Garavaglia. Christian. 2020. « The Emergence of Italian Craft Breweries and the Development of their Local Identity ». Dans *The Geography of Beer*, édité par Nancy Hoalst-Pullen et Mark Patterson, 135–147. Cham : Springer.

Gatrell, Jay, Reid, Neil et Steiger, Thomas L. 2018. « Branding Spaces: Place, Region, Sustainability and the American Craft Beer Industry ». *Applied Geography*, nº 90, p. 360-370.

Gaudin, François. 1993. *Pour une socioterminologie. Des problèmes sémantiques aux pratiques institutionnelles*. Rouen : Université de Rouen.

Gaudin, François. 2003. *Socioterminologie. Une approche sociolinguistique de la terminologie*. Louvain-la-Neuve : De Boeck Supérieur.

Gaudin, François. 2005. « La socioterminologie ». *Langages*, nº 157, p. 80-92.

Hede, Anne-Marie et Watne, Torgeir. 2013. « Leveraging the Human Side of the Brand Using a Sense of Place: Case Studies of Craft Breweries ». *Journal of Marketing Management*, vol. 29, nºs 1-2, p. 207-224.

Humbley, John. 2022. « The Reception of Wüster's General Theory of Terminology ». Dans *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge*, édité par Pamela Faber et Marie-Claude L'Homme, p. 15-36. Amsterdam : John Benjamins.

Kiegel-Keicher, Yvonne. 2022. « Bio- et éco-. Procédé de création lexicale dans la terminologie environnementale officielle française ». *Neologica*, nº 12, p. 129-149.

Konnelly, Lex. 2020. « Brutoglossia: Democracy, Authenticity, and the Enregisterment of Connoisseurship in “Craft Beer Talk” ». *Language & Communication*, nº 75, p. 69-82.

Légifrance. 2009. *Vocabulaire de l'environnement (liste de termes, expressions et définitions adoptés)*. Consulté le 20 juillet 2024. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000020506972

Légifrance. 2024. *Code général des impôts. Annexe III - Petites brasseries indépendantes (Articles 178-0 bis A à 178-0 bis B)*. Consulté le 20 juillet 2024. https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006069574/LEGISCTA000022911300

Lille. 2024. *39th EBC Congress and 6th Brewers Forum*. Consulté le 20 juillet 2024. https://lille2024.beer

Mac Cannel, Dean. 2013 [1976]. *The Tourist: A New Theory of the Leisure Class*. Berkeley : University of California Press.

Melewar, T. C. et Skinner, Heather. 2020. « Territorial Brand Management: Beer, Authenticity, and Sense of Place ». *Journal of Business Research*, nº 116, p. 680-689.

Mercurio, Nicla. 2023. « Terminologie(s) et territorialité : la description des bières artisanales en français et en italien ». Dans Christophe Roche (dir.), *TOTh 2022. Terminologie & Ontologie : Théories et Applications*, p. 139-159. Chambéry : Presses universitaires Savoie Mont Blanc.

Normattiva. 2016. *Disciplina igienica della produzione e del commercio della birra*. Legge 16 agosto 1962, n° 1354. Consulté le 20 juillet 2024. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:1962-08-16;1354~art2

Parizot, Anne. 2022. « Stratégies communicationnelles et marketing. Mise en scène discursive et storytelling des produits monastiques. Étude de cas ». *Revue de communication sociale et publique*, nº 34, p. 53-69.

Parizot, Anne, Verdier, Benoît et Leroyer, Patrick. 2024. « À travers le(s) verre(s) à bière : ceci n'est pas (qu')un verre ! ». Dans *Les terminologies professionnelles de la gastronomie et de l'œnologie : arts de la table. Dimensions créatives et culturelles*, édité par Audrey Moutat et Carine Duteil, p. 73-83. Limoges : Presses universitaires de Limoges.

Rice, Jeff. 2016. « Professional Purity: Revolutionary Writing in the Craft Beer Industry ». *Journal of Business and Technical Communication*, vol. 30, nº 2, p. 236-261.

Romagnoli, Elisa et Vecchio, Clara. 2015. « Les termes de la ville durable ». *Cahiers RAMAU*, nº 7, p. 52-66.

Sayhi, Sabri-Fabrice. 2012. « Traduire dans le domaine de l'économie écologique : les difficultés terminologiques ». *Traduire*, nº 227, p. 35-46.

Schnell, Steven M. et Reese, Joseph F. 2003. « Microbreweries as Tools of Local Identity ». *Journal of Cultural Geography*, vol. 21, nº 1, p. 45-69.

Shortridge, James R. 1996. « Keeping Tabs on Kansas: Reflections on Regionally Based Field Study ». *Journal of Cultural Geography*, vol. 16, nº 1, p. 5-16.

Temmerman, Rita. 2018. « Co-création et web 2.0. Noms de marques pour de nouvelles bières artisanales dans la ville multilingue de Bruxelles ». *Éla. Études de linguistique appliquée*, vol. 192, nº 4, p. 417-434.

Zanola, Maria Teresa. 2010. « La terminologie des énergies renouvelables entre communication institutionnelle et savoirs spécialisés ». *Dialogos*, vol. 11, nº 22, p. 83-99.

Liste des brasseries citées

- [1] La Mousse Bourbonnaise, www.la-mousse-bourbonnaise.fr [fermée].
- [2] Birracruda, https://www.birracruda.net/.
- [3] Birrificio Vecchia Orsa, https://www.vecchiaorsa.it/brewpub/.
- [4] 3:6:9 Bière Craft, https://www.369craft.beer/.
- [5] Brasserie La Boussole, https://www.brasserie-laboussole.fr/.
- [6] Crack Brewery, https://www.crakbrewery.com/.
- [7] Mudita Brewery, https://muditabrewery.com/.
- [8] Fabbrica della Birra Perugia, https://www.birraperugia.it/.
- [9] Ca' del Brado. Cantina brassicola, https://www.cadelbrado.it/.
- [10] La Leymentaise. Brasserie artisanale, https://www.laleymentaise.fr/.
- [11] Brasserie artisanale de Serre-Ponçon, https://basp05.com/.
- [12] La Brouche, http://www.biere-ariege.com/.
- [13] Brasserie Lézard Noir, www.bierelezardnoir.com.
- [14] Beautiful Rebels, https://beautifulrebels.it/.
- [15] La gorgée de Valserine, https://lagorgeedevalserine.fr/.
- [16] [17] La germoise, https://www.lagermoise.fr/.
- [18] Micro-brasserie du Séronais, https://www.microbrasero.com/.
- [19] Brasserie Volta, https://brasserievolta.fr/.
- [20] Birra di Fiemme, https://www.birradifiemme.it/.
- [21] [22] AgriBirrificio Fria, https://www.birrafria.it/.
- [23] Bulles de Provence, https://bullesdeprovence.com/.
- [24] Bàgolo. Birrificio sociale, https://www.birrificiobagolo.it/.

Posters

Ontologies for Digital Humanities: An example of utilisation in the research and study of Classics

Rafail Giannadakis

TALOS-AI4SSH Lab, University of Crete Research Center for the Humanities, the Social & Education Sciences, Greece

giannadakis.uni@gmail.com

Abstract. This poster presents a case study exploring the potential utilisation of an ontology in the research and study of Classics. Specifically, it aims to demonstrate whether SPARQL query graphs, which represent potential questions posed by classicists to an ontology concerning ancient crisis events, *i.e.* the "LACRIMALit Ontology", can assist in retrieving relevant data and addressing their needs. Initial results suggest that query graphs enable classicists to retrieve accurate and reliable data, expand the scope of their research, and promote logical and computational thinking through their construction.

Résumé. *Ce poster présente une étude de cas explorant l'utilisation potentielle d'une ontologie dans la recherche et l'étude des études classiques. Plus précisément, il vise à démontrer si les graphes de requêtes SPARQL — qui représentent les questions potentielles posées par les spécialistes des études classiques à une ontologie concernant les anciens événements de crise, c'est-à-dire l'« ontologie LACRIMALit » — peuvent aider à récupérer des données pertinentes et à répondre à leurs besoins. Les premiers résultats suggèrent que les graphes de requêtes SPARQL permettent aux spécialistes des sciences de l'antiquitéde récupérer des données précises et fiables, d'élargir le champ de leurs recherches, et de promouvoir une pensée logique et computationnelle grâce à la construction de ces graphes de requêtes.*

1. Introduction

Although it has been noted that a more systematic, community-driven and standards-based approach would significantly enhance the accuracy, consistency, and interoperability of ontologies in Humanities, their growing use in Digital Humanities is undeniable (Jansen, 2019). As such, the current poster

will explore the potential of such software artefacts, generated through this increasing use, to address cognitive and academic needs of students and researchers. More specifically, a case study will examine the application of an ontology with antiquarian content, illustrating its role in advancing research and study within the field of Classics.

This poster explores whether the data retrieval from an ontology related to Classics can be utilised for the research and study of this field. I will evaluate how query graphs help Classics students and researchers extract data from extensive datasets that align with their research goals or academic needs. Furthermore, I aim to assess how well experts can convert their questions into query graphs, whether these queries can address the complex inquiries of the experts and, finally, present the potential benefits and efficiencies of this method with a view to demonstrating its value in enhancing the Classics research and study.

2. Methodology

The ontology under search is the "LACRIMALit Ontology", a deliverable of the "LACRIMALit- Leaders and Crisis Management in Ancient Greek Literature" research project, which is being held in the Institute for Mediterranean Studies of the Foundation for Research and Technology Hellas from 2022-2025 (Papadopoulou *et al*., 2022). The latest available 1.2 version was constructed in Protégé (Papadopoulou & Roche, n.d.), *i.e.* the free software for editing ontologies (Musen, 2015).

As already noted, central to the evaluation of the current ontology's utilisation is the extraction of its data. Therefore, the extraction dictates the creation of query graphs according to "SPARQL Protocol and RDF Query Language", since it enables data retrieval from sources either stored or viewed as RDF (W3C, 2013). More specifically, it offers a set of keywords, query forms and patterns that allow the user to formally construct the query graphs. The following queries were addressed to the "SPARQLer-General purpose processor" endpoint, in which the query should be inserted, and the preferred format of the results should be specified (SPARQLer, n.d.).

For the constructed query graphs, two forms of results and three keywords were employed. Specifically, the "SELECT" form, which "returns variables and their bindings directly" (W3C, 2013), and the "ASK" form, which "tests whether or not a query pattern has a solution" (W3C, 2013), were used. The keywords included "UNION," represented as "{{A} UNION {B}}", which

syntactically specifies pattern alternatives (W3C 2013); "BIND", denoted as "BIND('value' AS ?variable)", which assigns literal values to variables (W3C, 2013); and the language filter, denoted as "FILTER (lang(?variable)= language_tag)" (W3C, 2013). Lastly, within the filter of language two ISO 639-1 language tags, "el" for post-1453 Modern Greek and "en" for English, were applied (ISO, 2002).

3. Implementation

As previously mentioned, the questions transformed into query graphs stemmed from possible research or cognitive needs of classicists, which could be addressed using the "LACRIMALit Ontology". Consequently, various standardized vocabularies were selected and applied based on the specific data needed for retrieval.

The first query graph addresses the cognitive need of researchers and students of antiquity to be aware of the sources from which they draw knowledge, focusing on retrieving the metadata of the ontology. Thus, the "SELECT" form of results, along with the "BIND" and "UNION" keywords and a language filter, were used. Additionally, four Dublin Core Metadata Initiative properties — "dc:creator", "dc:contributor", "dc:title" and "dc:subject" — were implemented to retrieve the ontology's creators, contributors, title and subject respectively (DCMI Usage Board, 2020).

The second query graph, presented in two versions, addresses the cognitive need of language experts to verify the morphology of terms, focusing on the validation of literals. Specifically, the formations "Swkratis" and "Socrates" were checked using the "ASK" form of results to determine whether these English literals, one in each version, matched the ontology's data. The "rdfs:label" property, which provides "a human-readable version of a resource's name" (W3C 2014) was employed for this purpose.

The last query graph addresses the cognitive need of classicists to understand terms, their definitions, their hypernym, and their hyponym(s). Specifically, the "SELECT" form of results, the "UNION" keyword and the "lac:Political_Action" class, defined by the ontology's creators, were used alongside several properties. The "rdfs:comment" property, which provides a "human-readable description of a resource" (W3C, 2014), was employed to retrieve the natural language definition of the "lac:Political_Action" class; while for its hypernym and hyponyms, the "rdfs:subClassOf" property was used, indicating the domain as subclass of the range class (W3C, 2014), along with the "rdfs:label" property.

In conclusion, the applications discussed in this case study highlight various cognitive and research needs of classicists that can be addressed through the adopted methodology—specifically, by using "SPARQL" query graphs to interact with an ontology containing relevant data.

4. Results & Conclusion

In conclusion, the "SPARQL" query graphs discussed effectively reflect possible needs of students and researchers that deal with antiquity. Specifically, concerns such as the source of information, correct morphology, hypernyms, hyponyms, and definitions of terms are all relevant to a classicist's work. Despite the demanded existence of ontologies in various fields and the challenges the humanist might encounter—such as the need for foundational knowledge in ontologies, standardized vocabularies, and "SPARQL", or as already discussed, the complexities of dealing with OWL, which require a background in Description Logic and generally in Logic for Computer Science (Roche and Papadopoulou, 2019)— this research or study approach has proven both efficient and valuable.

Building on this, leveraging query graphs within an ontology with antiquarian content offers numerous benefits to classicists. This method enhances the accuracy and precision of data retrieval by transforming research questions into query graphs using "SPARQL". Additionally, it allows classicists to quickly access reliable data, such as that found in a deliverable of a research project, significantly reducing research time and enabling a greater focus on analysis and interpretation. Query graphs also facilitate complex queries, expanding research capabilities and uncovering insights that might otherwise be difficult to obtain. Moreover, this approach fosters logical and computational thinking, as a result of addressing issues that may arise during the process of extracting the ontology's data. Overall, query graphs streamline the research process while broadening the scope for innovation and discovery.

I gratefully acknowledge the generous funding of the European Union in the context of the TALOS-Artificial Intelligence for the Humanities and the Social Sciences project (TALOS-AI4SSH) under the Horizon Europe Framework Programme (HORIZON-WIDERA-2022-TALENTS-01-01-ERA Chairs); Grant agreement ID: 101087269. I also want to thank my supervisors and my colleagues for their valuable feedback and encouragement.

References

DCMI Usage Board. 2020. "DCMI Metadata Terms." DCMI, January 20. Accessed September 22, 2024. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

ISO. "ISO 639 — Language Code." ISO. Accessed September 22, 2024. https://www.iso.org/iso-639-language-code

Jansen, L. n.d. "Ontologies for the Digital Humanities: Learning from the Life Sciences?" In *Proceedings of the WODHSA: First International Workshop on Ontologies for Digital Humanities and Their Social Analysis*, Part of the Fifth Joint Ontology Workshops (JOWO 2019) Episode V: The Styrian Autumn of Ontology, Graz, Austria, September 23–25, 2019. Accessed September 22, 2024. https://api.semanticscholar.org/CorpusID:209429361

Musen, M. A. 2015. "The Protégé Project: A Look Back and a Look Forward." *AI Matters* 1 (4): 4–12. Accessed September 22, 2024. doi.org/10.1145/2757001.2757003

Papadopoulou, M., and Roche, C. n.d. "LACRIMALit-O4DH." *O4DH – Ontologies for Digital Humanities*. Accessed September 22, 2024. http://o4dh.com/lacrimalit

Papadopoulou, M., Roche, C., and Tamiolaki, E. M. 2022. "The LACRIMALit Ontology of Crisis: An Event-Centric Model for Digital History." *Information* 13 (8): 398. Accessed September 22, 2024. doi.org/10.3390/INFO13080398

Roche, C., and Papadopoulou, M. 2019. "Mind the Gap: Ontology Authoring for Humanists." *Joint Ontology Workshops*. Accessed September 22, 2024. https://api.semanticscholar.org/CorpusID:209430626

"SPARQLer". n.d. Accessed September 22, 2024. http://sparql.org/sparql.html

W3C. 2013. "SPARQL 1.1 Query Language." Edited by S. Harris and A. Seaborne. Accessed September 22, 2024. https://www.w3.org/TR/sparql11-query/

W3C. 2014. "RDF Schema 1.1." Edited by D. Brickley and R. V. Guha. Accessed September 22, 2024. https://www.w3.org/TR/rdf11-schema/

Ontologies for Digital Humanities: An example of utilisation in the research and study of Classics

Rafail Giannadakis
Division of Classical Studies (Dept. of Philology, UoC) & TALOS-AI4SSH (ERA Chair, UoC)

Introduction

An increasing use of ontologies in Digital Humanities has been observed in recent years (Jansen 2019) and this poster discusses the potential utilisation of an ontology with antiquarian content in Classics research and study.

More specifically my research questions are:

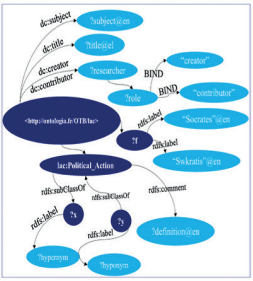
- Can query graphs render in the querying ontology the research questions of classicists?
- How effectively can Classics researchers and students convert their research questions into query graphs?
- Is this utilisation of query graphs beneficial and effective for Classics research?

Methodology

For my case study the 1.2 version of the "LACRIMALit Ontology" is utilised, a deliverable of the "LACRIMALit - Leaders and Crisis Management in Ancient Greek Literature"— an IMS/FORTH research project (Papadopoulou et al. 2022), which was built in Protégé (Papadopoulou & Roche n.d.)– the free software for editing ontologies (Musen 2015).

To query "LACRIMALit Ontology" via the "SPARQLer-General purpose processor" graphs (Pic. 1) were constructed according to "SPARQL Protocol and RDF Query Language", since it enables us to query data sources that are either stored or viewed as RDF (W3C 2013). From "SPARQL" the following were used:

- query forms ("ASK", "SELECT")
- keywords ("UNION", "BIND")
- filter of language (with the "el" referring to Modern Greek and the "en" referring to English ISO 639-1 language tags)

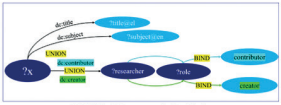


Pic 1. The query graphs visualized.

Implementation

The questions formed and constructed into query graphs arose from research or cognitive needs that a researcher or student of Classics may have, and which can be addressed by the "LACRIMALit Ontology".

1. The first query graph (Pic. 2) involves retrieving the ontology's metadata (Pic. 3). That is why DCMI properties along with the "SELECT" form of results, the "BIND" and the "UNION" keywords and the filter of language were implemented to retrieve the title, the subject and the research team of the ontology (DCMI Usage Board 2020).

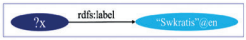


Pic 2. The 1st query graph visualized.

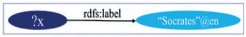
title	subject	researcher	role
"Ontologia LACRIMALit" @el	"LACRIMALit ontology describes events of crises in Greco-Roman Antiquity." @en	"Christophe Roche (USMB, LISTIC)"	"creator"
		"Maria Papadopoulou (USMB, LISTIC)"	"creator"
		"Markus Zimmermann (University of Bayreuth, IMS-FORTH)"	"contributor"
		"Roberta Dainotto (UoC, IMS-Forth)"	"contributor"
		"Anna-Maria Milaeu (UoC, IMS-FORTH)"	"contributor"
		"Panayotis Andreakis (UoC, IMS-FORTH)"	"contributor"
		"Eleni-Melina Tsesiokli (UoC, IMS-FORTH)"	"contributor"

Pic 3. The results of the 1st query graph in xml via "SPARQLer- General Purpose Processor" (retrieved 22/09/2024).

2. The second query graph (Pic. 4, 5) involves validating literals. More specifically, "Swkratis" and "Socrates" formations were checked, i.e. whether these English literals are correct based on the ontology's data. That is why the "rdfs:label" property defined by the W3C (W3C 2014) along with the "ASK" form of results were used.

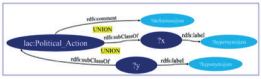


Pic 4. The invalid literal of the 2nd graph.



Pic 5. The valid literal of the 2nd graph.

3. The third query graph (Pic. 6) involves retrieving the hypernym, the hyponyms and the definition of the "lac:Political_Action" class defined by the creators of the ontology (Pic. 7). Thus, the "rdfs:comment", the "rdfs:subClassOf" and the "rdfs:label" properties along with the "SELECT" form of results, the "UNION" keyword and the filter of language were used.



Pic 6. The 3rd query graph visualized.

hypernym	hyponym	definition
"Events" @en		
	"harangue" @en	
	"truce" @en	
	"peace treaty" @en	
	"armistice" @en	
	"elections" @en	
	"alliance" @en	
	"hegemony" @en	
	"amnesty" @en	
		"action designed to attain a purpose by the use of political power or by activity in political channels" @en

Pic 7. The results of the 3rd query graph in xml via "SPARQLer- General Purpose Processor" (retrieved 22/09/2024).

Results & Conclusion

These questions address possible needs of researchers and students of antiquity, i.e. for those interested in:

- the source of retrieved information
- the correct morphology of terms
- the hypernym and the hyponyms of a term
- the definition of a term

Despite the:

- difficulties→ demanded existence of ontologies in different disciplines and fields
- basic knowledge required→ ontologies, SPARQL and standardized vocabularies

By exploiting this research method classicists can:

- retrieve accurate and reliable information (i.e. querying the "LACRIMALit Ontology")
- enhance and broaden the research possibilities
- develop logical and computational thinking skills

Contact info

Rafail Giannadakis
◦ BA student of Classical Studies, Dept. of Philology, UoC.
◦ Assistant Researcher and Social Media Manager of TALOS-AI4SSH funded by the European Union, UoC.
E-mail: giannadakis.uni@gmail.com
Website:
https://crete.academia.edu/RafailGiannadakis

Bibliography/ webliography

DCMI Usage Board. (2020, January 20). DCMI: DCMI Metadata Terms. DCMI. Accessed on September 22, 2024, from https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

ISO - ISO 639 — Language code. ISO. https://www.iso.org/iso-639-language-code

Jansen, L. (n.d.). Ontologies for the Digital Humanities: Learning from the Life Sciences? In Proceedings of the WODHSA. First International Workshop on Ontologies for Digital Humanities and Their Social Analysis. Part of the Fifth Joint Ontology Workshops (JOWO 2019) Episode V: The Styrian Autumn of Ontology, Graz, Austria, 23–25 September 2019. Accessed on September 22, 2024, from https://api.semanticscholar.org/CorpusID:209429361

Musen, M. A. (2015). The Protégé Project: A Look Back and a Look Forward. AI Matters, 1(4), 4–12. Accessed on September 22, 2024, from https://doi.org/10.1145/2757001.2757003

Papadopoulou, M., & Roche, C. (n.d.). LACRIMALit- 04DH. 04DH - Ontologies for Digital Humanities. Accessed on September 22, 2024, from http://o4dh.com/lacrimalit

Papadopoulou, M., Roche, C., & Tzevelekaki, E. M. (2022). The LACRIMALit Ontology of Crisis: An Event-Centric Model for Digital History. Information 2022, Vol. 13, Page 398, 13(8), 398. Accessed on September 22, 2024, from https://doi.org/10.3390/INFO13080398

SPARQLer - General purpose processor. (n.d.). sparql.org. Accessed on September 22, 2024, from http://sparql.org/sparql.html

W3C. (2013). SPARQL 1.1 Query Language. In S. Harris & A. Seaborne (Eds.), W3C. w3.org. Accessed on September 22, 2024, from https://www.w3.org/TR/sparql11-query/

W3C. (2014). RDF Schema 1.1. In D. Brickley & R. V. Guha (Eds.), W3C. w3.org. Accessed on September 22, 2024, from https://www.w3.org/TR/rdf11-schema/

La traduction des sigles dans le TAL : étude de cas dans le domaine de la santé environnementale

Maria Grazia Massimo

Université de Naples L'Orientale
m.massimo1@unior.it

Résumé. Le présent travail se propose d'analyser le traitement des sigles dans la traduction automatique pour la paire de langues français-italien. Le but est d'examiner les défis traductifs à travers la comparaison d'un groupe d'outils disponibles en ligne, tels que DeepL, Google Translate et Microsoft Bing. La recherche sera effectuée dans le domaine de la santé environnementale dans lequel l'utilisation des sigles ne se limite pas à la communication entre experts du domaine, mais peut également être présente dans la communication avec le grand public. À partir de ce constat, la sélection de deux corpus (spécialisé et non spécialisé) nous a permis de retenir 140 sigles dont l'analyse linguistique et traductologique nous amènera à dégager les choix traductifs proposés par les outils automatiques.

Abstract. *This work aims to analyze the processing of acronyms in automatic translation for the French-Italian language pair. The aim is to examine translation challenges through a comparison of a group of tools available online, such as DeepL, Google Translate and Microsoft Bing. Research will be conducted in the field of environmental health where the use of acronyms is not limited to communication between experts in the field, but may also be present in communication with the general public. From this observation, the selection of two corpora (specialized and non-specialized) allowed us to retain 140 acronyms whose linguistic and translational analysis will lead us to identify the translation choices of automatic tools.*

1. Introduction

Dès l'apparition des premiers logiciels de traduction automatique en 1950, les chercheurs visent la création de logiciels capables de produire des traductions comparables à celles du traducteur humain.

En particulier, grâce à l'évolution technologique, l'apprentissage profond, appliqué aux systèmes de traduction automatique, a amélioré au fil du temps la qualité des traductions générées, réduisant ainsi l'intervention d'un traducteur humain (Monti, 2019 : 16). Or, malgré les progrès évidents dans ce domaine, certains phénomènes de créativité lexicales, dont les sigles, restent problématiques pour le traitement automatique des langues.

2. Méthodologie

Notre étude propose d'évaluer la performativité des traducteurs automatiques dans le traitement des sigles appartenant au domaine de la santé environnementale. La constitution de deux corpus nous a permis de retenir, après validation manuelle, 140 sigles examinés pour comparer les choix traductifs proposés par les outils automatiques : les deux corpus comprennent dix textes de spécialité (manuels de gestion de l'environnement, articles scientifiques, textes législatifs) et dix textes informatifs (articles de presse et blogs) dans la période qui va de 2010 à 2020. Le choix de représenter les sigles lexicalisés à la fois dans les contextes spécialisés et non spécialisés repose sur le paramètre de leur fréquence dans le domaine traité.

Du point de vue traductologique, nous analyserons la traduction des 140 sigles, du français vers l'italien dans le traitement automatique des langues. Les logiciels de référence, Google Translate, DeepL et Microsoft Bing, ont été choisis en fonction de leur large diffusion.

Notre recherche repose sur trois phases :

- **2a.** traduction automatique des sigles en l'absence de leur forme développée et de cotexte, c'est-à-dire sans l'unité phrastique ni autre environnement lexical ;
- **2b.** traduction automatique des sigles accompagnés de sa forme développée mais en l'absence d'un cotexte ;
- **2c.** traduction automatique des sigles avec cotexte (SVO : Sujet-Verbe-Objet) et avec leur forme développée.

3. Résultats

À partir de la comparaison des traducteurs automatiques, nous pouvons remarquer que les outils Google Translate et Microsoft Bing ont du mal à reconnaître les sigles avec une fréquence de 60 % lorsque la fonction informative du texte est réduite. À titre d'exemple, le sigle le plus connu dans le domaine traité, **OMS** (Organisation Mondiale de la Santé), présenté sans cotexte et en absence de sa forme développée, est traduit de façon automatique par le pronom relatif (**chi**), parce que l'outil a traduit en italien le sigle anglais **WHO** (World Health Organization). Cela vaut pout Google Translate et Microsoft Bing, qui utilisent l'anglais comme langue de support à partir de laquelle ils proposent les traductions en italien, alors que DeepL basé sur le système neuronal, fournit une traduction exacte en respectant l'équivalent italien, à savoir **OMS** (Organizzazione Mondiale della Sanità). Un autre exemple proposant les mêmes résultats est le sigle **MATE** (Ministère de l'Aménagement du Territoire et de l'Environnement) qui a été reconnu en tant que terme anglais *mate*, traduit en italien par « copain ».

FR (Google Translate - Microsoft Bing)	IT (Google Translate - Microsoft Bing)
OMS	CHI
MATE	COMPAGNO/ACCOPPIARSI

Tab. 1. Résultat de l'hypothèse 2a.

Lorsqu'un sigle est accompagné de sa forme étendue entre parenthèses, tous les trois outils automatiques proposent la traduction correcte :

Google Translate, DeepL et Microsoft Bing Translator
Exemple : FR : OMS (Organisation Mondiale de la Santé) → **IT** : OMS (Organizzazione Mondiale della Sanità)

Tab. 2. Résultat de l'hypothèse 2b.

La performativité des traducteurs automatiques dépend en grande partie de la présence de la forme développée associée aux sigles ou de la présence d'un cotexte, qu'il soit spécialisé ou non, permettant à l'outil d'améliorer sa compréhension.

Cependant, un grand nombre de sigles ne possèdent pas d'équivalents italiens disponibles. À ce propos, nous remarquons une influence de la langue anglaise dans le choix des traductions des sigles indépendamment de la

présence de la forme développée du sigle, du cotexte ou du choix de l'outil. Dans ce cadre, le sigle **LCEE** (Loi canadienne sur l'évaluation environnementale), accompagné de sa forme étendue, a été traduit par son équivalent anglais **CEAA** (Canadian Environnemental Assessment Agency), qui est effectivement utilisé en langue italienne.

FR	IT
La LCEE relève de l'Agence canadienne d'évaluation environnementale (Agence **CEE**)	La CEAA rientra nella giurisdizione della Canadian Environmental Assessment Agency (CEE Agency)

Tab. 3. Résultat de l'hypothèse 2c issu d'un texte non spécialisé.

La présence d'équivalents anglais ne correspond pas toujours à la formulation exacte des sigles. En mars 2024, le sigle **MGE** (Manuel de Gestion de l'Environnement), examiné dans un cotexte et accompagné de sa forme étendue, a été traduit de façon différente par les trois outils proposant des sigles anglais appartenant à des domaines différents. Microsoft Bing Translator propose **EMM** (Entreprise Mobility Management), DeepL propose **EMP** (ElectroMagnetic Pulse), alors que pour Google Translate le sigle reste invariable.

Aucun outil n'a proposé son équivalent italien à savoir **SGA** (Sistema di Gestione Ambientale). En juillet 2024, il est possible de remarquer une variation dans les traductions au cours du temps, ce qui témoigne que le même sigle propose une traduction différente au fil du temps.

FR	IT
« Le MGE (Manuel de Gestion de l'Environnement) contient des méthodes précises et des mesures de protection »	L'EMM (Manuale di gestione ambientale) contiene metodi specifici e misure di protezione (**Microsoft Bing Translator**) Il Manuale di Gestione Ambientale (EMP) contiene metodi precisi e misure di protezione (**DeepL**) Il MGE (manuale di gestione ambientale) contiene precise modalità e misure di tutela (**Google Translate**)

Tab. 4. Résultat de l'hypothèse 2c issu d'un texte spécialisé.

4. Conclusions

L'analyse comparée de la paire de langues français-italien permet de constater les limites posées par la traduction automatique des sigles. En particulier, les trois logiciels utilisés (DeepL, Google Translate et Microsoft Bing) ont montré des résultats différents qui reflètent la performativité des outils. DeepL, en tant que système de traduction automatique plus récent et basé sur le système neuronal, présente les meilleurs résultats.

En l'absence d'un traduisant disponible, l'équivalent choisi par les traducteurs automatiques est, dans la plupart des cas, en langue anglaise, considérée comme langue pivot même s'ils adoptent d'autres domaines.

Nul doute que la recherche, avec le but d'ouvrir d'autres perspectives, vise à démontrer les limites des logiciels automatiques en tenant compte de l'aspect traductif et de la cohérence terminologique et référentielle.

Références

Altmanova, Jana, Centrella, Maria, Russo, Katherine Elizabeth (éds.). 2018. « Terminology & Discourse/Terminologie et discours », dans *Linguistic Insignts, Studies, Language and Communication*, vol. 241, Berne, Peter Lang.

Cabré, Maria Teresa. 1992. *La terminologie : théorie, méthode et applications*, Ottawa/Paris, Les Presses de l'Université d'Ottawa/Armand Colin.

Gérin, Michel, *et al.* 2003. *Environnement et santé publique. Fondements et pratiques*, Canada, Edisem.

Mayaffre, Damon, Vanni, Laurent. 2021. *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Paris, Honoré Champion.

Monnier, Philippe. 1994. « Usages et formations de sigles, une application dans l'industrie spatiale », dans *Linx*, n° 30.

Monti, Johanna. 2019. *Dalla Zairja alla traduzione automatica*, Napoli, Loffredo Editore.

Mortureux, Marie-Françoise. 1994. « Siglaison-acronymie et néologie lexicale », dans *Linx*, n° 30, « Les sigles », sous la direction de Marc Plénat.

Śmigielska, Beata. 2007. « Remarques sur la traduction automatique et le contexte », *Neophilologica*, n° 19, p. 253-267.

Wolosin, Claudia. 1996, « Problèmes de traduction posés par la siglaison dans le domaine des nouvelles technologies de l'information et de la communication », *ASp*, nos 11-14 (actes du 17e colloque du GERAS), p. 147-160. 10.4000/asp.3468

La traduction des sigles dans le TAL : étude de cas dans le domaine de la santé environnementale

Maria Grazia Massimo m.massimo1@unior.it
Université de Naples L'Orientale

INTRODUCTION

Dès l'apparition des premiers exemples de traduction automatique en 1950, les chercheurs visent la création de logiciels capables de produire des traductions comparables au traducteur humain. En particulier, grâce à l'évolution technologique, la valeur de la qualité des traductions s'est améliorée au fil du temps en adoptant les modèles neuronaux et en disposant des quantités de données multilingues nécessaires à l'apprentissage du système → produire des traductions finales sans intervention d'un traducteur humain (Morel, 2019 : 16). Or, malgré les progrès évidents dans ce domaine, certains phénomènes linguistiques, dont les **sigles**, restent problématiques pour le traitement automatique des langues.

MÉTHODOLOGIE

La recherche repose sur l'analyse linguistique et traductologique de deux types de corpus qui comprennent **dix textes de spécialité** et **dix textes informatifs** afin de répertorier 140 sigles, sélectionnés selon la fréquence d'usage.

1. Le choix des textes **techniques** (manuels de gestion de l'environnement) **de spécialité,** (textes législatifs) **et de vulgarisation** (articles) repose sur le paramètre de fréquence des sigles dans le domaine traité, indépendamment du genre textuel.
2. Du point de vue **traductologique**, l'analyse vise à démontrer les défis/limites dans la traduction des sigles dans le TAL. Les logiciels de référence employés sont **Google Translate, DeepL et Microsoft Bing** dans la combinaison linguistique **FR-IT**.

Analyse systématique du corpus reposant sur trois phases :

- traduction automatique des sigles en l'absence de contexte et sans recours à la forme développée ;
- traduction automatique des sigles avec leur développement complet mais en l'absence d'une phrase ;
- traduction automatique des sigles avec contexte et avec leur forme développée issus à la fois du corpus spécialisé et du corpus non spécialisé.

RÉSULTATS

1er résultat : mauvaise fonctionnalité des traducteurs automatiques due à la réduction de la fonction informative du texte (emploi des sigles)

Ex : le sigle OMS (Organisation Mondiale de la Santé), présenté sans contexte, est traduit de façon automatique par le pronom relatif (qui), parce que le traducteur automatique a considéré l'équivalent anglais WHO (World Health Organization).

2ème résultat : meilleure efficacité de la traduction automatique due à la présence de la forme étendue du sigle

Ex : OMS (Organisation Mondiale de la Santé) → OMS (Organizzazione Mondiale della Sanità)

3ème résultat : la langue anglaise comme langue pivot

Ex : les sigles LCEE (Loi canadienne d'évaluation environnementale), issu du corpus non spécialisé et MGE (Manuel de gestion de l'environnement) issu du corpus spécialisé, sont traduits de façon différente avec une efficacité distincte

CONCLUSIONS

- La recherche vise à démontrer les limites posées par la traduction des sigles dans le TAL. En particulier, les trois logiciels automatiques utilisés ont recensé trois résultats différents qui reflètent la performativité de l'outil. DeepL, en tant que système de traduction automatique plus récent, présente les meilleurs résultats.
- En l'absence d'un **traduisant disponible**, l'équivalent choisi par les traducteurs automatiques est, dans la plupart des cas, **l'anglais**, en supposant que beaucoup de secteurs adoptent des **standards internationaux (+ d'autres domaines)**.
- Nul doute que la recherche, avec le but d'ouvrir d'autres perspectives, vise à démontrer une approche multilingue qui peut observer le domaine de santé environnementale en tenant compte de l'aspect traductif et de la cohérence terminologique et référentielle.

RÉFÉRENCES

- Abramovici L., Dardaillon M., Russi M. T. (éds.), Terminologie d'discursive/ Terminologie et discours , dans «Linguistic Insights, Studies in Language and Communication», Vol. 141, Berne, Peter Lang, 2018.
- Cabré M. T., *La terminologie : Théorie, méthode et applications*, Ottawa/Paris, Les Presses de l'Université d'Ottawa/Armand Colin, 1998.
- Galan M. (et. al), *Environment et Santé Publique. Fondements et pratiques*, Canada, Edisem, 2003.
- Maaffre D. et Varol L., L'intelligence artificielle des textes. Des algorithmes à l'interprétation, Paris, Honoré Champion, 2021.
- Monnier P., Usages et formations de sigles, une application dans l'industrie spatiale, dans « Ling », n°20, 1994.
- Morel L., *Dalla lingua alla traduzione automatica*, Napoli, Jeffredo Editori, 2019.
- Montreuil M. F. Sigle(on acronymie et néologie lexicale, dans « Ling », n°30, Les sigles, sous la direction de Marc Plénat, 1994.
- Watson C., *Problèmes de traduction posés par la sigléson dans le domaine des nouvelles technologies de l'information et de la communication*, MSc, DO 10.34, 1995, DOI : https://doi.org/10.4000/ijcol.1668.

Assessing Comparability and Idiomaticity using Tertium Comparationis

Sarah Theroine*, Christophe Cruz**, Laurent Gautier***

* Université Bourgogne Europe, centre interlangues Textes, Images, Langages (TIL, EA 4182), Dijon
Sarah.theroine@ube.fr

**Université Bourgogne Europe, CNRS, laboratoire interdisciplinaire Carnot de Bourgogne (ICB UMR 6303), Dijon
Christophe.cruz@ube.fr

*** Université Bourgogne Europe, centre interlangues Textes, Images, Langages (TIL, EA 4182), Université Bourgogne Europe, Dijon
Laurent.gautier@ube.fr

Abstract. Comparable corpora consist of thematically related texts that have not undergone translation (McEnery, 2003; Zannetin, 2013; Tufis, 2012; Teubert, 1996: 247; Loock, 2016). Their authenticity makes them valuable for studying idiomaticity, as they are free from the linguistic distortions found in parallel corpora (Neveu, 2003; Wulff, 2019). However, their non-linear nature complicates alignment, leading to their underutilization in linguistic studies (Ebeling & Ebeling, 2020). To address this issue, we redefine *Tertium Comparationis* (Teubert, 1996; Chesterman, 1998; Moreno, 1998; Loock, 2016) as structured, indexed, and dynamic framework that functions as a measure of comparability by introducing five new evaluation criteria. Beyond the theoretical perspective, empirical validation is necessary. We introduce five new criteria—complementing four existing ones—to quantitatively assess corpus comparability and facilitate idiomaticity analysis.

Résumé. *Les corpus comparables sont des collections de textes partageant un même thème, mais n'ayant pas été traduits (McEnery, 2003 ; Zannetin, 2013 ; Tufis, 2012 ; Teubert, 1996 : 247 ; Loock, 2016). Leur authenticité les rend précieux pour l'étude de l'idiomaticité, car ils sont exempts des distorsions linguistiques présentes dans les cor-*

pus parallèles (Neveu, 2003 ; Wulff, 2019). Cependant, leur nature non linéaire complique leur alignement, entraînant une sous-utilisation dans les études linguistiques (Ebeling & Ebeling, 2020). Pour pallier ce problème, nous redéfinissons le Tertium Comparationis *(Teubert, 1996 ; Chesterman, 1998 ; Moreno, 1998 ; Loock, 2016) comme un cadre structuré et dynamique agissant comme une mesure de comparabilité grâce à l'introduction de cinq nouveaux critères d'évaluation. Ces derniers s'ajoutent en complément de quatre critères existants, afin d'évaluer quantitativement la comparabilité des corpus et de faciliter l'étude de l'idiomaticité.*

1. Introduction

Idiomaticity, as a fundamental characteristic of natural language, poses significant challenges in cross-linguistic analysis due to its non-compositional nature and cultural specificity (Sinclair, 1992). Traditional approaches often rely on parallel corpora, which—being composed of translated texts—introduce linguistic distortions that can affect the authentic representation of a language. Conversely, comparable corpora—thematically similar but non-translated, texts—provide a more reliable foundation for studying idiomaticity (McEnery, 2003; Zannetin, 2013; Loock, 2016). However, their absence of direct alignment complicates comparative analysis, necessitating a structured analytical framework, which leads to a new approach using *Tertium Comparationis* (TC). This paper redefines TC as a formalized, indexed, and dynamic framework, positioning it as a *measure of comparability* through the introduction of five new criteria.

2. State-of-the-Art

Traditionally, TC has been an implicit, qualitative concept in corpus linguistics and contrastive studies, often described as a common ground (Connor & Moreno, 2005). Despite its significance, it lacks precise definition and formalization. TC serves as a comparative basis for framing a comparable corpus (Teubert, 1996; Chesterman, 1998; Fix, 2008; Loock, 2016; Ebeling & Ebeling, 2020), ensuring homogeneity and authenticity through extra-linguistic criteria. Four criteria are widely applied in practice: i) topic, ii) discourse register, iii) document type, and iv) document structure. However, these remain subjective and insufficient for in-depth linguistic analysis. To address this issue, we propose the integration of language-based criteria, transforming TC into a measurable framework of comparability.

3. A new theoretical Framework

Literature suggests that TC plays an active role in corpus compilation prior to contrastive analysis. Our approach expands TC's function beyond a preparatory step, establishing it as an analytical tool. This perspective arises from three key observations: i) existing TC criteria are not clearly defined, ii) they are not quantifiable, making empirical verification difficult and iii) they are based on alignment principles relevant to parallel corpora, where source and target texts are naturally aligned through translation.

3.1. New Measurable Criteria

Our approach introduces five interdependent, quantifiable criteria designed to enhance contrastive corpus analysis ensuring both linguistic coverage and flexibility. These include: i) syntactic complexity, measuring sentence structure (length, clause dependency); ii) lexical richness, assessing vocabulary diversity; iii) terminological density, evaluating domain-specific terms; iv) information recognition, analyzing distribution and extraction; and v) semantic frames, identifying conceptual structures for cross-linguistic analysis. By integrating these refinements, TC evolves into a dynamic, multidimensional tool, combining qualitative and quantitative measures for more rigorous corpus comparability.

3.2. Methods

Our approach addresses the distinct dimensions of each criterion. The first criterion calculates corpus volume as a variability rate between corpora using a two-step method. This volume is defined as:

$$m_a = \frac{\sum_{i=1} Wai}{L_a} = \frac{W_a}{L_a}$$

This is reproduced in language B and the variability ratio between corpora is given by:

$$v = 1 - \frac{m_b}{m_a}$$

The second step involves POS-tagging and syntactic dependency trees using SpaCy. The second criterion is measured by the Type-Token Ratio (TTR) for the same volume, or the average TTR if volumes differ.

$$mTTR = \frac{1\sum_{i=1} v_i}{L\sum_{i=1} N_i}$$

The third criterion is analyzed using SketchEngine and TermoStats to study term frequency, collocations, and lexicon-grammatical patterns, enabling specialized term extraction for a multilingual dictionary. The fourth extracts evaluate information distribution using Kullback-Leibler divergence.

$$\mathrm{D_{KL}}(\theta_{\mathrm{ws}} \parallel \theta_{\mathrm{wt}}) = \sum_{i=1}^{|A|} p\left((w_{si}|w_s)\ log\ \frac{p(w_{si}|w_s)}{p(w_{si}|w_t)}\right)$$

The final criterion clusters words with similar actantial structures and usage contexts to build semantic frames. A thematic and conceptual analysis using Tropes is conducted to identify thematic scenarios or slots related to the extracted terms. This process structures conceptual slots, which define key semantic roles for each slot or Frame Element (FE). FEs are identified by analyzing recurrent lexicon-grammatical patterns, establishing semantic relationships between concepts. This approach builds domain-specific semantic frames, providing a structured representation of meaning.

4. Conclusion

This study redefines *Tertium Comparationis* as a quantifiable measure of comparability, integrating new linguistic-based criteria to complement existing extralinguistic factors. Furthermore, we hypothesize that these criteria are directly relevant to idiomaticity: the syntactic and lexical criterion is directly tied to language structure and fixed or idiomatic expressions; semantic frameworks, the most closely related, help analyze figurative meaning, context, and cultural dimensions. Given the cultural variations in idiomaticity across languages, contrastive analysis of comparable corpora offers a valuable approach for linguistic study. Furthermore, while the informational criterion indirectly influences idiomaticity by shaping discourse construction, the terminological criterion provides insights into domain-specific idiomaticity, or what we might call *professional idiomaticity.*

References

Chesterman, Andrew. 1998. *Contrastive Functional Analysis*. Amsterdam: John Ben-jamins.

Connor, Ulla M., and Moreno, Ana I. 2005. "Tertium Comparationis: A Vital Component in Contrastive Research Methodology." In *Directions in Applied Linguis-tics: Essays in Honor of Robert B. Kaplan*, edited by Paul Bruthiaux, Dwight Atkinson, William G. Eggington, William Grabe, and Vaidehi Ramanathan, 153–164. Clevedon, England: Multilingual Matters.

Ebeling, Jarle, and Ebeling, Olaf. 2020. "Contrastive Analysis, Tertium Comparationis and Corpora". *Nordic Journal of English Studies* 19(1): 97–117.

Fillmore, Charles J. 1985. "Frames and the Semantics of Understanding." *Qua-derni di Semantica: Rivista Internazionale di Semantica Teorica e Applicata* 6: 222-254.

Assessing Comparability with Tertium Comparationis: A linguistical and semantical Approach

Sarah THEROINE – Christophe CRUZ – Laurent GAUTIER
Laboratoire ICB, UMR 6306, CNRS, Université de Bourgogne 21000 DIJON, France
Laboratoire TIL, UR 4182 , Université de Bourgogne, 21000, DIJON, France

Summary : In this work, we present a review on the concept of Tertium Comparationis and then we introduce a novel dimension and application of TC as a measure of comparability applied to Corpus Linguistics.

Introduction

- The notion lacks of a proper definition. [1-2]
- TC is mostly seen as a common ground for building comparable corpora.
- Comparable corpora must shared a certain degree of similarity.
- Four established criteria: **topic, discourse register, document type** and **document structure.**

Discusion

- Said established criteria which are commonly used yet they rely on the structure of parallel corpora and may not be well adapted to comparable corpora.
- Established criteria cannot be quantified.
- Thus, we proposed five new criteria that may allow users to quantify the quality of their comparable corpora.

Figure 1: Hierarchy of the constant criteria of Tertium Comparationis

Figure 2 : New measurable criteria

Methods

- **Quantifying syntactic density and complexity:**
  - Calculating the average number of words per sentences and then number of clauses and phrases per text.
  - Generating dependency trees
- **Quantifying lexical richness:**
  - Calculating Standardized Type-Token Ration on the entire corpus.
- **Quantifying terminological density:**
  - Extracting the number of occurrences for each specialized terms.
  - Prompting models to evaluate the level of technicity according to their occurrences in the text.
- **Quantifying the level of information:**
  - Using DKL divergences [3] to evaluate the pourcentage of differences between corpus A and corpus B or the information of one document to another accross the same corpus.
- **Detecting and recognizing semantic frames:**
  - Identifiying recurrent lexico-grammatical patterns
  - Transferring said patterns into ontologies.

Dépressions 1007 hPa au large d'Ouessant se creuse et se décale vers le sud-est prévue 1001 hPa sur les pays de la Loire ce soir. Elle se comble sur le centre de la France demain dimanche.
Dépressions à 1007 hPa près des côtes bretonnes se creusant et se déplaçant vers le sud-est dans les prochaines heures, prévue 1001 hPa sur la région des Pays de la Loire ce soir. Elles s'affaiblissent progressivement sur la région lundi matin.
Tempêtes à 983 hPa au large de la Manche ce matin se déplaçant vers l'est dans les prochaines heures, prévue à 980 hPa sur la côte normande ce soir. Elles perdent en intensité sur le nord de la France demain matin.

<phenomenon> <mesure> <location> <time_period> <move> <time_period> <"prévue"> <mesure> <location> <time_period> <phenomenon> <move> <location> <time_period>

Conclusion

- Study aims to underline a novel understanding of Tertium Comparationis in theory and in practice.
- Study helps quantifying the quality and usability of comparable corpora
- Future work should consider creating an evaluation metric aiming to calculate the percentage of similarity between two comparable corpora and facilitating their alignment.

References

[1]Connor, Ulla M. & Moreno, Ana I. (2005). Tertium Comparationis: A vital component in contrastive research methodology. In P. Bruthiaux, D. Atkinson, W. G. Eggington, W. Grabe, & V. Ramanathan (eds), Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan. Clevedon, pp. 153-164. England: Multilingual Matters.
[2] Ebeling, O., & Ebeling, J. (2020). Contrastive analysis, tertium comparationis and corpora. Nordic Journal of English Studies, 19(1), 97-117.
[3] Brigitte Bigi. Using Kullback-Leibler Distance for Text Categorization. Advances in Information Retrieval, 2633, Springer Berlin Heidelberg, pp.305-319, 2003, ff10.1007/3-540-36618-0_22ff. ffhal01392500

Keywords

Tertium Comparationis, Comparable Corpora, Applied linguistic,

Analyse des désignations vocales du baroque français dans les guides d'écoute numériques

Sergio Piscopo

Università di Napoli L'Orientale
spiscopo@unior.it

Résumé. Cette étude examine la réutilisation de désignations vocales issues du répertoire baroque français, telles que *haute-contre*, *basse-taille* et *dessus*, dans des contextes numériques contemporains. Ces termes, historiquement ancrés dans les traités musicaux et les partitions du XVII^e^ et XVIII^e^ siècles, connaissent une résurgence *via* des guides d'écoute, des vidéos pédagogiques sur YouTube et les sites d'ensembles spécialisés dans le répertoire baroque. Nous analysons dans quelle mesure cette redistribution terminologique favorise une revitalisation sémantique ou s'inscrit dans une démarche de récupération philologique. Une fois que nous aurons isolé manuellement les termes liés aux tessitures vocales les plus répandues dans le contexte numérique, nous adopterons une approche qualitative à l'avenir afin d'explorer la compréhension et la perception de ces termes par le public, en distinguant les amateurs éclairés des novices. Nos résultats préliminaires suggèrent que le numérique agit comme un catalyseur, permettant une diffusion accrue de ces désignations, bien qu'elles puissent manquer de transparence pour le grand public. Cependant, pour les passionnés, cette revitalisation terminologique contribue à enrichir l'expérience et la compréhension du patrimoine musical baroque, affirmant ainsi leur pertinence dans un contexte contemporain.

Abstract. *This study examines the reuse of vocal terms from the French Baroque repertoire, such as* haute-contre*,* basse-taille *and* dessus*, in contemporary digital contexts. These terms, historically rooted in musical treatises and scores from the 17^th^ and 18^th^ centuries, are experiencing a resurgence through listening guides, educational videos on YouTube, and the websites of ensembles specialising in Baroque repertoire. We analyse to what extent this terminological redistribution favours a semantic revival or is part of a philological recovery process. Once we have manually isolated the terms associated with the most common vocal tessituras in the digital context, we*

will use a qualitative approach to explore the public's understanding and perception of these terms, distinguishing between enlightened amateurs and novices. Our preliminary findings suggest that digital technology is acting as a catalyst for the proliferation of these terms, although they may lack transparency for the public. For enthusiasts, however, this revival of terminology helps to enrich the experience and understanding of the Baroque musical heritage, and to reaffirm its relevance in a contemporary context.

1. Introduction

L'évolution du langage musical constitue un défi complexe, notamment en ce qui concerne la réutilisation de termes tombés en désuétude et devenus moins productifs dans le discours contemporain. Cette problématique s'intensifie dans le contexte numérique, en particulier dans les guides d'écoute numériques, où la réintroduction de désignations vocales du baroque français, telles que *haute-contre*, *basse-taille* et *dessus*, soulève des interrogations intéressantes du point de vue sémantique (Piscopo, 2024).

La motivation sous-jacente à cette étude réside dans la nécessité de comprendre la viabilité et la pertinence de la tentative de résurgence de cette terminologie à laquelle on assiste depuis quelques années, notamment dans le contexte numérique. La question est de savoir si cette entreprise représente une véritable revitalisation de ces désignations ou si elle s'apparente davantage à une récupération philologique dépourvue de toute adhérence sémantique actuelle. Nous mettons l'accent sur le domaine des désignations liées à la tessiture vocale de la période baroque française dans une perspective numérique, où l'on assiste à une reprise de ces désignations, en particulier dans les guides d'écoute. Ce genre de texte a un caractère polymorphe en raison de la présence de plusieurs types de texte aux degrés de spécialité variés en fonction du sujet. Le but pragmatique du guide est de donner au locuteur-écouteur un aperçu historico-musical par de courts textes et par l'usage de la terminologie qui lui est propre (Campos, 2005).

Ces guides étant consultables en ligne sur les sites officiels des ensembles consacrés à la musique baroque française, leur diffusion se généralise. Ils sont également relayés par les chaînes YouTube de ces ensembles ou de certains vidéastes, voire de *youtubeurs* qui réalisent des vidéos avec les compléments d'information nécessaires à l'écoute. Cela permet de contribuer ainsi à la diffusion de ces désignations à une plus large échelle. Nous nous sommes donc

demandé si ce renouveau pouvait en quelque sorte ressusciter cette terminologie désuète grâce au potentiel offert par le numérique dans le contexte du Web 2.0. La réponse, certes non exhaustive, impose un critère d'analyse qui tient compte de l'interprétation de ces désignations dans une perspective diachronique. Celles-ci, perdant leur valeur « dénominationable » (Fraith, 2015, 39) au niveau sémantique, risqueraient de ne pas être pleinement comprises par le locuteur d'aujourd'hui. Cependant, cette résurgence terminologique (Rousseau, 2019, 2021) à l'ère du numérique est susceptible d'intéresser à la fois les spécialistes et les amateurs éclairés, qui sont bien conscients de ce qu'ils lisent et recherchent. Bien que le contexte numérique soit largement utilisé par tous, les désignations retenues dans cette étude pourraient se répandre et toucher à un public plus vaste, contribuant ainsi à la rediffusion de ces désignations.

2. Méthodologie

Dans un premier temps, après avoir consulté les sites officiels des ensembles, nous avons recherché manuellement, dans la discographie et les publications, les désignations considérées. Puis, compte tenu de leur présence prédominante dans les programmes de salle et les guides de musique sacrée, nous les avons isolées une nouvelle fois manuellement pour vérifier leur occurrence dans les vidéos YouTube des chaînes qui publient des contenus multimédias consacrés à la musique de la période baroque. Nous avons donc constitué un corpus numérique contenant des articles, des vidéos, des transcriptions ainsi que des discussions issues de forums ou de réseaux sociaux. Nous avons constaté que les désignations les plus fréquentes concernant la tessiture vocale baroque dans les informations de ces chaînes YouTube étaient les suivantes : *dessus*, *bas-dessus*, *haute-contre* et *contre-ténor*. Nous nous sommes particulièrement intéressés aux ensembles Le Concert d'Astrée, Ensemble Antiphona, Folies Françaises, La Chapelle Harmonique et Le Poème Harmonique. Malgré la prolifération d'ensembles baroques, nous nous sommes concentrés sur ces derniers parce qu'ils accordent plus d'attention que d'autres à l'utilisation des désignations vocales. En plus des sites web officiels des ensembles, où il est possible de naviguer et de récupérer les désignations examinées, un certain nombre de chaînes YouTube participent également à leur diffusion. Sachant que la plateforme YouTube compte des millions d'abonnés et qu'autant de vidéos sont visionnées chaque jour, nous estimons que les chaînes YouTube consacrées au répertoire classique et baroque constituent un puissant vecteur de diffusion de cette terminologie à des fins quantitatives.

Les sites d'ensembles ainsi que les guides d'écoute s'adressent certainement à des amateurs éclairés. Le public cible de YouTube, quant à lui, est varié, mais son algorithme peut d'une certaine manière favoriser le processus de résurgence terminologique. À titre d'exemple, on peut citer les cas suivants : E.V, Midnight-Première, SP's score videos, etc.

Dans une deuxième phase de l'étude, nous souhaitons vérifier comment cette résurgence terminologique est accueillie par les amateurs et les néophytes. Après avoir mesuré la fréquence d'apparition des termes de tessiture vocale sur diverses plateformes numériques via le logiciel AntConc à partir de différents types de texte (guides d'écoute, transcriptions, articles savants et culturels, etc.), nous aimerions étudier les discours des youtubeurs ou des experts dans des vidéos, notamment la manière dont ils expliquent ces termes dans une perspective qualitative. Par conséquent, nos objectifs reposent sur une visualisation des contextes numériques où les termes sont le plus fréquemment utilisés, sur une catégorisation des utilisateurs (amateurs éclairés vs novices) et sur une évaluation de la pertinence sémantique contemporaine des termes.

3. Résultats

Dans le sillage de ce que l'on pourrait qualifier de rediffusion terminologique concernant les résultats préliminaires, l'étude des guides d'écoute retenus laisse entrevoir que ces désignations ne sont utilisées qu'en référence à la musique sacrée produite principalement en France au XVII^e^ siècle et non à d'autres genres opératiques tels que la cantate ou l'opéra. En effet, les messes ou les motets enregistrés ne comportent que des chanteurs ayant chacun leur propre tessiture[1], ou « noyau vocal », selon Marié de l'Isle (1873, 4), alors que les cantates ou les opéras sont dénominatifs et que les guides ne donnent généralement pas plus de détails à ce sujet.

Les programmes officiels des ensembles consacrés à la musique baroque reflètent également cette tendance. La persistance de l'utilisation des désignations vocales traditionnelles est notable, en particulier dans les publications en ligne sur les sites officiels des ensembles. Le Poème Harmonique, dirigé par Vincent Dumestre, en est un exemple illustratif ; cet ensemble intègre régulièrement des communiqués de presse et des articles dans la rubrique

1 « Partie du registre d'une voix qui est couverte avec un maximum d'aisance » (TLFi, *ad vocem*).

« Programmes ». Ces documents annoncent les concerts à venir et détaillent la distribution des rôles, soulignant ainsi l'utilisation continue des désignations de tessitures vocales. Par exemple, le communiqué du 6 septembre 2023[2] présente un concert consacré aux *Te Deum* de Marc-Antoine Charpentier et de Jean-Baptiste Lully, avec des rôles explicitement désignés. Nous soulignons en italiques les tessitures considérées : Amélie Raison, soprano ; Parco Garcia, *haute-contre*.

Cette pratique est principalement observée dans le cadre de programmes mettant en scène des œuvres sacrées, comme évoqué plus haut, où la distribution des voix se penche sur la tessiture vocale conçue à l'époque baroque en France. En contraste, les programmes annonçant des exécutions de musique sacrée italienne, comme le *Nisi Dominus* d'Antonio Vivaldi, adoptent une terminologie vocale différente, correspondant à la typologie vocale en usage : Eva Zaicik, mezzo-soprano ; Serge Goubioud et Benoît Joseph Meier, ténors ; Marie Theoleyre, soprano ; Geoffroy Buffière, baryton. On peut ainsi émettre l'hypothèse que les ensembles musicaux différencient leur terminologie en fonction du répertoire interprété. Dans ce dernier cas, l'ensemble accorde une attention particulière au traitement des désignations vocales baroques en fonction des contextes considérés. Une telle attention philologique est également commune à d'autres ensembles. Par conséquent, l'utilisation de la terminologie française pour la musique sacrée française et de la terminologie italienne pour le répertoire italien vise à préserver l'authenticité stylistique et à respecter les conventions linguistiques grâce à une adhérence sémantique. Cette approche permet de respecter les conventions linguistiques propres à chaque répertoire, assurant ainsi une meilleure adéquation sémantique entre les termes employés et les concepts musicaux qu'ils désignent.

4. Conclusion

Cette recherche, encore en cours de développement, s'inscrit dans la continuité d'une étude contrastive franco-italienne précédente sur les tessitures vocales à l'ère du baroque français[3]. Bien que nos résultats préliminaires montrent que le numérique agit comme un catalyseur pour la rediffusion de cette

2 On peut consulter la ressource à l'adresse suivante : https://lepoemeharmonique.fr/spectacle/te-deum/ (consulté le 10 septembre 2024).

3 Nous renvoyons à : Piscopo, Sergio. 2024. "Étude des désignations de tessiture vocale dans la musique baroque française et italienne : une approche terminographique". Dans *Testi e Linguaggi*, nº 18. Rome : Carocci Editore S.p.A.

terminologie, leur compréhension semble rester limitée à un public averti, tandis que le grand public peut en percevoir l'opacité sémantique. À ce stade, nous aimerions élargir notre corpus en intégrant davantage de données issues de plateformes et d'ensembles baroques internationaux variés. Par ailleurs, nous cherchons à approfondir l'étude qualitative, notamment en explorant de manière plus ciblée la perception et la compréhension de ces termes par des amateurs novices et spécialisés. Ces étapes supplémentaires permettront d'évaluer plus précisément si cette rediffusion correspond à une revitalisation sémantique pertinente ou à une simple récupération philologique. En poursuivant cette démarche, nous espérons offrir une perspective renouvelée sur l'impact du numérique dans la préservation et la réinterprétation de la terminologie musicale baroque liée à la tessiture vocale.

Références bibliographiques

Altmanova, Jana, Centrella, Maria, et Russo, Katherine E. (dir.) 2018. *Terminology & Discourse/Terminologie et discours*. Berne : Peter Lang.

Bonnard, Alain. 2004. *Le Lexique : annotations et termes musicaux*. Paris : BG Éditions.

Campos, Rémy, et Donin, Nicolas. 2005. « La musicographie à l'œuvre : écriture du guide d'écoute et autorité de l'analyste à la fin du dix-neuvième siècle ». Dans *Acta Musicologica*, vol. 77, nº 2, p. 151-204.

Frath, Pierre. 2015. « Dénomination référentielle, désignation, nomination ». Dans *Langue française*, vol. 4, nº 15, p. 33-46.

Kleiber, Georges. 1984. « Dénomination et relations dénominatives ». Dans *Langages*, nº 76, p. 77-94.

Piscopo, Sergio. 2024. « Étude des désignations de tessiture vocale dans la musique baroque française et italienne : une approche terminographique ». Dans *Testi e Linguaggi*, nº 18. Rome : Carocci Editore S.p.A.

Rousseau, Delphine-Anne. 2019. « Un cas de résurgence terminologique : la terminologie musicale en usage en France et en Angleterre à la seconde moitié du XVII^e siècle ». PhD diss., Université Lumière Lyon 2.

Marié de l'Isle, Claude Marie Mécène. 1873. *Formation de la voix, vocalises et exercices de prononciation*. Paris : Ménestrel.

Segonzac, Charlotte. 2023. *Écrivains et livrets d'opéra au tournant du* XX^e *siècle*. Paris : Honoré Champion Éditeur, coll. « Dialogue des Arts ».

Analyse des désignations vocales du baroque français dans les guides d'écoute numériques

Sergio PISCOPO
Université de Naples L'Orientale

INTRODUCTION

Points clés

- L'évolution du langage musical soulève des questions complexes concernant la description d'un lexique spécialisé qui a atteint un point de non-productivité.
- Ce phénomène est particulièrement visible dans le contexte numérique.
- L'intégration des tessitures vocales du baroque français soulève des questions sémantiques.
- Cette recherche explore la nature de cette résurgence terminologique.

Questions à discuter

- La réintroduction des tessitures baroques est-elle une véritable revitalisation ou une simple récupération philologique ?
- Quelles sont les implications de ce phénomène pour la compréhension dans le contexte numérique ?

MÉTHODOLOGIE

Points clés

- Repérage des désignations vocales baroques sur les sites web d'ensembles et les chaînes YouTube ;
- Analyse de l'utilisation de termes tels que *dessus*, *bas-dessus*, *haute-contre* et *contre-ténor* ;
- Focus sur les ensembles connus pour leur attention à la terminologie baroque : Le Concert d'Astrée, Ensemble Antiphona, Folies Françaises, La Chapelle Harmonique et Le Poème Harmonique ;
- Examen des chaînes YouTube dédiées à la musique baroque et de leur impact : E.V, Midnight-Première et SP's score videos.

Méthodologie de recherche

- Analyse de contenu des sites web d'ensembles et des chaînes YouTube.
- Identification des désignations vocales pertinentes et de leur utilisation.
- Analyse comparative de la portée et de l'engagement du public.

RÉSULTATS

Les désignations françaises sont principalement utilisées pour la musique sacrée française du XVII^{e} siècle

↓

Ces désignations sont moins fréquentes dans d'autres genres musicaux tels que la cantate et l'opéra

↓

Les ensembles musicaux différencient leur terminologie en fonction du répertoire interprété

↓

L'utilisation de la terminologie française pour la musique sacrée française et de la terminologie italienne pour le répertoire italien vise à préserver l'authenticité stylistique et à respecter les conventions linguistiques à l'aide d'une adhérence sémantique

CONCLUSIONS

Cette recherche prolonge une étude antérieure sur les tessitures vocales à l'époque du baroque français. Elle souligne les défis posés par l'adaptation de la terminologie musicale historique aux exigences contemporaines, notamment dans un cadre numérique. Si la réintroduction des dénominations vocales du baroque français peut approfondir notre compréhension de la musique de cette époque, elle demande une évaluation rigoureuse de sa pertinence et de sa faisabilité dans le contexte musical actuel.

RÉFÉRENCES

Altmanova, Jana, Maria Centrella, et Katherine E. Russo, eds. 2018. *Terminology & Discourse / Terminologie et discours*. Berne : Peter Lang.

Bonnard, Alain. 2004. *Le Lexique : annotations et termes musicaux*. Paris : BG Éditions.

Campos, Rémy, et Nicolas Donin. 2005. « La musicographie à l'œuvre : écriture du guide d'écoute et autorité de l'analyste à la fin du dix-neuvième siècle ». In *Acta Musicologica* 77, no. 2 : 151-204.

Frath, Pierre. 2015. « Dénomination référentielle, désignation, nomination ». In *Langue française* 4, no. 15 : 33-46.

Kleiber, Georges. 1984. « Dénomination et relations dénominatives ». In *Langages* 76 : 77-94.

Delphine-Anne, Rousseau. « Un cas de résurgence terminologique : La terminologie musicale en usage en France et en Angleterre à la seconde moitié du XVII^{e} siècle ». (PhD diss., Université Lumière – Lyon 2, 2019). (NNT :). (tel-02366882).

Marié de l'Isle, Claude Marie Mécène. 1873. *Formation de la voix, vocalises et exercices de prononciation*. Paris : Ménestrel.

Segonzac, Charlotte. 2023. *Écrivains et livrets d'opéra au tournant du XX^{e} siècle*. Paris : Honoré Champion Éditeur, Dialogue des Arts

Machine Translation Quality Evaluation: The Role of Terminology in Assessing Technical Texts

Andrea Fernández Vivanco

School of Translation and Information Sciences, Salamanca, Spain
afermamdezviv@usal.es

Abstract. This study examines the significance of terminology in the evaluation of machine translation (MT) quality for technical texts. While neural machine translation (NMT) has advanced considerably, challenges persist in the accurate translation of domain-specific terminology. This research analyzes translation errors through a case study focusing on aerospace manufacturing texts. The findings reveal that terminology-related errors account for the majority of inaccuracies, surpassing fluency and accuracy, thereby emphasizing the necessity of exploring domain-specific evaluation models into MT quality assessment and incorporating high-quality terminological resources into translation and assessment models

Résumé. *Cette étude examine la pertinence de la terminologie dans l'évaluation de la qualité de la traduction automatique dans les textes techniques. Bien que la traduction automatique neuronale ait fait des progrès significatifs, des défis en matière de traduction terminologique sont toujours présents. Notre recherche est basée sur une étude de cas de textes de fabrication aéronautique. Les résultats montrent que les erreurs liées à la terminologie sont plus fréquentes que les erreurs de fluidité et de contenu. Il est donc nécessaire d'explorer des modèles d'évaluation automatique adaptés au domaine et d'incorporer des ressources terminologiques de qualité dans les modèles de traduction et d'évaluation.*

1. Introduction

Machine Translation (MT) quality assessment has emerged as a pivotal area of research within translation studies and computational linguistics. The rapid advancement of Neural Machine Translation (NMT) technologies has significantly improved the fluency and general accuracy of translated texts. However,

challenges persist, particularly in the accurate translation of domain-specific terminology, which is crucial for technical texts. The evaluation of MT output quality, therefore, necessitates a comprehensive theoretical framework that adequately addresses both linguistic and domain-specific considerations.

2. Theoretical framework

2.1. Some Notes on Machine Translation Quality Assessment

The theoretical underpinnings of MT quality assessment are grounded in the interplay between human evaluation methods and automatic evaluation metrics. Human assessment is traditionally viewed as the gold standard due to its ability to capture nuanced linguistic and contextual subtleties that automated systems may overlook (Moorkens, 2018). Human evaluators can assess factors such as semantic accuracy, stylistic appropriateness, and pragmatic relevance, which are essential for ensuring that translations meet the intended communicative purpose.

In contrast, automatic evaluation methods offer scalability and consistency, utilizing computational models to assess translation quality based on predefined criteria. Metrics such as BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) have been widely adopted to evaluate MT outputs by comparing them to reference translations (Papineni *et al.*, 2002; Banerjee & Lavie, 2005). While these metrics provide rapid assessments, they often correlate poorly with human judgments in domain-specific contexts due to their limited consideration of semantic and terminological accuracy.

The primary distinction between human and automatic assessment lies in their respective capacities to handle linguistic complexities. Human assessment involves subjective evaluation by linguists or domain experts who can interpret context, detect nuanced errors, and assess the appropriateness of terminology within a specific field. This method is adaptable and can prioritize different quality aspects based on the text's purpose and audience (Castilho *et al.*, 2018). Automatic assessment relies on algorithmic evaluation using statistical models. While efficient, these methods often prioritize surface-level similarities and may not accurately reflect semantic equivalence or terminological precision, especially in specialized domains (Specia *et al.*, 2010). The limitations of automatic metrics in capturing terminological accuracy underscore the necessity for domain-specific evaluation approaches.

Recent studies have highlighted the importance of integrating both human and automatic evaluation methods to achieve a comprehensive assessment of MT quality. For instance, Rivera-Trigueros (2021) conducted a systematic review of MT systems and quality assessment procedures, emphasizing the need for combined evaluation approaches. Additionally, Wang and Chen (2023) reviewed current translation quality assessment research and suggested that static assessment models should be replaced with dynamic models that better meet the needs of translation teachers and learners.

2.2. Towards Domain-Specific Quality Assessment Models

The evaluation of MT quality in domain-specific texts requires specialized assessment models. These models should incorporate evaluation criteria that prioritize terminology accuracy over general linguistic fluency. The use of specialized corpora enhances the reliability of domain-specific translation assessments. Additionally, comparing MT outputs against validated terminological databases provides a more objective measure of translation quality. The remodeling of human assessment frameworks to integrate domain-specific linguistic variables is crucial for enhancing evaluation accuracy.

Our study proposes an adaptation of traditional MT assessment models and was based on the MQM (Lommel *et al*., 2015). However, together with accuracy and fluency, we separated terminology-related errors in order to account for the specificness of the evaluated text. Terminology errors were divided into three forms: inconsistency, semantic errors, and pragmatic errors. Inconsistency accounted for the fact of a term being translated in different ways within the same text, which can lead to confusion. Semantic errors arise when the translated term does not align with its intended meaning in the source text, while pragmatic errors occur when a term is not used in accordance with domain-specific conventions.

3. Case Study: MT for a Corpus in Aerospace Manufacturing

The aerospace industry is highly specialized, requiring precise translations due to its interdisciplinary nature, which incorporates fields such as manufacturing and physics. English functions as the dominant language in aerospace communication, yet MT plays an increasingly significant role in ensuring multilingual accessibility. This study assesses DeepL's performance in translating an English-language technical document corpus that describes advancements in satellite manufacturing for the Meteosat project. The primary

objective is to identify prevalent error types and evaluate the role of terminology in translation quality.

The findings, as exposed in Image 1, indicate that terminology represents the most significant challenge in MT for the used corpus in aerospace manufacturing. A quantitative analysis of errors reveals that terminology issues account for 57.6% of all errors, whereas fluency-related issues constitute 36.3%, and accuracy errors make up only 6.1%.

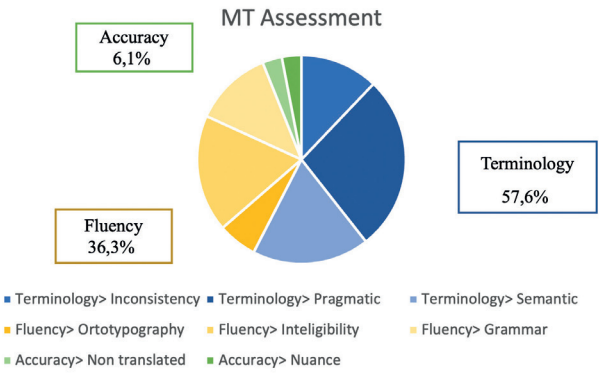


Figure 1.

3.1. Implications for MT Quality Evaluation

The prominence of terminology errors in domain-specific translations necessitates a reassessment of MT evaluation methodologies. Domain-specific quality assessment models must integrate both quantitative and qualitative measures to ensure comprehensive evaluation. Quantitative metrics should include the measurement of untranslated terms as an indicator of domain loss, as well as the analysis of term instability within a translation. Qualitative analysis should involve an in-depth examination of errors categorized as "terminology > pragmatic" to determine deviations from established domain usage conventions.

Research suggests that incorporating dynamic assessment models can better address the evolving needs of translation quality assessment. Additionally, integrating specialized terminological databases into NMT models has been shown to enhance term consistency and accuracy. Terminology-aware MT

models, which implement constraint-based decoding mechanisms, have the potential to improve term consistency and accuracy. Furthermore, human evaluation frameworks should be refined to incorporate weighted scoring systems that emphasize terminology accuracy over general fluency aspects.

4. Conclusions

This study underscores the critical role of terminology in MT quality evaluation for technical texts. The findings demonstrate that terminology-related errors significantly surpass fluency and accuracy issues in domain-specific translations. As a result, evaluation models must prioritize terminology consistency and incorporate domain-adapted validation mechanisms to enhance MT reliability. The case study on aerospace manufacturing highlights the necessity for refined evaluation methodologies that align with domain-specific requirements. Future research should explore hybrid models that combine automatic validation with expert human assessment to optimize translation quality in specialized fields.

References

Banerjee, S., & Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). Michigan: ACL.

Bentivogli, L., Bisazza, A., Cettolo, M., & Federico, M. (2018). "Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French". *Computer Speech & Language*, 52–70.

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). "Approches to Human and Machine Translation Quality Assessment". In *Translation Quality Assessment* (pp. 9–38). Dublin: Springer.

Forcada, M. (2017). "Making sense of neural machine translation". *Translation Spaces* 6 (2), 291–309.

Hasler, E., de Gispert, A., Iglesias, G. & Byrne, B. (2018). "Neural Machine Translation Decoding with Terminology Constraints". arXiv:1805.03750

Lommel, A., Burchardt, A., & Uszkoreit, H. (2015). "Multidimensional Quality Metrics (MQM) Definition". Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI). http://www.qt21.eu/mqm-definition/

Moorkens, J. (2018). "What to expect from Neural Machine Translation". *The Interpreter and Translator Trainer*, 375—387.

Papierini, K., *et al*. (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia.

Rivera-Trigeros, I. (2021). "Machine translation systems and quality assessment: A systematic review". *Language Resources & Evaluation*.

Wang, L., & Chen, Q. (2023). "Review of Translation Quality Assessment Research: Current Studies". *Scholars International Journal of Linguistics and Literature*, 473–477.

THE ROLE OF TERMINOLOGY ASSESSING DOMAIN-SPECIFIC MACHINE TRANSLATION

TOTh 2024
Terminologie & Ontologie : Théories et applications

Andrea Fernández Vivanco **University of Salamanca (Spain)**

Assessing the quality of machine translation (MT) output remains a challenge, although the last decade has broadened the variety of methodologies and models in human, automatic and hybrid evaluation. However, no model has been designed specifically for domain-specific texts and a scant few incorporate explicit weightings for terminology errors.

DOMAIN-SPECIFIC QA MODELS

Avenues for developing domains-specific automatic or hybrid MT quality assessment models:

- Domain-specific evaluation criteria
- Use of specialized corpora
- Automatic contrast with validated databases

Inconsistency: Different denominations given to the same term in the source text. It includes other types of inconsistency, such as writing an acronym with an article and then the same acronym without an article.

Semantic: Term in target is not in accordance with the meaning in source. Use of the term either refers to another signifier or is unintelligible.

Pragmatic: The term is not used in accordance with the conventions of the specialised language. E.g.: A term with certain denominative fixation used differently.

Term in source (EN)	MT (ES)	Fixed term (ES)
light beam	haz de **la** luz	haz de luz

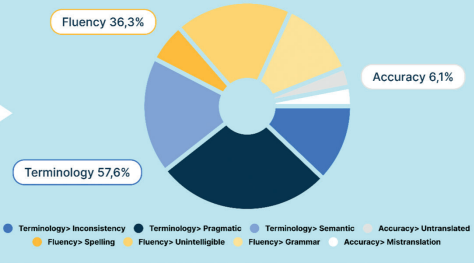
(*) Adapted from the MQM model, German Research Center in Artificial Intelligence.

CASE STUDY

Machine translation English-Spanish for a text on aerospace manufacturing

The aerospace industry necessitates the collaboration of professionals hailing from diverse global regions, although English has served as a predominant lingua franca. Consequently, a discernible domain loss in other languages is well documented. However, aerospace is also interdisciplinary and occasionally intersects with other fields characterized by well-established and validated terminologies, such as manufacturing or physics.

In this context, we set out to evaluate machine tranlation. We translated a text explaining the advancements in design and fabrication of the European satellite Meteosat. The results underscore that terminological-related-errors outweight any other category, and therefore should be paid closer attention.



Among the key insights of this case study: Several errors are due to term identification problems in the source text. That is, the MT tool fails to determine that two or more words are a domain-specific term and therefore should be translated as a unity.

Evaluation of machine translated texts offers new approaches for terminolists. When examining domain loss, we can evaluate machine translated texts in order to:

- assessing the number of untranslated terms.
- analysing lack of stability in the terms used MT tools.
- qualitatively analysing errors tagged as “terminology>pragmatic”.

Terminologie de la couleur bleue dans le corpus diachronique Frantext

Agnieszka K. Kaliska

Institut des langues et littératures romanes,
Université Adam Mickiewicz, Poznan, Pologne
agnieszka.kaliska@amu.edu.pl

Résumé. La présente étude se situe au croisement entre la lexicographie, la linguistique de corpus, et l'histoire de l'art, et vise à présenter les résultats d'analyses consistant à examiner l'évolution de la fréquence du bleu et de ses nuances — étudiés à travers 46 termes français, noms de couleurs, de pigments et de colorants, par exemple : *bleu céleste*, *bleu de Prusse*, *bleu Nattier*, *bleu turquin*, *smalte* — dans le corpus diachronique Frantext. Les chiffres obtenus ont été ajustés par centenaires et par genres textuels, et puis, comparés avec ceux d'autres ressources permettant également d'observer l'évolution des mots au fil du temps.

Abstract. *This study is situated at the crossroads of lexicography, corpus linguistics and art history, and aims to present the results of an analysis examining the popularity of the color blue and its various shades—studied through various French terms, names of colors, pigments, and dyes, including* i.a. bleu céleste*,* bleu de Prusse*,* bleu Nattier*,* bleu turquin*,* indigotine*,* smalte*—in Frantext diachronic corpus. Subsequently, the results are compared with those of other resources that also allow for the observation of word evolution over time.*

1. Introduction

L'objectif du présent article est de présenter les résultats d'analyses visant à examiner l'évolution de la fréquence du bleu et de ses nuances — étudiés à travers différents termes français, noms de couleurs, de pigments et de colorants, par exemple : *bleu céleste*, *bleu de Prusse*, *bleu Nattier*, *bleu turquin*, *indigotine*, *smalte* — dans le corpus Frantext qui réunit non seulement des textes littéraires mais aussi des textes appartenant aux domaines des sciences, des arts et des techniques (10 % du corpus entier).

Le recours au corpus Frantext est motivé par le fait qu'au cours des siècles, autrement qu'aujourd'hui, pour introduire et définir un terme dans un dictionnaire de langue, il fallait consulter un ou plutôt des experts, peut-être aussi des ressources écrites. Toujours est-il que le travail de lexicographes anciens fut surtout un travail d'introspection. Or, dans nos recherches précédentes (Kaliska *et al.*, 2022), en nous basant sur neuf éditions du *Dictionnaire de l'Académie Française* (id. depuis 1694 jusqu'à nos jours), nous avons démontré que le nombre des termes techniques se référant aux couleurs et matériaux bleus augmentait d'une édition à l'autre. Deux éditions se sont avérées particulièrement révolutionnaires en ce qui concerne l'introduction des mots techniques et scientifiques dans le dictionnaire de bel usage, à savoir celle de 1835 et la plus récente, la neuvième édition. Nous nous proposons alors de continuer la recherche entamée, en nous focalisant cette fois-ci sur la présence des termes collectés dans le corpus de langue périodisé.

La présente étude s'articule ainsi au croisement triple entre la lexicographie, la linguistique de corpus, et l'histoire de l'art. Ce dernier axe trace l'histoire de la croissante popularité du bleu au cours des siècles. Longtemps réservée à la Vierge Marie, la couleur bleue est devenue au fil du temps un composant essentiel de la mode vestimentaire de l'aristocratie (*cf.* Pastoureau, 2000, 15). Les teinturiers et les chimistes rivalisaient pour trouver de nouveaux pigments, à la fois intenses et de longue tenue, ce qui favorisait le développement d'une terminologie nuancée se référant à de vastes matériaux et différentes tonalités du bleu. Le bleu devenait naturellement de plus en plus populaire. L'histoire de son succès, ici très brièvement esquissée, constitue donc une de trois axes d'orientation pour la présente étude.

2. Méthodes d'analyse

46 termes dont la présence a été confirmée dans au moins une des neuf éditions du DAF (*cf.* Gostkowska *et al.*, 2022) ont été recherchés dans le corpus Frantext, sur une période de 423 ans (c'est-à-dire de 1600 à 2023), afin de vérifier si les termes notés par les lexicographes du DAF apparaissent dans le corpus permettant d'observer l'usage de mots spécialisés, devenus courants, au fil des siècles.

Voici une liste complète de termes que nous avons sélectionnés pour la présente étude : *bleu ardoise, bleu argent, bleu azur, bleu barbeau, bleu canard, bleu céleste, bleu ciel, bleu d'azur, bleu d'émail, bleu d'empois, bleu d'inde, bleu de ciel, bleu de cobalt, bleu de méthylène, bleu de molybdène,*

bleu de montagne, bleu de nuit, bleu de Prusse, bleu de roi, bleu de safre, bleu de Saxe, bleu de Sèvres, bleu de Vermeer, bleu horizon, bleu marine, bleu mourant, bleu Nattier, bleu nuit, bleu pers, bleu pervenche, bleu pétrole, bleu roi, bleu turquin, bleu turquoise, carmin d'indigo, cendres bleues, céruléen, indigo des Indes (occidentales), indigo des Indes (orientales), indigo minéral, indigotine, lazurite, outremer, outremer artificiel, safre, smalte.

Les chiffres obtenus ont été ajustés par centenaires et par genres textuels, et adaptés de manière à illustrer les résultats obtenus en forme de tableaux et d'historiogrammes. Nous les avons comparés ensuite aux résultats obtenus récemment dans une autre étude, notamment celle consacrée à la terminologie de la couleur bleue en diachronie longue selon Google Livres Ngram Viewer (Kaliska, 2025).

3. Résultats

3.1. Statistiques des termes par époques et par genres textuels

Le corpus Frantext est échelonné du X^e^ au XXI^e^ siècle. Les textes qui datent du XX^e^ siècle ont la plus grande représentation dans Frantext, il en est de même pour le nombre de termes qui y ont été retrouvés (*cf.* Fig. 1).

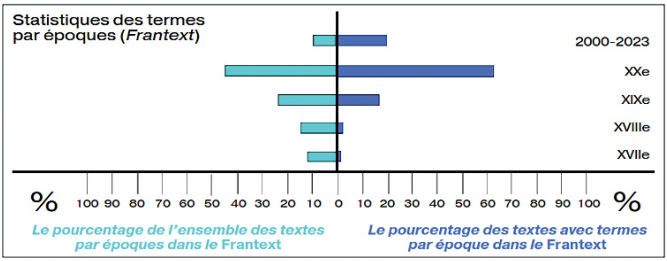


Fig. 1. Textes et termes par époque dans le corpus Frantext.

Un pourcentage important de termes apparaît dans les textes publiés au XXI^e^ siècle. Ceci peut, d'un côté, être la conséquence de l'expansion des sciences qui accélère nettement depuis l'époque industrielle, mais aussi d'une démocratisation des savoirs techniques qui en découle. Les termes sont devenus de plus en plus courants tout en se faufilant dans différents types de contextes. La répartition d'emplois retrouvés par genres textuels montre que les termes

étudiés sont deux fois plus représentés dans les écrits *fictionnels* que dans les écrits *non fictionnels* et les écrits dont le genre textuel figure comme *non renseigné* (*cf.* Fig. 2). Notons cependant que les textes de cette dernière catégorie appartiennent souvent au domaine technique.

Et puisque nous avons affaire aux termes naturellement polysémiques, il n'est guère étonnant que ce soit la valeur "*couleur (tonalité)*" qui soit plus volontiers actualisée dans un texte littéraire que la valeur "*couleur (matériau)*" qui se réalisera dans un texte appartenant au domaine technique, par exemple :

> (1) Texte technique : *C'est une liqueur âcre, extraordinairement caustique, qui verdit le sirop de violettes, détruit promptement le tissu cellulaire, les cheveux, les fragmens d'os et autres substances animales solides, saponifie les huiles grasses, décolore le* ***bleu de Prusse****, dissout le soufre par l'ébullition, et se mêle à l'alcool sans manifestation de précipité.* (J.-B. Caventou, *Manuel des pharmaciens et des droguistes*, t. 2, 1821, p. 575. Source : Frantext)

> (2) Texte littéraire : *Comme c'était le dimanche, les bœufs étaient à l'étable et les laboureurs sur le pas de la porte, dans leurs habits de fête, c'est-à-dire en gros drap* ***bleu de Prusse****, de la tête aux pieds*. (G. Sand, *Le Meunier d'Angibault*, 1845, p. 69. Source : Frantext)

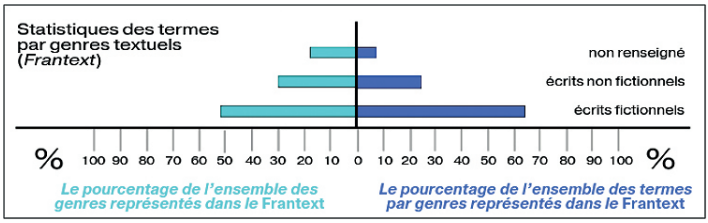


Fig. 2. Textes et termes par époque dans le corpus Frantext.

3.2. Termes les plus fréquents

Nous observons une correspondance entre les chiffres obtenus pour les termes les plus fréquents, à savoir *bleu marine* (576 occ.) et *bleu ciel* (375 occ.), ce dernier avec pour variante *bleu de ciel* (187 occ.), et les résultats obtenus à l'aide de Google Ngram Viever que nous avons utilisé pour explorer la sous-collection française du corpus Google Livres, également périodisée (*cf.* Kaliska, 2025).

Selon le *Glossaire des matériaux de la couleur* de Guineau (2005, p. 134), les termes *bleu ciel* et *bleu de ciel* sont attribués, respectivement, au pigment artificiel composé d'un mélange coprécipité de stannate de cobalt et de sulfate de calcium, et à la couleur bleu clair qui, en peinture, peut d'ailleurs être obtenue par différentes méthodes, tel un mélange de bleu de Prusse et de blanc de plomb. Cette deuxième signification a certainement prévalu en faveur de *bleu ciel* dans le corpus des textes majoritairement fictionnels, d'autant plus que c'est un terme très iconique qui évoque tout de suite la couleur claire du ciel. Cette nuance tellement appréciée peut également être désignée par d'autres variantes comme *bleu céleste*, autrefois plus fréquent, selon Frantext (*cf.* Tab. 1).

	1600-1699	**1700-1799**	**1800-1899**	**1900-1999**	**2000-2023**
bleu céleste	4	18	26	15	7
bleu de ciel	0	0	138	45	4
bleu ciel	0	0	27	269	86

Tab. 1. Fréquence de bleu ciel, bleu de ciel et bleu céleste dans Frantext.

Le terme *bleu marine* désigne une couleur obtenue en peinture à l'aide d'un mélange de bleu d'outremer et de noir d'ivoire, ou de bleu d'indigo en teinture (Guineau, 2005, p. 138). Les résultats de Frantext témoignent d'une popularité indéniable de *bleu marine* depuis le XX^e^ siècle, cette nuance de bleu foncé caractéristique qui a été adoptée par les corps de marine (*cf.* Tab. 2).

	1600-1699	**1700-1799**	**1800-1899**	**1900-1999**	**2000-2023**
bleu marine	0	0	15	364	197

Tab. 2. Fréquence de bleu marine dans Frantext.

3.3. Les absents

Certains termes notés dans le DAF n'ont pas été trouvés dans Frantext : *bleu d'empois*, *bleu d'inde*, *bleu de molybdène*, *bleu de safre*, *bleu de Saxe*, *indigo minéral*, *outremer artificiel*.

4. Conclusion

Si les dictionnaires d'autrefois se basaient essentiellement sur un travail d'introspection, les grands corpus diachroniques permettent aujourd'hui de confronter les décisions des lexicographes aux discours d'époques. Quant à la présence des termes étudiés dans le corpus général, on peut remarquer que dans les textes littéraires, les emplois *couleur* dominent, contrairement aux textes techniques et scientifiques où dominent les emplois *matériau* (*i.e. pigment*, *colorant*). La popularité de ces derniers est limitée à certaines époques et due au succès des propriétés colorantes de matériaux désignés (par exemple : *bleu marine*). Quant aux noms de couleurs, le caractère iconique des appellations telles que *bleu ciel* semble primordial. Certaines correspondances qui se laissent observer entre les résultats obtenus à l'aide de différentes ressources et différents outils permettant d'observer le fonctionnement des mots au fil du temps confirment que les unités terminologiques sont des unités dynamiques (Cabré, 2005) dont la signification — se construisant dans le discours et en fonction de celui-ci — privilégie certaines valeurs au détriment des autres.

Références

Cabré, Teresa. 2005. « La terminologie, une discipline en évolution : le passé, le présent et quelques prospectives » [trad. par M.-C. L'Homme], *Debate Terminologico*, nº 1.

Dictionnaires de l'Académie française *:* https://www.dictionnaire-academie.fr/.

Frantext : https://www.frantext.fr/.

Google Livres Ngram Viewer : https://books.google.com/ngrams/.

Gostkowska, Kaja, Kaliska, Agnieszka. 2022. « L'histoire d'une couleur vue à travers un dictionnaire. Sur l'exemple des termes de couleur BLEU dans les neuf éditions du *Dictionnaire de l'Académie française* ». Dans *SHS Web of Conferences*, nº 138 (Actes du 8e congrès mondial de Linguistique française).

Guineau, Bernard. 2005. *Glossaire des matériaux de la couleur et des termes techniques employés dans les recettes de couleurs anciennes*. Turnhout : Brepols Publishers.

Kaliska, Agnieszka, Gostkowska, Kaja. 2022. « La présence des termes de couleurs et l'évolution de leurs définitions dans les neuf éditions du *Dictionnaire de l'Académie* », *Les Cahiers du dictionnaire*, nº 14. Paris : Classiques Garnier, p. 471-491.

Kaliska, Agnieszka. 2025. « Terminologie de la couleur bleue en diachronie longue selon Google Livres Ngram Viewer », *Corpus*, nº 26. doi.org/10.4000/13653

Pastoureau, Michel. 2000. *Bleu. Histoire d'une couleur*. Paris : Éditions du Seuil.

Sorba, Julie, Kraif, Olivier, Renwick, Adam et Denoyelle, Corinne. (2024). « La constitution de corpus en diachronie longue : méthodologies, objectifs et exploitations linguistiques et stylistiques », *Corpus*, nº 24. doi.org/10.4000/corpus.8461

Agnieszka K. Kaliska

Terminologie de la couleur bleue dans les corpus diachroniques de la langue française

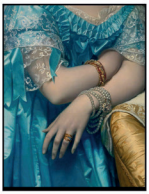
Agnieszka Kaliska
Université Adam Mickiewicz à Poznan

INTRODUCTION

Longtemps réservé à la Vierge Marie, symbolisant le deuil de la Mère de Jésus, le bleu est peu à peu devenu la couleur favorite de l'aristocratie (cf. Pastoureau, 2000, 15). De nos jours, souvent associé à l'amitié et à la sympathie, il est la couleur emblème de l'Union Européenne.

L'objectif du présent poster est de présenter les résultats d'analyses consistant à **examiner l'évolution de la fréquence du bleu et de ses nuances** – étudiés à travers différents termes français, noms de couleurs, de pigments et de colorants, p. ex. : ***bleu céleste, bleu d'outremer, bleu de Vermeer, bleu turquin, indigotine, bleu de safre, smalte*** – **dans le corpus Frantext** qui réunit non seulement des textes littéraires mais aussi **des textes appartenant aux domaines des sciences, des arts et des techniques** (10% du corpus entier) ; ensuite de comparer les résultats obtenus avec ceux d'autres ressources.

MÉTHODES

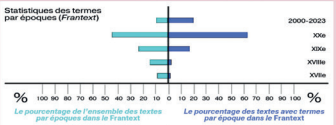


- **46 termes** dont la présence a été confirmée dans au moins une des neuf éditions du DAF (cf. Gostkowska & Kaliska, 2022) **ont été recherchés dans le corpus Frantext, sur une période de 373 ans (c.-à-d. de 1650 à 2023)**, afin de vérifier si les termes notés par les lexicographes du DAF apparaissent dans un corpus permettant d'observer l'usage de mots spécialisés, devenus courants, au fil des siècles ;
- **les chiffres ont été ajustés par centenaires et par genres textuels**, et adaptés de manière à illustrer les résultats obtenus en forme de **tableaux**, d'**historiogrammes** et de **nuages de points** ;
- **les résultats ainsi illustrés ont été comparés** à ceux obtenus précédemment pour les dictionnaires, le DAF en particulier (Gostkowska & Kaliska, 2022), et ceux obtenus récemment dans une étude consacrée à la terminologie de la couleur bleue en diachronie longue selon Google Livres Ngram Viewer (Kaliska, 2025 – à paraître).

Phot. *Extrait du portrait de Joséphine-Éléonore-Marie-Pauline de Galard de Brassac de Béarn (1825–1860), Princesse de Broglie par Jean-Auguste-Dolminique Ingres.*

RÉSULTATS

Termes les plus fréquents dans le Frantext (total: 1838 emplois)	
bleu marine	576
bleu ciel	375
*bleu de ciel***	187
outremer	199
*bleu nuit**	146
*bleu de nuit***	51
bleu roi	99
*bleu de roi***	59
bleu de Prusse	93
*bleu horizon**	90
bleu céleste	74
bleu turquoise	70
*bleu de turquoise***	10
bleu de méthylène	57
bleu d'azur	38
bleu azur	35
*bleu pétrole**	32
céruléen	30
*bleu pervenche**	27
*bleu de pervenche***	7
*bleu ardoise**	23
*bleu d'ardoise***	10
*bleu ardoisé***	3
*bleu canard**	16
bleu de cobalt	15
*bleu cobalt***	14

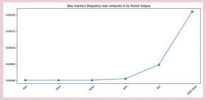


Termes notés dans le DAF qui n'ont pas été trouvés dans Frantext : *bleu d'empois, bleu d'Inde, bleu de molybdène, bleu de safre, bleu de Saxe, indigo minéral, outremer artificiel.*

VARIANTES : Le corpus note, en revanche, de nombreuses variantes, p.ex.: *smalt* (3) vs. *bleu de smalt* (1) vs. *smalte* (1). Ce dernier est noté dans le DAF et le Glossaire de Guineau (article: *smalte*) ; *bleu de smalt* apparaît dans l'article *bleu azur* (Guineau, 2005, p. 132).

* termes en bleu ont été trouvés dans le DAF, ils n'apparaissent pas pour autant dans le *Glossaire* terminologique de Guineau (2005) ;
** variantes en rouge magenta ne sont notés ni dans le DAF, ni dans le Glossaire, mais ils ont été trouvées dans le corpus

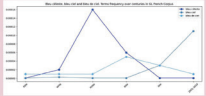
Exemple (source: **texte technique**): *C'est une liqueur âcre, extraordinairement caustique, qui verdit le sirop de violettes, détruit promptement le tissu cellulaire, les cheveux, les fragments d'os et autres substances animales solides, saponifie les huiles grasses, décolore* ***le bleu de Prusse****, dissout le soufre par l'ébullition, et se mêle à l'alcool sans manifestation de précipité.* **(1821)** Exemple (source: **texte littéraire**): *Comme c'était le dimanche, les boeufs étaient à l'étable et les laboureurs sur le pas de la porte, dans leurs habits de fête, c'est-à-dire en gros drap* ***bleu de Prusse****, de la tête aux pieds.* **(1845)**



1600-1699	1700-1799	1800-1899	1900-1999	2000-2024
0	0	15	364	197

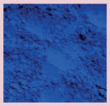
Les plus hauts pics de croissance pour un siècle donné (fréq. relative) selon GL Ngram Viewer *vs.* le nombre de textes avec *bleu marine* dans le Frantext.

Bleu marine*, couleur* obtenue en peinture à l'aide d'un mélange de bleu d'outremer et de noir d'ivoire, ou de bleu d'indigo en teinture (Guineau, 2005: 138).



	1600-1699	1700-1799	1800-1899	1900-1999	2000-2024
bleu céleste	4	18	26	15	7
bleu ciel	0	0	27	269	86
bleu de ciel	0	0	138	45	4

Les plus hauts pics de croissance pour un siècle donné (fréq. relative) selon GL Ngram Viewer *vs.* le nombre de textes du Frantext avec *bleu céleste, bleu ciel* et *bleu de ciel* (pigments différents de couleur similaire selon Guineau).



DISCUSSION & CONCLUSIONS

Si les dictionnaires anciens basaient essentiellement sur un travail d'introspection, les corpus diachroniques permettent aujourd'hui de confronter les décisions des lexicographes aux discours d'époques.

travail d'introspection → *dictionnaires* → *corpus*

glossaires terminologiques et/ou consultations avec des experts

- Quant à la présence de la *terminologie de la couleur bleue* dans le corpus général :
  - dans les textes **littéraires**, les emplois *couleur* dominent, contrairement aux textes **techniques et scientifiques** où dominent les emplois *pigment, colorant* et *substance* ;
  - les plus hauts pics de fréquence témoignent d'une grande popularité des termes étudiés à certaines époques ; pour les noms de pigments, cette popularité serait due au succès de leurs propriétés colorantes (*bleu marine, bleu de Prusse*) ; quant aux noms de couleurs, le caractère iconique des appellations telles que *bleu céleste, bleu ciel* ou *bleu de ciel* semble primordial ; à l'usage, différentes variantes se laissent d'ailleurs observées (y compris celles qui n'ont pas été notées dans le *Glossaire*) : *N (de)E*) *N, N Adj* ;
- une corrélation entre les résultats obtenus sur le Frantext et ceux obtenus sur le GL Ngram Viewer n'est
- que partielle mais les mesures employées sont un peu différentes (dépendent des fonctionnalités de l'outil).

« Les unités terminologiques ne peuvent plus être perçues uniquement comme des unités de représentation et de transmission d'un savoir précis homogènes et totalement contrôlées mais plutôt comme des unités dynamiques qui contribuent à la construction des connaissances par leur utilisation en discours. » (Cabré, 2005)

RÉFÉRENCES

Cabré T. (2005). La terminologie, une discipline en évolution : le passé, le présent et quelques prospectives [trad. par M.-C. L'Homme]. *Debate Terminológico*, 1.

Guineau B. (2005). *Glossaire des matériaux de la couleur et des termes techniques employés dans les recettes de couleurs anciennes*. Turnhout: Brepols Publishers.

Google Livres Ngram Viewer – en ligne : https://books.google.com/ngrams/

Gostkowska K., Kaliska A. (2022). « La présence des termes de couleurs et l'évolution de leurs définitions dans les neuf éditions du *Dictionnaire de l'Académie* », *Les Cahiers du dictionnaire* 2022/14 : 471-491. Paris : Classiques Garnier.

Kageura K. (2015). Terminology and lexicography. In: Theoretical Perspectives on Terminology (eds. P. Faber, M.-C. L'Homme). John Benjamins Publishing Company. DOI: 10.1075/hoct.l.04ker2

Pastoureau M. (2000). *Bleu. Histoire d'une couleur*. Paris: Éditions du Seuil.

Sorba J., Kraif O., Renwick A. et Denoyelle C. (2024). « La constitution de corpus en diachronie longue : méthodologies, objectifs et exploitations linguistiques et stylistiques », *Corpus* [En ligne]. DOI : https://doi.org/10.4000/corpus.8461 [consulté le 06 mai 2024]

TOTh 2024

Achevé d'imprimer en décembre 2025
Imprimerie Présence Graphique
2 rue de la Pinsonnière
F - 37260 MONTS