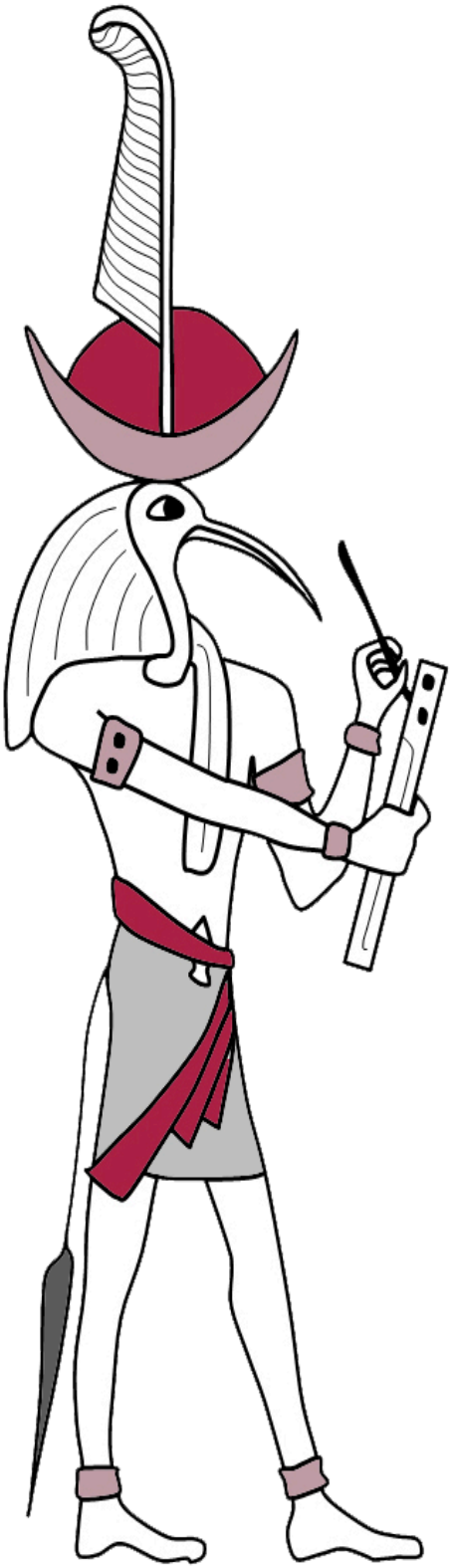


TOTh 2026

Terminology & Ontology: Theories and Applications

20th Anniversary Edition of the Longest-Running International Conference on Terminology

Book of Abstracts

Dates:
June 4 & 5, 2026

Location:
University Savoie Mont-Blanc (France) &
Online

Chairman:
Christophe Roche, University of Crete & Université Savoie Mont-Blanc

Organizing Committee:
Maria Papadopoulou, University of Crete
Silvia Piccini, Italian National Research Council
Rafail Giannadakis, University of Crete

Organized by:
University of Crete – TALOS AI for SSH
Project number 101087269

With the support of:
University of Crete
Université Savoie Mont-Blanc – Polytech Annecy Chambéry
Ass.I.Term – Associazione Italiana per la Terminologia

Table of Contents

Session 3

Session 4

Poster Session 1

Poster Session 2

Foreword

The 20th International Conference TOTh – Terminology & Ontology: Theories and Applications is a special anniversary edition of the longest-running international conference dedicated to terminology. Held onsite and online on 4 and 5 June 2026 at Université Savoie Mont-Blanc, France, TOTh 2026 celebrates twenty years of exchange on terminology, ontology, language and knowledge.

This Book of Abstracts presents the contributions selected for TOTh 2026. The programme reflects the breadth of current research in the field, from terminology, translation, corpus linguistics and natural language processing to ontology, knowledge engineering, semantic technologies, digital humanities, cultural heritage and artificial intelligence. The opening talk, "On the 20th Anniversary of the TOTh Conference" by Dr Danielle Candel (CNRS, Université Paris Cité), invites participants to reflect on two decades of TOTh while considering future directions for research and practice. The conference is organized in the context of the TALOS AI for SSH project, funded by the European Union under Horizon Europe (Grant Agreement No. 101087269), with the support of Polytech Annecy Chambéry, the University of Crete, and Ass.I.Term – Associazione Italiana per la Terminologia.

Avant-propos

Organisée en présentiel et en ligne les 4 et 5 juin 2026 à l'Université Savoie Mont-Blanc (France), TOTh 2026, la plus ancienne conférence internationale dédiée à la terminologie, sera l'occasion de célébrer vingt années d'échanges autour de la terminologie, de l'ontologie, de la langue et de la connaissance.

Ce livre des résumés présente les contributions sélectionnées pour TOTh 2026. Le programme reflète la diversité des recherches actuelles dans le domaine, de la terminologie, de la traduction, de la linguistique de corpus et du traitement automatique des langues à l'ontologie, l'ingénierie des connaissances, les technologies sémantiques, les humanités numériques, le patrimoine culturel et l'intelligence artificielle. La conférence d'ouverture, « On the 20th Anniversary of the TOTh Conference », par Danielle Candel (CNRS, Université Paris Cité), invite les participants à revenir sur deux décennies de TOTh tout en envisageant les perspectives futures de la recherche et des pratiques. La conférence est organisée dans le cadre du projet TALOS AI for SSH, financé par l'Union européenne au titre d'Horizon Europe (convention de subvention no 101087269), avec le soutien de Polytech Annecy Chambéry, de l'Université de Crète et d'Ass.I.Term – Associazione Italiana per la Terminologia.

Program Committee

Manuel Alcántara-Plá, Universidad Autónoma de Madrid – Spain
Amparo Alcina, Universitat Jaume I – Spain
Xiaomi An, Renmin University – China
Albina Auksoriūtė, The Institute of the Lithuanian Language – Lithuania
Mohamed Sahbi Baazaoui, Al Wasl University – United Arab Emirates

Bruno Bachimont, Université de technologie de Compiègne – France
Andrea Bellandi, Italian National Research Council – Italy
Harry Bunt, Tilburg University – Netherlands
Danielle Candel, CNRS, Université Paris Cité – France
Sylviane Cardey, Université de Franche-Comté – France
Stéphane Chaudiron, Université de Lille – France
Valérie Delavigne, Université Paris 3 – France
Sylvie Despres, Université Paris 13 – France
Juan Carlos Díaz Vásquez, EAFIT University – Colombia
Giorgio Maria Di Nunzio, Università degli Studi di Padova – Italy
Caroline Djambian, Université Grenoble Alpes – France
Antoine Doucet, La Rochelle Université – France
Iris Eshkol-Taravella, Université Paris Nanterre – France
Pamela Faber, Universidad de Granada – Spain
Christiane Fellbaum, Princeton University – USA
Ágota Fóris, Károli Gáspár University – Hungary
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
François Gaudin, Université de Rouen – France
Laurent Gautier, Université de Bourgogne – France
Teodora Ghiviriga, Alexandru Ioan Cuza University – Romania
Gernot Hebenstreit, University of Graz – Austria
John Humbley, Université Paris Cité – France
Yangli Jia, Liaocheng University – China
Olha Kanishcheva, University of Jena – Germany
Barbara Inge Karsch, BIK Terminology – USA
Hendrik Kockaert, University of Leuven – Belgium
Natalie Kübler, Université Paris Cité – France
Héba Lecocq, Université Sorbonne Nouvelle – France
Hélène Ledouble, Université de Toulon – France
Patrick Leroyer, Aarhus University – Denmark
Georg Löckinger, University of Applied Sciences Upper Austria – Austria
Elpida Loupaki, Aristotle University of Thessaloniki – Greece
Rodolfo Maslias, Vice-President of TermNet
Christine Michaux, Université de Mons – Belgium
Ouafae Nahli, Italian National Research Council – Italy
Fidelma Ní Ghallchobhair, Foras na Gaeilge, Irish-Language Body – Ireland
Henrik Nilsson, C.A.G. Mawell – Sweden
Vincent Nyckees, Université Paris Cité – France
Mavina Pantazara, National and Kapodistrian University of Athens – Greece

Maria Papadopoulou, University of Crete – Greece
Mojca Pecman, Université Paris Cité – France
Sandrine Peraldi, University College Dublin – Ireland
Platon Petridis, National and Kapodistrian University of Athens – Greece
Silvia Piccini, Italian National Research Council – Italy
Aurélie Picton, Université de Genève – Switzerland
Suzanne Pinson, Université Paris Dauphine – France
Marina Platonova, Riga Technical University – Latvia
Alain Polguère, Université de Lorraine – France
María Pozzi, El Colegio de México – Mexico
Bihua Qiu, China National Committee for Terms in Sciences and Technologies – China
Renato Reinau, Université de Genève – Switzerland
Christophe Roche, University of Crete – Greece / Université Savoie Mont Blanc – France
Mathieu Roche, CIRAD – France
Laurent Romary, INRIA & HUB-ISDL – Germany
Micaela Rossi, Università degli Studi di Genova – Italy
Antonio San Martín Pizarro, Université du Québec à Trois-Rivières – Canada
Elsabe Taljard, University of Pretoria – South Africa
Rita Temmerman, Vrije Universiteit – Belgium
Anne Theissen, Université de Strasbourg – France
Wei Tong, Peking University – China
Katerina Toraki, ELETO – Greece
Federica Vezzani, Università degli Studi di Padova – Italy
Kara Warburton, Université du Québec à Trois-Rivières – Canada
Lotte Weilgaard Christensen, University of Southern Denmark – Denmark
Tanja Wissik, Austrian Academy of Sciences – Austria
Maria Teresa Zanola, Università Cattolica del Sacro Cuore – Italy

Opening Talk

On the 20th Anniversary of the TOTh Conference

Danielle Candel
CNRS group HTL (Histoire des théories linguistiques),
Université Paris Cité—France
danielle.candel@u-paris.fr – https://htl.cnrs.fr/equipe/danielle-candel/

Abstract

In this brief overview of the TOTh conference and its twenty-year history we intend to show how TOTh has innovated in highlighting a field that can be described as « multifaceted, » « multi-phenomenal, » and constantly evolving, that of ontology. To do this, it is informative to examine the specialized areas that have been covered, along with their main characteristics, and also to consider the tools that have been used. It is also enlightening to discuss the languages that were used to extract the elements analyzed in the TOTh published chapters to show how they enrich the analysis. The selected data are useful for illustrating the definitions of ontology, ontoterminology, and terminology. It is of course important to first deal with these definitions and more specifically consider the term « ontoterminology », that deserves to be understood. One may look at the definition provided at https://www.lalanguefrancaise.com in 2024, but it is also fit to refer to Christophe Roche's definition given below. An attempt will be made to characterize specificities of ontoterminology and terminology, whether by language or by specialty. Terminology itself also needs to be defined. In these fields, the growing role of artificial intelligence has to be highlighted. This retrospective, covering a period of twenty years – a relatively short timeframe – will allow, among other things, to highlight progress in a field as rich as ontology, while also emphasizing the difficulty of defining this concept. The study will revisit links between language and knowledge and the conceptual advances in this area. It will make extensive use of some reported experiences, highlighting lessons learned, problems described, and solutions developed. One type of analysis will be invaluable in this synthesis: that of the opening lectures of the TOTh conferences and in this respect it is logical to begin with the first one (given by Christophe Roche in 2007) which sets the ground for TOTh « basics »: « The term ontoterminology reflects this need to place the concept and its name at the center of terminology (…) » (Ch. Roche). As for the disciplines associated with terminology and ontology, they include translation, lexicology, knowledge engineering, information science. All of which are both long-established concepts and more recent items.

Biography

Danielle Candel (danielle.candel@u-paris.fr – https://htl.cnrs.fr/equipe/danielle-candel/) is now an Associate Member at the CNRS group HTL (Histoire des théories linguistiques), Université Paris Cité. She was first a lexicographer at the Trésor de la langue française dictionary, CNRS. Then she was asked to create and lead a team which worked for all domains covered by the French official Terminology framework ("Délégation Générale à la langue française et aux langues de France"), for about fifteen years, with an outcome published in the "FranceTerme" database. She

is still engaged in official Terminology as a linguist expert. Her research is mostly connected with terminological and lexicographical issues described in two recent chapters: “Terminology in France: evolution of its Official framework”, Benjamins 2025; “General principles of Wüster’s general Theory of terminology”, Benjamins 2022. She codirected or cosigned for instance « Eugen Wüster et la terminologie de l’École de Vienne », HEL Livres 2023 ; « Prescriptions en langue », HEL XL1,2, 2019; « La prescription linguistique : applications et réactions », ELA 191, 2018.

Session 1

Parallel Wisdom:
Modeling Ancient Greek and Chinese Philosophers

Rafail Giannadakis, giannadakis.uni@gmail.com; TALOS Lab, University of Crete, Greece
Maria Papadopoulou, maria.papadopoulou@uoc.gr; TALOS Lab, University of Crete, Greece
Hui Liu, luisaliu0339@gmail.com; College of Foreign Languages, Nanjing University of Aeronautics and Astronautics, China

Keywords

ancient Greek philosophy; ancient Chinese philosophy; ontoterminology; RDF graph; multilingual terminology

Table of Contents

Extended Abstract

1. Introduction & Motivation

This paper presents an interoperable dataset in which ancient Greek and early Chinese philosophical traditions are modeled within a shared formal structure, enabling comparative study to become findable, accessible, interoperable, reusable, and queryable (Berners-Lee 2006; Wilkinson et al. 2016). The dataset covers philosophers, works, places, schools, and time periods from both traditions. Its aim is to provide a structured basis for comparing aspects of ancient Greek and Chinese philosophy through shared entity types and explicitly modeled relations.

The modeled domain of comparative Greek – Chinese scholarship is well established, yet it remains difficult to operationalize at scale because of challenges of cross-cultural comparability, categorization, and terminology. The two traditions are mediated by different naming conventions, chronologies, and textual transmission histories. Because Greek and Chinese traditions organize knowledge differently and are mediated by distinct terminological systems, even simple research tasks become difficult when we attempt to compare them systematically or model them in formal data structures. Work on Greece – China comparison has repeatedly highlighted the methodological risks of projecting categories across cultures, while still arguing for the value of controlled, explicit comparison (Lloyd 2004; 2018).

To address this, the project adopts a multilingual approach to philosophical inquiry. Entities are enriched with multilingual labels and terminological variants, enabling cross-lingual discovery while preserving philological specificity. This approach supports plural interpretive entry points into the data and reflects broader calls within comparative philosophy for more inclusive and globally grounded frameworks of analysis. Comparative analysis often relies on local, manual harmonizations that are hard to inspect, reuse, or extend. A knowledge-graph approach addresses this tension by making relations visible and testable without presupposing that the underlying concepts are equivalent.

The innovation of the project is methodological. It translates domain knowledge into structured and linked data within a Semantic Web framework (Berners-Lee, Hendler, and Lassila 2001). A knowledge graph can represent, and allow users to query, relations such as who authored which work, which school a thinker is associated with, where and when a person is active, which texts mention which thinkers, and how intellectual lineages and influences are recorded. Such queries are not an alternative to interpretation. They are a way to surface patterns, gaps, and competing attributions as explicit statements that can be checked and revised.

Methodologically, the work maintains a strict distinction between conceptual modeling and instantiation. Conceptual modeling defines the classes and properties that structure the domain. Instantiation populates that structure with named individuals and links them to sources and identifiers. This separation supports transparency and long-term maintenance, and aligns with ontoterminology, which couples a formal conceptual system with a controlled linguistic layer while keeping concept definition distinct from term usage (Roche 2007; Roche et al. 2009; Roche 2012). Implementation is carried out in TEDI, a tool designed to support domain experts by separating conceptual and linguistic dimensions and by supporting definition through essential characteristics, in line with terminology standards (Roche & Papadopoulou 2019; Roche 2025; Milio, Giannadakis, & Lourentzaki 2025).

The output is designed for both machine and human use. For machine processing, data are published as RDF and aligned with SKOS and OntoLex-Lemon, with SPARQL used for evaluation and scholarly retrieval (W3C 2005; W3C 2013; W3C 2016). For human-facing access, the same model supports HTML exports and teaching-oriented visualizations such as OntoDictionaries, maps, object collections, and graphs. Development is iterative: an initial phase prioritized broad population of core entities; subsequent phases prioritize deeper and more balanced comparative modeling across Greek and Chinese sides (Giannadakis, Papadopoulou, & Roche 2024).

2. Related work

Related work in humanities ontologies largely emphasizes cultural heritage description and metadata interoperability (objects, places, agents, events), which is essential for aggregation and preservation but usually treats "philosophy" as topical metadata rather than as a domain whose internal structures can drive comparative research questions.

Within philosophy-specific infrastructures, the Internet Philosophy Ontology (InPhO) is of key importance. InPhO targets the discipline of philosophy as a dynamic, computational ontology that is iteratively built from large-scale text resources, especially the Stanford Encyclopedia of Philosophy, combined with expert feedback and automated reasoning (Murdock and Allen 2011). In their published description, InPhO's workflow begins with a small manually crafted seed lexicon, then uses statistical inference over SEP content to propose relations among topics and synthesizes expert evaluations via logic-based reasoning (notably answer set programming) to generate a usable taxonomy for navigation and tooling (Murdock and Allen 2011). InPhO is also infrastructure-forward: it exposes entities as web resources with stable URIs and provides a RESTful API that returns both human-facing HTML and machine-friendly JSON, while still supporting OWL exchange for archival and interoperability (Murdock and Allen 2011).

Earlier work frames the ontology as a philosophy-oriented semantic infrastructure and articulates its core modeling layers (e.g., Thinker, Idea, Document, Organization), aimed at organizing and navigating philosophy resources at scale (Buckner, Niepert, and Allen 2007). InPhO's scope, however, is not antiquity-first and not cross-tradition by design. Our contribution targets the remaining gap: a single model engineered for antiquity-focused comparison, where ancient Greek and ancient Chinese materials are typed in parallel and queried through shared patterns. Concretely, we implement a shared backbone (e.g., Person, Philosopher, Place, Period, Philosophical Work) with tradition-specific subtyping (ancient-Greek-Philosopher *vs* ancient-Chinese-Philosopher, corresponding places, works, and periods) and aligned relations (e.g., pupilOf/teacherOf, birthPlace/deathPlace, authorOf/mentions) so that comparable competency questions can be asked symmetrically across traditions.

3. Resources, workflow, and versioning

The ontology and dataset are built through an iterative workflow purpose-built for comparative humanities. This cycle moves from conceptualization, i.e., defining a unified structure for these two distinct philosophical traditions to population with instantiated data on philosophers, works, and places. Each cycle is validated against competency questions using SPARQL to ensure consistency and completeness. The process culminates in dual-format publication: machine-ready RDF/OWL and human-readable HTML for accessible scholarly review without specialist tooling.

The project's evolution reflects a strategic shift from broad initial coverage to deep comparative modeling.

– Early Version (KELKIP): Prioritized a wide population of both ancient Greek and Chinese entities to establish the shared backbone and test the viability of the upper-level categories.
– Version 1.0 (TALOS): Refined the schema and substantially deepened the Ancient Greek component. This phase focused on improving the consistency of relations and the density of structured links (e.g., school affiliations and intertextual references) required for rigorous prosopography (Giannadakis, Papadopoulou, and Roche 2024 – openly accessible via Zenodo). Detailed dataset metrics are also provided in the same repository, as well as on the TALOS website (https://talos-ai4ssh.uoc.gr/resources/open-datasets/).
– Version 2.0 (July 2026) (TALOS & NUAA): Currently in development, this phase retains the Greek refinements while undertaking a major enrichment and normalization of the Chinese component. The goal is to rebalance the graph so that both branches can be queried through stable, parallel patterns with improved granularity.

This development is supported by a distributed collaboration structure. Contributors from TALOS lead the ontological modeling and structural refinements, while the team at NUAA focuses on the expansion of the Chinese side of the dataset. As the project matures, the collaborative emphasis moves progressively from quantity—simply populating more entities—toward structural symmetry, ensuring that cross-cultural comparisons are grounded in consistent modeling decisions rather than superficial parallels.

4. Modeling the domain

In this section, we present the core conceptual structure of the Philosophers ontology. Developed within TEDI, the ontoTerminology EDItor, and exported as an RDF graph, the base formalizes the prosopography, intellectual geography, and textual transmission of ancient philosophical traditions. The section below represents a transitional stage between Version 1.0, which established a dense prosopography of Ancient Greek thinkers, and the forthcoming version focused on the enrichment of the Chinese part of the base. The design prioritizes logical consistency, utilizing a lightweight upper model that enforces disjointness between agents, creations, periods, and locations.

4.1. Core Classes

The ontology is organized around five primary concepts.

- Philosopher: This is the central concept, currently populated with over 700 individuals. To facilitate comparative analysis without collapsing distinct traditions, it is divided into two disjoint subconcepts: <Ancient-Greek-Philosopher> (e.g., *plato*, *aristotle*) and <Ancient-Chinese-Philosopher> (e.g., *confucius*, *yan ying*).
- Philosophical Work: Intellectual output is modeled as a distinct entity rather than a data property. Containing over 230 works, this class is similarly subclassed into Greek and Chinese variants (e.g., *<Ancient-Greek-Philosophical-Work>*).
- Place and School: The spatial dimension is captured by the Place class, containing over 200 locations from major city-states to smaller settlements. Intellectual affiliation is modeled through <School-of-Philosophy>, which categorizes 35 major traditions (e.g., pythagorean, stoic, neo-taoism).
- Period: Chronology is reified as an entity class (Period) populated with century-level instances (e.g., *5th_century_BCE*), allowing temporal assertions to be treated as object relations.

4.2. Core Relations

The research potential of the dataset lies in the object properties connecting these classes.

- Textual Production: Authorship is encoded via authoredBy. In addition, the model also captures doxography through mentions and isMentionedIn, creating graph edges between texts and the philosophers they discuss and/or mention (e.g. Platonic dialogues or Diogenes Laertius's work).
- Geography and Mobility: Biological lifecycles are mapped through birthPlace and deathPlace. Intellectual mobility is captured via stayedIn (and its inverse residentedBy).
- Time and Lineage: We employ a dual temporal strategy: centuryOfLiving links to <Period> instances for grouping, while the floruit data property stores specific active dates as literals, where available. Pedagogy and intellectual relationships are formalized via relationships such as pupilOf, hasInfluenced, representativeOf, memberOfPhilosophicalSchool, and their inverses.

4.3. Multilingual and Comparative Strategy

Although the resource remains under development, entities are enriched with multilingual literals, allowing a single entity to be queryable across languages (e.g., *Confucius* / 孔子). The shared backbone enables parallel querying; a query targeting the superclass <Philosopher> functions identically for both traditions.

4.4. Instantiation Patterns

The populated graph follows a strict metadata pattern. A standard <Philosopher> instance (e.g., *tauriscus*) includes type, centuryOfLiving, floruit (e.g., "2nd_century_BCE"), and lineage links. Greek place records (e.g., *nicomedia*) are enriched with pleiadesCoordinates for visualization and rdfs:seeAlso links to external resources, serving as geospatial hubs in the network.

5. Validation and Utilization

5.1 Validation

Validation is implemented as query-driven checking: we translate competency questions that reflect concrete research tasks and translate them into SPARQL queries. This serves two goals at once. It tests internal coherence (whether the model behaves as intended) and it demonstrates retrieval value for humanities workflows.

In the current version, the competency questions are deliberately grounded in the Greek side, because version 1.0 is more densely populated and normalized there. The same question-types are being carried over to the Chinese (zh) subset, where ongoing work focuses on aligning classes, places, schools, temporal entities, and naming conventions so that identical query patterns can run comparatively.

CQ1 Prosopography by place and time: Which philosophers are associated with a given place (as birthplace and or residence or stay), in which century did they live, and which school (if any) are they linked to.

CQ2 Prosopography by school and geography: For a given philosophical school, which philosophers belong to it, where were they born, where did they stay, and how are they distributed across centuries.

CQ3 Author–work completeness: For each philosopher, which philosophical works are attributed to them, and for each work what basic metadata is available (title label, associated resources or identifiers when present).

CQ4 Mentions and reception: Which philosophical works mention a given philosopher, and which other philosophers are co-mentioned in the same works.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ont:  <http://www.ontologia.fr/OTB/Philosophers#>
PREFIX otv:  <http://www.ontologia.fr/OTB/otv#>

SELECT DISTINCT ?name ?century ?workTitle
FROM <http://ontologia.fr/OTB/Philosophers.rdf>
WHERE {
  ?x otv:instanceOf ont:Ancient-Greek-Philosopher ;
     rdfs:label ?name ;
     ont:centuryOfLiving ?t ;
     ont:authorOf ?w .
  ?t rdfs:label ?century .

  ?w otv:denotedByProperName ?pn .
  ?pn otv:properName ?workTitle .

  FILTER (lang(?name) = "en")
  FILTER (lang(?century) = "en")
  FILTER (CONTAINS(STR(?pn), "_en"))
  FILTER (CONTAINS(LCASE(STR(?workTitle)), "on nature"))
}
ORDER BY ?name
```

name	century	workTitle
"anaxagoras_of_clazomenae" @en	"5th_century_BCE" @en	"On nature (Anaxagoras)"
"diogenes_of_apollonia" @en	"5th_century_BCE" @en	"On nature (Diogenes)"
"empedocles_of_acragas" @en	"5th_century_BCE" @en	"On nature (Empedocles)"
"epicurus" @en	"3rd_century_BCE" @en	"On nature (Epicurus)"
"epicurus" @en	"4th_century_BCE" @en	"On nature (Epicurus)"
"gorgias_of_leontini" @en	"4th_century_BCE" @en	"On nature (Gorgias)"
"gorgias_of_leontini" @en	"5th_century_BCE" @en	"On nature (Gorgias)"
"heraclitus_of_ephesus" @en	"5th_century_BCE" @en	"On nature (Heraclitus)"
"heraclitus_of_ephesus" @en	"6th_century_BCE" @en	"On nature (Heraclitus)"
"parmenides" @en	"5th_century_BCE" @en	"On nature (Parmenides)"
"philolaus" @en	"5th_century_BCE" @en	"On nature (Philolaus)"

FIG 1: SPARQL query retrieving Ancient Greek philosophers who authored a work titled "On Nature" and their centuries of living, with the resulting author–century–title table shown on the right.

5.2 Utilisation

Utilisation is ensured through multi-format publication from TEDI. TEDI supports exports in RDF/OWL for Protégé (Musen 2015) and SPARQL (for sample SPARQL questions see the TALOS SPARQL endpoint: https://tools.talos-lab.eu/sparql/), HTML (Onto-Dictionary and object/instance collections), TBX (ISO 30042:2019; terminology exchange), and CSV (lightweight reuse and inspection). These outputs support non-technical access through structured concept- and term-level navigation, alongside instance-level HTML pages for philosophers, works, places, and schools, enabling effective use in teaching contexts and scholarly browsing environments.

FIG 2: "Confucius" proper name entry in the TEDI OntoDictionary HTML export.

Current population already supports core humanities use cases: prosopography (grouping by school, place, century), intellectual geography (birthplace clusters and residence mobility), intertextual retrieval (works mentioning philosophers; philosophers discussed by works), and

network perspectives (student–teacher chains and influence relations where present). The comparative framing builds on established cross-cultural scholarship on Greece and China and makes it operational as repeatable query patterns over aligned subsets.

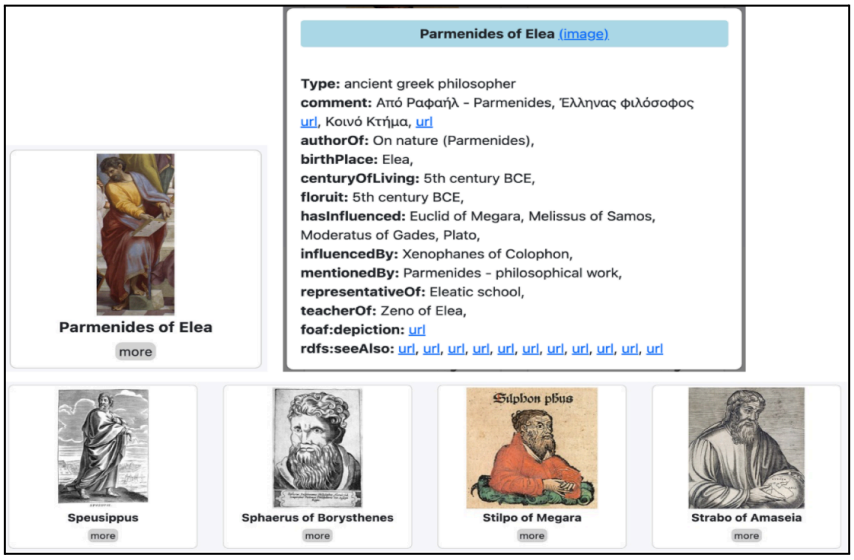


FIG 3: "Parmenides of Elea" entry in the TEDI Object Collection HTML export.

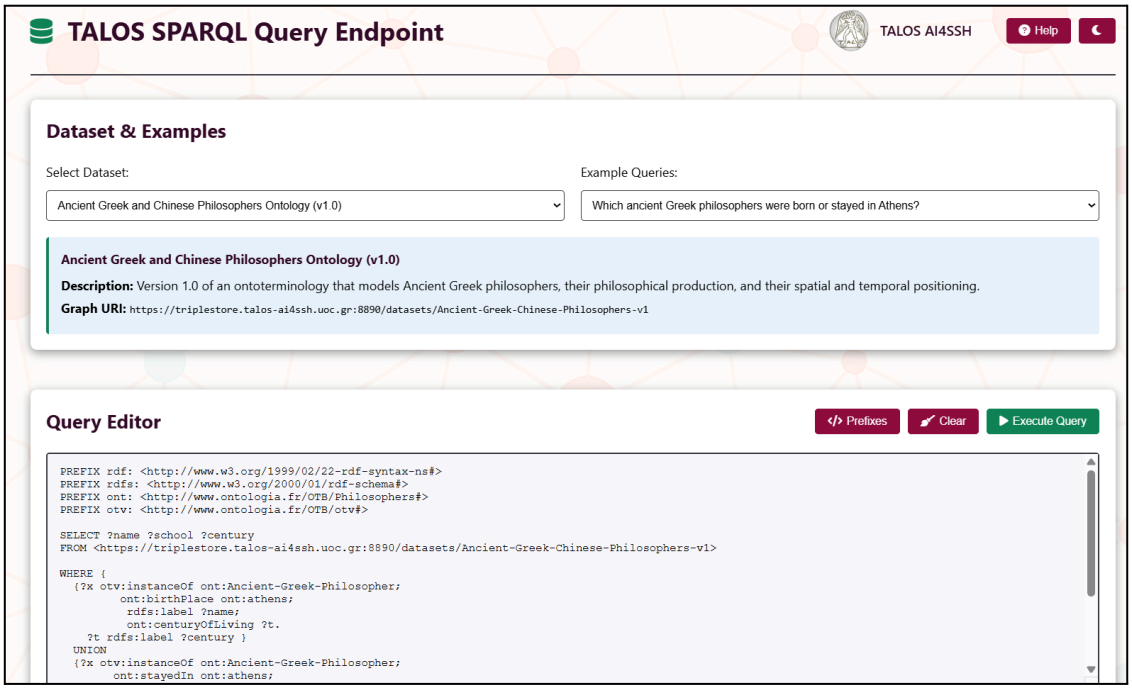


FIG 4: Indicative SPARQL queries can be executed via the TALOS SPARQL endpoint (https://tools.talos-lab.eu/sparql/)

6. Results and Future work

The current iteration of the knowledge graph represents a populated ontology at substantial scale, characterized by a consistent class structure and well-defined relations. Its unified

backbone successfully accommodates both Greek and Chinese philosophical traditions within a single schema, even where depth of population currently differs across the two sides. This structure demonstrates that comparative study can be operationalized through parallel typing, shared relation sets, and query shapes repeatable across traditions. However, several limitations must be explicitly acknowledged: uneven granularity between the Greek and Chinese populations at this stage; historical uncertainties in dates, attributions, and influence claims; and translation and name-variant issues that require normalization without loss of philological nuance.

Future development will prioritize strengthening comparative modeling by adding explicit "comparison hooks"—including lineage structures, philosophical school analogues, and shared problem spaces—that support aligned queries across traditions. On the Chinese side, efforts will enrich and normalize coverage of philosophers, works, places, and schools while harmonizing transliterations and name variants and preserving multilingual labels. Concurrent refinements to the Greek data will focus on improving school membership coverage, influence edges, and work-mention linking. Finally, evaluation will expand through a larger set of competency questions (including explicitly comparative ones) demonstrating that repeated query templates can yield meaningful cross-tradition comparisons.

7. References

Berners-Lee, T. 2006. "Linked Data." *World Wide Web Design Issues*. Retrieved from https://www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, T., Hendler, J., and Lassila, O. 2001. "The Semantic Web." *Scientific American* 284 (5): 34–43. https://www.scientificamerican.com/article/the-semantic-web/.

Giannadakis, R., Papadopoulou, M, and Roche, C. 2024. "Ontoterminology of Ancient Greek and Chinese Philosophers, V1.0" *Zenodo*. https://doi.org/10.5281/zenodo.13372136; https://talos-ai4ssh.uoc.gr/resources/open-datasets/open-dataset-5-ancient-greek-and-chinese-philosophers-ontology/.

ISO. 2019. *Management of Terminology Resources, TermBase eXchange (TBX)*. ISO 30042:2019.

Lloyd, G. E. R. 2004. *Ancient Worlds, Modern Reflections: Philosophical Perspectives on Greek and Chinese Science and Culture*. Oxford University Press.

Lloyd, G. E. R. 2018. *Ancient Greece and China Compared*. Cambridge: Cambridge University Press.

Milio, R., Giannadakis, R., & Lourentzaki, A. (2025). "Creating ontoterminologies for antiquity: Workflow, challenges and solutions." In *Proceedings of the 4th International Conference on Multilingual Digital Terminology Today: Design, Representation Formats and Management Systems* (MDTT 2025). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3990/paper4.pdf.

Murdock, J., and Allen, C. 2011. "InPhO for All: Why APIs Matter." *Journal of Digital Humanities in the Cognitive Sciences* 1 (3). https://www.inphoproject.org/papers/2011JDHCSMurdockAllen.pdf.

Musen, M. A. 2015. “The Protégé Project: A Look Back and a Look Forward.” *AI Matters* 1 (4): 4–12. https://dl.acm.org/doi/pdf/10.1145/2757001.2757003.

Niepert, M., Buckner, C., and Allen, C. 2007. “InPhO: The Indiana Philosophy Ontology.” *APA Newsletter on Philosophy and Computers* 7 (1). https://www.inphoproject.org/papers/2007NiepertBucknerAllen.pdf.

Roche, C. 2007. “Le terme et le concept: Fondements d’une ontoterminologie.” In *TOTh 2007: Terminologie & Ontologie*, 1–22.

Roche, C. 2012. “Ontoterminology: How to Unify Terminology and Ontology into a Single Paradigm.” In *Proceedings of LREC 2012*, 2626–2630. https://aclanthology.org/L12-1321/.

Roche, C., Calberg-Challot, M., Damas, L., and Rouard, P. 2009. “Ontoterminology: A New Paradigm for Terminology.” In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 321–326. Madeira, Portugal, October 2009. https://hal.science/hal-00622132v1.

Roche, C., and Papadopoulou, M. 2019. “Mind the Gap: Ontology Authoring for Humanists.” In *Proceedings of WODHSA*, 65–72. https://ceur-ws.org/Vol-2518/paper-WODHSA7.pdf.

Roche, C. “Tedi.” Software Website. Distributed freely by the University of Crete. https://ontoterminology.com/tedi.

W3C. 2009. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. https://www.w3.org/TR/skos-reference/.

W3C. 2013. *SPARQL 1.1 Query Language*. W3C Recommendation. https://www.w3.org/TR/sparql11-query/.

W3C Ontology-Lexica Community Group. 2016. *OntoLex-Lemon: Lexicon Model for Ontologies*. W3C Community Group Report. https://www.w3.org/2016/05/ontolex/.

Wilkinson, M. D., et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Anaïs Guillem, Violette Abergel, Miled Rousset

Managing the Thesaurus Data Cycle as a FAIR Semantic Artefact for Heritage Science

Short Abstract:

Thesauri play a central role in the structuring, mediation, and dissemination of domain knowledge, yet their lifecycle is often treated as a series of disconnected technical operations rather than as a coherent semantic process. In heritage science and cultural heritage, this fragmentation is amplified by the diversity of actors, tools, and infrastructures involved in thesaurus creation, maintenance, and reuse. This paper addresses this gap by combining a conceptual framing of the thesaurus as a FAIR semantic artefact with its practical implementation and deployment within operational research infrastructures of E-RIHS and Espadon.

The thesaurus data life cycle is managed through two complementary environments. Opentheso supports thesaurus edition, governance, and maintenance through collaborative editorial workflows. OntoPortal provides the infrastructure for publication, sharing, and reuse, ensuring interoperability, visibility, and long-term accessibility of semantic artefacts. Several instances of Opentheso are in use in the HS community: Frollo, Opentheso-Humanum, Opentheso-Espado. The Heritage Science Portal (HS Portal) is a deployed instance of Ontoportal dedicated to heritage science and supported by Espadon, E-RIHS-fr, E-RIHS-it, and E-RIHS-eu. It is proposed as a semantic artefact catalogue in the Catalogue of Services of E-RIHS DIGILAB.

By articulating conceptual principles with concrete tools and infrastructures, this contribution demonstrates how FAIRness can be operationalized for thesauri in real-world contexts. It highlights thesaurus management as a socio-technical and epistemic process and positions the thesaurus data cycle as a core methodological concern for terminology and ontology in applied domains.

Keywords:
Thesaurus
Opentheso
Ontoportal
HSPortal
Heritage science
FAIR
Thesaurus edition

Thesaurus sharing
Thesaurus data cycle
Semantic artefact
FAIRness

Topics:
Terminology

- Construction of Terminological Resources, Terminography
- Terminological Resource Design, Quality and Evaluation

Linguistics
Artificial Intelligence

- Knowledge Acquisition and Maintenance: Ontology Building, Alignement, Quality and Evaluation
- Semantic Web, Metadata, Linked and Open Data, FAIR Data

Applications

- Digital Humanities and Cultural Heritage
- Information Retrieval
- Information Sciences, Digital Libraries, Thesauri
- ICT Applications

Extended abstract:

1. Introduction and Problem Statement

Thesauri occupy a central position as instruments for structuring domain knowledge, mediating meaning, and supporting semantic interoperability. Beyond their function as controlled vocabularies, thesauri formalize conceptual systems by making explicit the relationships, hierarchies, and terminological choices that underlie knowledge organization within a given domain. As such, they play a key role in sharing understanding across communities, disciplines, and information systems.

Despite this theoretical grounding, the operational reality of thesaurus management often remains disconnected from terminological principles. In practice, thesauri are frequently treated as *static resources* or technical by products of information systems, with limited attention paid to their evolution, governance, and conditions of reuse. The

data life cycle of thesauri, from conceptual modelling and editorial decisions to publication, maintenance, and reuse, is rarely made explicit.
This fragmentation is particularly visible in heritage science, a field characterized by heterogeneous data, interdisciplinary collaboration, and research infrastructures with long history traditions. Thesaurus practices are distributed across projects, institutions, and tools, leading to inconsistent editorial policies, and limited interoperability between terminological resources. These issues directly affect the intelligibility, durability, and reusability of thesauri as shared semantic references.

The growing adoption of FAIR principles further aims at mitigating these limitations. While *FAIRness* is increasingly invoked as a requirement for semantic resources, it cannot be achieved through format compliance alone. It instead requires explicit management of processes, responsibilities, and knowledge transformations throughout the thesaurus data life cycle: a thesaurus is not merely as a vocabulary, but a managed *semantic artefact* whose lifecycle must be made explicit.

2. Thesauri as FAIR Semantic Artefacts

A *semantic artefact* can be understood as a structured, intentional representation of conceptual knowledge that is designed to mediate meaning between humans and information systems (Corcho et al., 2024). Thesauri belong to this category insofar as they encode conceptual distinctions, semantic relations, and terminological choices within a formalized structure. A thesaurus is not a neutral resource, but a constructed artefact whose intelligibility depends on the traceability of its production and use.

Applying *FAIR principles* to thesauri highlights this artefactual dimension. *Findability* requires the use of persistent identifiers at both thesaurus and concept levels, as well as inclusion in catalogues that allow semantic resources to be discovered and referenced beyond their original context of creation. *Accessibility* concerns the conditions under which thesauri can be consulted and retrieved, including the availability of standard access mechanisms, clear usage licenses, and stable interfaces that support both human and machine access. *Interoperability* raises questions of semantic alignment and formal representation, requiring the use of shared models, standards, and representation languages, such as SKOS, that enable thesauri to be integrated within broader knowledge organization and data infrastructures. *Reusability* depends on the availability of contextual information that allows users to assess the scope, validity, and applicability of a thesaurus, including documentation of provenance, versioning practices, and governance structures.

These requirements make clear that FAIRness is not an intrinsic property of thesauri. A thesaurus may comply with a given format or standard while remaining opaque in terms of editorial decisions, conceptual assumptions, or lifecycle management. FAIRness instead emerges from the explicit organization of processes through which thesauri are created, maintained, published, and reused. It is therefore a property of workflows, responsibilities, and governance as much as of technical representations. Understanding thesauri as FAIR semantic artefacts thus requires shifting attention from formats alone to the socio technical processes that sustain thesaurus resources over time.

3. The thesaurus data life cycle

Understanding thesauri as FAIR semantic artefacts requires making their life cycle explicit and actionable at every step. Rather than a sequence of technical steps, the thesaurus data life cycle encompasses a set of interrelated phases through which conceptual knowledge is progressively formalized, stabilized, disseminated, and potentially transformed through reuse. Making these phases visible is essential for ensuring the intelligibility, durability, and FAIRness of terminological resources.

The life cycle begins with concept modelling and editorial work, during which domain experts and terminologists define concepts, select preferred terms, establish semantic relations, and document scope and usage. These activities involve interpretative decisions that directly shape the conceptual structure of the thesaurus. Governance and validation mechanisms intervene to ensure consistency, quality, and alignment with agreed editorial policies, often through collective review processes and role based responsibilities. Over time, thesauri evolve as domains change, new knowledge emerges, or conceptual frameworks are revised. Versioning and evolution therefore constitute a critical phase of the life cycle, requiring mechanisms to document changes, preserve earlier states, and maintain referential stability for users and dependent systems.

Publication and dissemination mark the transition from editorial environments to broader infrastructures where thesauri become accessible as shared semantic resources in an ontology library (Ding and Fensel, 2001) or an ontology repository (Hartmann and Gomez-Pérez, 2009), or a semantic artefact catalogue (Jonquet and Grau, 2024). Reuse and feedback complete the cycle by introducing new actors and use cases, ranging from data annotation and information retrieval to semantic alignment and knowledge graph construction. These uses may generate feedback that informs

subsequent editorial decisions, thereby reinforcing the cyclical nature of thesaurus management.

Across all phases, the *thesaurus data life cycle* (Figure 1) involves both human and technical actors whose interactions are often implicit. Domain experts, terminologists, developers, and infrastructure providers contribute at different moments, mediated by tools, standards, and institutional frameworks. When transitions between phases are undocumented or poorly articulated, important contextual information is lost, leading to reduced transparency, limited reusability, and weakened trust in the resource. Making the data life cycle explicit thus provides a backbone for thesaurus intelligibility and reuse, and forms a prerequisite for operationalizing FAIR principles in practice.

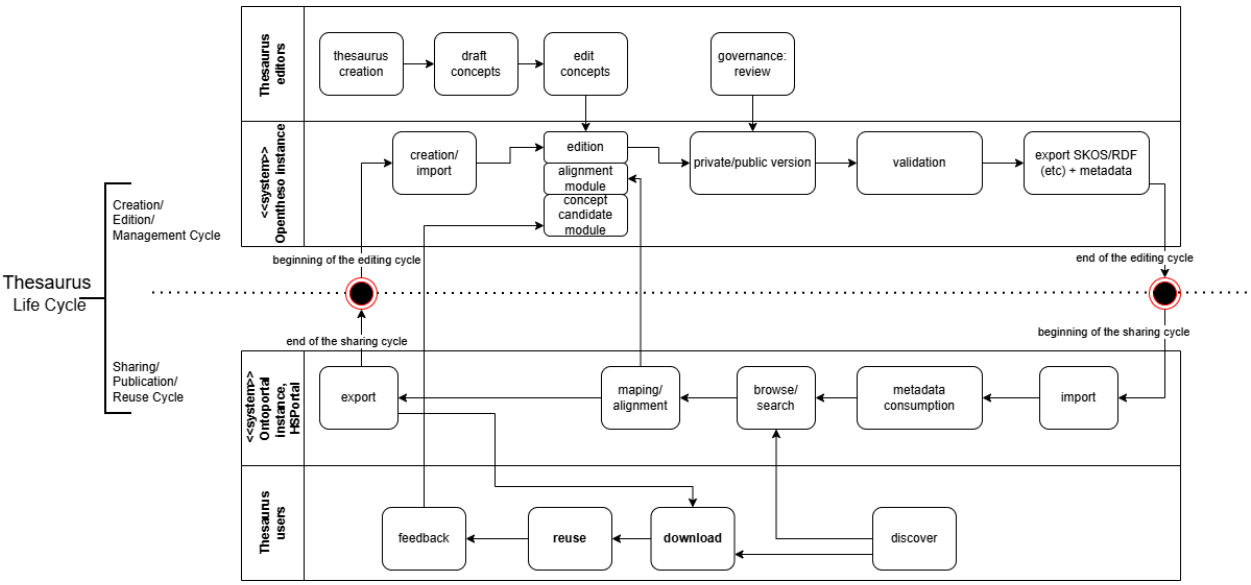


Figure 1: thesaurus life cycle.

4. Operational implementation and deployment in heritage science infrastructures

The conceptual framing of thesauri as FAIR semantic artefacts is operationalized through a set of complementary tools and infrastructures that support the full thesaurus data life cycle, from collaborative edition and governance to publication, dissemination, and reuse. Rather than acting as neutral containers, these environments embed editorial, governance, and sustainability principles that directly shape how thesauri are produced, maintained, and mobilized within heritage science research contexts.

4.1 Thesaurus edition, governance, and management

Thesaurus edition and governance are supported through Opentheso, an environment dedicated to the collaborative construction and long-term management of thesaurus resources (Figure 2). While Opentheso is primarily designed to support editorial workflows, its scope extends beyond the core edition and maintenance phases of the thesaurus life cycle (Figure 1). It also integrates a substantial part of the sharing and dissemination functions, including publication services, persistent identification, and access mechanisms that enable thesauri to circulate as reusable semantic artefacts within and across communities.

Editorial workflows are designed to reflect established terminological practices by structuring concept creation, semantic relations, and documentation within a shared and controlled framework. Role-based permissions and validation mechanisms allow multiple contributors to participate in thesaurus development while preserving editorial coherence, accountability, and governance. Collaborative practices are central to this approach, as thesaurus construction in heritage science typically involves domain experts, terminologists, and technical actors. Opentheso supports this collaboration by making editorial actions explicit and traceable, thereby facilitating shared responsibility and quality control.

These principles are concretely implemented through several deployed instances serving heritage science communities. The Frollo instance supports the Notre-Dame de Paris research programme by providing a shared environment for terminological coordination across disciplines (Opentheso-Frollo). Opentheso Humanum is operated within the Huma-Num infrastructure (Opentheso-Huma-Num), while Opentheso-Espadon is deployed in the Espadon infrastructure to support thesaurus management aligned with national and European heritage science initiatives (Opentheso-Espadon). Together, these instances illustrate how a common editorial framework can be adapted to diverse institutional and disciplinary contexts while maintaining coherent lifecycle management principles.

Opentheso operates as a hybrid environment that bridges editorial control and controlled dissemination, while remaining deliberately focused on a single class of semantic artefacts: thesauri. This specialization contrasts with OntoPortal, which concentrates on the sharing and reuse cycle but supports a broader range of semantic artefacts, including ontologies, vocabularies, and knowledge organization systems.

4.1 Thesaurus edition, governance, and management

Thesaurus edition and governance are supported through Opentheso, an environment dedicated to the collaborative construction and long-term management of thesaurus

resources (Figure 1 and 2). Editorial workflows are designed to reflect established terminological practices by structuring concept creation, semantic relations, and documentation within a shared and controlled framework. Role-based permissions and validation mechanisms allow multiple contributors to participate in thesaurus development while preserving editorial coherence and accountability.

Collaborative practices are central to this approach, as thesaurus construction in heritage science typically involves domain experts, terminologists, and technical actors. Opentheso supports this collaboration by making editorial actions explicit and traceable, thereby facilitating shared responsibility and quality control. Provenance, versioning, and change-tracking mechanisms document the evolution of concepts over time, preserve earlier states, and support the long-term intelligibility of the thesaurus as a managed semantic artefact rather than a static resource.

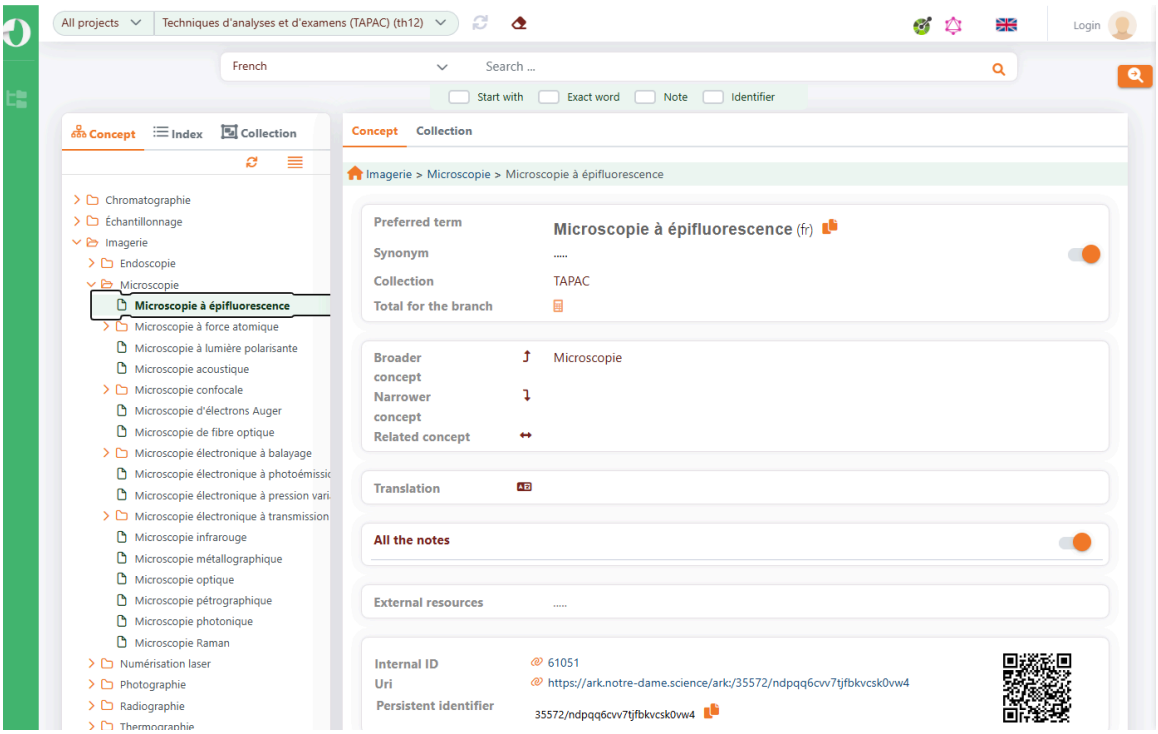


Figure 2: Frollo, instance of Opentheso (thesaurus edition and management tool): example of the concept “Microscopie à épifluorescence” (fr) in the thesaurus th12-TAPAC.

These principles are concretely implemented through several deployed instances serving heritage science communities. The Frollo instance supports the Notre-Dame de Paris research programme by providing a shared environment for terminological coordination across disciplines (Opentheso-Frollo). Opentheso Humanum is operated within the Huma-Num infrastructure (Opentheso-Huma-Num), while

Opentheso-Espadon is deployed in the Espadon infrastructure to support thesaurus management aligned with national and European heritage science initiatives (Opentheso-Espadon). These instances illustrate how Opentheso can be adapted to diverse institutional and disciplinary contexts while maintaining coherent thesaurus lifecycle activites.

4.2 Publication, dissemination, and reuse of thesauri

Publication, sharing, and reuse are addressed through OntoPortal, which provides an infrastructure for disseminating thesauri as semantic artefacts beyond their editorial context (Jonquet et al., 2023). Integration within such an environment supports findability by cataloguing thesauri alongside other semantic resources and by assigning persistent identifiers that ensure stable referencing. Interoperability is ensured through standardized representations and access mechanisms, enabling reuse across heterogeneous systems, applications, and research workflows.

OntoPortal also contributes to long-term accessibility and visibility by offering stable access points and services for both human users and machines in a broader ecosystem of semantic resources for comparison, alignment, and reuse at scale.

In the heritage science domain, the sharing cycle is realized through the Heritage Science Portal, a dedicated instance of OntoPortal tailored to community needs (Figure 3) (Guillem et al., 2026a). Supported by Espadon, E-RIHS France, and E-RIHS Europe, the portal provides a shared access point for thesauri and other semantic artefacts, facilitating discovery, comparison, and reuse within and beyond heritage science. Its integration into the E-RIHS DIGILAB service catalogue positions thesauri as first-class semantic resources alongside other digital research services.

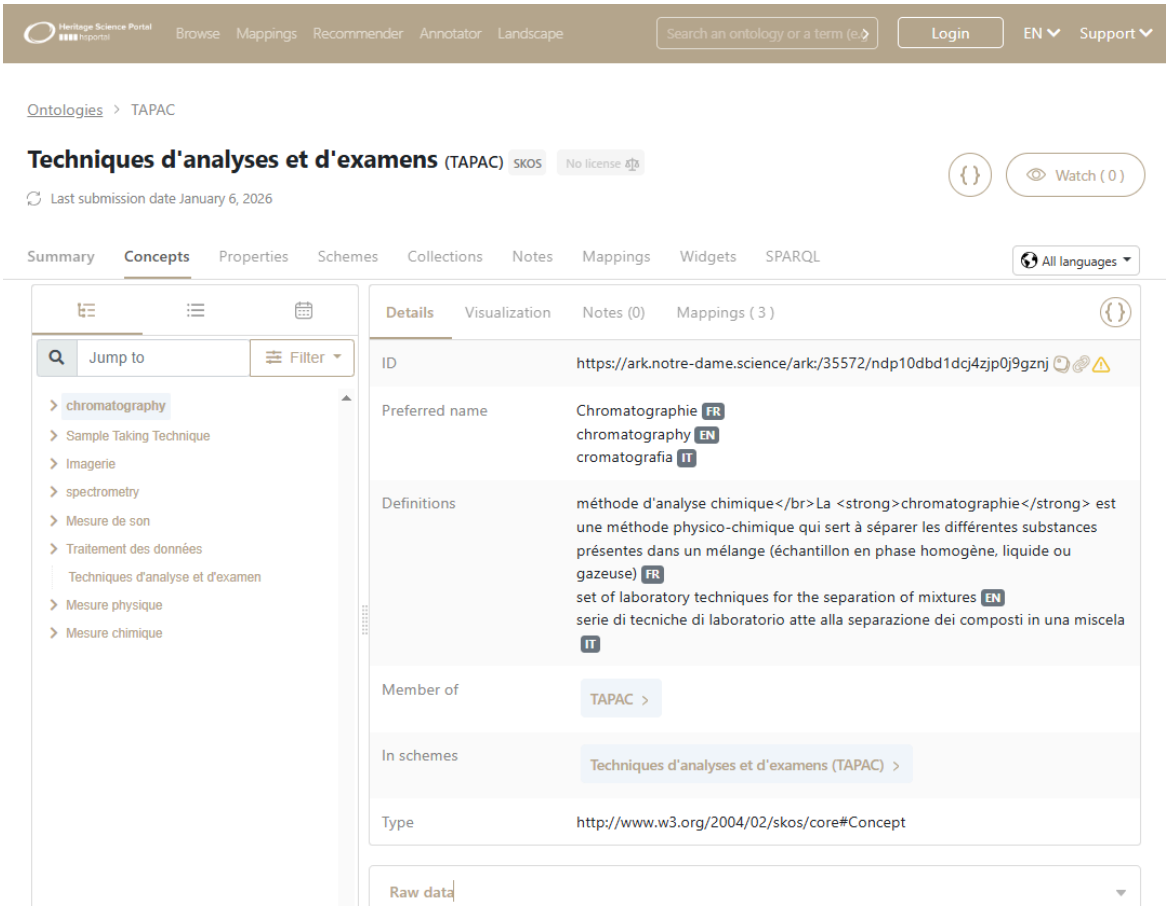


Figure 3: Heritage Science Portal, instance of OntoPortal (semantic artefact catalogue): example of the concept “Microscopie à épifluorescence” (fr) in the thesaurus th12-TAPAC.

5. Discussion and conclusion

The approach presented in this contribution highlights thesaurus management as a fundamentally socio-technical process in which conceptual modelling, editorial practices, technical infrastructures, and institutional governance are tightly interwoven. This articulation between editorial and dissemination environments embeds FAIR principles, governance requirements, and lifecycle considerations directly into operational infrastructures, rather than treating them as abstract guidelines. Treating thesauri as FAIR semantic artefacts brings into focus the human decisions, responsibilities, and knowledge transformations that underpin terminological resources,

and challenges views that reduce thesaurus management to a purely technical task. From this perspective, lifecycle management becomes a central methodological concern for thesaurus and ontology, as it directly conditions the intelligibility, trustworthiness, and reusability of semantic resources.

These findings have important implications for terminology governance. Explicit lifecycle management supports transparency in editorial decision making, clarifies roles and responsibilities, and enables collective validation processes that extend beyond individual projects. At the same time, the work underlines the limits of purely technical approaches to FAIR compliance. While standards, formats, and interfaces are necessary, they are insufficient in the absence of documented processes, governance structures, and provenance information. FAIRness emerges not from compliance alone, but from the sustained organization of practices across the lifecycle of thesauri.

Although the contribution is grounded in heritage science, the proposed approach is transferable to other domains where terminological resources are produced collaboratively and reused across heterogeneous systems. Fields such as digital humanities, healthcare, education, and industry face similar challenges related to thesaurus evolution, interoperability, and long term sustainability. The articulation of conceptual principles with operational infrastructures therefore provides a reusable methodological framework rather than a domain specific solution.

In conclusion, this work combines a conceptual framing of thesauri as FAIR semantic artefacts with their concrete implementation and deployment within research infrastructures. It demonstrates how thesaurus lifecycle management can support cross domain reuse, alignment with existing standards, and future integration with knowledge graphs and emerging LLM aware environments (Guillem et al., 2026b). By positioning thesauri as durable and governable semantic resources, this contribution argues for an infrastructure oriented understanding of terminology work. Sustainable semantic artefacts require both theoretical grounding and operational infrastructures capable of supporting their evolution over time.

Bibliography

Corcho, Oscar, Fajar J. Ekaputra, Ivan Heibi, et al. 2024. “A Maturity Model for Catalogues of Semantic Artefacts.” _Scientific Data_ 11 (1): 479.

Ding, Y. & Fensel, D. Ontology library systems: The key to successful ontology reuse. In *Proceedings of the Semantic Web Working Symposium (SWWS 2001)*, 93–112.

https://files.ifi.uzh.ch/ddis/iswc_archive/iswc/ih/SWWS-2001/program/full/paper58a.pdf (2001).

Guillem, Anais, Erica Scarpa, Riccardo Valente, et al. 2026a. “HSPortal, Instance of OntoPortal for Heritage Science and Semantic Artefact Catalog Service of the DIGILAB Platform in the E-RIHS Research Infrastructure.” under preparation.

Guillem, Anaïs, Kévin Réby, John Samuel, et al. 2026b “LLMtheso: Curating Heritage Science Thesaurus from ‘Dirty’ Data Using a Neuro-Symbolic LLM–SKOS Workflow.” under review.

Hartmann, J., Palma, R. & Gómez-Pérez, A. Ontology repositories. In Staab, S. & Studer, R. (eds.) *Handbook on Ontologies*, 551–571, https://doi.org/10.1007/978-3-540-92673-3_25 (Springer, 2009).

Jonquet, Clement, John Graybeal, Syphax Bouazzouni, et al. 2023. “Ontology Repositories and Semantic Artefact Catalogues with the OntoPortal Technology.” In *The Semantic Web – ISWC 2023*, edited by Terry R. Payne, Valentina Presutti, Guilin Qi, et al. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47243-5_3.

Jonquet, Clement, and Nina Grau. 2024. *M4.4 - Review of Semantic Artefact Catalogues and Guidelines for Serving FAIR Semantic Artefacts in EOSC*. M4.4. Zenodo. https://doi.org/10.5281/zenodo.12799796.

“Opentheso-Frollo.” n.d. Accessed February 8, 2026. https://frollo.notre-dame.science/.
“Opentheso-Huma-Num.” n.d. Accessed February 8, 2026. https://opentheso.huma-num.fr/.

Critères pour le choix d'un vocabulaire contrôlé en terminologie

Pierre Lerat
Professeur honoraire (Paris 13-Villetaneuse)
pierre.lerat@wanadoo.fr

Résumé :

« Dans la réalité il n'y a pas de termes » (Wüster 1985 :6). Il y a des objets (matériels et immatériels) et des conceptualisations plus ou moins partagées (une machine-outil, c'est quoi ?). Les conceptualisations partagées dans une communauté de travail sont techniques au sens d '« ensemble de savoir-faire permettant d'obtenir des résultats conformes à des projets » (Simondon 2008 :269). Ces projets relèvent de la pensée, non de la langue.

La terminologie est une science humaine parce que les connaissances y sont lexicalisées, et aussi parce que les dénominations techniques servent à communiquer ces connaissances.

Pour que la terminologie soit formalisable, une étape nécessaire préalablement semble être l'usage d'un vocabulaire contrôlé. Est-ce possible ? Les critères favorables semblent être les suivants : la hiérarchie des concepts opérationnels, un domaine bien circonscrit, le choix d'une dénomination motivée morphologiquement, la distribution dans un corpus spécialisé, les connexités conceptuelles, la traductibilité.

Mots-clés : terminologie, lexicographie, vocabulaire contrôlé, traduction, ingénierie des connaissances

Abstract:

"In reality, there are no terms" (Wüster 1985: 6). There are objects (material or immaterial) and conceptualizations that are more or less shared (what is a machine tool?). The conceptualizations shared within a working community are technical in the sense of "a set of skills enabling result to be obtained in line whith projects" (Simondon 2008: 269). These projects are a matter of thought, not language.

Terminology is a human science because knowledge is lexicalized in it, and also because technical terms are used to communicate this knowledge.

In order for terminology to be formalized, a necessary preliminary step seems to be the use of a controlled vocabulary. Is it possible? The following criteria seem to be favourable: the hierarchy of operational concepts, a well-defined domain, the choice of a morphologically motivated name, distribution in a specialized corpus, conceptual connections, translatability.

Key words: terminology, lexicography, controlled vocabulary, translation, knowledge engineering

Evolving Terminology Standards: a critical look at ISO 704:2022

Claudia Brunini, Istituto Nazionale di Statistica (Istat) - Italia, claudia.brunini@istat.it
Eugenio Concetti, Istituto Nazionale di Statistica (Istat) - Italia, eugenio.concetti@istat.it
Stefania Pantoni, Istituto Nazionale di Statistica (Istat) - Italia, stefania.pantoni@istat.it

1. Introduction

In July 2022, the fourth edition of ISO 704 was published, succeeding ISO 704:2009 and its predecessor, ISO 704:2000. The long interval since the second edition reflects the substantial conceptual evolution that has occurred in terminology studies over the past two decades. Advances in linguistics, cognitive science, ontology, and terminology theory have reshaped how concepts are analyzed, defined, and represented. ISO 704:2022 incorporates these developments into a more comprehensive and methodologically rigorous standard, offering clear guidance for the management of terminological resources and positioning terminology work at the forefront of contemporary knowledge organization and standardization.

Among the most significant updates, the 2022 edition:

- Aligns the document's structure and content with ISO 1087:2019;
- Introduces concept models in accordance with ISO 24156-1:2014;
- Expands the treatment of associative concept relations and designation–concept relations;
- Provides a more systematic approach to appellations, nomenclature and proper names;
- Revises or replaces examples as needed and adds new ones to illustrate key concepts.

The contribution aims to highlight the developments that most directly impact the management of terminological resources, showing how ISO 704:2022 supports a more integrated, interoperable, and methodologically robust approach to contemporary terminology work.

2. Concept systems and terminological practice

The process of conceptualization marks the shift from the world of objects to the level of linguistic representation (Clause 4). In ISO 704:2022 this phase is described with much greater clarity: it shows how lexemes originate from the abstraction of properties into characteristics, and it explains the function of the definition as the device that delimits the scope of a concept (its extension) by specifying only the essential characteristics that constitute its intension.

The standard also clarifies the role of terminology, which consists in defining the function that terms serve within the specialized language under consideration. The new version of the standard also gives greater emphasis to a clear definition of terminological activity (Clause 5.4.2).

In the process of conceptualization, the relevant properties of all objects in the category are abstracted into characteristics. The relevant characteristics must be identified on the basis of content related to the domain of interest. This type of activity requires careful investigation and an in-depth knowledge of the domain. The new edition of ISO 704 sets out a systematic procedure that involves: identifying the properties possessed by objects within the domain, determining which of these properties are abstracted into characteristics, and establishing how the characteristics combine to form a concept.

After identifying the characteristics, it is possible to establish the definition. Not all characteristics are equally important. It is necessary to list all the essential ones, otherwise there is a risk of representing a different concept, corresponding to a different set of objects. Non-essential characteristics, on the other hand, may be added if needed. Having carefully identified the list of characteristics makes it easier to define the differences with similar concepts and eventually to determine the relationships that exist between them.

Concepts do not exist as isolated units of knowledge but always in relation to one another (Clause 5.5.1). The types of relations considered remain the same as in the previous editions: hierarchical relations, comprising both generic and partitive relations, and associative relations. However, the 2022 edition places much greater emphasis on the representation of concept systems. Relations between concepts are no longer presented merely as lists or traditional concept diagrams, but also through UML-based concept models in accordance with ISO 24156-1 (Clause 5.5.3).

With regard to the generic hierarchical relation, the connection between the set of characteristics of the subordinate concept and that of the superordinate concept is illustrated with greater clarity, introducing the so-called inheritance principle (Clause 5.5.4.2.1), which is a way of testing and validating the generic relations. These are valuable indications for formulating definitions that are clear, complete, and coherent. Another important element introduced is the role of the individual concept within a hierarchy. If such individual concepts are present, they occupy the bottom rung (i.e., the last item) of the concept ladder. An extension consisting of a single object cannot be subdivided into a more specific concept (Clauses 5.5.4.2.2 and 5.5.4.3.2).

For partitive hierarchies, the 2022 edition emphasizes that concepts connected by a partitive relation do not inherit the characteristics of the superordinate concept. The parts that constitute the whole may be similar in nature (such as ‘atoms’ in an ‘oxygen molecule’) or distinctly different from one another. In Clause 5.5.4.3.1 it is also highlighted that, in order to correctly identify a partitive relation, one must first determine the position of the comprehensive concept within the overall hierarchy of the domain. Once this position—and thus the characteristics the comprehensive concept inherits along its generic lineage—has been established, its partitive concepts and their respective extensions can be identified in a consistent manner.

The standard also highlights an interaction between generic and partitive hierarchies. When two or more coordinate concepts share the same part, that part originates from the generic concept above them and is therefore not a part of each coordinate concept individually, but a part inherited from the generic concept. It must be analysed accordingly.

On associative relations, the revised ISO adopts a more structured and systemic perspective. They are organized into a detailed typology of types and subtypes, with explicit directionality and more rigorous criteria for conceptual modelling. This development makes associative relations not only descriptive but also operationally usable within terminological and semantic models, marking a clear methodological advancement over the previous editions.

The section on the conceptual system outlines the purpose and methods of terminological work. The construction of the conceptual system is presented as the foundation for uniform and standardized terminology, and the operative steps required to build it are now described with greater clarity (Clause 5.6.2). The revised edition also assigns much greater importance to the graphical representation of the concept system, which may be produced using the traditional concept diagram or, alternatively, UML-based concept models in accordance with ISO 24156-1.

The types of concept system (Clause 5.6.3) are expanded by retaining the three basic categories and adding a fourth (mixed systems), while also introducing two further classification criteria: the number of subdivision criteria (monodimensional vs. multidimensional systems) and the number of immediate superordinate concepts (monohierarchical vs. polyhierarchical systems). The result is a multidimensional typology that captures not only the nature of the relations involved but also the structural complexity of the system, marking a shift from a simple list of system types to a more comprehensive and analytically robust classification model.

The work on concepts finishes with the drafting definitions. The steps involved in developing concept systems and defining concepts are closely related (Clause 5.6.2). The description of the task of writing definitions is significantly expanded. It is emphasized that a definition must provide only the essential characteristics, clearly distinguishing it from encyclopaedic descriptions. The 2022 edition introduces guidance specific to standardization, allows complementary media such as graphics and formulas, and explicitly warns against replacing definitions with synonyms, abbreviations, or full forms. It also requires attention to the needs of different target audiences and acknowledges that definitions may need amplification for non-experts. Overall, the 2022 edition shifts from a brief logical description of the step to a comprehensive set of methodological, communicative, and practical principles for crafting definitions in terminological resources.

The 2022 edition adds also a clear emphasis on citing authoritative sources to ensure the credibility, traceability, and legal correctness of terminological information. It highlights the need to respect copyright, avoid citation errors, and follow established standards such as ISO 690, ISO 12615, and ISO 23185 (Clause 6.7).

3. Designations and relations with concepts

The revision of ISO 704 culminating in the 2022 edition highlights a profound shift in perspective in the way linguistic designations are conceived and managed within terminological resources. While previous editions of the standard were still largely rooted in a tradition oriented towards the production of relatively static terminological repertories—primarily intended for human consultation and linguistic normalization in documentary contexts—the 2022 revision is explicitly situated within a framework oriented towards knowledge modelling, the management of terminological databases, and semantic interoperability between information systems. This paradigm shift reflects a broader reorientation of terminological work, in which terminological resources are no longer conceived merely as auxiliary linguistic tools, but as integral components of complex information environments designed to interact with databases, thesauri, ontologies and multilingual systems.

From a structural point of view, this evolution is reflected in the reorganization introduced in ISO 704:2022, which consolidates content that was previously distributed across multiple chapters into a single chapter devoted to designations. In earlier editions of ISO 704, the treatment of terms and terminological relations was clearly separated from activities such as standardization, harmonization and deprecation, which were conceived as subsequent phases and, to a large extent, external to conceptual analysis. The 2022 edition abandons this sequential approach and integrates these activities directly into the analysis of relations between designations and concepts, framing terminological work as a unified and continuous process, consistent with the requirements of digital resource management and long-term maintenance.

Clause 7.1 of ISO 704:2022 introduces a significant shift by explicitly requiring consideration of the intended user groups—such as experts, non-experts, specialist or general users—in the selection and

management of designations. This pragmatic dimension is complemented by an innovation of particular relevance for interoperability: the introduction of a temporal dimension. The standard requires the documentation of the historical sequence of designations associated with a concept, as well as changes in their status over time. This dynamic perspective, absent from earlier editions of ISO 704, is essential for ensuring traceability, versioning and semantic coherence in evolving terminological resources.

The classification of designations, addressed in Clause 7.2, represents a further step towards rigorous formalization. ISO 704:2022 establishes a systematic correspondence between types of designations (terms, proper names and symbols), types of concepts (general or individual), and the multiplicity of the entities they denote. Compared to the primarily descriptive classifications found in earlier editions of the standard, this matrix makes explicit relations that can be directly implemented in data models and conceptual schemas, thereby aligning ISO 704 with approaches oriented towards knowledge engineering and formal concept modelling, as commonly adopted in knowledge organization systems.

Clause 7.3, devoted to terms in the strict sense, particularly highlights the transition from a predominantly form-based approach to a functional and semantic one. Whereas earlier editions of ISO 704 focused mainly on morphological distinctions, such as that between simple and complex terms, the 2022 edition explicitly acknowledges that terms may belong to different parts of speech (nouns, verbs, adjectives), shifting attention from linguistic form to functional role. This reflects actual usage in natural languages and specialized communication. The introduction of criteria for term classification based on acceptability, structure and formation method also allows for a more articulated representation of terminological variation, which is essential in multilingual and multi-domain contexts.

Of particular relevance, also from the perspective of interoperability, is the redefinition of the concept of appellation. In earlier editions of ISO 704, appellations constituted the category used to designate individual concepts. The 2022 edition reorganizes this category by treating appellations as a type of term referring to groups of identical objects (for example products, services or living organisms), while reserving proper names for the function of unique identifiers of single entities. This distinction, formalized in Clauses 7.3.3 and 7.4, aligns the standard with the requirements of information retrieval systems and with the approach adopted in ISO 25964-1:2011, which clearly distinguishes between general concepts and instances. Within the typology of terms, Clause 7.3.4 also formally introduces nomenclatures, understood as designation systems structured and governed by international scientific communities (e.g. IUPAC for chemical nomenclature), a category that had not been explicitly addressed in earlier editions.

Clause 7.5, dealing with symbols, does not introduce major conceptual changes, but updates references to contemporary standards for safety symbols and public information symbols. It also clarifies the status of alphanumeric codes, which may be considered symbols when they do not function as linguistic abbreviations.

Clause 7.6 of ISO 704:2022, devoted to the formation of terms, largely retains the framework established in previous editions, while extending its scope to include appellations and proper names. A subclause on transliteration and transcription is explicitly included, integrating content that had previously been treated separately. The principle formerly referred to as “linguistic economy” is reformulated as “concision” (7.6.2.4), without substantial changes in content, while the principle of appropriateness (7.6.2.3) maintains its emphasis on terminological neutrality and the avoidance of negative connotations, introducing explicit examples aimed at preventing discriminatory usage.

The treatment of relations between designations and concepts, developed in Clause 7.7, constitutes one of the most significant areas of innovation. In addition to the established distinction between monosemy (one term for one concept) and homonymy (the same term for different concepts), ISO 704:2022 introduces mononymy (one concept rappresented by one designation), clearly distinguishing between a concept-oriented and a designation-oriented analytical perspective. The standard also explicitly differentiates polysemy from homonymy: while homonymy concerns identical terms designating completely unrelated concepts, polysemy refers to identical terms designating distinct but conceptually related concepts that share part of their meaning. Within the broader framework of homonymy, homophony and homography are also addressed, referring respectively to phonetic and graphic identity with different meanings. Synonymy and quasi-synonymy are likewise treated, accounting for fully or partially interchangeable terms. Another innovative element of ISO 704:2022 is the explicit inclusion of antonymy, covering contrary and contradictory concepts. Even more relevant from the perspective of interoperability is the formal introduction of equivalence (7.7.3), conceived as a fundamental relation for multilingual work and as a structured relation potentially enabling alignment between distinct linguistic resources, in line with the requirements of multilingual thesauri and controlled vocabularies.

The departure from a rigidly prescriptive approach becomes evident in the transition from deprecation to acceptability rating (7.7.7). The rating system introduced in ISO 704:2022 (with preferred, admitted, deprecated and obsolete designations) guides users without imposing normative choices, reflecting contemporary practices in terminological data management and facilitating integration with heterogeneous information systems.

4. Conclusions

The analysis of ISO 704:2022 shows a clear shift toward a more articulated, methodologically grounded, and interoperable framework for terminology work.

The standard strengthens the link between conceptualization and definition writing, clarifying how concepts emerge from domain-specific properties and how their essential characteristics determine both intension and extension. It also refines the treatment of concept relations, introducing inheritance principles, a more rigorous handling of individual and partitive concepts, and a structured typology of associative relations that transforms them into operational modelling tools.

The representation of concept systems is expanded through UML-based models and a multidimensional classification of system types, reflecting the growing need for formal, machine-processable structures. Definition writing is reframed as a communicative and methodological task, with explicit guidance on essential characteristics, audience needs, and the proper use of designations. Finally, the emphasis on authoritative sources introduces a level of traceability and legal clarity before absent.

ISO 704:2022 introduces a decisive innovation in the treatment of designations. Designations are no longer conceived as static linguistic labels, but are explicitly associated with a temporal dimension that records changes in status and usage over time. Within the category of terms, the formal introduction of nomenclatures acknowledges structured designation systems governed by international scientific communities, while the reorganization of appellations and the explicit distinction between appellations and proper names clarify the representation of groups of identical objects and individual entities. At the same time, the expansion and systematization of relations between designations and concepts, particularly through the explicit treatment of equivalence, along

with a refined handling of synonymy, polysemy and homonymy, strengthen the methodological framework for multilingual and interoperable terminological resources. This evolution is further supported by the introduction of acceptability rating, which replaces rigid prescriptive mechanisms with a graded system capable of guiding usage while accommodating terminological variation.

The strong alignment of ISO 704:2022 with ISO 1087:2019, which provides the standardized vocabulary for terminological work, its methodological coherence with ISO 25964-1:2011 for thesauri and interoperability, and its connection with ISO 24156-1:2014 used for concept modelling based on the Unified Modeling Language (UML), which integrates and/or replace traditional diagrams, definitively confirm the transition towards a fully integrated, concept-centric approach to terminology management. This shift situates work on designations no longer at the level of linguistic description alone, but within the broader framework of information architecture and shared knowledge, where terminological resources are conceived as structured, interoperable components of larger semantic ecosystems.

Session 2

When Fuzziness is in the Mind, Not in the Concept: Terminology, Prototypes, and Ontological Modeling

Giorgio Maria di Nunzio[1], Federica Vezzani[2]
[1] Department of Information Engineering
[2] Department of Linguistic and Literary Studies
University of Padua
Padova, Italy
{giorgiomaria.dinunzio,federica.vezzani}@unipd.it

Abstract

In recent decades, terminology studies have increasingly engaged with findings from cognitive psychology, corpus linguistics, and discourse analysis in order to account for variation in term usage, historical change, and context-dependent meaning (Cabré, 1999; Temmerman, 2000). This development has led to the emergence of socio-cognitive approaches to terminology, which challenge classical Wüsterian models of concepts traditionally associated with standardization-oriented terminology work. In particular, prototype-based theories of categorization have been mobilized to argue that many concepts lack clear boundaries and that classical Aristotelian definitions based on necessary and sufficient characteristics are inadequate for describing real-world terminological practice (Lakoff, 1990; Rosch, 1975, 1973). These socio-cognitive insights have contributed significantly to the descriptive analysis of terminological variation and to the discussion of the possible pitfalls of the general theory of terminology. Nevertheless, this paper argues that the theoretical implications of socio-cognitive approaches for concept modeling, ontology design, and terminology management have often been overstated. The central claim we try to defend here is that graded categorization behavior and prototype effects pertain to cognitive access and usage patterns, not to the ontological structure of concepts themselves. Failure to maintain this distinction has often led to an unnecessary dichotomy (classical vs modern terminology, right vs wrong, etc.) with a consequent conceptual confusion and practical difficulties in applied terminology and ontology engineering.

1. Introduction

From an applied perspective, terminology work operates under constraints that require conceptual precision, stability, and interoperability, particularly in domains such as science, technology, medicine, engineering, and any other domain for which we (researchers) experience a high degree of change and innovation. Terminology resources are increasingly integrated into digital infrastructures, including databases, knowledge graphs, semantic web applications, and AI-driven decisionsupport systems (Di Nunzio, 2025; Gruber, 1993; Guarino et al., 2009; Khemakhem et al., 2025). In these contexts, concepts must be modeled in a way that supports logical inference, data integration, and consistent interpretation across systems and communities. This data modeling approach where concepts in an ontology (in a computer science sense) exist and are distinguishable one from the other is almost perfectly matched by the classical view of terminology, particularly in the tradition associated with Eugen Wüster. This approach, in fact, sees concepts as units of knowledge defined by characteristic sets that allow them to be clearly distinguished within a concept system. From an applied standpoint, its strength lies precisely in its normative orientation: it provides principles for constructing terminology resources that are operational, reusable, and

aligned with domain knowledge rather than with individual perspectives of what a concept could be Importantly, classical terminology does not deny that users may experience difficult in applying concepts; rather, it treats such difficulty as a matter of education, expertise, or interface design, not as evidence of conceptual indeterminacy. More recently (Di Nunzio et al., 2025), this distinction has been explicitly addressed in applied work showing how disagreements among domain experts about the definition or scope of a concept often reflect differences in perspective, interpretation, or contextual emphasis rather than genuine indeterminacy at the conceptual level. In such cases, the observed "fuzziness" arises from the interaction between experts and the concept, shaped by disciplinary background, task-specific goals, or evidential priorities, rather than from the absence of a determinate underlying concept. From this viewpoint, divergent expert definitions do not necessarily call for weakening conceptual boundaries in terminological or ontological models; instead, they motivate the explicit representation of alternative viewpoints.

Socio-cognitive terminology, most prominently articulated by Temmerman (2000), emphasizes that many terminological categories, especially in the life sciences and other rapidly evolving domains, do not exhibit the stability presupposed by classical definitions. Drawing on prototype theory and discourse analysis (Rosch, 1975), Temmerman highlights the role of typicality, metaphor, historical development, and community specific practices in shaping terminological meaning. These observations are particularly relevant for applied terminology management, where terminologists must often deal with competing definitions, overlapping concepts, and evolving domain knowledge. In this sense, on the one hand it is true that the classical approach is often criticized as overly rigid or disconnected from actual language use. On the other hand, it is also true that treating concepts as inherently fuzzy or prototype structured at the ontological level poses serious challenges for applications that use knowledge bases and ontologies for mapping terminological data. This has often resulted in an unnecessary opposition between classical and socio-cognitive terminology, framed as a choice between rigid formalism and descriptive adequacy, rather than as a difference in explanatory level.

2. Our proposal

With the same spirit of our previous works dedicated to finding positive similarities among different approaches which would help (at least in theory) to allow for integration of different resources, in this paper we argue that socio-cognitive terminology tends to conflate three distinct levels of analysis that should be kept separate for both theoretical clarity and practical effectiveness:

- First, the ontological level of concepts as elements of a domain,
- Second, the definitional level of terminological descriptions and standards, and
- Third, the cognitive level of how users access, interpret, and apply concepts in practice.

Prototype effects and graded typicality belong primarily to the third level. When they are projected onto the first level, they introduce unnecessary vagueness into conceptual models that are meant to support precise reasoning and interoperability. The practical consequences of this conflation become especially evident when terminology is integrated with ontology engineering and artificial intelligence (Smith, 2001). In AI and knowledge representation, a clear distinction is maintained between ontologies, which provide formal, explicit specifications of concepts and relations, and user models, which capture how agents reason about, misunderstand, or prioritize information. Ontologies are evaluated with respect to consistency, coverage, and inferential adequacy, while

user models are designed to accommodate variability, uncertainty, and heuristic reasoning. Importantly, inaccuracies or hesitations in user classification do not motivate changes to the ontology itself; instead, they motivate improvements in user support, explanation facilities, and interface design. This separation offers a powerful applied framework for terminology studies. Concepts can be modeled ontologically in a determinate and theory-driven manner, while prototype effects can be addressed through annotations, usage notes, corpus-based evidence, and adaptive user interfaces. For example, terminological databases can maintain precise concept definitions while also recording typical examples, frequency patterns, or domain-specific salience information to support users with different backgrounds and levels of expertise. Such an approach preserves the operational strengths of classical terminology while incorporating the descriptive insights emphasized by socio-cognitive approaches.

From this applied standpoint, the often-cited opposition between Aristotelian and prototype-based views appears less as a fundamental theoretical conflict and more as a misunderstanding of explanatory scope. Prototype theory does not demonstrate that concepts are inherently fuzzy; it demonstrates that human interaction with concepts is mediated by cognitive heuristics (Fodor, 1998; Murphy, 2002; Rosch, 1973). When this insight is properly situated at the level of cognitive access rather than concept structure, it becomes compatible with classical terminology and with contemporary ontology engineering practices. The paper therefore proposes a layered model for applied terminology and ontology work, consisting of:

1. Ontologically grounded concepts defined within a domain theory;
2. Terminological definitions that normatively describe these concepts for communication and standardization as well as empirical patterns of term usage observed in corpora and discourse;
3. Cognitive prototypes that influence how users process terms and apply concepts.

Each layer serves a distinct function and should be modeled using appropriate methods and tools. Adopting this layered approach has concrete benefits for applied terminology and AI-enabled systems. It supports the development of interoperable terminology resources, facilitates integration with formal ontologies and knowledge graphs, and enables the design of user-centered applications without compromising conceptual rigor. It also provides a principled way to incorporate socio-cognitive insights into terminology management while avoiding the erosion of conceptual clarity that is essential for standardization and automated reasoning.

3. Final Considerations

In this paper, we argue that the question is not whether terminology should account for cognitive and discursive variation, it should, but how this variation should be represented within terminological and ontological systems. By restoring a clear distinction between user-related cognitive processes and the formal structure of concepts, and by aligning terminology theory with established practices in ontology engineering and AI, this approach avoids conceptual confusion while preserving practical effectiveness. We discuss an applied and methodologically coherent framework for addressing fuzziness, variation, and typicality in contemporary terminology work. Importantly, this position does not deny the usefulness of representing uncertainty or vagueness in applied systems; rather, it aligns with existing work in ontology engineering that treats fuzziness as a modeling or reasoning layer, not as a property of concepts themselves.

References

Teresa Cabré. Terminology. Theory, Methods and Applications. John Benjamins, 1999. ISBN 978-90-272-9865-2. doi: 10.1075/tlrp.1.

Giorgio Maria Di Nunzio. Terminology-Augmented Generation (TAG): Foundations, Use Cases, and Evaluation Paths. Journal of Digital Terminology and Lexicography, 1(1):97–104, July 2025. ISSN 3103-360. doi: 10.25430/pupj.jdtl.1752566034.

Giorgio Maria Di Nunzio, Lucia Manni, Emanuela De Lisa, and Federica Vezzani. The Elephant in the Room: The Impact of Domain-Expert (Dis)agreement. A Case Study on Cryptic Species. In Federica Vezzani, Giorgio Maria Di Nunzio, Elpida Loupaki, Georgios Meditskos, and Maria Papoutsoglou, editors, Proceedings of the 4rd International Conference on Multilingual Digital Terminology Today (MDTT 2025), volume 3990 of CEUR Workshop Proceedings, Thessaloniki, Greece, June 2025. CEUR. URL https://ceur-ws.org/Vol-3990/\#short25.

Jerry A. Fodor. Concepts: Where Cognitive Science Went Wrong. Clarendon Press, February 1998. ISBN 978-0-19-151906-2.

Thomas R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, June 1993. ISSN 1042-8143. doi: 10.1006/knac.1993.1008.

Nicola Guarino, Daniel Oberle, and Steffen Staab. What Is an Ontology? In Steffen Staab and Rudi Studer, editors, Handbook on Ontologies, pages 1–17. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-92673-3. doi: 10.1007/978-3-540-92673-3\ 0.

Mohamed Khemakhem, Cristina Valentini, Natascia Ralli, Sérgio Barros, Georg Löckinger, Federica Vezzani, Ana Salgado, Zhenling Zhang, Sabine Mahr, Sara Carvalho, Klaus Fleischmann, and Rute Costa. Terminology Management Meets AI: The ISO/TC 37/SC 3/WG 6 Initiative. In Proceedings of the 5th Conference on Language, Data and Knowledge: TermTrends 2025, pages 16–24, Naples, Italy, September 2025. Unior Press. ISBN 978-88-6719-334-9.

George Lakoff. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. University of Chicago Press, Chicago, IL, April 1990. ISBN 978-0-226-46804-4.

Gregory L. Murphy. The Big Book of Concepts. The MIT Press, July 2002. ISBN 978-0-262-28035-8. doi: 10.7551/mitpress/1602.001.0001.

Eleanor Rosch. Cognitive representations of semantic categories. Journal of Experimental
Psychology: General, 104(3):192–233, 1975. ISSN 1939-2222. doi:10.1037/0096-3445.104.3.192.

Eleanor H. Rosch. Natural categories. Cognitive Psychology, 4(3):328–350, May 1973. ISSN 0010-0285. doi: 10.1016/0010-0285(73)90017-0.

Barry Smith. Beyond Concepts: Ontology as Reality Representation. In Barry Smith and Christopher Welty, editors, Formal Ontology in Information Systems (FOIS), pages 1–12. ACM Press, 2001.

Rita Temmerman. Towards New Ways of Terminology Description. John Benjamins Publishing Company, 2000. ISBN 978-90-272-2326-5. doi: 10.1075/tlrp.3.

An OWL Ontology and Ontoterminology for Classical Athenian Legal Events

Rachel Milio
TALOS-AI4SSH, University of Crete, Greece
rachel.milio@outlook.com
https://orcid.org/0009-0000-0420-4711

Abstract

This paper proposes an ontological and ontoterminological model for the representation of Classical Athenian Legal Events. This model combines the OWL ontological approach and the ISO 1087 and 704-compliant ontoterminological approach to define the central class of Legal Event. While other ontological models exist for modern legal systems, this paper asserts that Athenian Law is alien to modern systems and therefore requires a unique modelling approach. This model links Legal Events, involved people, and the surviving texts which serve as a primary resource for the study of Classical Athenian Law as a part of broader research towards modelling the corpus of Attic Oratory. This paper situates the Athenian Law model within the field of ontology, more specifically event ontologies and case-based legal ontologies, and outlines key domain constraints of Athenian Law which differ from modern law. This model can then be used to structure knowledge of legal events in Classical Athens as a machine-readable, queryable knowledge graph.

1. Introduction

An inescapable consideration for scholars of Classical Athenian Law is that it is "substantially alien" to modern systems[1] (Todd, 1993, p. 68). Although there are surface level similarities in regards to the presence of formal legal events, domain experts caution against filling in gaps with assumptions drawn from modern legal systems (see Loomis (1972) for a discussion of modern homicide categories i.e. first-degree, second-degree, manslaughter in US law in contrast to the Classical Athenian distinctions). This challenge is not unique to scholars of Athenian law, but rather is indicative of a broader challenge faced by all scholars of the past who seek to model their domains while avoiding the biases and assumptions of the present day. In order to avoid such assumptions, this paper proposes a unique ontological model for Classical Athenian Legal Events as part of a doctoral research project towards modelling the corpus of Attic Oratory (speeches written in Classical Athens for oral delivery, most often in a court setting). The model uses the central class of Legal Event to link between events, people, and speeches.

This paper uses two paradigms for structuring data: ontology and ontoterminology. Formal ontology in computer science is traditionally defined as "an explicit specification of a

[1] In this paper, "modern law" and "modern legal systems" refer both to civil law, practiced in much of the world including Europe, as well as common law, practiced in the US and British Commonwealth.

conceptualization" (Gruber, 1993, p. 199). An ontology written in OWL (the Web Ontology Language) provides a formal model of a domain using Description Logics (DLs) to define classes and properties within said domain (Guarino et al., 2009). Meanwhile, ontoterminology is "a terminology whose conceptual system is a formal ontology," meaning it implements ISO 1087 and 704 standards for terminology as an OWL ontology (Roche, 2012, p. 2626). OWL ontologies use *extensional* class definitions (set inclusion and property restrictions) while ontoterminologies use Aristotelian, *intensional* concept definitions, in which a concept is defined as a unique combination of essential characteristics (Roche, 2023). This research uses an ontoterminology to define types of legal processes (Milio et al., 2024, updated version forthcoming), and an OWL ontology to define Legal Events and their relations with people and speeches. The combination of both paradigms enables two types of definitions: specialized language, such as ancient Greek and English terms for legal processes, can have human-friendly but machine-readable terminological definitions; and the superclass of Legal Event can be defined according to its relations with other classes in the ontology.

2. Related Work

In Cultural Heritage domains, the standard event model, which is also an ISO standard since 2006, is CIDOC CRM (Bekiari et al., 2021). The CIDOC CRM class E5_Event defines events as "distinct, delimited and coherent processes and interactions of a material nature, in cultural, social or physical systems."[2] Because of CIDOC CRM's broad, general nature, other domain ontologies extend CIDOC CRM to ensure interoperability. For example, the LACRIMALit ontology of ancient crisis events extends the CIDOC CRM class E7_Activity, a subclass of E5_event that are carried out by an E39_Actor (Papadopoulou et al., 2022). The LINCS Event Vocabulary also extends CIDOC CRM, connecting its Event class to E7_Activity through the skos:related property, in order to describe biographical events within the domain of British and Canadian history and literature (Warren et al., 2023). In addition to CIDOC CRM, Papadopoulou et al. (2019) provides an overview of other conceptualizations of events developed for the Semantic Web, including ArCo, the Italian Cultural Heritage Knowledge Graph (Carriero et al., 2019); and the Linking Open Description of Events (LODE) (Shaw et al., 2009).

Also important to this research are legal ontologies, which model some aspect of a legal system such as laws, jurisdiction, or cases. Especially relevant to this work is the OWL ontology for legal cases proposed by Wyner and Hoekstra (2012). The purpose of this ontology is to annotate cases for the sake of a queryable database. The central class of the ontology is Case, which Wyner and Hoekstra "under-define" as "the class comprised of individuals which have a defendant, a plaintiff, and a huge; the class comprised of the union of decided and undecided cases" (2012, p. 92). Other classes include Decision, Jurisdiction, Participant (including witnesses, judges, solicitors, defendants, and plaintiffs), and Evidence.

[2] http://www.cidoc-crm.org/cidoc-crm/E5_Event

3. Features of Athenian Law

This research combines the event-driven approach of models such as CIDOC CRM with the needs of a legal ontology. However, in modelling Classical Athenian Legal Events, there are a few features of the domain of Athenian law as opposed to modern law which pose modeling constraints. These constraints relate both to the system of Athenian law itself, as well as the nature of extant sources describing this system.

1) Incomplete sources: Forensic speeches were often preserved throughout antiquity for stylistic purposes, rather than for the reconstruction of the legal system (Todd, 1990, p. 165). We cannot always attach a speech to a type of legal event, understand the exact charges, or know the names of the litigants.
2) Lack of decisions: As compared to models for modern legal cases such as Wyner and Hoekstra (2012), there are few instances in which the outcome of a legal event is known. Some decisions can be speculated using epigraphical evidence[3], and very rarely outcomes are stated in other speeches.[4] Therefore, a model of Athenian legal events cannot be focused on the outcomes of cases, most of which are unknowable.
3) Non-uniform litigants: In contrast to modern cases which consistently have a plaintiff and defendant, not every Athenian legal event has these participants. Rather, the lawsuit for adjudication (διαδικασία), well-represented in the Attic Oratory corpus, can have any number of claimants who all put forward their claim to an inheritance (Todd, 1993, p. 112).

4. Legal Event Model

Based on these key constraints of Athenian Law, the model must fully represent incomplete, heterogenous legal event data. The central class in the data model, atticlaw:Legal_Event, is a subclass of CIDOC CRM's E5_Event class, in order to facilitate interoperability with other Linked Open Data projects (Figure 1). Other notable classes include atticlaw:Agent, which is a subclass of E39_Actor and itself has subclasses for People and Legal Bodies, and orators:Speech, defined in an ontoterminology for types of rhetorical speeches in Classical Athens (publication forthcoming). The Legal Event ontology links to standards such as CIDOC CRM through properties such as rdfs:subClassOf rather than directly reusing these classes. Using existing vocabularies can force knowledge engineers to think in terms of that vocabulary rather than their own domain. Therefore, in order to maintain the specificity of the Legal Event domain while still enabling data sharing and interoperability, the domain model tailors its classes to the domain and then links them to external vocabularies.

---

[3] An example of this is the suit for damages between two half-brothers, both named Manitheus, over whether the defendant brother (also known as Boeotus) has the right to use the name Mantitheus in public affairs, argued in Dem. 39 (Scafuro, 2011). Both half-brothers appear in a later naval inscription as Mantitheus (*IG* II^2 1622.442–443), leading scholars to assume that the defendant Boeotus won his case.

[4] For example, we know that Aeschines won his case against Timarchus, argued in Aes. 1, from another speech delivered by Demosthenes (Dem. 19.257) for a separate legal dispute against Aeschines (Carey, 2011).

The atticlaw:Legal_Event class is comprised of individuals with at least two people as participants, and argued in at least one forensic speech. People are connected to Legal Events through the property :hasParticipant, which has subproperties for when participants are plaintiffs, defendants, claimants, or witnesses. Due to the non-uniformity of litigants, a Legal Event is not required to have a plaintiff and defendant. In cases where a plaintiff/defendant/claimant are not named, an individual is still created, labelled "Unnamed Plaintiff" etc. Furthermore, only Legal Events substantiated by a speech within the corpus of Attic Oratory are recorded, meaning every Legal Event is linked to a speech.[5]

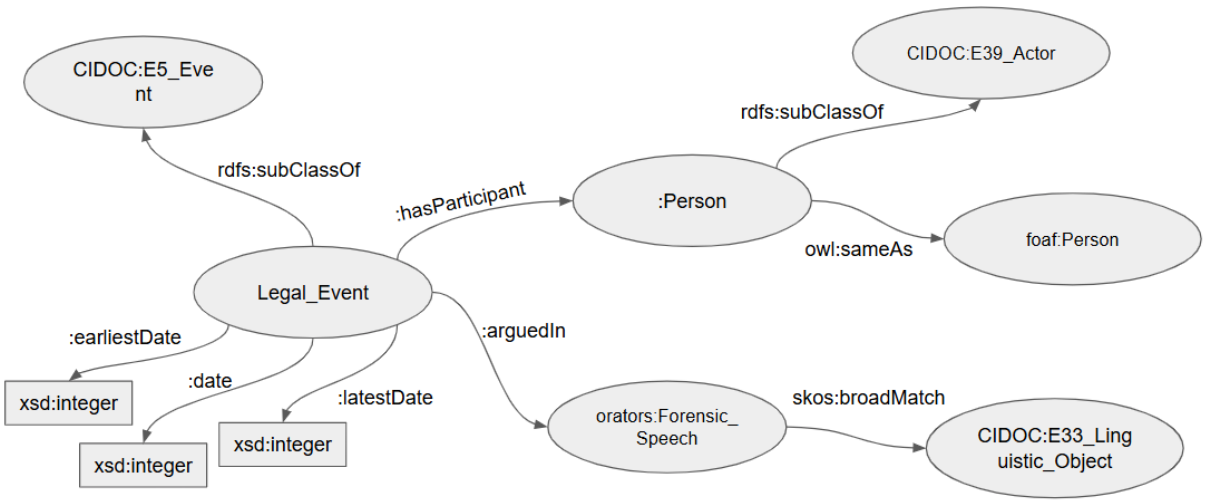


Figure 1. Legal_Event model as an extension of CIDOC CRM[6]

A unique aspect of this event model is the use of an ontoterminology to define the subclasses of Legal_Event. The Legal Processes ontoterminology (Milio et al., 2024) defines the types of disputes handled in the Athenian legal system, using essential characteristics to form concept definitions. Types of Legal Events are robustly defined and aligned with English and ancient Greek designations. The Legal Processes ontoterminology was created in the ontoTerminology EDItor (TEDI), which enables the user-friendly creation of ISO 1087 and 704 compliant ontoterminologies (Roche, 2026). The value of using the paradigm of ontoterminology to define specific legal event types, rather than creating the entire model within the standard ontology editor Protégé, is the streamlined creation of Aristotelian definitions which use essential characteristics. Definitions in OWL ontologies are *extensional*, based on classical first-order logics and property restrictions (Roche, 2023, p. 18). Meanwhile, Aristotelian definitions based on essential characteristics use second-order logic, meaning that the creation of these definitions is not directly possible in DLs. Therefore, in order to incorporate definitions based on essential

[5] Other Legal Events could be extrapolated based on in-text references, but these events are out of the scope of this research for now.

[6] For the sake of clarity, some aspects of the event model have been condensed in this visualization.

characteristics, which more closely resemble the thinking of humanists, this model uses ontoterminology.

A concept definition for a Legal Process created in TEDI is based on several axes of analysis, including whether a process was public or private (i.e. whether a process could be initiated by any willing male citizen, or only by the injured party), whether a process had a fixed or an assessable penalty, which magistrate would supervise the trial, and for which charges the process could be initiated. A resulting term and concept definition (in this case, the definition for the term 'suit for damages') as seen in the TEDI HTML dictionary export feature can be seen in Figure 2.

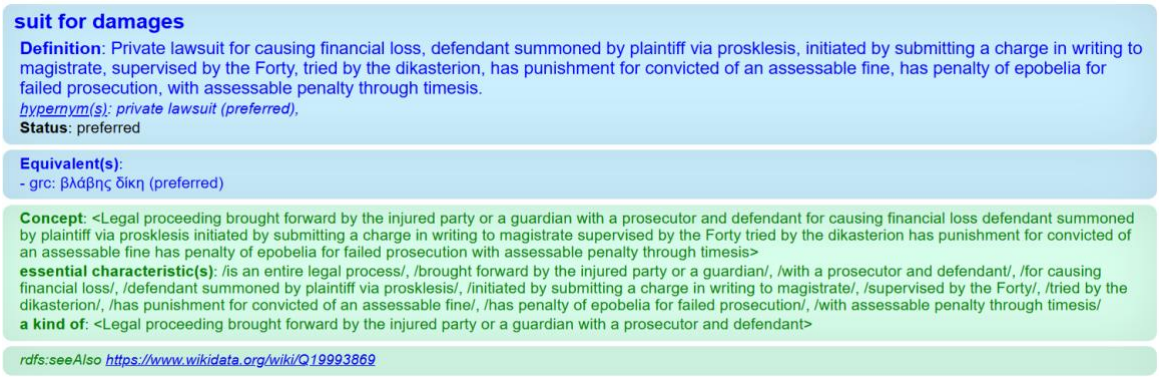


Figure 2. HTML dictionary entry for 'suit for damages'

However, the entire domain model for Legal Events could not be created in TEDI. This is because "unlike Protégé, TEDI is not a universal system for concept modelling" (Roche and Papadopoulou, 2019, p. 3). While legal processes are defined through their essential characteristics, the generic Legal Event is defined through the involved participants and speeches (the other relevant aspects of the domain). Therefore, an instance of a Legal Event of the type designated by the term 'suit for damages' also must have at least two participants and be argued in at least one speech (Figure 3).

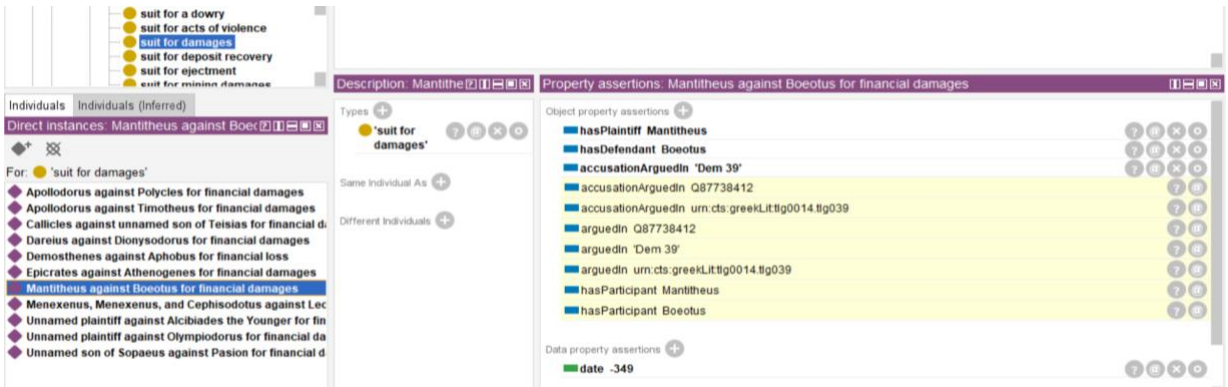


Figure 3. An instantiation of a Legal_Event, the suit by Mantitheus against Boeotus/Mantitheus for damages

5. Conclusions

This paper presents a model for Classical Athenian Law which adapts the unique features of the domain for representation as an OWL ontology as well as an ontoterminology. Due to the alienness of the Athenian legal system and the incomplete or fragmentary nature of our sources for Athenian legal events, models for modern legal systems are not suitable, necessitating the development of this new model centered around the class Legal_Event. This model incorporates the extensional definitions of ontology to define its central class, as well as the intensional, Aristotelian definition of ontoterminology to define the types of legal processes in Classical Athens. Through this model, Legal Events can be queried based on their type, date, participants, and relevant texts in the Attic Oratory corpus. In the future, this model will be populated with other participants, such as witnesses. Furthermore, the model will be part of a broader knowledge graph for Attic Oratory, including a prosopography of Athenians mentioned in the corpus. Through this, the domain of Attic Oratory and Classical Athenian Law can be standardized as Linked Open Data on the Semantic Web.

References

Bekiari, C., Bruseker, G., Doerr, M., Ore, C.-E., Stead, S., & Velios, A. (2021). *Volume A: Definition of the CIDOC Conceptual Reference Model* (Version 7.1.1). https://cidoc-crm.org/sites/default/files/cidoc_crm_v.7.1.1_0.pdf

Carey, C. (2011). Aeschines 1.: Against Timarchus. In M. Gagarin (Ed.), Speeches from Athenian Law (pp. 183–244). University of Texas Press. https://www.jstor.org/stable/10.7560/723627.17

Carriero, V. A., Gangemi, A., Mancinelli, M. L., Marinucci, L., Nuzzolese, A. G., Presutti, V., & Veninata, C. (2019). ArCo: The Italian Cultural Heritage Knowledge Graph. *The Semantic Web – ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II*, 36–52. https://doi.org/10.1007/978-3-030-30796-7_3

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/knac.1993.1008

Guarino, N., Oberle, D., & Staab, S. (2009). What Is an Ontology? In S. Staab & R. Studer (Eds.), *Handbook on Ontologies* (pp. 1–17). Springer. https://doi.org/10.1007/978-3-540-92673-3_0

Loomis, W. T. (1972). The Nature of Premeditation in Athenian Homicide Law. *The Journal of Hellenic Studies*, *92*, 86–95. https://doi.org/10.2307/629975

Milio, R., Papadopoulou, M., & Roche, C. (2024). Ancient Greek Oratory Ontology v.1.0 (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13379144

Papadopoulou, M., Roche, C., & Tamiolaki, E.-M. (2022). The LACRIMALit Ontology of Crisis: An Event-Centric Model for Digital History. *Information*, *13*(398). https://www.mdpi.com/2078-2489/13/8/398

Roche, C. (2012). How to unify terminology and ontology into a single paradigm. *Eighth International Conference on Language Resources and Evaluation*, 2626–2630.

Roche, C. (2023). Ontology for terminology: A passport to the digital world. *14th Conference "Hellenic Language and Terminology,"* 14–24.

Roche, C. (2026). Tedi. Retrieved February 7, 2026, from http://ontoterminology.com/tedi

Scafuro, A. C. (2011). Against Boeotus I. In M. Gagarin (Ed.), *Demosthenes, Speeches 39-49* (pp. 33–58). University of Texas Press. https://www.jstor.org/stable/10.7560/725560.8

Shaw, R., Troncy, R., & Hardman, L. (2009). LODE: Linking Open Descriptions of Events. In A. Gómez-Pérez, Y. Yu, & Y. Ding (Eds.), The Semantic Web (Vol. 5926, pp. 153–167). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10871-6_11

Todd, S. C. (1993). *The Shape of Athenian Law*. Oxford University Press. https://doi.org/10.1093/oso/9780198148944.001.0001

Todd, S. (1990). The Use and Abuse of the Attic Orators. *Greece & Rome*, *37*(2), 159–178.

Warren, R., Brown, S., Stacey, D., Lemak, A., Drudge-Wilson, J., Jakacki, D., Cummings, J., Martin, K., & Long, J. (2023). *Event Vocabulary*. LINCS. https://vocab.lincsproject.ca/Skosmos/event/en/

Wyner, A., & Hoekstra, R. (2012). A legal case OWL ontology with an instantiation of Popov v. Hayashi. *Artificial Intelligence and Law*, *20*(1), 83–107. https://doi.org/10.1007/s10506-012-9119-6

Encoding Linguistic Memory: A Semantic Web Approach to Pre-Soviet and Diaspora Ukrainian Feminine Occupational and Status Terms

Olena Synchak

Department of Slavic Studies, University of Klagenfurt
Olena.Synchak@aau.at, o_synchak@ucu.edu.ua

Extended Abstract

This paper presents an ontology-driven, Semantic Web–based approach to the analysis of feminine occupational and status designations as attested in pre-Soviet urban usage in L'viv and in Ukrainian diaspora speech in Munich. The study is conducted within the framework of the LUFEMM project, which examines the evolution and transmission of Ukrainian feminine personal nouns in the twentieth century, focusing on these two historically and sociolinguistically interconnected sites. The project's overarching aim is the preservation of pre-Soviet linguistic heritage. Against this background, the paper demonstrates how Semantic Web technologies can be used to formally represent and analyze historically grounded and socially embedded terminological variation, while remaining methodologically sensitive to different elicitation and documentation contexts.

The study examines feminine occupational and status designations largely absent from contemporary Ukrainian codification (e.g. *фризьєрка* 'hairdresser.F'*, касьєрка* 'cashier.F') but normative for speakers shaped by pre-Soviet or anti-Soviet Western Ukrainian codificatory regime, whose status changed through Soviet language planning that promoted alternative standardized forms (e.g. *перукарка* 'hairdresser.F', *касирка* 'cashier.F'). The coexistence of pre-Soviet and Soviet codificatory norms constitutes a central methodological challenge of the study. Preserved in the linguistic memory of elderly speakers, these designations emerge primarily in experiential contexts and can be described as historically layered, culturally dependent lexical material. Their current marginal position reflects codificatory and ideological interventions during the Soviet period rather than inherent instability within the lexical system.

Soviet language policy reconfigured established concept–lexeme mappings without eliminating the underlying occupational role concepts. While feminine designations were often replaced in codification by masculine default forms, the corresponding gender-specific sub-concepts lost terminological recognition but did not disappear from cognitive or social practice. These forms were preserved in linguistic memory through intergenerational transmission and reinforced in post-war migration contexts. This distinction between conceptual persistence and codificatory suppression can be explored through Semantic Web–based modeling of historically competing codificatory layers, informed by a comparison between data collected in L'viv and Munich with entries from both diaspora and Soviet-era official dictionaries.

Research background and objectives

Framed by ideologically driven language standardization in the Soviet period [3], the LUFEMM project examines feminine occupational and status designations at the intersection of language standardization, ideology, and language contact, with a focus on their historical transmission. Previous research has shown that Soviet language planning and Russification policies significantly disrupted pre-Soviet feminization practices in Ukrainian, generally in lexicography [6] and particularly in the legal [4] and military-related domains [5]. At the same time, Ukrainian diaspora communities—while often characterized by domain-restricted

language use [1] —have functioned as reservoirs of pre-Soviet linguistic norms [4], albeit under conditions of reduced terminological activation.

This study addresses two interrelated research questions. First, how do different elicitation methods activate competing terminological strata in feminine occupational naming among elderly speakers in L'viv and Munich? Second, how can Semantic Web technologies be used to model this variation in a way that preserves historical depth, social embedding, and methodological context rather than flattening them into a single "normative" layer?

Data and elicitation design

The dataset analyzed in this paper is based on visual-stimulus–based surveys and written questionnaires collected in L'viv and Munich. The sample consists of elderly Ukrainian-speaking women born before 1950 (ten respondents in each city), many of whom experienced post–World War II displacement and long-term multilingual contact. This demographic profile is crucial, as it captures speakers whose linguistic repertoires were shaped by pre-Soviet or early Soviet norms and later maintained under diaspora conditions.

The methodological design deliberately combines two elicitation modes with distinct cognitive and sociolinguistic affordances. Visual elicitation activates experiential, socially embedded lexical knowledge by prompting respondents to name professions or social roles depicted in images. Written questionnaires, by contrast, tend to elicit norm-oriented, metalinguistically controlled responses shaped by literacy practices and awareness of codified standards. The contrast between these modes allows for the identification of competing lexical strata within the same terminological domain.

A recurring observation in the Munich data concerns explicit metalinguistic commentary on limited opportunities for terminological use. Respondents frequently note that they "have no one to use these words with" or resort to descriptive paraphrases (e.g., "a woman who cooks") instead of established occupational terms. In some cases, respondents report that they can imagine the term hypothetically but do not employ it in everyday communication. These responses underscore the fragility of occupational terminology in diaspora contexts and highlight the importance of elicitation design for accessing latent terminological knowledge.

Terminological findings: competing lexical strata

The analysis reveals the systematic coexistence of at least two competing terminological strata. The first consists of Soviet standardized literary forms promoted through twentieth-century codification and language planning, as documented in the *Explanatory Dictionary of the Ukrainian Language* (SUM, 1970–1980). The second comprises non-codified but stable feminine designations rooted in a Western Ukrainian urban koine, as attested in the *Ukrainian–German Dictionary* by Kuzelia and Rudnytsky (Leipzig, 1943), and preserved through intergenerational transmission in L'viv and Munich, including contact-induced variants shaped by sustained interaction with Polish and German.

A recurrent pattern concerns the preferential use of Western Ukrainian urban koine forms in response to visual stimuli, observed under different elicitation conditions in Munich and L'viv. In Munich, respondents frequently produced urban koine forms (e.g. *кельнерка* 'waitress') largely without apparent awareness of the contemporary standardized alternative (*офіціантка*), suggesting continuity of pre-Soviet lexical norms within diaspora usage. In L'viv, by contrast, respondents were generally aware of the codified standard forms but

deliberately supplied urban koine designations in accordance with the interview task, which explicitly aimed at eliciting pre-Soviet linguistic practices. In both settings, such choices are more plausibly interpreted as socially embedded occupational naming practices shaped by urban experience, linguistic memory, and task-sensitive elicitation, rather than as random dialectal variation.

From a terminological perspective, these findings suggest the need to view terminology beyond strictly codified standards, taking into account historically layered and competing codificatory frameworks. Instead, they support a socioterminological view in which terminological units are embedded in social practice, historical experience, and domain-specific usage [2]. Visual elicitation appears to preferentially activate what may be termed *experiential terminological units*— lexical items grounded in lived work-related and occupational contexts —rather than normatively sanctioned equivalents [8].

Methodological framework

The analytical framework integrates Frame-Based Terminology, socioterminology, and elements of the Lexical Grammar Model. Frame-Based Terminology provides the conceptual backbone for modeling occupational domains as structured frames with roles such as Agent, Activity, Status, Domain. Socioterminology foregrounds the role of social embedding, register variation, and historical stratification, while the Lexical Grammar Model supports the analysis of recurrent grammatical and collocational patterns. The analysis also incorporates a language contact perspective to account for contact-induced variation in feminine occupational designations, both in L'viv urban koine and diaspora contexts.

Data annotation follows a multi-layer scheme that captures normative–chronological affiliation (*pre-Soviet, Soviet, post-Soviet*), register *(literary standard, urban koine, dialectal*), morphological status (*availability* or *absence of a feminine form*), elicitation type (*spontaneous, prompted, repetition*), response certainty (*confident, uncertain*) and response timing (*immediate, hesitant, delayed*). This layered annotation is essential for preserving analytical distinctions between usage-based normativity and codified normativity.

Semantic Web modeling and ontology construction

The complexity of the data—characterized by historical layering, elicitation sensitivity, and speaker-specific metadata—makes it particularly suitable for Semantic Web technologies. The paper describes the construction of a lightweight domain ontology for feminine occupational and status terminology using RDF and OWL to model historically competing codificatory layers. Occupational roles are modeled as conceptual frames, while individual lexical realizations are represented as lexical entities linked to shared concepts rather than treated as isolated variants.

Semantic relations such as register affiliation, chronological layer, elicitation mode, and speaker background are encoded as formally defined properties. This enables the representation of terminological stratification as an explicit, machine-readable structure. Elicitation events (visual vs. written) are modeled as integral components of the data, allowing for systematic analysis of correlations between elicitation mode and terminological choice.

Analytical affordances and expected results

Ontology-based representation enables complex querying and hypothesis testing through formal query languages. Researchers can systematically explore, for example, which terminological strata are preferentially activated by visual stimuli, or how response uncertainty correlates with register and migration history. More broadly, the approach transforms a heterogeneous dataset into a structured knowledge graph that supports transparency, reproducibility, and cumulative research.

The expected outcome of the study is a principled demonstration that non-codified urban koine forms may function as legitimate terminological units within specific social and historical contexts, particularly when contrasted with Soviet-era standardized forms that redefined normativity through centralized codification. By distinguishing between pre-Soviet and Soviet standards as historically competing codificatory regimes, and by taking linguistic memory into account, this paper explores a methodological perspective that complements normative models of terminology.

Challenges and implications

The paper also addresses methodological challenges, including limited sample size, uneven terminological activation in diaspora contexts, and the risk of overformalizing socially nuanced data. The methodological challenge of sampling will be addressed at later stages of the research, when two subcorpora—L'viv and Munich—will be constructed from the collected oral biographical interviews and analyzed using established model of lexical growth by Herdan-Heaps. Other challenges are discussed not as shortcomings but as inherent features of research on linguistic memory and displacement.

In conclusion, the paper argues that applying Semantic Web technologies to historically and socially complex lexical data enables a more fine-grained analysis of the relationship between conceptual persistence and codificatory suppression across competing codificatory regimes. The proposed approach has implications for terminology description, lexicography, and the digital humanities, demonstrating that forms traditionally treated as non-normative—particularly those ideologically marginalized through codification—can be modeled as historically dependent components of terminological systems.

Sources

[1] Kuzelia, Zenon, Rudnytskyy, Yaroslav. *Ukrayinsko-nimetskyy slovnyk.* [Ukrainian-German dictionary]. Leipzig, 1943.

[2] *Slovnyk ukrainskoyi movy* [A Dictionary of the Ukrainian Language]. In 11 Volumes. Kyiv: Naukova Dumka, 1970-1980. Vol. 1-11. URL: http://sum.in.ua.

References

1. Azhniuk, B. (1999). *Movna iednist' natsiï: Diaspora i Ukraïna*. Vydavnychyĭ dim Dmytra Buraho.
2. Faber, P. (Ed.). (2012). *A cognitive linguistics view of terminology and specialized language*. De Gruyter Mouton.
3. Masenko, L., Kubaichuk, V., & Demska-Kul'chyts'ka, O. (eds.). (2005). *Ukraïns'ka mova u XX storichchi: istoriia linhvotsydu. Dokumenty i materialy*. Kyïv.
4. Moser, M. (2016). The 'mirror from overseas': The history of Modern Standard Ukrainian as reflected in the North American Ukrainian newspaper Svoboda (The early years: from 1893 to the 1930s). In M. Moser, *New contributions to the history of the Ukrainian language* (pp. 411–444). Canadian Institute of Ukrainian Studies Press.

5. Synchak, O. (2025). Standardizing legal feminine terms in the Web Dictionary of Ukrainian Feminine Personal Nouns. In *Terminology & Ontology: Theories and applications – Proceedings of the 18th TOTh International Conference* (6–7 June 2024) (pp. 199–218). TOTh 2024.
6. Synchak, O. (2025). War, women, and language change: A corpus-based study of Ukrainian military-related feminine terms – Decolonial implications. *Sprawy Narodowościowe: Seria nowa*, *2025*(57), Article 3422. https://doi.org/10.11649/sn.3422
7. Synchak, O. (2023). Vplyv movnoï polityky na leksykohrafuvannia feminityviv. *Ucrainica X: Současná ukrajinistika problémy jazyka, literatury a kultury*. Univerzita Palackého v Olomouci, Olomouc, 2023. c. 63-71. https://openarchive.nure.ua/entities/publication/47657457-8374-440d-a96a-82c2b94ce0d5
8. Temmerman, R. (2000). *Towards New Ways of Terminology Description. The Sociocognitive-Approach*. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Ontological Instability and Statistical Amplification: The Paradox of "Humanizing" LLM-Generated Text

Claudiu Creanga[a,*], Liviu Dinu[a]

[a] *Faculty of Mathematics and Informatics, University of Bucharest, Strada Academiei 14, Bucharest, 010014, Romania*

Abstract

Large Language Models (LLMs) challenge the assumption that linguistic fluency serves as a proxy for human authorship. Current detection methods rely on two primary paradigms: supervised statistical classifiers (e.g., RoBERTa) and structural methods based on latent event extraction. In this study, we empirically evaluate the stability of these features against semantic and structural perturbations using the M4 dataset ($N = 10{,}000$). We document a counter-intuitive phenomenon we term **Statistical Amplification**: adversarial rewriting intended to "humanize" text inadvertently increases Verb Diversity (from 0.76 to 0.92), amplifying the statistical artifacts used by classifiers. Simultaneously, we find that structural detection methods are remarkably brittle: adversarial paraphrasing disrupts 87% of extracted event sequences (Jaccard similarity 0.067), rendering event-based features unstable. These findings have implications for the analysis of specialized languages: current tools conflate the *terminological complexity* of scientific writing with the *statistical complexity* of AI generation, leading to high false positive rates in academic domains.

1. Introduction

In specialized domains such as terminology and knowledge engineering, the reliability of a text is often inferred from its linguistic precision and structural coherence. However, the emergence of generative AI necessitates

*Corresponding author

Email address: claudiu.creanga@fmi.unibuc.ro (Claudiu Creanga)

a rigorous re-evaluation of these markers. Does the "complexity" of a text indicate expert human authorship, or merely high-entropy machine generation?

Current state-of-the-art detection approaches generally fall into two categories: supervised classifiers fine-tuned on transformer architectures ([4]) and zero-shot methods that exploit probability curvature ([2]). While these models report impressive benchmark performance (often exceeding 99% accuracy), their mechanisms remain opaque, and they frequently exhibit brittle performance in deployment. In practice, they struggle with out-of-distribution data, domain shifts, and simple adversarial perturbations ([10]). Recent work has proposed "Structural Detection" ([5]) as a more robust alternative, hypothesizing that while AI can mimic surface syntax, it cannot replicate the coherent "Event Sequences" characteristic of human reasoning.

This paper investigates the boundaries of these paradigms through the lens of conceptual stability. We stress-test a fine-tuned RoBERTa-based detector and a Latent Space structural detector against three distinct attack surfaces: semantic rewriting, structural perturbation, and tokenizer manipulation. Our objective is to determine whether these systems detect genuine *semantic artificiality* or merely superficial *statistical artifacts*.

2. Methodology

We utilize the M4 dataset ([8]), a multilingual, multi-domain benchmark that includes diverse registers ranging from narrative (WikiHow) to highly specialized academic discourse (ArXiv, PeerRead). To evaluate feature stability, we introduce two primary metrics:

1. **Verb Diversity (VD):** The Type-Token Ratio of verbs, serving as a proxy for lexical variation and statistical complexity.
2. **Event Sequence Preservation (ESP):** The Jaccard similarity between extracted event triggers before and after semantic perturbation.

We employ Mistral-7B-Instruct to generate adversarial paraphrases designed to "humanize" the text. The prompt explicitly instructs the model to vary sentence structure and vocabulary, testing whether the underlying semantic features survive synonymous transformation. Additionally, we conduct a controlled "Persona Stress Test" ($N = 300$) to analyze how varying levels of requested complexity (Child vs. PhD personas) impact detection scores.

3. Results: The Disconnect Between Form and Meaning

Our analysis reveals that current detection methods struggle to distinguish between lexical variation (a stylistic choice) and ontological structure (a semantic invariant).

3.1. Statistical Amplification and the "Academic Mimicry" Problem

Standard adversarial theory suggests paraphrasing should evade detection by smoothing out generator-specific statistical artifacts. However, our results contradict this assumption. We found that instruction-tuned rewriting *amplifies* these artifacts rather than removing them.

Rewriting increased Verb Diversity from 0.762 to 0.924. This 21% increase in lexical diversity resulted in the detection score remaining high (0.949). The detector relies on high lexical diversity as a proxy for "machine-generated" text. This dependency on complexity creates a conflict with academic discourse. Formal human writing (e.g., ArXiv) naturally exhibits high lexical density and complex terminology. Because the detector overfits to complexity, it yields a **76.3% False Positive Rate** on human scientific abstracts. The tool effectively penalizes the precise terminology of the expert, mistaking it for the statistical noise of the machine.

3.2. The Entropy Paradox: Lexical vs. Semantic Complexity

To provide a theoretical explanation for this amplification, we compared two distinct measures of entropy: Gzip compression ratios (a proxy for lexical/character-level entropy) and Model Perplexity (a proxy for semantic/predictive entropy).

Table 1: The Entropy Paradox: Machine text is semantically predictable but lexically hyper-diverse.

Text Type	**Gzip Ratio (Lexical)**	**Perplexity (Semantic)**
Human (M4)	0.541 ± 0.03	40.0 ± 19.3
Machine (Bloomz)	0.587 ± 0.04	30.2 ± 12.2

As shown in Table 1, machine text exhibits a paradox. It is significantly more predictable to a language model ($PPL_{machine} = 30.2$ vs. $PPL_{human} = 40.0$), confirming that LLMs gravitate toward high-probability semantic paths.

Conversely, machine text is *less compressible* ($Ratio_{machine} = 0.587$ vs. $Ratio_{human} = 0.541$), indicating higher character-level information density.

This confirms our **Statistical Amplification** hypothesis: while LLMs minimize **semantic risk** (choosing "safe" concepts), they express these concepts using an optimized, **hyper-diverse** vocabulary that minimizes character-level redundancy.

3.3. Persona Analysis: The Mimicry Ceiling

To further investigate the link between complexity and detection, we prompted Mistral-7B with three personas: "Child," "Neutral," and "PhD." This allows us to test the model's ability to modulate its register.

- **PhD Persona:** Maximized diversity (0.948 ± 0.06) and achieved a perfect detection score (1.000).
- **Neutral Persona:** Maintained high diversity (0.818 ± 0.10) with a perfect detection score (1.000).
- **Child Persona:** Lowered diversity to 0.764 ± 0.14. Even this **simplified text** failed to reach the natural human baseline of 0.572. The detector still identified 98.8% of these samples.

This demonstrates a **"Complexity Ceiling"**: current LLMs struggle to simulate the low-complexity repetition typical of natural human cohesion. Even when explicitly asked to be simple ("Explain to a 5-year-old"), they generate a statistical signature that is "too rich" to be human.

3.4. Structural Brittleness in Knowledge Extraction

We evaluated the Latent Space detector to see if extracting "events" offers a more stable anchor than surface statistics.

Table 2: Impact of Semantic Paraphrasing on Event Structures

Metric	Value	Outcome
Original Detection Score	1.000	Detection Success
Post-Paraphrase Score	0.130	Detection Failure
Event Sequence Preservation	**0.131**	**Structural Collapse**
Jaccard Similarity	0.067	Near-zero overlap

As shown in Table 2, adversarial rewriting destroys 87% of extracted event sequences. While the narrative meaning is preserved for a human reader, the specific verbs used to trigger "events" in the extraction pipeline are altered (e.g., "analyzed" becomes "examined"). This confirms that automatically extracted event sequences are not invariant properties of the text. They are highly sensitive to surface realization, making them unreliable candidates for mapping the "knowledge structure" of AI discourse.

4. Tokenizer Blindness: The Semiotic Gap

Finally, we examined the reliance of detectors on specific token encodings versus visual signs. We employed a Homoglyph Attack, replacing Latin characters with visually identical Cyrillic counterparts.

- **Result:** A 30% character substitution rate reduced detection accuracy by 41%.
- **Implication:** The detector does not process the *term* (the visual signifier) but rather the *token ID*. This "Tokenizer Blindness" reveals a disconnect between the semantic layer intelligible to humans and the processing layer of the model.

5. Conclusion

Supervised detectors currently conflate **linguistic complexity** with **artificiality**. This poses a barrier to the analysis of Language for Special Purposes (LSP), as the statistical signature of expert terminology overlaps with the signature of high-temperature AI generation. The low Jaccard similarity (0.067) under paraphrasing suggests that current event extraction methods are insufficiently robust for building stable knowledge graphs from AI-generated text. A relation extracted from one surface form may vanish in a synonymous paraphrase, indicating that the "knowledge" captured is often a surface artifact rather than a deep semantic invariant.

Declaration of Generative AI Use

In accordance with the conference guidelines, the authors certify that no generative AI tools were used for the writing, editing, translation, or idea generation of this manuscript.

Generative AI models (specifically *Mistral-7B-Instruct*) and detection models (based on *RoBERTa*) were utilized exclusively as **experimental tools** and objects of study to generate the adversarial datasets and perform the classifications described in the Methodology section.

6. Acknowledgements

This research is supported by:

- the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416;
- a grant from Accenture Lab;

References

[1] Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., et al. (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

[2] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C.D., & Finn, C. (2023). DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 24950–24962). PMLR.

[3] Gritsai, G., Voznyuk, A., Grabovoy, A., & Chekhovich, Y. (2024). Are AI Detectors Good Enough? A Survey on Quality of Datasets with Machine-Generated Texts. *arXiv preprint arXiv:2410.14677*.

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

[5] Tian, Y., Pan, Z., & Peng, N. (2024). Detecting Machine-Generated Long-Form Content with Latent-Space Variables. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 10619–10635). Association for Computational Linguistics.

[6] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., et al. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

[7] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., Geiping, J., & Goldstein, T. (2024). Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. *arXiv preprint arXiv:2401.12070*.

[8] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Afzal, O.M., Mahmoud, T., Sasaki, T., Arnold, T., Aji, A.F., Habash, N., Gurevych, I., & Nakov, P. (2024). M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1369–1407). St. Julian's, Malta: Association for Computational Linguistics.

[9] Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. In *Advances in Neural Information Processing Systems*, 36.

[10] Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? *arXiv preprint arXiv:2303.11156*.

[11] Wolff, M., Wolff, S., Attacking neural text detectors, *arXiv preprint arXiv:2002.11768*, 2020.

[12] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

[13] 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 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics.

The More the Better? Terminological Data as Knowledge Base for a RAG-based Question Answering

Christian Lang (Leibniz Institute for the German Language (IDS), Mannheim)
lang@ids-mannheim.de
Karolina Suchowolec (TH Köln)
karolina.suchowolec@th-koeln.de
Vanessa Jochum (TH Köln)
vanessa.jochum@th-koeln.de

1 Introduction

Prescriptive terminological databases (TDBs) as proposed by the Vienna School (Drewer & Schmitz, 2017) are designed to ensure consistent language use, particularly in functional texts such as manuals. While these TDBs can be considered part of the so-called knowledge paradigm (e.g. L'Homme, 2021), their primary function is to control word forms rather than meanings. However, a prescriptive TDB also imposes an application-specific view on specialized concepts. Building on the idea of a knowledge paradigm and community's proposals to use TDBs for application-specific knowledge management (Drewer et al., 2017), we examine the use of a traditional, prescriptive TDB as a source of application-specific knowledge for RAG-based question answering (for RAG, cf. Lewis et al., 2020; Shuster et al., 2021). In particular, we address the question which subset(s) of the fields of a TDB yield the best results, when used as a knowledge base in a RAG- based question-answering system. We explicitly do not address the question on the performance of an LLM-based question answering system with no additional knowledge base compared to a system employing a RAG pipeline.

Our use case is a chatbot for prospective students or freshmen at a German university that answers questions on a specific academic program. In our experiment, we systematically vary different LLMs, question types as well as the amount and type of information in the terminological knowledge base. The quality of generated answers is first assessed automatically using the LLM-as-Judge approach; then, selected answers are assessed by human evaluators.

With this study, we address the common practice of using prescriptive TDBs as sources of application-specific knowledge in RAG pipelines, broadening our understanding of reusing existing language resources in AI applications.

2 Related Work

Using TDBs as an external knowledge base for LLMs (outside the prompt) is a new field. Beside positional papers and conceptual work (Di Nunzio, 2025; Fleischmann & Lang, 2025), some studies already exist. Lackner, Vega-Wilson, and Lang (2025) query a commercial TDB via an API to retrieve preferred terms for terminology revision and machine translation. Although no separate terminological knowledge base is created, the prompt is augmented by

the query result. In another terminology-revision approach,[1] the prompt is typically modified directly – by replacing non-preferred terms – rather than being augmented. Prompt augmentation is only required when revising ambiguous terms. Further, Hamm (2025) touches upon the question, which fields of a TDB are needed for term disambiguation in a RAG-based machine translation, but no systematic answer is given. All these applications for terminology revision and translation focus on TDBs as a source of linguistic knowledge.

Terminological data as a source of domain-specific knowledge were used in the study by Lang et al. (2024). Here, terminological data were part of a larger knowledge base in a naive RAG pipeline for question answering on linguistics. However, a fixed-size chunking was used, which might not be optimal for structured data such as concept-oriented TDBs (cf. Lackner et al., 2025).

The evaluation of the quality of the automatically generated texts poses a major challenge. While evaluation by human experts remains a central, albeit time-consuming and costly, method, a variety of automatic evaluation approaches exist. These range from statistical measures such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), which provide useful quantitative indicators but may not fully capture human judgments (cf. e.g. Krishna et al., 2021), to the use of LLMs for evaluating AI-generated answers within the LLM-as-Judge approach (e.g. Chiang et al., 2023; Kim et al., 2024; Pombal et al., 2025).

For the above-mentioned terminology-related text generation tasks, BLEU and COMET are already in use (cf. Lackner et al., 2025). For manual assessment, Remy et al. (2023) propose a framework consisting of three dimensions – factuality, insight and fluency – each rated on a 1–5 scale. Aggregated ratings result in six levels (high–good–usable–useful–useless–hurtful) for the overall quality assessment of the generated text.

3 Method

To address the research question, we conducted an experiment where 27 questions, covering three different types, were answered using a naive RAG workflow (cf. Gao et al., 2024).[2] Here, we systematically varied five answer-generating LLMs and 28 different terminological knowledge bases. This results in a 3 × 5 × 28 design. Then, we took an LLM-as-Judge approach for automatically assessing the quality of the generated answers using criteria adapted from Remy et al. (2023). Finally, for the top-performing LLM, we selected answers generated using three knowledge bases for human evaluation. Because we do not compare a RAG-based question answering system with a system with no additional knowledge base, we did not use the performance of a 'bare' LLM as a baseline.

3.1 Implementation

For implementation, we used a combination of local and server-based infrastructure. Knowledge bases were stored in a local vector database (ChromaDB).[3] As embedding model for knowledge base texts and questions we used multilingual-e5-large (Wang et al., 2024).

[1] Thiemo von Gillhaußen, Congree Language Technologies GmbH, personal communication, October 23rd, 2025

[2] Although there are more sophisticated approaches to RAG (cf. Gao et al., 2024), we chose naive RAG for this study in order to maintain a controlled experimental design and keep the focus on comparing the knowledge bases.

[3] https://github.com/chroma-core/chroma

For answer generation, we used LLMs hosted on the GWDG infrastructure (Doosthosseini et al., 2024). The evaluation of the answers using the LLM-as-Judge method was then performed again on local hardware. All procedures were implemented in Python.

3.2 Knowledge Bases

The original terminological data is a small set of 84 concepts – general concepts related to the German higher education system (e.g. "ECTS", "vorlesungsfreie Zeit" ["lecture-free period"]) and concepts specific to the programs offered by the institute (e.g. "Studienrichtung A" ["specialization A"], "Flexelement" ["elective subject"]). The TDB was created following the principles of the Vienna School (ISO 26162-1, 2019) and it is a lightweight version of a database used in an academic course on terminology, with some administrative and metalinguistic fields excluded. The TDB is stored in a concept-oriented terminology tool and includes the following fields:

- Concept level: academic *degree* [values: "Bachelor's", "Master's"], academic *program* [values: list of academic programs offered by the institute that are relevant to the concept], concept *definition* and subject *area* [top-down classification of the concept by a human, example values: "academic administration", "graduation"]
- Term level: *term* itself, *normative status* [values: "preferred", "non-preferred"], *usage status* [values: "in use", "obsolete"], and *context* sentence.

The fields "academic degree" and "academic program" seem to be specific to our application, but their meaning is, in fact, similar to "domain" or "product line", typical of prescriptive TDBs.

In our study, we examine which subsets of the database fields yield the best results as knowledge bases in our terminological RAG setup, and therefore considered the subsets shown in table 3 in the Appendix. We used these 28 subsets to create individual knowledge bases, chunking terminological data by concepts (cf. Lackner et al., 2025).

3.3 Questions and Question Types

In order to assess how useful terminological data can be to answer domain-specific questions, we identified three types of suitable terminology-related questions: 1) definition questions that use a preferred term; 2) definition questions that use a non-preferred term; this non-preferred term can either be stored in the TDB or it can be missing in the TDB. 3) Miscellaneous usage questions on concepts contained in the TDB.

First, we randomly sampled 18 concepts from the TDB. For nine concepts, we formulated nine definition questions using the preferred term, and parallel nine questions using a non-preferred term (of these, 5 used a term stored in the TDB, and 4 used a missing term). For the remaining nine concepts, we created nine usage questions. While definition questions consistently followed the same syntactic pattern (*What does X mean?*), usage questions lacked a consistent syntactic structure. This provided 27 questions for RAG-based

question answering.[4]

Question Type	Example
definition (PT)	*Was bedeutet "Studienrichtung"?* [*What does "specialization" mean?*]
definition (non-PT)	*Was bedeutet "Schwerpunkt"?* [*What does "subject area" mean?*]
usage	*Welche Möglichkeiten gibt es, im Ausland zu studieren?* [*What are the options for studying abroad?*]

Table 1: Question types and example items

3.4 Language Models

In total, we used five LLMs to answer the questions – Google Gemma 3 27B Instruct, Meta Llama 3.1 8B Instruct, Meta Llama 3.3 70B Instruct, Mistral Large Instruct, and OpenGPT-X Teuken 7B Instruct Research. We considered different sizes of LLMs to gain insight as to whether smaller LLMs also provide satisfactory results when combined with a terminological knowledge base. All models were hosted on the GWDG infrastructure and accessed via the same OpenAI-compatible API.

For question answering, we used a system prompt that instructed the model to act as an expert advisor for prospective students and freshmen of a specified undergraduate program. The model was asked to provide concise German-language answers (maximum five sentences) based solely on retrieved contextual information, to cite the corresponding source metadata, to flag outdated or prohibited terms when present, and to indicate when an answer was unknown. Consistent with established practice, all system prompts were written in English (cf. Schulhoff et al., 2025) although the questions were written in German. The relevant chunks were retrieved by a similarity search (cosine similarity). A representative excerpt of the system prompt is shown below:

> You are an expert counseling prospective students and freshmen of <program>. IMPORTANT: Use ONLY the following pieces of retrieved context to answer the question.
> If and only if the question contains a word that is marked 'veraltet' and / or ' verboten' in the context, point this out and provide the corresponding word marked 'erlaubt" and / or 'aktuell'. Use five sentences maximum and keep the answer concise.

Note that we requested the LLM to give a hint if a non-preferred term was used in the question, linking linguistic and knowledge-based applications.

Each LLM generated answers to all 27 questions using all 28 knowledge bases, resulting in 756 answers for an LLM and 3,780 answers in total.

3.5 LLM-as-Judge

3,780 answers cannot be fully evaluated by human experts. Therefore, we applied an LLM-as-Judge approach. For practical reasons, we only chose one model – M-Prometheus 3B (Pombal et al., 2025) – for the first assessment. Prometheus is a multilingual open-weight LLM judge created by finetuning Qwen2.5-Instruct that compares AI-generated answers with a ground truth and generates a rating between 1 and 5.

[4] This low number of questions is due to the fact that a human evaluation of all automatically answered questions is an integral part of our experimental design.

Ground truths were provided by two experts – an academic counselor and a senior student. Hence, each automatically generated answer was assigned two Prometheus scores (one for each ground truth). We then combined these two scores into a single score (*comb_score*) by taking the arithmetic mean. This combined score, therefore, includes not only integer values between 1 and 5, but also half-point increments.

3.6 Human Evaluation

For an in-depth analysis, nine human experts who are well familiar with the academic program evaluated a sample of the answers generated by the LLMs based on the results in Section 4.1. As a dataset, we chose three sets of answers generated using three knowledge bases by the best performing LLM: 1) the best performing knowledge base, 2) a middle-scored knowledge base, and 3) the worst performing knowledge base.

Each expert rated answers to all 27 questions, nine for each knowledge base. To avoid fatigue and individual effects, the evaluation of answers generated using a given knowledge base was evenly distributed among all experts. Therefore, the experts were divided into three groups, with each group assigned its own list of 27 items to rate. Each list included selected items covering all three question types and all three knowledge bases. Hence, each member of a group rated the same set of answers, but the order of the items was randomized.

The experts used the full rating schema by Remy et al. (2023) (three dimensions, 1–5 scale) with wording adapted to the task. The evaluation was performed on a university-specific online platform.

Finally, we aggregated the evaluation results to assess the quality of the answers using six quality levels as in Remy et al. (2023) (*remy_factor*).

4 Preliminary Results

4.1 LLM-as-Judge

Table 2 shows the mean and median of the combined Prometheus scores (*comb_score*) for each model regardless of question type and knowledge base. Overall, the scores remain in a relatively low range, with the highest average reaching only 2.45 / highest median reaching 2.5. Mistral Large and LLaMA 3.3 70B achieve the highest performance, while Gemma stands out as the lowest performer.

	comb_score	
model	**mean**	**median**
mistral-large-instruct	2.45	2.5
llama-33-70b-instruct	2.31	2.0
teuken-7b-instruct-research	2.19	1.5
llama-31-8b-instruct	2.13	2.0
gemma-3-27b-it	1.57	1.0

Table 2: Mean and median of combined Prometheus scores (*comb_score*)

Figure 1 shows the distribution of *comb_score* for each model across individual knowledge bases. It indicates – with few minor exceptions – similar score patterns for knowledge bases across the models. However, we find for certain knowledge bases (e.g., 27, 23, and 12) a higher density of better scores.

To gain a better understanding of how different knowledge bases perform, we examined, for the two models with the highest performance according to Table 2 (Mistral Large and LLaMA 70B), the average *comb_score* across all question types. We found that knowledge base 27 consistently yields the highest performance. This knowledge base includes information on the academic degree and program, but most importantly, it contains both – definition and context sentences. This appears not to be limited to knowledge base 27 – all knowledge bases that rank among the top three performers for either model shown in Figure 2 (20, 23, 27, and 12) contain either definitions and/or context sentences (see Table 3). Conversely, all knowledge bases ranked among the lowest three performers for either model (13, 4, 22, 11, 4, 13) contain neither definitions nor context sentences. Notably, knowledge bases containing more information (cf. Appendix) do not necessarily score better.

4.2 Human Evaluation

Nine human evaluators assessed a subset of LLM-generated answers. For the evaluation dataset, we focused on Mistral Large as the best-performing LLM and selected answers generated using three knowledge bases: (1) the best-performing KB 27 (containing both – definitions and context sentences), (2) a medium-performing KB 30 (containing context sentences, but no definitions), and (3) the lowest-performing KB 11 (containing neither definitions nor context sentences), see Figure 2. As described in Section 3.6, the dataset was divided into three lists to ensure that each answer was rated by three evaluators (= one evaluator group), yielding three assessments per answer. In total, 27 questions × 3 knowledge bases × 3 lists = 243 answers were evaluated. Due to missing ratings in 15 cases, the final analysis included 228 data points.

Globally, we see a correlation between Prometheus scores and quality levels by Remy et al. (2023) (*remy_factor*) based on a Spearman rank correlation test which shows a significant positive association between c*omb_score* and *remy_factor* ($\rho = 0.28$, $p < 0.001$). However, due to the low agreement among the evaluators (see Appendix) the results of the human evaluation are not suitable for validating the Prometheus scores for individual items (answers).

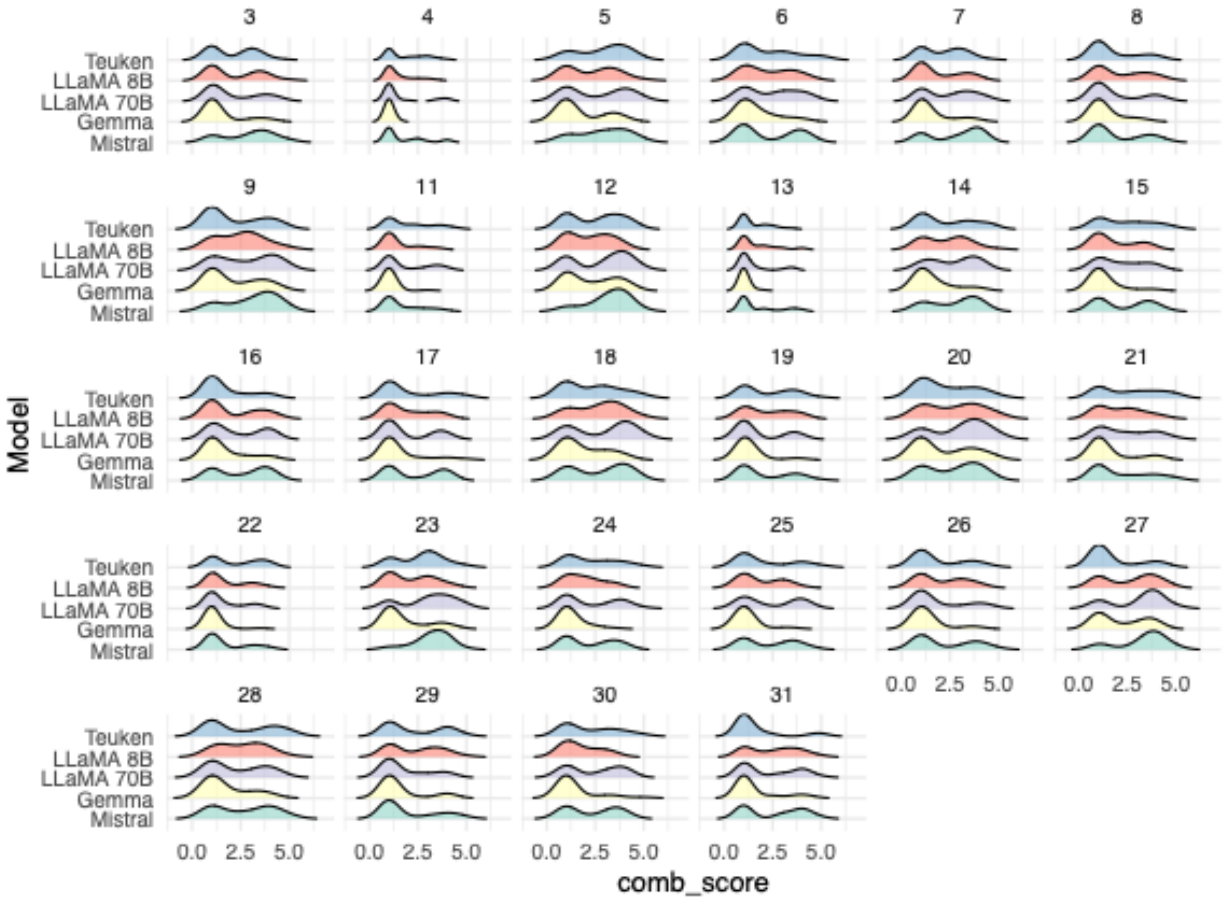
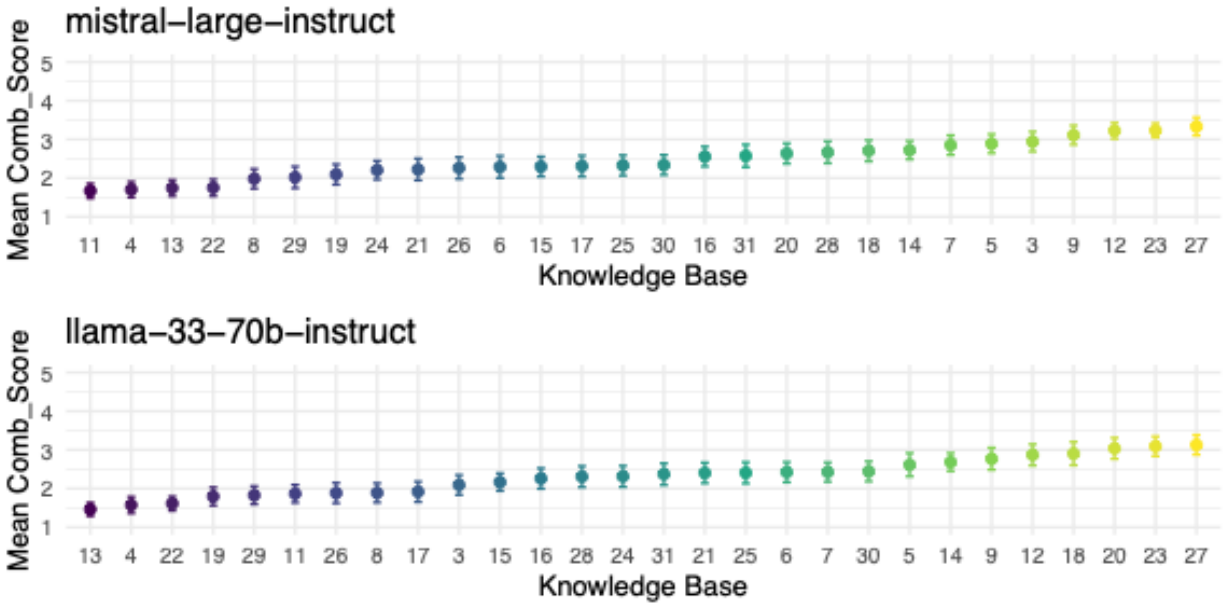


Figure 1: Distribution of combined Prometheus scores (*comb_score*) across different models and knowledge base combinations

Figure 2: Mean combined Prometheus scores by knowledge base for Mistral Large and



LLaMA 70B according to LLM-as-Judge (M-Prometheus 3B)

5 Discussion

We systematically varied five LLMs to gain insight into their performance in combination with a terminological knowledge base. Based on the LLM-as-Judge approach, the choice of an LLM seems to matter as some models were rated higher than others. It is possible that different LLMs might need different type of (terminological) data to achieve a better performance.

In quantitative terms, our findings show that the more is not necessarily the better. On the contrary, including all TDB data, which might be interpreted as our baseline, reduced answer quality in our application. Instead, it was the type of information that mattered most. Here, the LLM-as-Judge approach highlighted context sentences as key knowledge base elements – more than definitions. Context sentences in TDBs are usually well curated and reflect the application-specific language use better than a general source. In other words, context sentences may provide LLMs with richer in-context data for semantic search and probabilistic text generation. However, the better performance of knowledge bases with context sentences may also result from each concept having only one definition, but possibly multiple context sentences. As a result, the difference in performance may, therefore, be driven by quantitative rather than qualitative factors. This should be investigated systematically. Also, the literature identifies different types of context sentences, e.g. defining, associative or linguistic (TermWebPublish, 2025) A systematic study on the performance of these types may also clarify the significance of context sentences.

All in all, a further investigation of the LLM-as-Judge approach, e.g. using different sizes of different LLMs, is needed. Additionally, in light of the low inter-annotator agreement among human evaluators, more in-depth analyses are needed.

6 Conclusion

Our study provides empirical evidence for the relevance of context sentences and, to a lesser extent, definitions, when using prescriptive TDBs as sources of application-specific knowledge in RAG workflows. Hence, it is primarily the type of information – and only secondarily its amount – that matters. Therefore, we advocate for a closer look into so-called encyclopedic information in prescriptive TDBs if they are intended to be used for knowledge management purposes in AI applications.

Acknowledgements

This study was partially funded by the Ministry for Culture and Science of the State of North Rhine-Westphalia (MKW, Germany) as a part of the project KI:edu.nrw.
We would like to thank all colleagues who contributed to the study, either by creating the ground truths or by participating in the human evaluation.

References

A. Rapp, L. Di Caro, F. Meziane, & V. Sugumaran (Eds.), *Natural language processing and information systems* (pp. 161–171). Springer.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., ... Xing, E. P. (2023, March). *Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.* LMSYS.org. https://lmsys.org/blog/2023-03-30-vicuna/

Di Nunzio, G. M. (2025). Terminology-augmented generation (TAG): Foundations, use cases, and evaluation paths. *Journal of Digital Terminology and Lexicography, 1,* 97–104.

Doosthosseini, A., Decker, J., Nolte, H., & Kunkel, J. M. (2024). *Chat AI: A seamless Slurm-native solution for HPC-based services. arXiv.* https://arxiv.org/abs/2407.00110

Drewer, P., Massion, F., & Pulitano, D. (2017). *Was haben Wissensmodellierung, Wissensstrukturierung, künstliche Intelligenz und Terminologie miteinander zu tun?* DITeV. https://downloads.ditev.org/publikationen/DITeV_org_Terminologie_und_KI_2017_03_22_v2.pdf

Drewer, P., & Schmitz, K.-D. (2017). *Terminologiemanagement: Grundlagen, Methoden, Werkzeuge.* Springer.

Fleischmann, K., & Lang, C. (2025). *Terminologie in der KI: Wie mit Terminologie der Output von LLMs und GenAI optimiert werden kann, Terminologie in der KI – KI in der Terminologie: Akten des Symposions, Worms, 27.–29. März 2025* (pp. 83–95). DTT e.V.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Dai, Y. B. Y., ... Wang, H. (2024). *Retrieval-augmented generation for large language models: A survey. arXiv.* https://arxiv.org/abs/2312.10997

Hamm, J. (2025). *Terminologische Konsistenz und generative KI – ein perfect match? Produktiver Einsatz von Sprachmodellen im Terminologiemanagment und beim Post-Editing. Terminologie in der KI – KI in der Terminologie: Akten des Symposions, Worms, 27.–29. März 2025* (pp. 151–163). DTT e.V.

ISO 26162-1. (2019). *Management of terminology resources – Terminology databases – Part 1: Design.* ISO.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., ... Seo, M. (2024). *Prometheus: Inducing fine-grained evaluation capability in language models. arXiv.* https://arxiv.org/abs/2310.08491

Krishna, K., Roy, A., & Iyyer, M. (2021). *Hurdles to progress in long-form question answering. arXiv.* https://arxiv.org/abs/2103.06332

Lackner, A., Vega-Wilson, A., & Lang, C. (2025). *Terminology augmented generation: A systematic review of terminology formats for in-context learning in LLMs. Proceedings of the 4th International Conference on Multilingual Digital Terminology Today (MDTT 2025)* (s.p.). Thessaloniki, Greece: CEUR Workshop Proceedings. Retrieved 2025-10-15, from https://ceur-ws.org/Vol-3990/short10.pdf

Lang, C., Schneider, R., & Tu, N. D. T. (2024). *Automatic question answering for the linguistic*

domain – An evaluation of LLM knowledge base extension with RAG.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv.* Retrieved 2023-10-09, from https://doi.org/10.48550/arXiv.2005.11401 arXiv Version Number: 4

L'Homme, M.-C. (2021). *Lexical semantics for terminology: An introduction.* John Benjamins.

Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics. https://aclanthology.org/W04-1013/

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: A method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Pombal, J., Yoon, D., Fernandes, P., Wu, I., Kim, S., Rei, R., ... Martins, A. F. T. (2025). *M-Prometheus: A Suite of Open Multilingual LLM Judges. arXiv.* https://arxiv.org/abs/2504.04953

Remy, F., Demuynck, K., & Demeester, T. (2023, July). Automatic glossary of clinical terminology: A large-scale dictionary of biomedical definitions generated from ontological knowledge. In D. Demnerfushman, S. Ananiadou, & K. Cohen (Eds.), *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* (pp. 265–272). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bionlp-1.23

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... Resnik, P. (2025). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv.* https://arxiv.org/abs/2406.06608

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval augmentation reduces hallucination in conversation. Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved 2023-10-10 from https://doi.org/10.18653/v1/2021.findings-emnlp.320

TermWebPublish. (2025). *Datcatinfo* (Tech. Rep. Version v1.9.2 build 85). Interverbum Technology AB. version v1.9.2 (build 85). Retrieved from https://datcatinfo.termweb.net/en/dict/202/496450/1951663?lang=eng&target=zul&section=0&domain=0&term=context&config=0

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual e5 text embeddings: A technical report. arXiv.* https://arxiv.org/abs/2402.05672

Appendix

Subsets of TDB

3	{def}	4	{terms}
5	{context}	6	{def, degree program}
7	{def, area}	8	{def, terms}
9	{def, context}	11	{degree program, terms}
12	{degree program, context}	13	{area, terms}
14	{area, context}	15	{terms, context}
16	{def, degree program, area}	17	{def, degree program, terms}
18	{def, degree program, context}	19	{def, area, terms}
20	{def, area, context}	21	{def, terms, context}
22	{degree program, area, terms}	23	{degree program, area, context}
24	{area, terms context}	25	{degree program, terms, context}
26	{def, degree program, area, terms}	27	{def, degree program, area, context}
28	{def, degree program, terms, context}	29	{def, area, terms, context}
30	{degree program, area, terms, context}	31	{def, degree program, area, terms, context}

Prompt for Prometheus-Scores

###Score Rubrics:
[Factuality, Insight, Fluency]
Score 1: There are severe mistakes in the information provided in the answer and it fails to give a clear overall impression of the issue at hand, regardless of its fluency.
Score 2: At most, there are one or two minor mistakes in the information provided in the answer. Still, the answer provides at least some key elements that help to understand the issue at hand. While the answer may appear to follow a template, it remains clear and is grammatically well-structured.
Score 3: All the information in the answer is correct, and it provides at least some key elements that help to understand the issue at hand. While the answer may appear to follow a template, it remains clear and is grammatically well-structured.
Score 4: All the information in the answer is correct and sufficient to understand the issue at hand. The answer is well written and may even be found as-is in the official FAQs of the university.
Score 5: All the information in the answer is correct, and it includes all the significant details needed to understand the issue at hand. The answer is well written and may even be found as-is in the official FAQs of the university.

Inter-Annotator Agreement of Human Evaluators

	kappa			
eval. group	**qual.**	**corr.**	**ins.**	**fl.**
1	0.067	0.096	0.164	-0.069
2	0.054	0.115	-0.047	-0.019
3	0.177	0.153	0.189	-0.036

Inter-annotator agreement (κ) for each evaluation dimension (qual. = overall quality level; corr. = correctness; ins. = insight; fl. = fluency) and evaluator group.

Outils d'annotation linguistique de l'incertitude : du discours scientifique à sa vulgarisation

Ianis Pontier[a] and Iana Atanassova[a,b]

[a]Université Marie et Louis Pasteur, CRIT (UR 3224), F-25000 Besançon, France
[b]Institut Universitaire de France (IUF), France
ianis.pontier@univ-fcomte.fr, iana.atanassova@univ-fcomte.fr

Extended Abstract, Toth 2026

Keywords: Uncertainty, Science dissemination, Popular science, Scientific research, Scientific articles, Corpus, Epistemology

Topics: Medical Translation and Interpreting, Artificial Intelligence, Discourse Analysis, Telemedicine and communication,

Abstract

Cet article explore la relation entre publications scientifiques et articles de vulgarisation, en interrogeant plus spécifiquement la manière dont l'incertitude est exprimée et mobilisée au sein de ces deux genres. Dans cette perspective, nous proposons une évaluation de deux méthodes d'appariement de phrases appliquées à partir d'un sous-corpus de 20 articles provenant du corpus *SciNews*. L'enjeu de ces méthodes est d'automatiser l'identification des correspondances entre les phrases extraites des articles de recherche et leurs éventuelles paraphrases dans les textes de vulgarisation associés. Une telle approche nous permet finalement de proposer une ontologie des manières dont les incertitudes formulées dans les sources primaires sont retransmises et transformées dans le discours second. À terme, la production d'outils automatiques fiables pour cette tâche permettra de contrôler la qualité des textes de vulgarisation scientifique.

1 Introduction

Les avancées récentes en Traitement Automatique des Langues (TAL) et l'essor de la Science Ouverte permettent d'exploiter de grands corpus d'articles scientifiques pour analyser les propriétés intrinsèques de la science. Cette approche offre de nouvelles perspectives pour aborder des problématiques en épistémologie grâce aux outils du TAL. Dans ce contexte, nous nous intéressons au concept d'incertitude scientifique, qui comprend, au-delà des simples probabilités, de nombreux composants du processus de recherche, tels que la formulation d'hypothèses de travail, les marges d'erreur des instruments de mesure, les raisonnements abductifs et inductifs ou l'utilisation de modèles incomplets. L'
incertitude

s'exprime alors dans les publications scientifiques, vecteur principal de communication des résultats de la recherche, mais également, et de manière distincte, dans des textes de vulgarisation destinés au grand public (JENSEN 2008).

Par ailleurs, l'incertitude scientifique fait partie intégrante du processus de recherche, aussi bien en sciences humaines qu'en sciences expérimentales, comme démontré dans les travaux récents (REY 2022; NINGRUM et al. 2025). NINGRUM and ATANASSOVA (2024) ont proposé un cadre d'annotation de l'incertitude scientifique dans les articles de recherche comprenant cinq dimensions : le temps (incertitude au passé/présent/futur), la source (l'incertitude provient de l'auteur ou d'anciens travaux ou les deux), la nature (incertitude épistémique ou aléatoire ou les deux), le contexte (la position de l'incertitude par rapport à la structure de l'article) et l'expression (quantifiée ou non quantifiée). De plus, NINGRUM and ATANASSOVA (2024) ont établi des corpus de référence et des outils pour l'identification et l'annotation automatique des incertitudes dans les publications scientifiques.

2 Jeux de données : corpus SciNews

Nous utilisons un sous-corpus de SciNews (LIU et al. 2024) qui est un corpus interdisciplinaire mettant en parallèle 41 872 publications scientifiques associées à leurs résumés journalistiques. Ces derniers, issus de la plateforme *Science X*, proviennent de sources académiques reconnues et sont rédigés par des chercheurs ou par des journalistes professionnels. L'ensemble du corpus couvre la période de 2004 à 2024, et comprend les domaines suivants : *Médecine, Biologie, Espace, Planète Terre, Chimie, Physique, Ordinateur, Nano* et *Autres* (cinéma, société, etc.). Ce corpus est disponible au format JSON à travers la plateforme *Huggingface*.

3 Méthodes

Ce travail utilise comme point de départ l'outil *UnScientify* (NINGRUM 2025), qui permet d'annoter les phrases exprimant de l'incertitude dans les articles scientifiques de différents domaines. Cet outil repose sur des règles linguistiques explicites, adaptées aux textes scientifiques. Les évaluations sur une tâche de classement binaire montrent que *UnScientify* surpasse les modèles de langage utilisant les réseaux de neurones, y compris les techniques d'adaptation de modèles pré-entraînés (*fine-tuning*) et les IA génératives avec plusieurs techniques de *prompting* (GUTEHRLÉ, NINGRUM, and ATANASSOVA 2026).

À travers notre recherche, nous cherchons à voir si et comment l'expression de l'incertitude en sciences est altérée lors du passage à des niveaux de moindre complexité langagière, c'est-à-dire quand les résultats de la recherche sont présentés à des publics plus larges. C'est par exemple le cas dans la vulgarisation scientifique. Nous nous intéressons alors à la reprise de l'incertitude exprimée par la recherche dans les textes de vulgarisation. À cette fin, nous avons essayé d'automatiser l'appariement d'une phrase exprimant de l'incertitude dans un article scientifique avec la phrase correspondante dans une vulgarisation de cette recherche, autrement dit avec une phrase présentant les mêmes informations.

Dans un premier temps, nous proposons et évaluons deux méthodes d'association de phrases issues du corpus *SciNews*. Plus précisément, ces méthodes visent à automatiser l'appariement de phrases extraites des articles de recherche avec leurs éventuelles paraphrases dans les textes de vulgarisation associés.

Dans un second temps, nous considérons des appariements validés de phrases recueillies dans un sous-corpus de 20 paires de textes. À partir de ces données, nous faisons émerger une ontologie des manières de reprendre l'incertitude dans la vulgarisation : reprise *verbatim*, transformation et disparition.

3.1 Approche explicite

Premièrement, nous avons développé un algorithme d'association semi-automatique basé sur l'utilisation des embeddings *fastText* à 300 dimensions disponibles sur le web[1]. Cet algorithme comprend plusieurs étapes détaillées ci-dessous :

- Sélection d'un ensemble de paires de phrases candidates, composées d'une phrase d'un article de recherche et d'une phrase d'un article de vulgarisation qui contiennent au moins un Groupe Nominal (GN) identique. Les GNs sont identifiés à l'aide de la librairie *Spacy* et au modèle *en_core_web_lg*. Deux GNs sont considérés comme identiques s'ils ont la même forme ou s'ils ont des embeddings *fastText* avec une similarité cosinus supérieure à 0,9.
- Elimination des paires pour lesquelles la phrase de recherche n'exprime pas d'incertitude. L'identification de l'incertitude est faite par la librairie *UnScientify* (NINGRUM 2025).
- Identification des paraphrases parmi les paires de phrases candidates : sélection des paires de phrases ayant une similarité cosinus supérieure à $0,925$. La similarité est calculée à partir des embeddings *fastText*, pondérées par TF-IDF. Cette méthode de calcul permet d'augmenter le poids des mots ayant un apport sémantique significatif, et au contraire, de réduire ceux des mots dont l'apport est moindre (GAUTAM 2013). La similarité cosinus a été choisie parmi un ensemble de similarités qui ont été évaluées, dont Jaccard, Dice et Bray-Curtis (BUSCALDI et al. 2020).
- Evaluation manuelle des paraphrases identifiées. Toutes les paires ont été annotées selon les catégories : *Pertinent, Non-pertinent* ou *Non déterminé*.

3.2 Comparaison avec un grand modèle de langage (LLM)

Pour le corpus de 20 paires d'articles, nous évaluons les performances d'un LLM (Gemini 3 Flash) pour la même tâche, à savoir l'identification des paires de phrases entre articles de recherche et textes de vulgarisation. Chaque paire d'articles a été soumise au LLM à trois reprises, en utilisant le même prompt. Les résultats du LLM sont comparés, après post-traitement, aux paires de phrases identifiées par notre algorithme.

[1] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.en.300.bin.gz

4 Résultats

4.1 Evaluation de UnScientify pour le traitement de textes de vulgarisation

UnScientify étant un outil développé principalement pour le traitement d'articles scientifiques, nous proposons une évaluation de cet outil pour l'annotation des textes de vulgarisation. Nous avons annoté manuellement un échantillon de 400 phrases, dont 100 identifiées comme porteuses d'incertitude et 300 phrases qui n'expriment pas d'incertitude. Cette évaluation montre 16 faux positifs et 33 faux négatifs, soit une précision ($P = 0{,}84$) et un rappel ($R = 0{,}72$) qui donnent un F_1-score de $0{,}77$.

Ces résultats peuvent être améliorés par l'adaptation des ressources linguistiques de *UnScientify* pour prendre en compte la spécificité des textes de vulgarisation. Cependant, dans la présente étude, nous utilisons la version originale de *UnScientify* qui n'est pas optimisée pour les textes de vulgarisation. L'évaluation, avec un score de F1 de $0{,}77$, montre que cette approche est pertinente.

4.2 Identification de paires de phrases

Le corpus d'évaluation comprend 20 paires d'articles de recherche et de vulgarisation. Le tableau 1 montre les résultats produits par les deux approches décrites ci-dessus. Nous voyons que le LLM Gemini 3 Flash surpasse nos deux approches tant quantitativement : plus de paires détectées, que qualitativement : une pertinence des paires détectées largement supérieure. Cela signifie que le rappel et la précision pour le LLM sont supérieurs aux approches explicites.

Table 1: Evaluation des méthodes d'appariement des phrases. (N = nombre de paires de phrases identifiées.)

Méthode	**N**	**Pertinent**	**Non-pertinent**	**Non déterminé**
Approche explicite, sans TF-IDF	87	37,9 %	62,1 %	0,0 %
Approche explicite, avec TF-IDF	34	47,1 %	52,9 %	0,0 %
Gemini (Passe 1)	129	57,4 %	39,5 %	3,1 %

4.3 Ontologie

L'analyse manuelle du corpus de paires de phrases nous permet de proposer une ontologie de la reprise de l'incertitude dans la vulgarisation. Dans ce qui suit, nous listons une série d'exemples canoniques pour illustrer chacune des catégories de notre ontologie.

Les paraphrases où la reformulation est transparente sont les reprises *verbatim* des termes du texte original :

(1) a. The association between body mass index (BMI) in late-life and dementia risk *remains unclear*.(Scientific Article 4)

b. However, the association between body mass index (BMI) in late-life and dementia risk *remains unclear*. (News Article 4)

Dans d'autres cas, comme (2), (3) ou (4) l'incertitude est reformulée :

(2) a. *We hypothesized that* antibiotic treatment during the first days of life *may exert* a long-lasting effect on childhood growth by disrupting the natural gut microbial colonization process. (Scientific Article 2)
b. These findings *suggest a potential link* between neonatal antibiotic exposure and impaired childhood growth, which *may be* a result of alterations caused by antibiotics in the composition of the gut microbiome. (News Article 2)

Cette reformulation relève parfois de contraintes propres à la vulgarisation, ici le passage à l'hyperonyme, à savoir d'un terme générique (*factors*) pour éviter les termes techniques (*the high heterogeneity*, *potential recall*, *misclassification biases*) :

(3) a. However, the high heterogeneity, and *potential recall* and *misclassification biases* on fried-food consumption from the original studies *should be considered* when interpreting the findings of this meta-analysis. (Scientific Article 12)
b. The design of the included studies varied considerably, added to which, they all relied on memory—factors *that should be taken into consideration* when interpreting the results, caution the researchers. (News Article 12)

La reformulation accompagne parfois des transformations morphologiques, comme la nominalisation du verbe dans la vulgarisation :

(4) a. For diagnostic ability testing purposes, we also examined as predictors phenotypic measures as apps (for suicide risk (CFI-S, Convergent Functional Information for Suicidality) and for anxiety and mood (SASS, Simplified Affective State Scale)) by themselves, as well as in combination with the top biomarkers (the combination being our a priori primary endpoint), to provide context and *enhance precision of predictions*.(Scientific Article 13)
b. Researchers have developed *a more precise way* of diagnosing suicide risk, by developing blood tests that work in everybody, as well as more personalized blood tests for different subtypes of suicidality that they have newly identified, and for different psychiatric high-risk groups.(News Article 13)

Enfin, dans d'autres cas les informations relatées sont davantage problématiques, comme en (5) où l'incertitude semble absente de la vulgarisation. En (6), la mise en garde sur les conséquences à long terme est absente de la vulgarisation :

(5) a. Although the ability to deliver epinephrine for anaphylaxis using the L-SOMA *might be hindered by swallowing difficulties* in serious cases, the capsule could still be used to deliver the drug for slower-onset allergic reaction. (Scientific Article 8)
b. Traverso and his colleagues have been working on many strategies to deliver such drugs orally, and in 2019, they developed a capsule that could be used to inject up to 300 micrograms of insulin. (News Article 8)

(6) a. Recently, antibiotic therapy in the first week of life was reported to be associated with decreased growth during the first year of life 13 , but longer-term outcomes *remain unknown*. (Scientific Article 2)

b. Exposure to antibiotics in the first days of life *is thought* to affect physiological aspects of neonatal development. (News Article 2)

Parfois l'élision de l'incertitude est à l'initiative du chercheur. En effet, la vulgarisation mobilise souvent l'autorité du chercheur, à travers la citation d'un passage de l'article ou de propos recueillis lors d'entretiens directs comme en (7).

(7) a. Our data *indicate that* maternal diet influences the infant gut microbiome and that these effects differ by delivery mode. (Scientific Article 35)
b. Sara Lundgren, lead author of the study said : " Our study demonstrates an association of a readily modifiable factor, maternal diet, with the infant gut microbiome. (News 35)

Dans l'exemple (7), l'usage d'un *hedge* dans l'article source, tel que *indicate that*, relève d'une stratégie de *face-saving* visant à anticiper une éventuelle contradiction par les pairs (HYLAND 1996). En exprimant ses affirmations avec prudence, la chercheuse protège sa crédibilité dans le milieu académique. Par nature, cette forme d'incertitude est indissociable de la posture rhétorique du chercheur face à ses collègues ; elle tend donc logiquement à disparaître dans la vulgarisation, car les acteurs et leurs enjeux relationnels ne sont plus les mêmes.

5 Conclusion

Notre travail présente deux contributions majeures. Dans un premier temps, nous proposons un outil pour apparier des phrases provenant des publications scientifiques qui expriment de l'incertitude à leurs reformulations dans les textes de vulgarisation. Nous évaluons ces performances sur le corpus SciNews relativement à un LLM.

Dans un second temps, l'analyse des tendances de reformulation observées dans les textes de vulgarisation est menée à partir des données recueillies. Cette étude nous permet de modéliser le traitement des informations initialement présentées comme incertaines dans la vulgarisation. Alors, nous établissons une ontologie des différentes tendances de reformulation de l'incertitude lors de sa reprise dans la vulgarisation.

L'approche que nous développons ici ouvre des perspectives pour le "fact-checking" appliqué aux sciences. En permettant la comparaison systématique de l'expression de l'incertitude entre deux documents traitant du même objet, notre outil pourrait aider à déceler d'éventuelles distorsions d'information ou des simplifications excessives lors du transfert de connaissances en sciences.

Parmi les limites de l'approche proposée : nous ne prenons pas en compte les variations idiosyncrasiques des auteurs et nous considérons les vulgarisateurs comme un ensemble homogène. Également, nous ne discriminons pas nos résultats selon les disciplines, alors que NINGRUM (2025) a montré qu'elles ont chacune leur propre rapport à l'incertitude.

Bibliography

BUSCALDI, Davide, Ghazi FELHI, Dhaou GHOUL, Joseph LE ROUX, Gaël LEJEUNE, and Xudong ZHANG (June 2020). "Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? (Sentence Similarity : a study on similarity metrics with words and character strings)". In: *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*. JEP/TALN/RECITAL 2020. Ed. by Rémi CARDON, Natalia GRABAR, Cyril GROUIN, and Thierry HAMON. Nancy, France: ATALA et AFCP, pp. 14–25.

GAUTAM, Jyoti (2013). "An Integrated and Improved Approach to Terms Weighting in Text Classification". In: *International journal of Computer science issues* 10, pp. 310–314.

GUTEHRLÉ, Nicolas, Panggih Kusuma NINGRUM, and Iana ATANASSOVA (2026). "A large-scale multi-disciplinary analysis of uncertainty in research articles". In: *Data & Knowledge Engineering* 163, p. 102561. ISSN: 0169-023X. DOI: 10.1016/j.datak.2026.102561.

HYLAND, Ken (1996). "Writing Without Conviction? Hedging in Science Research Articles". In: *Applied Linguistics* 17.4, pp. 433–454. ISSN: 0142-6001. DOI: 10.1093/applin/17.4.433.

JENSEN, Jakob D. (2008). "Scientific Uncertainty in News Coverage of Cancer Research: Effects of Hedging on Scientists and Journalists Credibility". In: *Human Communication Research* 34.3, pp. 347–369. ISSN: 1468-2958. DOI: 10.1111/j.1468-2958.2008.00324.x.

LIU, Dongqi, Yifan WANG, Jia LOY, and Vera DEMBERG (2024). *SciNews: From Scholarly Complexities to Public Narratives – A Dataset for Scientific News Report Generation*. DOI: 10.48550/arXiv.2403.17768. arXiv: 2403.17768[cs].

NINGRUM, Panggih Kusuma (2025). "Identifying and annotating scientific uncertainty in scholarly texts: methods, frameworks, and applications". PhD thesis. Besançon, Université Marie et Louis Pasteur.

NINGRUM, Panggih Kusuma and Iana ATANASSOVA (2024). "Annotation of scientific uncertainty using linguistic patterns". In: *Scientometrics* 129.10, pp. 6261–6285. ISSN: 0138-9130, 1588-2861. DOI: 10.1007/s11192-024-05009-z.

NINGRUM, Panggih Kusuma, Philipp MAYR, Nina SMIRNOVA, and Iana ATANASSOVA (2025). "Annotating scientific uncertainty: A comprehensive model using linguistic patterns and comparison with existing approaches". In: *Journal of Informetrics* 19.2, p. 101661. ISSN: 1751-1577. DOI: 10.1016/j.joi.2025.101661.

REY, François-C. (2022). "Ontologie de l'incertitude et annotation sémantique d'un corpus autour du changement climatique". Issue: 2022UBFCC007. Theses. Université Bourgogne Franche-Comté.

Exploration de requêtes syntaxiques pour l'extraction de contextes riches en connaissances : le cas des contextes définitoires

Rim Abouwarda, Cécile Frérot, Olivier Kraif

nom.prenom@univ-grenoble-alpes.fr

Univ. Grenoble Alpes, LIDILEM, F-38000 Grenoble, France

1. Introduction

L'analyse des contextes constitue la « pierre de touche » du travail terminologique (Dubuc, 1985, p. 62) et dans une optique terminographique, les contextes riches en connaissances (CRC) permettent de repérer des relations conceptuelles et contribuent à guider le lecteur dans sa compréhension du concept (Meyer, 2001). C'est dans cette perspective terminographique que nous nous situons. Notre étude porte sur les contextes définitoires que Meyer (*ibid.*) dote d'un intérêt particulier pour cette application : *(i)* ils constituent une amorce de définition et *(ii)* les terminographes doivent s'en emparer pour rédiger des définitions.

Dans une approche à partir de corpus, la confrontation aux données met en évidence que les contextes tendent à « dévier » de la formule aristotélicienne (Meyer, *ibid.*), correspondant à la définition introduite par le marqueur *is_a* / *est_un* où X = Y + caractéristiques spécifiques (X correspondant au terme à définir et Y à l'hypéronyme le plus proche). Cette « déviance » a des conséquences pour l'extraction outillée. Elle implique notamment la définition de marqueurs pour la relation d'hypéronymie (Lefeuvre, 2017) ou de marqueurs introduisant des énoncés définitoires (Auger, 1997 ; Rebeyrolle & Tanguy, 2000 ; Fener & Dahdouh, 2021) utilisés dans la construction de requêtes. Cette construction repose sur des patrons lexico-syntaxiques qui permettent d'extraire à partir de corpus des fragments textuels dont on fait l'hypothèse qu'ils contiennent des CRC. Actuellement, ces patrons sont mis en œuvre au sein d'outils d'interrogation exploitant des corpus annotés au niveau grammatical et morphosyntaxique qui s'appuient sur le langage d'expression de requêtes CQL (Sketch Engine (León-Araúz & Martín, 2018), TXM (Condamines et al., 2022)). Les requêtes représentent des séquences lexico-grammaticales et morphosyntaxiques définies sur le plan syntagmatique. Leur projection en corpus vise à « capter » les fragments textuels correspondants. Il s'agit de requêtes de surface qui n'exploitent aucune relation syntaxique entre les éléments de la requête (sujet-verbe, verbe-objet). D'autre part, la formulation de ces requêtes dépend de l'outil utilisé, une limite majeure étant que les requêtes doivent être adaptées si les outils d'analyse varient (Condamines et al. 2022).

2. Question de recherche

Dans ce contexte, nous souhaitons explorer l'apport d'une « véritable analyse syntaxique » (Rebeyrolle & Tanguy, 2000) dans l'extraction des CRC et la mise en œuvre de requêtes qui s'appuient sur les relations syntaxiques (notamment sujet et objet, avec une vision élargie des objets (arguments et circonstants)). Cette mise en œuvre passe par la construction de requêtes exploitant des dépendances syntaxiques au sein d'un patron *X(sujet)-marqueur verbal-Y(objet)*. Elle interroge dans quelle mesure des requêtes basées sur des dépendances peuvent contribuer à l'extraction de contextes définitoires en « captant » les éléments du patron (par exemple un verbe et son complément d'objet) apparaissant dans des séquences non contigües (présence d'adverbe par exemple).

Par ailleurs, dans un domaine où il n'existe pas à l'heure actuelle d'outil d'extraction automatique de CRC « prêt à l'emploi » (outils *off-the-shelf*, Marshman, 2022) et librement accessibles, nous cherchons à proposer des requêtes qui, en répondant au critère de l'utilisabilité (ou facilité d'usage (Barcenilla & Brangier, 2024)) permettent aux terminologues d'extraire des connaissances spécialisées à partir de leur propre corpus. C'est dans le cadre de la science ouverte avec l'outil Lexicoscope (Kraif, 2019) que nous menons cette première étude.

3. Démarche de construction des requêtes syntaxiques pour l'extraction de contextes définitoires

Le Lexicoscope[1] est un outil d'analyse de corpus annotés[2] qui exploite les dépendances syntaxiques (par exemple, les relations sujet-verbe, ou verbe-objet) pour explorer les profils combinatoires de mots ou d'expressions dans de vastes corpus textuels. Dernièrement, il a fait l'objet d'adaptations à la terminologie en proposant une fonctionnalité d'extraction automatique de terminologie, qui s'intègre dans le cadre plus vaste de la linguistique de corpus outillée.

Une spécificité du Lexicoscope est de mettre en jeu un système de requêtes basées sur l'exemple (Augustinus et al., 2016). A partir d'exemples (concordances) choisis par l'utilisateur, l'outil génère automatiquement des requêtes avancées[3] intégrant des dépendances syntaxiques. Une série de requêtes syntaxiques a été testée pour extraire des CRC à partir d'un corpus en français dans le domaine de la psychiatrie (Abouwarda & Kraif, 2024). L'expérience a porté sur le terme « dépression », l'approche étant terminologique avec comme point de départ le terme.

Dans cette étude, nous utilisons ce même corpus composé de textes spécialisés extraits d'ouvrages et de revues dans le domaine de la psychiatrie, branche de la médecine qui étudie les processus biologiques associés aux processus mentaux. Dans le Lexicoscope, le corpus est divisé en deux sous-corpus :

Sous corpus	Nb. de documents	Nb. de phrases	Nb. de tokens	Nb. de formes
Ouvrages	8	107 373	2 069 886	1 708 067
Revues	3 736	691 097	17 000 000	14 493 055

Tableau 1: Données statistiques sur le corpus

Dans notre démarche de construction de requêtes, la sélection d'exemples repose, en amont, sur une première requête en langage naturel qui sert d'amorce. Cette requête comprend un verbe (ou marqueur verbal[4]) associé à un ou plusieurs items lexicaux (ex. *on entend ; est défini comme*). Les requêtes d'amorce sont guidées par un ensemble de travaux sur les marqueurs de relations conceptuelles (Lefeuvre, 2017 ; León-Araúz & Martín, 2018 ; Fener & Dahdouh, 2021) et les verbes introducteurs de contextes définitoires (Frérot & Valentini, 2020). Nous déclinons certains marqueurs en différentes requêtes d'amorce. Par exemple, pour le marqueur *définir*, nous définissons trois requêtes *on définit, est défini comme* et *se définit comme*[5]. A partir

[1] http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/

[2] L'analyseur Stanza est utilisé pour l'étiquetage, la lemmatisation et l'analyse en dépendances (modèle GSD).

[3] Le langage TQL est utilisé (Tree Query Language).

[4] Nous avons notamment étudié *désigner, définir, entendre, appeler* et *nommer*.

[5] Ces requêtes correspondent respectivement aux constructions agentive, passive et pronominale passive (Rebeyrolle & Tanguy, 2000).

des concordances produites par une requête, nous sélectionnons une partie de concordance qui correspond au patron ciblé. C'est le cas de *par sélection, on entend des mécanismes*[6] qui correspond au patron que nous cherchons à « capter » en corpus (*par X, on entend Y*) ou bien encore le cas de *L'éthologie se définit comme la science* qui correspond au patron ciblé (*X se définit comme Y*). Ces exemples montrent que nous cherchons à extraire des contextes dans lesquels X et Y sont présents, sachant que nous ne ciblons pas l'extraction des caractéristiques spécifiques. Le Lexicoscope génère ensuite automatiquement des requêtes avancées associées à des arbres mettant en évidence les relations syntaxiques. Avant de lancer une requête, une étape de « simplification » est nécessaire : elle consiste à supprimer les lemmes associés aux étiquettes grammaticales (correspondant aux X et Y[7]). Nous conservons uniquement les lemmes qui entrent dans l'élaboration du patron syntaxique (prépositions notamment) et les marqueurs verbaux. C'est à ce stade que nous adaptons les requêtes sur le plan syntaxique. Par exemple, pour la requête d'amorce *on définit* et l'exemple sélectionné *Elles définissent la toxicomanie comme,* nous ajoutons l'étiquette NOUN comme sujet dans la requête[8].

4. Evaluation et analyse des contextes extraits

Dans cette étude exploratoire, nous avons évalué les contextes produits à partir de requêtes qui contraignent sur le plan syntaxique la présence d'un sujet et d'un objet. C'est le cas de la requête suivante qui correspond au patron *X(sujet) désigner Y(objet)* :
<c=DET,#1>&&<c=NOUN,#2>&&<l=désigner,c=VERB,#3>&&<c=DET,#4>&&<c=NOUN,#5>::(det,2,1) (det,5,4) (nsubj,3,2) (obj,3,5)

Le principal critère qui a guidé l'étape de validation humaine est lié à l'application terminographique : le contexte constitue-t-il une « amorce de définition » (Meyer, 2001) et permet-il de « dégager une image précise de la notion » (Dubuc, 2002, p. 61) pour l'utilisateur d'une base de données terminologiques qui consulte le champ Contexte d'une fiche ? Nous avons utilisé une mesure de précision[9], indicateur privilégié dans les évaluations (Marshman, 2022). L'extraction du contexte élargi aux phrases voisines[10] nous a permis de valider des contextes dans lesquels la dimension interphrastique est essentielle au terminologue pour interpréter le contexte et « reconstruire » les connaissances ; c'est le cas des ellipses (1) et des reprises (notamment avec *terme,* qui renvoie à *psychothérapie* (2)).

(1) *(…) survient un temps organisateur pendant lequel l'enfant va avoir à s'approprier les propres éléments de sa psyché. Ce temps se définit comme le moment où l'évolution du groupe familial*

(2) *La psychanalyse est-elle une psychothérapie ? Si par ce terme, on désigne tout ce qui vise à améliorer le fonctionnement psychique, il est évident que oui*

Les contextes que nous avons analysés montrent tout d'abord que grâce à une analyse en dépendances, les requêtes syntaxiques captent *i)* les séquences non contiguës dans lesquelles le

---

[6] Concordance : *Par sélection, on entend des mécanismes qui permettent à une ou plusieurs représentations de « dominer » le traitement.*

[7] Exemples :
<~~l=par,~~c=ADP,#1>&&<~~l=sélection,~~c=NOUN,#2>&&<~~l=on,~~c=PRON,#3>&&<l=entendre,c=VERB,#4>&&<~~l=un,~~c=DET,#5>&&<~~l=mécanisme,~~c=NOUN,#6>::(case,2,1) (det,6,5) (nsubj,4,3) (obj,4,6) (obl:mod,4,2)

[8] <c=PRON|NOUN,#1>&&<l=définir,c=VERB,#2>&&<c=DET,#3>&&<c=NOUN,#4>&&<l=comme,c=ADP,#5>&&<c=DET,#6>&&<c=NOUN,#7> ::(case,7,5) (det,4,3) (det,7,6) (nsubj,2,1) (obj,2,4) (obl :mod,2,7)

[9] La précision correspond au nombre de concordances dans lesquelles un contexte définitoire est avéré divisé par le nombre total des concordances extraites (et multiplié par 100 pour obtenir un pourcentage).

[10] Paramétrable dans Lexicoscope avec l'affichage KWIC en sélectionnant *Intégrer les phrases voisines dans l'empan de concordance.*

marqueur verbal et X/Y sont à distance ; *ii)* les contextes dans lesquels X est un terme complexe (Nom Adjectif, Nom de Nom). Les exemples suivants illustrent des contextes avec insertion d'adverbe (3), apposition (4), groupe prépositionnel (5) et termes complexes (6) :

(3) *Par « nécessités », j'entends non seulement les commodités qui (…)*
(4) *l'empathie, dans sa définition stricto sensu, se définit comme*
(5) *la psychophysique se définit de façon générale comme l'étude de la perception en lien avec les paramètres physiques des stimuli ; le symptôme désigne en médecine un phénomène observable lié à un état*
(6) *On entend essentiellement par psychologique clinique une étude approfondie basée sur l'examen des cas individuels / Les rythmies du sommeil désignent des mouvements rythmés qui surviennent pendant le sommeil*

D'autre part, une seule requête dans le Lexicoscope permet d'accéder aux différentes variantes syntaxiques d'un patron par le biais des « réalisations[11] ». Ainsi, pour le marqueur *entendre (par)*, les contextes extraits à partir d'une même requête comportent le patron *par X, entendre Y* (7) et la variante *entendre X par Y* (8) :

(7) *Par hypothétisation, nous entendons la formulation par le thérapeute d'une hypothèse basée sur des informations qu'il possède*
(8) *On entend actuellement par cures de sommeil, les méthodes qui permettent d'obtenir un sommeil discontinu*

La série de requêtes générées automatiquement par la sélection d'exemples, puis éventuellement adaptées comme nous l'avons expliqué, a produit des contextes définitoires fiables. Par exemple, pour les requêtes associées aux marqueurs *entendre, définir* et *désigner,* la précision avoisine 95%. Les contextes non validés proviennent d'analyses syntaxiques erronées[12] ou de paramètres discursifs (dans ces cas, la définition du terme n'est pas amorcée[13]).

Toutefois, le nombre de contextes extraits[14] nous amène à nous interroger sur les possibilités d'augmenter la couverture des requêtes. Il s'agit alors de « relâcher » une contrainte syntaxique ou d'en placer une autre. Les premiers tests réalisés montrent des résultats variables selon les marqueurs. Par exemple, la suppression de la relation sujet dans la requête d'amorce *se définit comme* dégrade les résultats : la requête produit 125 contextes (contre 45) mais la précision chute à 66,4% (de nombreux contextes impliquent un nom propre ou un pronom personnel dans lesquels *se définit comme* est synonyme de « se qualifie de », « se considère comme »).

5. Conclusions et perspectives

Les requêtes contraintes sur le plan syntaxique (sujet et objet du verbe) produisent des résultats encourageants. Elles ne sont toutefois qu'une première étape dans l'exploration des contextes extraits car la couverture reste limitée. Cette étape doit s'accompagner d'un travail d'élaboration de requêtes plus complexes qui implique une connaissance affinée des relations

[11] Une des fonctionnalités de Lexicoscope est de proposer les différentes « réalisations » lexicales et syntaxiques d'un patron syntaxique (lorsque certains éléments de la requête sont variables).
[12] C'est le cas de l'exemple suivant où *entendre* n'a pas de valeur définitoire (*entendre* est un verbe polysémique et c'est ici le sens de *percevoir par l'ouïe* qui est activé) : *Le Child Study Center a formé des centaines de personnes et l'on entend en permanence, de par le monde, les étudiants (…) parler de ce qu'ils ont pu ici recevoir (…).*
[13] Exemple : *nous seulement il nous faudrait préciser ce que les uns et les autres entendent par analyse du transfert mais encore on risquerait de réduire (…)*
[14] A titre d'exemple, 77 contextes pour la requête d'amorce *on entend,* 45 pour la requête *se définit comme* ou 128 contextes pour la requête *est défini comme*.

syntaxiques prises en charge par l'analyseur. Nous envisageons d'autre part d'étendre cette étude aux énoncés définitoires indirects impliquant d'autres marqueurs.

Par ailleurs, nous allons porter nos réflexions sur la sélection des requêtes à diffuser (faudra-t-il privilégier les requêtes correspondant aux patrons les plus productifs et fiables dans notre corpus d'étude ?) et sur leurs modalités de diffusion. La diffusion de requêtes « prêtes à l'emploi » vise à créer des conditions favorables aux études en terminologie sur la « portabilité » (Marshman et al., 2008) des CRC en facilitant leur extraction sur un ensemble de corpus. Cette « portabilité » constitue un enjeu dans l'extraction de connaissances spécialisées, questionnant la généricité des patrons pour chaque nouveau projet terminologique portant sur des corpus de genres textuels et de domaines variés.

Bibliographie

Abouwarda, R., & Kraif, O. (2024). Utilisation de requêtes syntaxiques pour la terminologie : Une étude de cas dans les domaines de la psychologie et de la psychiatrie. *SHS Web of Conferences*, *191*, 11005. https://doi.org/10.1051/shsconf/202419111005

Auger, A. (1997). *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles* [Université de Neuchâtel]. https://doi.org/10.35662/unine-thesis-1615

Augustinus, L., Vandeghinste, V., & Vanallemeersch, T. (2016). Poly-GrETEL: Cross-Lingual Example-based Querying of Syntactic Constructions. *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*.

Barcenilla, J., & Brangier, É. (2024). Psychologie du Travail et des Organisations. In *Psychologie du Travail et des Organisations* (p. 650-655). Dunod. https://doi.org/10.3917/dunod.valle.2024.01.0650

Condamines, A., Escoubas Benveniste, M.-P., & Federzoni, S. (2022). Apport d'un corpus de presse spécialisé parallèle français/italien à l'analyse des marqueurs et de la relation de méronymie: *Éla. Études de linguistique appliquée*, *N° 208*(4), 429-446. https://doi.org/10.3917/ela.208.0049

Dubuc, R. (1985). *Manuel pratique de terminologie*. Linguatech.

Fener, P., & Dahdouh., C. (2021). *Repérage automatique d'énoncés définitoires avec Unitex pour l'aide à l'enrichissement de ressources terminologiques : Retour d'expérience. 2*.

Frérot, C., & Valentini, C. (2020). Constitution d'un corpus de contextes définitoires dans le domaine de la propriété intellectuelle : Vers la définition de structures linguistiques dans les brevets. *Terminologica, C. Roche (Ed.)*, 283-306.

Kraif, O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope. In Max Silberstein (dir.), *Langue française*, N° 203, Armand Colin, p. 67-82.

Lefeuvre, L. (2017). *Analyse des marqueurs de relations conceptuelles en corpus spécialisé : Recensement, évaluation et caractérisation en fonction du domaine et du genre textuel*. Toulouse 2.

León-Araúz, P., & Martín, A. S. (2018). *The EcoLexicon Semantic Sketch Grammar : From Knowledge Patterns to Word Sketches* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1804.05294

Marshman, E. (2022). Knowledge patterns in corpora. In P. Faber & M.-C. L'Homme (Éds.), *Terminology and Lexicography Research and Practice* (Vol. 23, p. 291-310). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.13mar

Marshman, E., L'Homme, M.-C., & Surtees, V. (2008). Portability of cause–effect relation markers across specialised domains and text genres : A comparative evaluation. *Corpora*, *3*(2), 141-172. https://doi.org/10.3366/E1749503208000130

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography : A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Éds.), *Natural Language Processing* (Vol. 2, p. 279-302). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.2.15mey

Rebeyrolle, J., & Tanguy, L. (2000). *Repérage automatique de structures linguistiques en corpus : Le cas des énoncés définitoires. Cahiers de Grammaire*, *25*, 153-174.

De la *fiche termontologique* au réseau de connaissances : vers un modèle de construction d'une structure virtuelle de documentation de l'administration publique camerounaise

Samuel Benga*, Jérémie Fifen-Moluh**, Kelly Hapi-Nzepang***, Kidio Rolland Samni****

*LingaTech Consulting, Cameroun samuel.benga@gmail.com
**MINESEC, Cameroun fifenmoluh@gmail.com
***MINEPDED, Cameroun hapikelly@gmail.com
****ECONTRANS LANGUAGE SERVICES, Cameroun kidiorolland@gmail.com

Résumé. À l'heure de la mutation numérique des États, les administrations subsahariennes font face à une densification documentaire dont la complexité défie désormais les outils de gestion conventionnels (Mban, 2007). Si le travail terminologique demeure le pivot de cette transition, la traditionnelle « fiche terminologique » (Pavel & Nolet), par sa nature statique et cloisonnée, s'avère insuffisante pour orchestrer l'agilité des flux d'informations contemporains. Cette contribution entend démontrer que l'activité terminologique, bien au-delà d'une simple entreprise nomenclaturale, s'érige en un pilier méthodologique capable de fonder une vaste structure virtuelle de documentation. En s'inscrivant dans le sillage des travaux de Depecker (2002) sur la mesure de la pensée, cette étude déploie un modèle où l'analyse sémantique permet de cartographier les interrelations conceptuelles, métamorphosant des données éparses en un réseau de connaissances dynamique. Les résultats mettent en lumière une structuration transversale des actes administratifs, faisant de la terminologie pratique le moteur de l'interopérabilité des services. *In fine*, ce changement de paradigme garantit une circulation fluide des savoirs et assure une modernisation pérenne de la mémoire institutionnelle au sein de l'architecture virtuelle de l'État.

Mots-clés : *terminologie textuelle, réseau de connaissances, administration publique, structure virtuelle de documentation, interopérabilité sémantique, ingénierie des connaissances, mémoire institutionnelle.*

Session 3

Cultural Bias in Ontologies: Formation and Containment

Antonia Lourentzaki
tonialouren@gmail.com
TALOS Project in Artificial Intelligence for the Humanities and Social Sciences
University of Crete

Abstract

This proposal critically examines ontological approaches to modelling and managing cultural heritage data, with particular attention to the formation, reproduction, and mitigation of cultural bias. Drawing on research in Semantic Web technologies, formal ontologies, AI-driven knowledge graphs, and metadata standards, it investigates how bias becomes embedded in digital heritage systems through modelling choices, training data, and computational practices. By synthesizing technical, methodological, and archaeological perspectives, and engaging with documented case studies and computational examples from contemporary heritage datasets, the study aims to identify gaps in current practices and contribute to more critically informed and inclusive heritage modelling.

From an archaeological perspective, ontologies function as epistemic structures that shape classification, interpretation, and authority. Archaeological engagement with typology and interpretive practice underscores the risks of embedding cultural and ideological assumptions into formal knowledge systems, emphasizing the need for reflexivity and interpretive plurality in digital representations of the past.

The research reviews past and current ontological frameworks, examines tools and methods through a critical lens, and evaluates approaches to bias detection and mitigation. It is structured in six parts: (1) Introduction – ontologies in cultural heritage and archaeological epistemology; (2) Tools and methods – Semantic Web, knowledge graphs, digital twins, visualizations and ontological modelling; (3) AI in archaeology – spatial and remote analyses and digital reconstruction; (4) Challenges and bias – formation, impact, and mitigation; (5) Case studies – application to selected datasets and assessment of bias containment; (6) Conclusions – recommendations for more inclusive and reflexive heritage modelling.

Keywords: ontology, terminology, ontoterminology, epistemology, standards, cultural bias, cultural heritage, linked open data, semantic web

1. Introduction

The study of cultural heritage in the digital age presents unprecedented opportunities alongside significant epistemological challenges. While ontologies, Semantic Web technologies, and AI systems provide powerful frameworks to structure, analyze, and interpret complex heritage data, they are not neutral instruments. Bias is systematically embedded during data collection,

categorization, modelling, and computational interpretation, actively shaping scholarly and public understanding of the past.

Cultural heritage encompasses not only physical artefacts but also events, practices, social relations, and the interpretive processes through which the past acquires meaning. Following UNESCO's cultural heritage classification[1], Doerr (2009) distinguishes heritage into tangible, both movable (paintings, sculptures, furniture, wall paintings, documents) and immovable (historical buildings, monuments, archaeological sites), intangible heritage (oral traditions, knowledge systems, skills, and practices related to nature and society), and digital heritage, comprising records, representations, and metadata. Critically, as Doerr (2009) notes, such classifications are epistemic acts; they structure knowledge and foreground certain perspectives, making ontological design a primary site for mediating—or marginalizing—different viewpoints.

This research conducts a critical review of existing ontological frameworks and AI applications in heritage through a lens of bias-aware computational practice. It investigates how bias arises, is measured, and can be mitigated, bridging theoretical critique with technical assessment. Applying a multi-layered methodology grounded in archaeological theory, the study combines computational evaluation, bias metrics, and case studies to advance more transparent, inclusive, and reflective practices for digital heritage.

2. Ontologies, AI, and the Technical Mediation of Bias

The convergence of ontologies and Artificial Intelligence (AI) in cultural heritage creates powerful socio-technical systems where bias is systematically embedded or mitigated through formal engineering choices.

2.1 Ontological Engineering and Epistemic Constraints

Ontologies like the CIDOC Conceptual Reference Model (CRM) provide an event-centric framework (*E5 Event*, *E39 Actor*) for semantic interoperability (Doerr, 2009). However, their high-level abstractions can epistemically constrain modelling if applied a priori. Bias arises from decisions on *class granularity*, *hierarchy depth*, and *property design* (Nafis et al., 2019). Without specialized extensions, models may lack constructs for documenting provenance and plural interpretations. Extensions like **CRMtex** or **HiCO** (*hico:InterpretationAct*) enable the modelling of interpretive acts and assertions, supporting plural representations of historical knowledge (Bikakis et al., 2021). Ontology selection and design are thus primary bias intervention points, requiring evaluation against domain coverage and the need to separate evidence from assertion.

2.2. AI Pipelines: Amplification Loops and Semantic Countermeasures

AI and machine learning (ML) models ingest data already structured by ontologies, creating a technical pipeline where bias is amplified. Training on heritage datasets inherits and reproduces

[1] Preparing World Heritage Nominations (Second edition, 2011). Published in November 2011 by the United Nations Educational, Scientific and Cultural Organization.

existing biases in cataloguing and representation (Dallas, 2016). Mitigation requires targeted technical countermeasures. At the data and model level, this includes employing fairness metrics—such as CLIP embedding analysis—and techniques like adversarial learning to minimize reliance on spurious, biased features (Radford et al., 2021; Zhou et al., 2024). At the semantic integration level, ontologies serve as validation frameworks. Knowledge graphs populated via standards like CIDOC-CRM allow AI outputs (e.g., an automated classification) to be cross-checked against curated semantic relationships, flagging statistical outliers for expert review (Bobasheva et al., 2022). A critical ontological distinction is modelling the AI itself not as an intentional CIDOC:E39_Actor but as a prov:Activity or software process, thereby correctly attributing agency and responsibility for interpretive decisions (Felicetti et al., 2025).

2.3. AI and Bias-Aware Practice in Digital Archaeology

In digital archaeology, AI supports pattern recognition and data analysis at scale, from classifying artifacts in imagery to identifying settlement patterns in LiDAR data (Papadopoulos, 2023). These computational practices extend traditional methods but introduce specific bias risks. Convolutional Neural Networks (CNNs) trained on imbalanced datasets privilege well-documented artifact types or regions, while AI-assisted digital reconstructions (e.g., of the Villa of the Papyri or the Temple of Zeus at Olympia) risk presenting a single hypothesis as an objective truth (Zarmakoupi, 2010; Patay-Horváth, 2012). Mitigation here is inherently technical: frameworks like Heritage Digital Twins integrate multimodal data into dynamic, queryable models where assumptions remain visible (Felicetti et al., 2025). Furthermore, ontological integration is crucial; detected entities must be embedded in knowledge graphs (e.g., via CIDOC-CRM) for semantic querying, but the AI tools themselves must be correctly modeled as prov:Activity rather than intentional E39_Actor to avoid misattributing agency and obscuring responsibility for interpretive bias.

2.4. Technical Exemplars: Ontologies for Bias Documentation

Specialized ontologies provide technical blueprints for formally documenting bias and context. For instance, the IAS (Image Annotation Situations) Ontology applies the DnS (Descriptions and Situations) pattern (Gangemi & Mika, 2003) to model ImageAnnotationSituation as a context-bound event. It captures annotator, time, and purpose, enabling SPARQL queries to compare divergent labels, thus formalizing cultural subjectivity in tagging (Pandiani & Presutti, 2022). Similarly, the Doc-BiasO Ontology documents bias in ML pipelines by linking bias:RepresentationBias to specific mls:MLTask and dcat:Dataset instances via properties like bias:quantifiedBy, creating an audit trail integrated with PROV-O for lineage (Russo & Vidal, 2024). The SEBI (Scholarly Evidence Based Interpretation) Framework uses RDF to model authenticity claims (sebi:Authentic, sebi:Forgery). It links claims via hico:InterpretationAct to prov:Agent and evidence scored via forgont:hasEvaluationScore, making evaluative bias traceable (Pasqual, 2025). Complementing these, techniques for Embedding-Knowledge Graph Mapping project vectors from models like Word2Vec onto semantic networks (WordNet, Framester). Analyzing cosine similarities within the graph structure exposes stereotypical associations for systematic debiasing (Marinucci et al., 2023).

In summary, bias is a structural feature engineered into heritage informatics and digital archaeology systems. Effective mitigation is a technical imperative, achieved through modular ontology design, semantic validation of AI outputs, and the implementation of formal ontologies that document interpretive provenance and bias.

3. Case Studies: Bias Mitigation in Historical and Archaeological Ontologies

This chapter presents two ontology-based case studies drawn from the author's own research in historical and archaeological knowledge representation. Rather than treating bias solely as a post hoc analytical problem, these case studies integrate bias awareness directly into the design of classes, properties, and modelling decisions. Particular attention is given to inclusivity, the formal separation between empirical evidence and interpretation, and the avoidance of implicit assumptions related to gender, social status, ethnicity, and cultural representation. They operationalize core design principles: separating entities from roles, distinguishing empirical evidence from interpretation, and maintaining minimal commitment to avoid overdetermination. The modelling prioritizes conceptual clarity and inclusivity before alignment to standards like CIDOC CRM.

3.1.Case Study I: Bias Mitigation in the Hellenistic Events Ontoterminology

The Hellenistic Events dataset (Lourentzaki, Papadopoulou, Roche, 2024; Lourentzaki, Papadopoulou, 2025), models military, political and social events of the Hellenistic period through an event-centric ontology. Bias mitigation is addressed through explicit representation of agency, role diversity, and historical plurality.

All historical agents are instantiated under a neutral <Human-Agent> concept, decoupled from gender, status, or ethnicity. Social and political functions—such as <Monarch>, <Princess>, <Governor>, or <General>—are modelled as roles assigned within specific events. Women identified in historical sources as influencing or shaping military and political events are modelled as instances of the concept <Human-agent> and assigned explicit roles as monarchs, princesses, or governors rather than being treated explicitly as peripheral figures or implicit relations to male actors. This modelling choice counters the traditional marginalization of women in event-based historical narratives and ensures that agency is not inferred solely from gendered assumptions embedded in historical sources.

Furthermore, agents involved in events, as well as historians documenting them, are modelled with explicit references to ancestry or cultural affiliation—Greek, Roman, Jewish, Illyrian, Persian, and so on — reflecting the plural cultural landscape of the Hellenistic world. In this way, the dataset distinguishes between events and their documentation.

Crucially, events are not treated as self-evident historical facts but are explicitly linked to their documentary attestations. Each event is connected to primary textual sources structured at the level of individual chapters and passages, incorporating the original text, modern translations, and bibliographic references drawn from the Perseus Digital Library. This design choice distinguishes between historical occurrence and historiographical narration, foregrounding the

fragmentary, selective, and ideologically conditioned nature of ancient sources. By modelling sources as first-class entities, the ontology exposes how bias enters historical representation through transmission, authorship, and survival of evidence.

To address the systematic underrepresentation of non-elite actors, efforts have been made to include marginalised groups. Enslaved individuals are represented through a dedicated sub-concept <Slave> under a broader <Role> hierarchy, rather than being subsumed under generic or undefined social categories. This modelling strategy makes visible social groups that are often structurally invisible in both contemporary historical narratives and digital heritage datasets. These modelling strategies function as a corrective to elite- and male-centered historiography, demonstrating how ontological design can actively challenge inherited narrative biases.

3.2. Case Study II: Interpretive Bias and Gender Representation at Göbekli Tepe

A comparable, though contextually distinct, strategy is applied in the Göbekli Tepe ontotermonology (Lourentzaki, Papadopoulou, 2025; Milio, Giannadakis, Lourentzaki, 2025). Bias arises primarily from interpretive uncertainty and ongoing scholarly debate rather than textual omission. The archaeological record at Göbekli Tepe has generated divergent interpretations concerning anthropomorphic representation and gender attribution, particularly due to the predominance of male anthropomorphic figures and the near absence of clearly identifiable female representations[2] [3] [4].

To avoid reinforcing speculative or culturally loaded assumptions, the ontology distinguishes strictly between observable features and inferred meanings. Anthropomorphic reliefs are modelled using specific artefact-type concepts, such as <Type of Artefact Relief of Anthropomorphic Being Male> and <Type of Artefact Relief of Anthropomorphic Being Female>, with gender attribution applied only when iconographic evidence is unambiguous. Female representations, rare in the material record, are represented explicitly when attested, while male representations are not treated as default or normative. Importantly, animal representations are not assigned gender, even where later symbolic interpretations might suggest it, in order to avoid projecting human categories onto symbolic motifs or modern specific gender frameworks onto prehistoric material culture.

Symbolic interpretations —such as ritual functions or cosmological meanings— associated with architectural features, reliefs, or iconography are modelled as possible interpretations

[2] Human figures at Göbekli Tepe are rare, and the symbolism of anthropomorphic pillars and animals reflects broader social and ideological meanings rather than straightforward gender representation. Interpretations are therefore highly contingent on theoretical framing. *Referenced in:* Mithen, S., Richardson, A., & Finlayson, B. (2023). The flow of ideas: shared symbolism during the Neolithic emergence in Southwest Asia: WF16 and Göbekli Tepe. *Antiquity, 97(394),* 829–849.

[3] Analyses of the iconographic and sculptural material at Göbekli Tepe indicate that male-associated imagery — including depictions of animals with phallic symbolism — is far more frequent than clear female representation. Distinctly feminine motifs are rare, suggesting a gender imbalance in the site's visual repertoire. This observation is based on archaeological interpretations rather than definitive biological sex assignments. Referenced in: Schmidt, K. (2006). Göbekli Tepe – The Stone Age Sanctuaries: New Results of Ongoing Excavations with a Special Focus on Sculptures and High Reliefs. Neo-Lithics, 1/06, 3–13.

[4] The scarcity of female-associated depictions highlights how the material record itself can reflect social, symbolic, or cultural priorities, and illustrates the importance of careful, inclusive ontological modelling when representing gender and social roles in heritage datasets.
Reference: Dietrich, O., Notroff, J., & Schmidt, K. (2012). Feasting, Monumentality, and the Emergence of Social Complexity at Göbekli Tepe. In Current Anthropology, 53(6), 1–23.

rather than definitive assertions. Rather than assigning meanings directly to artefacts, the ontology introduces a dedicated <Symbolic Meaning> concept, with sub-concepts including <Animal Symbolic Interpretation>, <Religious Interpretation>, <Ritual Interpretation>, and <Interpretive Status>. Symbolic meanings are modelled as interpretive constructs that are linked to specific artefacts or reliefs and systematically connected to supporting bibliographic sources, allowing multiple, even conflicting, scholarly positions to coexist within the same knowledge structure. This modelling strategy prevents the ontology from endorsing a single explanatory framework and instead foregrounds the interpretive character of archaeological knowledge.

These case studies demonstrate how bias mitigation can be embedded at the level of ontological design. Through deliberate conceptual structuring, precise definition of classes, and a clear separation between material evidence and scholarly interpretation, the datasets demonstrate how reflexivity, provenance awareness, and inclusivity can be operationalized within digital heritage systems. Rather than merely reflecting inherited narratives, such modelling practices enable more accountable, pluralistic, and critically engaged representations of the past.

4. Conclusions

The findings of this research underscore that bias in cultural heritage ontologies is not merely a technical issue but also shaped by human assumptions and the nature of the data itself. Curators, archaeologists, and scholars inevitably introduce disciplinary perspectives, cultural assumptions, and professional biases into the design of ontologies and knowledge graphs, while historical and archival sources themselves often reflect selective documentation and dominant narratives. When such materials are used to train AI models or populate ontological frameworks, these embedded biases can be amplified, influencing classification, retrieval, and analysis in ways that reinforce stereotypes or narrow historical interpretations.

This research adopts an ontological stance in which interpretation, rather than definition, is treated as the primary locus of meaning in cultural heritage modelling. Ontologies are understood not as neutral mirrors of reality but as epistemic instruments that structure knowledge through selective abstraction.

The case studies illustrate practical strategies for addressing these challenges, including transparent recording of interpretive choices, explicit differentiation between evidence and conjecture, and careful representation of underrepresented actors and social roles, with attention to race, gender, and geographical imbalances. Incorporation of these practices support counteraction of inherited biases and production of systems that are both methodologically comprehensive and ethically responsible. Ultimately, this research highlights the importance of reflexive design, critical evaluation of training data, and the active mitigation of both human- and data-driven biases to foster more balanced, inclusive, and reliable digital heritage infrastructures.

References

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M., (2022). The Values Encoded in Machine Learning Research. In Proceedings of the 2022 ACM Conference on Fairness, Accountability,

and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 173–184. Doi: https://doi.org/10.1145/3531146.3533083

Bikakis A., Hyvönen E., Jean S., Markhoff B., Mosca A., (2021). Editorial: Special issue on Semantic Web for Cultural Heritage. Semantic Web 12. IOS Press: 163–167.

Bobasheva, A., Gandon, F., and Precioso, F. (2022). Learning and Reasoning for Cultural Metadata Quality: Coupling Symbolic AI and Machine Learning over a Semantic Web Knowledge Graph to Support Museum Curators in Improving the Quality of Cultural Metadata and Information Retrieval. J. Comput. Cult. Herit. 15, 3, Article 40 (September 2022). doi: https://doi.org/10.1145/3485844

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research 81:77-91.

Caliskan, A. (2023, August). Artificial Intelligence, Bias, and Ethics. In IJCAI (pp. 7007-7013).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Carboni, N., & Luca, L.D. (2019). An Ontological Approach to the Description of Visual and Iconographical Representations. Heritage.

Dallas, C., (2016). Jean-Claude Gardin on Archaeological Data, Representation and Knowledge: Implications for Digital Archaeology. J Archaeol Method Theory 23, 305–330. https://doi.org/10.1007/s10816-015-9241-3

Doerr, M. (2009). Ontologies for Cultural Heritage. Handbook on Ontologies.

Estermann, B. (2014). Diffusion of Open Data and Crowdsourcing among Heritage Institutions: Results of a Pilot Survey in Switzerland. *Journal of Theoretical and Applied Electronic Commerce Research, 9(3),* 15–31.

Evans, Th. L., Daly, P., (Eds.), Digital archaeology: bridging method and theory. London; UK, New York; Routledge. 2004.

Felicetti, A., Himmiche, A., & Somenzi, M. (2025). Knowledge Graphs and Artificial Intelligence for the Implementation of Cognitive Heritage Digital Twins. Applied Sciences, 15(18), 10061. https://doi.org/10.3390/app151810061

Felicetti, A., & Niccolucci, F. (2025). Artificial Intelligence and Ontologies for the Management of Heritage Digital Twins Data. Data, 10(1), 1. doi: https://doi.org/10.3390/data10010001

Foka, A., & Griffin, G. (2024). AI, Cultural Heritage, and Bias: Some Key Queries That Arise from the Use of GenAI. Heritage, 7(11), 6125-6136. https://doi.org/10.3390/heritage7110287

Foka, A., Griffin, G., Ortiz Pablo, D. et al. (2025). Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice. AI & Soc 40, 5835–5847 https://doi.org/10.1007/s00146-025-02349-z

Gattiglia, G. (2025). Managing Artificial Intelligence in Archeology. An overview, Journal of Cultural Heritage, Volume 71: 225-233, ISSN 1296-2074, https://doi.org/10.1016/j.culher.2024.11.020.

Hitzler P, Janowicz K, Bikakis A, et al. (2021). Editorial: Special issue on Semantic Web for Cultural Heritage. Semantic Web: – Interoperability, Usability, Applicability.;12(2):163-167. doi:10.3233/SW-210425

Ignatowicz, J., Kutt, K., & Nalepa, G. J. (2025). Position Paper: Metadata Enrichment Model: Integrating Neural Networks and Semantic Knowledge Graphs for Cultural Heritage Applications. arXiv preprint arXiv:2505.23543.

Marinucci, L., Mazzuca, C. & Gangemi, A. (2023). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. AI & Soc 38, 747–761. doi: https://doi.org/10.1007/s00146-022-01474-3

Nafis, F., Yahyaouy, A., & Aghoutane, B. (2019). Ontologies for the classification of cultural heritage data. 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 1-7. https://doi.org/10.1109/WITS.2019.8723850

Pansoni, S., Tiribelli, S., Paolanti, M., Frontoni, E. and Giovanola, B., (2023)."Design of an Ethical Framework for Artificial Intelligence in Cultural Heritage," 2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS), West Lafayette, IN, USA, pp. 1-5, doi: 10.1109/ETHICS57328.2023.10155020.

Papadopoulos, N. (2023). ClepsYdra: translating submerged archaeological remains from shallow water to digital environment with geoinformatics. In Advances in On-and Offshore Archaeological Prospection: Proceedings of the 15th International Conference on Archaeological Prospection (pp. 61-64).

Sol Martinez Pandiani, D., Presutti, V. (2022). Coded Visions: Addressing Cultural Bias in Image Annotation Systems with the Descriptions and Situations Ontology Design Pattern. In Proceedings of the 6th International Conference of Graphs and Networks in the Humanities 2022: Technologies, Models, Analyses, and Visualizations.

Pavlidis, G. (2025). Agentic AI for Cultural Heritage: Embedding Risk Memory in Semantic Digital Twins. Computers, 14(7), 266. https://doi.org/10.3390/computers14070266

Schwartz, R., Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a standard for identifying and managing bias in artificial intelligence (Vol. 3). Gaithersburg, MD: US Department of Commerce, National Institute of Standards and Technology.

Schimmenti, A., Pasqual, V., Vitali, F., & van Erp, M. (2025). Knowledge Graphs Generation from Cultural Heritage Texts: Combining LLMs and Ontological Engineering for Scholarly Debates. arXiv preprint arXiv:2511.10354.

Shen, Y., Wang, Z., Sun, Q., Chen, A., Rushmeier, H., (2021). Reconstructing Dura-Europos From Sparse Photo Collection Using Deep Contour Extraction. In: Chalmers A. and Hulusie V., EUROGRAPHICS. Workshop on Graphics and Cultural Heritage.

Spennemann, D. H. R. (2024). Will Artificial Intelligence Affect How Cultural Heritage Will Be Managed in the Future? Responses Generated by Four genAI Models. Heritage, 7(3), 1453-1471. https://doi.org/10.3390/heritage7030070

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*.

Reyero Lobo P, Daga E, Alani H, Fernandez M. (2022). *Semantic Web technologies and bias in artificial intelligence: A systematic literature review.* Semantic Web: – Interoperability, Usability, Applicability.;14(4):745-770. doi:10.3233/SW-223041.

Russo, M., Vidal, M.-E., (2024). Leveraging Ontologies to Document Bias in Data. AEQUITAS 2024: Workshop on Fairness and Bias in AI, Santiago de Compostela, Spain.

Zarmakoupi, M., (2010). The Villa of the Papyri at Herculaneum: archaeology, reception, and digital reconstruction. Berlin; New York: De Gruyter.

Zhitomirsky-Geffet M, Kizhner I, Minster S (2023), "What do they make us see: a comparative study of cultural bias in online databases of two large museums". Journal of Documentation, Vol. 79 No. 2 pp. 320–340, doi: https://doi.org/10.1108/JD-02-2022-0047

Zhou, Z., Xi, Y., Xing, S., & Chen, Y. (2024). Cultural Bias Mitigation in Vision-Language Models for Digital Heritage Documentation: A Comparative Analysis of Debiasing Techniques. Artificial Intelligence and Machine Learning Review, 5(3), 28-40. https://doi.org/10.69987/AIMLR.2024.50303

Concept status as a category in legal terminology work
An analysis of concept management in the *JuriTerm NL-VL* project

Vince Liégeois
Dutch Language Institute
Vince.Liegeois@ivdnt.org

Keywords: comparative terminology – concept status – legal terminology – pluricentric languages – terminology management

1. Context – Legal terminology differs fundamentally from more traditional domains of terminology work, such as mechanics and medicine. It exhibits a high degree of what de Groot (2002, 223–224) terms *Systemgebundenheit*, meaning that legal concepts vary across legal systems: concepts may be fully equivalent, quasi-equivalent, partially equivalent, or exclusive to one legal system (Muhr & Peinhopf, 2025, 9). This characteristic poses a considerable challenge for multisystemic terminology work, both between systems operating in different languages (e.g., Germany and France) and between systems that share a pluricentric language (e.g., the Netherlands and Belgium [Flanders]). Consequently, numerous terminologists have sought to develop frameworks to address this issue (Lenci et al., 2009; Engberg, 2020; Nazarov, 2025), and some have pointed to the need for a dedicated ISO standard for legal terminology (Prieto Ramos, 2014; Chiocchetti et al., 2023). Within the *JuriTerm NL-VL* project, which aims to describe Dutch legal terminology from the Netherlands and Belgium (cf. 2), the data category CONCEPT STATUS is introduced as a means of addressing *Systemgebundenheit* in terminology management (cf. 3).

2. The *JuriTerm NL-VL* project – One of the tasks of the Language Union – the regulatory body for the Dutch language in the Netherlands, Belgium, and Suriname – is to promote greater harmonisation of governmental and legislative terminology between the Netherlands and Belgium. To this end, in 2025 it established the *JuriTerm NL-VL* project in collaboration with the Dutch Language Institute. Within this project, a concept-based terminology database is developed that documents conceptual differences and correspondences between the two legal systems and records and evaluates the terms used to denote the relevant concepts. The data are collected using an empirical micro-comparative law approach (Chiocchetti et al., 2013, 12–13; Husa, 2015) and processed by means of the Dutch Language Institute's NLP tool for legal terminology work, *JuriTermWerk*.

3. Concept status – A recurrent problem encountered in the work carried out within the *JuriTerm NL-VL* project concerns concepts that are exclusive to one legal system, yet display similarities to a concept in the other system (i.e. partial equivalence) or refer to a rule or practice that also exists there. To address these issues, and drawing on previous literature (Willems & Judo, 2023), a new data category was introduced: CONCEPT STATUS. This CONCEPT STATUS, implemented as a deductive drop-down list, records whether a concept is PRESENT or ABSENT

in a legal system, i.e., whether lexical evidence for the concept can be found. Where a concept is PRESENT, a further distinction is made between CODIFIED, NON-CODIFIED and DECODIFIED concepts. The first are mentioned or defined in legislation, the second are confined to legal doctrine or jurisprudence, and the third are concepts from previous legislation (i.e., repealed concepts). For ABSENT concepts, the data category records the extent to which they are applicable in the other system (NON-APPLICABLE, PARTIALLY APPLICABLE, FULLY APPLICABLE, APPLICABILITY UNKNOWN) or whether they are newly proposed concepts (PROPOSED). This notion of applicability is of particular relevance to the Language Union's harmonisation objectives.

PRESENT		CODIFIED
		NON-CODIFIED
		DECODIFIED
ABSENT		NON-APPLICABLE
		PARTIALLY APPLICABLE
		FULLY APPLICABLE
		APPLICABILITY UNKNOWN
		PROPOSED

4. Case studies – In this talk, the data category of the CONCEPT STATUS and its utility will be illustrated through case studies from labour law, social security law, and the law of obligations, which constitute the core domains of the *JuriTerm NL-VL* project. The talk will also address the challenges and limitations associated with this data category and introduce additional concept-level data categories developed within the *JuriTerm NL-VL* project: the COMPARATIVE LAW NOTE and INTERSYSTEMIC RELATION.

5. Bibliography

Chiocchetti, E., Heinisch-Obermoser, B., Löckinger, G., Lušicky, V., Ralli, N., Stanizzi, I. & Wissik, T. 2013. *Guidelines for collaborative legal/administrative terminology work*. EURAC research. https://cordis.europa.eu/docs/projects/cnect/7/270917/080/deliverables/001-D33Guidelinesforcollaborativelegaladministrativeterminologywork.pdf.

Chiocchetti, E., Lušicky, V. & Wissik, T. 2023. Terminology standards and their relevance for legal interpreters and translators. Results of an exploratory study in Austria and Italy. *Digital Translation*, 10(2), 156-179. https://doi.org/10.1075/dt.00006.chi.

de Groot, G. 2002. Rechtsvergleichung als Kerntätigkeit bei der Übersetzung juristischer Terminologie. In U. Haß-Zumkehr (Ed.), *Sprache und Recht*, pp. 222-239. De Gruyter. https://doi.org/10.1515/9783110622836-015.

Engberg, J. 2020. Comparative Law for Legal Translation. Through Multiple Perspectives to Multidimensional Knowledge. *Semiotics of Law*, 33, 263-282. https://doi.org/10.1007/s11196-020-09706-9.

Husa, J. 2015. *A New Introduction to Comparative Law*. Bloomsbury.

Lenci, A., Montemagni, S., Pirrelli, V., Venturi, G. 2009. NLP-based ontology learning from legal texts. A case study. In J. Breuker, P. Casanovas, M.C.A. Klein & E. Francesconi (Eds.),

Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood, pp. 75-94. IOS. https://doi.org/10.3233/978-1-58603-942-4-75.

Muhr, R. & Peinhopf, M. 2015. *Wörterbuch rechtsterminologischer Unterschiede Österreich-Deutschland*. Peter Lang.

Nazarov, W. 2025. *Frame-basierte Rechtsübersetzung. Frame-Semantik als ontologisches und rechtstranslatorisches Analyseinstrument am Beispiel französischer und bundesdeutscher Rechtstermini*. Peter Lang. https://doi.org/10.3726/b22413.

Prieto Ramos, F. 2014. Parameters for Problem-Solving in Legal Translation. Implications for Legal Lexicography and Institutional Terminology Management. In A. Wagner, K.-K. Sin & L. Cheng (Eds.), *The Ashgate Handbook of Legal Translation*, pp. 121-134. Ashgate.

Willems, D. & Judo, F. 2023. Noord-Zuidverschillen in de juridische terminologie. Zin en onzin. *Tijdschrift voor Wetgeving*, 24(4), 230-238.

Terminology Planning and Sustainable Development: The Role of the State Commission for Terminology in Mongolia

Munkhtsetseg Namsrai
Institute of Language and Literature, Mongolian Academy of Sciences
Email: munkhtsetsegn@mas.ac.mn
https://orcid.org/0000-0001-5285-0566

Keywords: terminology planning, State Commission for Terminology, sustainable development, digital transformation, and knowledge modeling.

1. Background and Rationale

Terminology concerns not only linguistic matters but also the advancement of science and education, as it functions as a fundamental gateway to scientific knowledge. A term therefore serves a dual role: it is a unit of language within the vocabulary and, at the same time, the smallest unit of scientific knowledge used to designate a concept. In the context of today's knowledge-based economy and innovation-driven era, systematic terminology work has become an urgent task across all fields in Mongolia, including political, legal, economic, scientific, technological, and educational domains.

Nowadays, sustainable development has become a comprehensive framework guiding national policies in the economic, social, cultural, and environmental spheres. A knowledge-based society increasingly depends on the precise formulation, management, and dissemination of concepts, particularly in the fields of science, technology, governance, and education. In this context, terminology planning is emerging as a strategic instrument for ensuring conceptual clarity, linguistic stability, and the effective transmission of knowledge.

For developing countries such as Mongolia, terminology planning functions not merely as a linguistic activity but as a fundamental element in the development of national knowledge infrastructures. The rapid expansion of new scientific and technological fields driven by digitalization, innovation, and global integration creates a growing need for standardized, transparent, and accessible terminological resources. Without coordinated planning, terminological inconsistency leads to long-term constraints on policy implementation, education, research, and public relations.

Within the framework of the "Sustainable Development Vision of Mongolia-2030" and GOAL 2.4 of the "Vision-2050 Long-Term Development Policy of Mongolia", national policy emphasizes strengthening the linkage between science and industry and fostering a knowledge-based society capable of competing internationally. These policy documents highlight the importance of developing a strong national system of science, technology, and innovation, as well as creating favorable conditions for public–private partnerships (PPP) with a multi-source system of R&D and innovation in order to transform knowledge into economic value. In this context, the development of terminology infrastructure can be considered a prerequisite if not a foundational condition for achieving the objectives outlined in these policy documents. The "Vision-2050 Long-Term Development Policy of Mongolia", for example, includes under GOAL 1.3 the objective of creating a knowledge-rich Mongolian language environment and establishing a lexical fund to preserve the richness of the Mongolian language, ensuring that every citizen masters and uses the Mongolian language as a component of national identity and values. More specifically, clause 1.3.1 states for enriching Mongolian scientific and

technological terminology and developing a knowledge-based Mongolian language. Accordingly, the systematic development, harmonization, and where appropriate standardization of scientific and technical terminology constitutes an essential foundation for the country's future development.

This abstract examines the role of the State Commission for Terminology, a central institutional actor in national terminology planning in Mongolia. It focuses on how terminological governance contributes to sustainable development goals and how digital transformation and knowledge modeling can enhance the effectiveness, accessibility, and interoperability of terminological resources. It addresses the following research question: *How can terminology planning, as implemented through the State Commission for Terminology in Mongolia, function as a component of national knowledge infrastructure supporting sustainable development?* Additionally, the study explores: (1) What are the institutional and methodological limitations of current terminology governance? (2) How can digital transformation and knowledge modeling enhance terminology planning effectiveness?

2. Methodology

This study adopts a qualitative research design combining policy analysis, comparative analysis, and conceptual modeling.

First, policy analysis is conducted based on key Mongolian legal and strategic documents (e.g., Law on the Mongolian Language, Vision-2050), focusing on how terminology is framed as part of knowledge and language policy. Analytical criteria include institutional roles, regulatory mechanisms, and implementation tools.

Second, a comparative framework is applied by examining selected international standards (ISO 704, ISO 15188, ISO 29383) and best practices from established terminology systems (e.g., Canada, Catalonia). The comparison focuses on institutional structure, workflow, and digital infrastructure.

Third, a conceptual workflow model is developed to represent the process of terminology approval, standardization, and dissemination in Mongolia. The model is based on existing institutional practices and is used to identify structural gaps and opportunities for digital transformation. Key arguments of the paper is terminological governance: the need for the State Commission to become not just a "filter", but a knowledge orchestrator.

3. Terminology Planning as a Component of Sustainable Development

Building on this framework, terminology planning contributes to sustainable development by ensuring conceptual clarity and supporting knowledge-based policy implementation.

Terminology planning supports the implementation of the Sustainable Development Goals by localizing new knowledge and technological developments in the national language environment, reducing conceptual gaps, and promoting mutual understanding between sectors. In particular, in cross-cutting issues such as climate change, digital transition, and innovation policy, clarity of terminology is the basis for increasing the effectiveness of policy implementation. Terminology planning is widely understood as the systematic development, implementation, and dissemination of domain-specific terminology to support knowledge communication and innovation (ISO 29383; Wright & Budin, 1997). Terminology planning

helps make knowledge societies (United Nations, 2015) and support the UN Sustainable Development Goals (SDGs). To empower knowledge societies, it is important to use and integrate terminology effectively in education, professional activities, in organizations, and throughout society (Galinski, 2024).

Terminology planning is traditionally understood as a systematic process involving the identification, analysis, standardization, and dissemination of specialized terms within and across domains. In the context of sustainable development, terminology planning plays several critical roles.

First, it supports policy coherence by ensuring that key concepts used in legislation, strategic documents, and development programs are consistently defined and interpreted. Ambiguity or variation in core terms—such as those related to environmental protection, public health, digital governance, or education—can lead to misinterpretation and ineffective implementation.

Second, terminology planning contributes to inclusive and equitable knowledge access. Clear and standardized terminology facilitates learning, professional training, and public understanding, thereby reducing knowledge gaps between experts and non-specialists. This is particularly important in multilingual or culturally diverse societies, where national languages must be continuously developed to express new scientific and technological concepts.

Third, terminology planning supports the sustainability of national languages. By systematically developing domain-specific terminology, states strengthen the capacity of their languages to function in advanced fields of knowledge, reducing reliance on uncontrolled borrowing and preserving linguistic identity in the era of globalization.

In this regard, the Mongolian experience demonstrates that terminology planning is closely aligned with broader sustainable development objectives, including quality education, innovation, institutional effectiveness, and cultural sustainability.

4. The Role of the State Commission for Terminology in Mongolia

Mongolia has a rich tradition of terminology work and modern terminology studies began in 1924 when The State Commission for Terminology (SCT) was established under the Committee of Sciences (presently Institute of Language and Literature, Mongolian Academy of Sciences) after the People's Revolution of Mongolia in 1921. Historically, the Commission has played a crucial role in developing Mongolian terminology in all fields such as law, public administration, science, and technology. SCT had been responsible for organizing and managing the terminology work of the country. As a result of the productive labor of SCT members, dozens of terminology dictionaries were published and used in different domains and enriched Mongolian vocabulary with new terms until 1990. Until this time, there was common regulation that new terms could be used only after SCT's approval. Thus, terminology work was well-organized and systemic. Since 1991, due to social transitions, the activity of the SCT was discontinued, and legal framework for organizing terms with a unified policy, support, and control were weakened, resulting in chaos in the use of terms. As a result of it, terminology work became inconsistent with many domain-specific dictionaries containing unstandardized and inconsistent terms formed by private companies, state organizations and individuals without any supervision. In other words, at that time there was no unified policy

and control of terminology work as well as any cooperation between subject field specialists, terminologists, linguists, translators and standardization organization.

Nowadays, the government of Mongolia has paid special attention on terminology work adopted "Mongolian language law" which regulates the use of the language and terminology. According to the law, the National Council for Language, the State Commission for Terminology and the Institute of Language and Literature, Mongolian Academy of Sciences are the main bodies who are responsible for terminology work. Especially, the State Commission for Terminology should function as the primary coordinating body responsible for the regulation, standardization, and approval of terminology used in official and specialized contexts. Its mandate includes reviewing proposed terms, harmonizing terminological usage across sectors, and providing authoritative recommendations for state institutions, educational organizations, and media.

Institutionally, the Commission operates under the framework of the Law on the Mongolian Language, the Law on Culture, the Law on the Protection of Cultural Heritage, and related language-policy regulations. Its chairperson is the head of the National Council for Language Policy, while the secretary is the head of the Council's service office. The Commission's membership includes representatives of research institutes of the Mongolian Academy of Sciences, ministries, universities, the Institute of Education, and other relevant research organizations, reflecting an inter-institutional structure that connects research, education, and public administration. According to the rule of the State Commission for Terminology, the Commission shall undertake the following duties:

2.1.1. Developing the study of scientific and technological terminology of the modern Mongolian language along with setting, standardizing, disseminating, and enforcing Mongolian terms in a scientifically reasonable manner, taking into account commonly used terms at the international level.

2.1.2. To study how the approved terminology meets the requirements and needs of the social and scientific development of the country and how it enters into life, make necessary changes in time, and promptly inform the public;

2.1.3. Regularly make assessment the work done on terminology from the point of view of theory and practice and develop the theory and methodology of terminology of the Mongolian language;

2.1.4. To widely involve professionals and journalists, translators, writers, and media workers in the work of defining, presenting, standardizing, popularizing, and implementing terminology it is needed to regularly receive their opinions and advice, academic research on the theoretical, practical, and methodological issues of terminology research and organize conferences, workshops, seminars, and round table discussions;

2.1.5. Adhering to a scientific and creative methodology in the terminology work trying not to create two extremes: translating foreign terms with improper or incorrect Mongolian words, or using foreign terms even if there are equivalent in Mongolian;

2.1.6. If it is necessary, to order the terminology database of some fields based on the agreement with the relevant organizations and people (National Council for Language Policy, 2017).

Thus, the State Commission for Terminology represents the institutional mechanism through which Mongolia implements terminology planning as part of national language policy and knowledge-infrastructure development. By coordinating research institutions, government bodies, and professional communities, the Commission contributes to the standardization and dissemination of scientific terminology necessary for education, research, and innovation.

However, today the State Commission for Terminology only performs the task of reviewing and approving domain terminology. Through expert committees and interdisciplinary collaboration, it has contributed to the stabilization of core terminological systems and supported the use of Mongolian as a language of governance and scholarship. For instance, in 2025, the SCT in collaboration with Institute of Language and Literature, Mongolian Academy of Sciences, reviewed and approved terminology of four different domains: forestry, movie, inclusive education, and museum through several discussions. This illustrates the current workflow. The process involves proposal submission, expert review, interdisciplinary discussion, and final approval. However, the absence of a unified digital tracking system limits transparency, version control, and broader dissemination.

The contemporary development landscape presents new challenges. The increasing speed of knowledge production, the interdisciplinary nature of emerging fields, and the expansion of digital communication require more flexible, scalable, and data-driven approaches to terminology planning. In this context, the Commission's role is evolving from a primarily normative authority to a strategic coordinator within a broader terminological ecosystem. This evolution highlights the need to integrate institutional expertise with modern technological tools, enabling the Commission to respond more effectively to the demands of sustainable development.

5. Digital Transformation and Terminological Infrastructure

Digital transformation has fundamentally changed the way terminological data are created, managed, and disseminated. Traditional print-based dictionaries and static terminological lists are increasingly insufficient for meeting the needs of dynamic knowledge environments. Instead, digital terminological platforms allow for continuous updating, version control, and wide accessibility.

For terminology planning in Mongolia, digital transformation offers several key opportunities. First, it enables the creation of centralized, searchable terminological databases that can serve multiple user groups, including policymakers, researchers, educators, translators, and the general public. Second, digital tools facilitate inter-institutional collaboration by allowing experts from different domains to contribute to and review terminological data in a structured environment.

Although the State Commission for Terminology functions as the formal decision-making body, the substantive research and organizational work of terminology development is primarily conducted by the Section of Applied Linguistics at the Institute of Language and Literature of the Mongolian Academy of Sciences. The Institute of Language and Literature, Mongolian Academy of Sciences in collaboration with the National Council for Language Policy launched a centralized virtual terminology platform in May 2023. It aims to ensure the sustainability and accessibility of terminology resources. Its primary focus is to foster long-term accessibility of terminology resources, knowledge sharing and collaboration.

The platform, accessible at www.terminology.mn, currently features:

- Digitized Mongolian-Russian Terminology Dictionary: A three-volume resource encompassing 110,000 terms.
- Standardized Terminology Dictionary: Containing 10,439 terms with definitions.
- Online Scientific Terminology Dictionary: Comprising 10,534 terms with definitions.

The digital platform enhances transparency and traceability in terminological decision-making. By documenting sources, definitions, usage contexts, and conceptual relations, it supports evidence-based terminology planning and increase user trust in authoritative resources.

The integration of digital strategies into the work of the State Commission for Terminology is therefore essential for ensuring the long-term sustainability, relevance, and usability of national terminological resources.

6. Knowledge Modeling and Terminology Planning

Beyond digitization, knowledge modeling represents a more advanced approach to managing terminological information. Knowledge modeling involves the formal representation of concepts, their attributes, and their relationships within a domain, often using structured models such as concept systems, ontologies, or knowledge graphs.

In terminology planning, knowledge modeling enables a shift from term-centered lists to concept-oriented systems, in line with international terminological principles. This approach supports conceptual consistency across domains and facilitates interoperability with other knowledge systems, including educational platforms, digital libraries, and information systems.

The adoption of knowledge modeling approaches by the State Commission for Terminology would strengthen Mongolia's terminological infrastructure and align national practices with global trends in knowledge management.

A possible conceptual model for terminology knowledge representation may include:
(1) Concept node (definition, domain, source),
(2) Term variants (preferred, admitted, deprecated),
(3) Semantic relations (hierarchical, associative),
(4) Metadata (approval status, date, authority).

Such a model can serve as a basis for ontology-based terminology systems and knowledge graphs, enabling interoperability with educational and research platforms.

7. Implications and Conclusion

From the perspective of terminology planning theory, the Commission's work currently emphasizes corpus-level decisions (term approval and standardization), while status planning, implementation mechanisms, and digital terminology infrastructure development remain comparatively limited**.** As a result, the broader policy goals outlined in national development documents such as building a knowledge-based society, strengthening innovation systems, and supporting digital transformation are only partially supported by terminology policy implementation.

Another challenge concerns terminology dissemination and enforcement**.** Although the Commission has the authority to provide methodological recommendations, publish bulletins, and monitor terminology usage, the institutional mechanisms required to ensure consistent nationwide implementation across education, research, administration, and media remain underdeveloped. This situation reflects a common issue in terminology policy in smaller

language communities, where terminology approval does not automatically lead to terminology adoption**.**

Furthermore, the rapid expansion of scientific knowledge, digital technologies, and multilingual communication requires terminology work to move beyond traditional approval procedures toward continuous terminology management, database development, and integration with knowledge systems. Without such infrastructure, terminology planning risks remaining disconnected from innovation policy and digital knowledge production.

In this context, the State Commission for Terminology can be seen as an institution with significant normative authority but limited operational capacity in terminology infrastructure development**.** Strengthening its role in coordination, research support, digital terminology resources, and cross-sector implementation would allow terminology planning to function not only as a linguistic activity but also as part of Mongolia's knowledge infrastructure and sustainable development strategy.

References

Galinski, C., Vesna, L (2024). Empowering knowledge societies using terminological approaches - how top-down policies can meet bottom-up application. *Proceedings of abstracts of the "Terminology and Development of Science and Technology" organized under the auspices of the President of Mongolian.* Ulaanbaatar: Naranzon print LLC. pp.6-8

International Organization for Standardization. (2022). *ISO 704: Terminology work - Principles and methods*. ISO.

International Organization for Standardization. (2001). *ISO 15188: Project management guidelines for terminology standardization*. ISO.

International Organization for Standardization. (2020). *ISO 29383: Terminology policies - Development and implementation*. ISO.

National Council for Language Policy. (2017). *Rule of State Commission for Terminology*. Retrieved from https://nclp.mn/content/73

Wright, S. E., & Budin, G. (Eds.). (1997). *Handbook of terminology management* (Vol. 1–2). John Benjamins Publishing.

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations. https://sdgs.un.org/2030agenda

Impact of Different Conceptualizations on the Representation of Specialized Medical Knowledge: A Case Study on Fecal Microbiota Transplantation

Vanessa Bonato[1], Federica Vezzani[2], Giorgio Maria Di Nunzio[3]

[1]Department of Linguistic and Literary Studies, University of Padua, Via E. Vendramini 13, Padua, 35137, Italy
vanessa.bonato@phd.unipd.it
[2]Department of Linguistic and Literary Studies, University of Padua, Via E. Vendramini 13, Padua, 35137, Italy
federica.vezzani@unipd.it
[3]Department of Information Engineering, University of Padua, Via G. Gradenigo 6b, Padua, 35131, Italy
giorgiomaria.dinunzio@unipd.it

1. Introduction

Within the terminological framework, the representation of specialized knowledge related to the biomedical domain presents a broad spectrum of challenges, stemming from the inherent complexity of the domain and its distinctive features. For instance, ongoing scientific advancements may generate variation that can involve both the linguistic and the conceptual dimensions of terminology (León-Araúz, 2017; Freixa, 2022; Vezzani & Costa, 2024). As a consequence, from a conceptual standpoint, examining medical terminology involves retracing and reflecting the different reconceptualizations of existing medical concepts that have developed diachronically and are shared by domain experts (Vezzani & Costa, 2024). The phenomenon of variation, however, may also impact the representation of concepts and concept relations as formalized within a concept system, the latter of which is envisaged as a "set of concepts structured in one or more related domains according to the concept relations among its concepts" (ISO 1087, 2019). Indeed, knowledge representation is influenced by multidimensionality, as the structure of concept systems is shaped by the foregrounding of some characteristics that constitute a concept over others (Bowker, 1997, 2022; Bowker & Meyer, 1993; Kageura, 1997; Wright, 1997; Rogers, 2004). Thus, selecting one classification of a concept over another directly affects the representation of concept-related knowledge, since different concept systems may represent knowledge about the same concepts in distinct ways. Moreover, multidimensionality is a phenomenon that also manifests in relation to terminological definitions (León-Araúz & San Martín, 2012).

2. Challenges in Gut-Brain Axis Terminology

In this broader context, terminology related to the gut-brain axis poses specific issues in the representation of specialized knowledge at the conceptual level. Within the last few years, this

field of study has experienced a marked increase in scientific interest, as evidenced by the substantial expansion of PubMed literature on the subject, demonstrating its current relevance within the biomedical domain (Martinelli et al., 2025). In particular, research is focusing extensively on the investigation of the potential associations between the gut microbiota and several neurodegenerative diseases, including Alzheimer's disease (Jiang et al., 2017; Zhou et al., 2025), Parkinson's disease (Vascellari et al., 2020; Wang et al., 2025), multiple sclerosis (Correale et al., 2022; Mohsen et al., 2025) and amyotrophic lateral sclerosis (Mazzini et al., 2018; Chen et al., 2024), as well as mental health-related conditions such as anxiety and depression (Simpson et al., 2021; Xiong et al., 2023; Cao et al., 2025). Against this backdrop of rapidly evolving research, concepts require constant terminological analysis to enable the state-of-the-art representation of medical knowledge. This, for example, leads to the necessity of drafting definitions that timely reflect the knowledge held by the community of experts. However, the representation of biomedical knowledge related to the gut-brain interplay must also account for the fact that, in some cases, there is no universal agreement on the definition of concepts, resulting in multiple definitions for a given concept. An illustrative example is the concept <Dysbiosis>[1] which, despite being a key concept in gut-brain axis research, is characterized by the absence of a standardized and universally accepted definition (Guarner et al., 2024; Almonte & Zitvogel, 2025).

Notably, the synchronous existence of multiple definitions of a concept relevant to the gut-brain axis cannot be attributed exclusively to a lack of consensus among domain experts. In some circumstances, indeed, it also reflects the considerable diversity of international regulatory frameworks. Both these factors have direct implications for the representation of the conceptual dimension of specialized biomedical terminology. In this respect, the concept <Fecal microbiota transplantation> emerges as a representative case study for exploring the challenges involved in specialized knowledge representation, when such knowledge lies at the intersection of two distinct specialized domains of human activity: the biomedical and the legal.

3. The Case Study of <Fecal microbiota transplantation>

According to the definition provided in MeSH Terms, fecal microbiota transplantation (FMT) entails the "transfer of gastrointestinal microbiota from one individual to another by infusion of donor feces to the upper or lower gastrointestinal tract of the recipient".[2] In particular, the efficacy of fecal microbiota transplantation has been investigated in relation to multiple medical conditions, among which Parkinson's disease (Dutta et al., 2019; Scheperjans et al., 2024), *Clostridium difficile* infection (Hocquart et al., 2018; Gupta et al., 2022) and psychiatric disorders (Chinna Meyyappan et al., 2020; Doll et al., 2022), and it is actively being explored in health conditions associated with gut dysbiosis (Hou et al., 2025). Notwithstanding this, multiple studies highlight that the concept is characterized by the lack of a definition

[1] In this abstract, we adopt a graphical notation to differentiate between concepts and objects in terminology. Concepts are capitalized and enclosed in single chevrons (as in <Concept>). Objects, instead, are written in lowercase and enclosed in apostrophes (as in 'object').

[2] https://www.ncbi.nlm.nih.gov/mesh/?term=fecal+microbiota+transplantation

universally agreed upon within the scientific community (Hoffmann et al., 2017; Merrick et al., 2020; Mullish et al., 2024). Consequently, issues arises, for instance, in the exact identification of the characteristics that make up the concept, with implications for concept-related knowledge representation and organization.

In the study by Hoffman et al. (2017), issues linked to the definition of <Fecal microbiota transplantation> are raised. Among these considerations, the degree of manipulation applied to transplanted material used in microbiota transplantation (MT) is discussed, highlighting the disagreement on the part of many experts on "whether there is - or should be - a well-defined limit on the degree of manipulation that would be appropriate to include within the definition of MT" (p. 210). In fact, both the formulation and the degree of manipulation of the material used in fecal microbiota transplantation may directly affect classification from a regulatory viewpoint (Hoffman et al., 2017). Taken together, these factors exemplify the challenges involved in defining the concept, as both the clinical and regulatory frameworks should be considered.

Furthermore, <Fecal microbiota transplantation> presents an additional layer of complexity. As a matter of fact, regulatory frameworks adopted by authorities vary substantially across countries (Merrick et al., 2020; Lodberg Hvas et al., 2023; Mullish et al., 2024; Rodriguez et al., 2025). For example, as pointed out by Mullish et al. (2024), fecal microbiota transplantation is considered as a medicinal product in the United Kingdom, while it constitutes a biological product in the USA. At the European level, it is regarded as a human tissue product in Italy (Mullish et al., 2024), whereas it is classified as a drug in France (Sintes et al., 2024). However, an evolution of regulatory frameworks concerning human microbiota can also be traced from a diachronic perspective. Indeed, a new European regulation will apply in August 2027, under which intestinal microbiota will be considered a SoHO, that is, a substance of human origin (European Union, 2024).

The existence of multiple classifications of fecal microbiota transplantation entails distinct consequences, both at the medical level and the terminological level, potentially affecting clinical practice on the one hand and impacting the conceptual dimension of terminology on the other. According to Merrick et al. (2020), indeed, "[r]egulation seeks to improve quality and safety, however, lack of standardisation creates confusion, and overly restrictive regulation may hamper widespread access and discourage research using FMT". From a terminological standpoint, in turn, differences in concept classification result in multiple definitions of the concept, that are directly tied and embedded to the respective regulatory system of reference.

From a terminological perspective, it is therefore possible to observe that multiple conceptualizations of the object 'fecal microbiota transplantation' emerge, specifically motivated by the existence of different regulations on the matter. Each conceptualization gives rise to a different concept, understood as a "unit of knowledge created by a unique combination of characteristics" (ISO 1087, 2019). Indeed, the set of characteristics that constitutes each concept respectively includes a different generic concept, defined as the "concept in a generic relation that has the narrower intension". Each of the concepts, that are abstraction of the same

object, can be situated within a distinct concept system, in which conceptual knowledge related to each regulatory system can be organized.

Adopting the theoretical approach proposed by Vezzani (2025), the relation established between these concepts can be framed as a relation of exact equivalence, situated within the conceptual dimension of terminological analysis. As a matter of fact, the three conditions outlined by the author as necessary and sufficient to determine the phenomenon of exact equivalence are concurrently met. Specifically, the first condition concerns a required relation between entities that share the same nature from an ontological viewpoint, which, in this circumstance, indeed occurs between multiple concepts. The second condition calls for these concepts to belong to distinct concept systems, which, in this case study, represent different regulatory systems respectively adopted in different countries. Finally, the third condition is likewise fulfilled, as these concepts constitute the abstraction of the same object.

As previously mentioned, however, the lack of a standardized definition as well as the coexistence of different conceptualizations of the same object give rise to fundamental questions regarding the representation of specialized biomedical knowledge at the conceptual level. For instance, it poses challenges for representing and formalizing knowledge within a single concept system that should reflect the knowledge collectively shared by the experts of the domain. Moreover, the lack of a standardized definition results in difficulties in identifying the characteristics that constitute the concept, whose determination is fundamental to the construction of the concept system. Particularly, a key issue arises in individuating the essential and delimiting characteristics of the concepts.

4. Objectives of the Work

Given these theoretical premises, in this work we will investigate at a finer-grained level of analysis the different conceptualizations of ‘fecal microbiota transplantation’ across various regulatory systems. The objectives of this study are: 1) to trace and compare the conceptualizations of ‘fecal microbiota transplantation’ across regulatory frameworks, and 2) to address the representation of analyzed concepts within a concept system specific to the gut-brain axis domain.

From a methodological standpoint, the analysis will focus on identifying the characteristics that make up the concepts under investigation, especially aiming at determining their respective essential and delimiting characteristics. To this end, regulations adopted in the countries under consideration will serve as a starting point for the terminological analysis.

Building on this, the work will examine the impact that the existence of multiple concepts that are the abstractions of the same object and are characterized by a relation of exact equivalence has on the representation of specialized knowledge within the gut-brain axis domain. Specifically, emphasis will be placed on the role that associative concept relations assume in this context, within a concept system purposefully developed to represent and systematically

organize concepts and concept relations related to the gut-brain interplay. In particular, the study illustrates and exemplifies how associative relations can be used to represent relations between concepts within a single concept system, which reflects the shared knowledge of domain experts on the gut-brain axis and transcends differences related to regulatory frameworks.

To conclude, the study will address the systematic representation of the analyzed terminological data within the FAIRterm terminology resource, in which terminological entries are characterized by their concept orientation (Vezzani, 2021).

Taking this work as a starting point, future research concerns the analysis of the representation of concepts situated at the crossroads between the biomedical and legal domains in other resources, such as specialized resources specifically focused on the representation of knowledge on the gut-brain axis. Moreover, it is also essential to consider that concepts relevant to the gut-brain axis pertain to an emerging domain that is continuously evolving. Taking this into account, another line of investigation to be considered for the study of these concepts concerns the phenomenon of conceptual variation in relation to conceptual and referential neologisms (Pelletier, 2012).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-4 in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

5. References

Almonte, A. A., & Zitvogel, L. (2025). Gut reactions: Harnessing microbial metabolism to fuel next-generation cancer immunotherapy. *Journal for ImmunoTherapy of Cancer*, *13*(4), e011540. https://doi.org/10.1136/jitc-2025-011540

Bowker, L. (1997). Multidimensional Classification of Concepts and Terms. In S. E. Wright & G. Budin (Eds.), *Handbook of Terminology Management: Volume 1: Basic Aspects of Terminology Management* (pp. 133–143). John Benjamins Publishing Company. https://doi.org/10.1075/z.htm1.16bow

Bowker, L. (2022). Multidimensionality. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge* (pp. 127–148). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.06bow

Bowker, L., & Meyer, I. (1993). Beyond 'textbook' concept systems: Handling multidimensionality in a new generation of term banks. In K.-D. Schmitz (Ed.), *Proceedings*

of the Third International Congress on Terminology and Knowledge Engineering (pp. 123-137). Indeks Verlag.

Cao, Y., Cheng, Y., Pan, W., Diao, J., Sun, L., & Meng, M. (2025). Gut microbiota variations in depression and anxiety: A systematic review. *BMC Psychiatry*, *25*(1), 443. https://doi.org/10.1186/s12888-025-06871-8

Chen, S., Cai, X., Lao, L., Wang, Y., Su, H., & Sun, H. (2024). Brain-gut-microbiota axis in amyotrophic lateral sclerosis: A historical overview and future directions. *Aging and Disease*, *15*(1), 74–95. https://doi.org/10.14336/AD.2023.0524

Chinna Meyyappan, A., Forth, E., Wallace, C. J. K., & Milev, R. (2020). Effect of fecal microbiota transplant on symptoms of psychiatric disorders: A systematic review. *BMC Psychiatry*, *20*(1), 299. https://doi.org/10.1186/s12888-020-02654-5

Correale, J., Hohlfeld, R., & Baranzini, S. E. (2022). The role of the gut microbiota in multiple sclerosis. *Nature Reviews Neurology*, *18*, 544–558. https://doi.org/10.1038/s41582-022-00697-8

Doll, J. P. K., Vázquez-Castellanos, J. F., Schaub, A.-C., Schweinfurth, N., Kettelhack, C., Schneider, E., Yamanbaeva, G., Mählmann, L., Brand, S., Beglinger, C., Borgwardt, S., Raes, J., Schmidt, A., & Lang, U. E. (2022). Fecal microbiota transplantation (FMT) as an adjunctive therapy for depression—Case report. *Frontiers in Psychiatry*, *13*, 815422. https://doi.org/10.3389/fpsyt.2022.815422

Dutta, S. K., Verma, S., Jain, V., Surapaneni, B. K., Vinayek, R., Phillips, L., & Nair, P. P. (2019). Parkinson's disease: The emerging role of gut dysbiosis, antibiotics, probiotics, and fecal microbiota transplantation. *Journal of Neurogastroenterology and Motility*, *25*(3), 363–376. https://doi.org/10.5056/jnm19044

European Union. (2024). Regulation (EU) 2024/1938 of the European Parliament and of the Council of 13 June 2024 on standards of quality and safety for substances of human origin intended for human application and repealing Directives 2002/98/EC and 2004/23/EC (Text with EEA relevance). *EUR-LEX*. http://data.europa.eu/eli/reg/2024/1938/oj

Freixa, J. (2022). Causes of terminological variation. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology: Explaining Terms, Concepts and Specialized Knowledge* (pp. 399–420). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.18fre

Guarner, F., Bustos Fernandez, L., Cruchet, S., Damião, A., Maruy Saito, A., Riveros Lopez, J. P., Rodrigues Silva, L., & Valdovinos Diaz, M. A. (2024). Gut dysbiosis mediates the association between antibiotic exposure and chronic disease. *Frontiers in Medicine*, *11*, 1477882. https://doi.org/10.3389/fmed.2024.1477882

Gupta, K., Tappiti, M., Nazir, A. M., Koganti, B., Memon, M. S., Aslam Zahid, M. B., Shantha Kumar, V., & Mostafa, J. A. (2022). Fecal microbiota transplant in recurrent Clostridium Difficile infections: A systematic review. *Cureus*, *14*(5), e24754. https://doi.org/10.7759/cureus.24754

Hocquart, M., Lagier, J.-C., Cassir, N., Saidani, N., Eldin, C., Kerbaj, J., Delord, M., Valles, C., Brouqui, P., Raoult, D., & Million, M. (2018). Early fecal microbiota transplantation improves survival in severe Clostridium difficile infections. *Clinical Infectious Diseases*, *66*(5), 645–650. https://doi.org/10.1093/cid/cix762

Hoffmann, D. E., Palumbo, F. B., Ravel, J., Rowthorn, V., & von Rosenvinge, E. (2017). A proposed definition of microbiota transplantation for regulatory purposes. *Gut Microbes*, *8*(3), 208–213. https://doi.org/10.1080/19490976.2017.1293223

Hou, S., Yu, J., Li, Y., Zhao, D., & Zhang, Z. (2025). Advances in fecal microbiota transplantation for gut dysbiosis-related diseases. *Advanced Science*, *12*(13), 2413197. https://doi.org/10.1002/advs.202413197

ISO 1087. (2019). *Terminology work and terminology science – Vocabulary*. International Organization for Standardization. https://www.iso.org/standard/62330.html

Jiang, C., Li, G., Huang, P., Liu, Z., & Zhao, B. (2017). The gut microbiota and Alzheimer's disease. *Journal of Alzheimer's Disease*, *58*(1), 1–15. https://doi.org/10.3233/JAD-161141

Kageura, K. (1997). Multifaceted/Multidimensional Concept Systems. In S. E. Wright & G. Budin (Eds.), *Handbook of Terminology Management: Volume 1: Basic Aspects of Terminology Management* (pp. 119–132). John Benjamins Publishing Company. https://doi.org/10.1075/z.htm1.15kag

León-Araúz, P. (2017). Term and concept variation in specialized knowledge dynamics. In P. Drouin, A. Francœur, J. Humbley & A. Picton (Eds.), *Multiple Perspectives on Terminological Variation* (pp. 213–258). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.18.09leo

León-Araúz, P., & San Martín, A. (2012). Multidimensional categorization in terminological definitions. In R. Vatvedt Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 578–584). Department of Linguistics and Scandinavian Studies, University of Oslo. https://euralex.org/publications/multidimensional-categorization-in-terminological-definitions/

Lodberg Hvas, C., Keller, J., Dahl Baunwall, S. M., Edwards, L. A., Ianiro, G., Kupcinskas, J., Link, A., Mullish, B. H., Satokari, R., Terveer, E., & Vehreshild, M. J. G. (2023). European academic faecal microbiota transplantation (EURFMT) network: Improving the safety and

quality of microbiome therapies in Europe. *Microbiota in Health and Disease, 5*, e954. https://doi.org/10.26355/mhd_202311_954

Martinelli, M., Silvello, G., Bonato, V., Di Nunzio, G. M., Ferro, N., Irrera, O., Marchesin, S., Menotti, L., & Vezzani, F. (2025). Overview of GutBrainIE@CLEF 2025: Gut-brain interplay information extraction. In G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), *CLEF 2025 Working Notes, Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September 2025* (pp. 65–98). CEUR-WS. https://ceur-ws.org/Vol-4038/paper_5.pdf

Mazzini, L., Mogna, L., De Marchi, F., Amoruso, A., Pane, M., Aloisio, I., Bozzi Cionci, N., Gaggìa, F., Lucenti, A., Bersano, E., Cantello, R., Di Gioia, D., & Mogna, G. (2018). Potential role of gut microbiota in ALS pathogenesis and possible novel therapeutic strategies. *Journal of Clinical Gastroenterology*, *52*, S68-S70. https://doi.org/10.1097/MCG.0000000000001042

Merrick, B., Allen, L., Zain, N. M. M., Forbes, B., Shawcross, D. L., & Goldenberg, S. D. (2020). Regulation, risk and safety of faecal microbiota transplant. *Infection Prevention in Practice*, *2*(3), 100069. https://doi.org/10.1016/j.infpip.2020.100069

Mohsen, E., Haffez, H., Ahmed, S., Hamed, S., & El-Mahdy, T. S. (2025). Multiple sclerosis: A story of the interaction between gut microbiome and components of the immune system. *Molecular Neurobiology*, *62*, 7762–7775. https://doi.org/10.1007/s12035-025-04728-5

Mullish, B. H., Merrick, B., Quraishi, M. N., Bak, A., Green, C. A., Moore, D. J., Porter, R. J., Elumogo, N. T., Segal, J. P., Sharma, N., Marsh, B., Kontkowski, G., Manzoor, S. E., Hart, A. L., Settle, C., Keller, J. J., Hawkey, P., Iqbal, T. H., Goldenberg, S. D., & Williams, H. R. T. (2024). The use of faecal microbiota transplant as treatment for recurrent or refractory Clostridioides difficile infection and other potential indications: Second edition of joint British Society of Gastroenterology (BSG) and Healthcare Infection Society (HIS) guidelines. *Journal of Hospital Infection*, *148*, 189–219. https://doi.org/10.1016/j.jhin.2024.03.001

Pelletier, J. (2012). La variation terminologique: un modèle à trois composantes. [Doctoral dissertation, Université Laval]. https://hdl.handle.net/20.500.11794/23488

Rodriguez, J., Cordaillat-Simmons, M., Pot, B., & Druart, C. (2025). The regulatory framework for microbiome-based therapies: Insights into European regulatory developments. *npj Biofilms and Microbiomes*, *11*, 53. https://doi.org/10.1038/s41522-025-00683-0

Rogers, M. (2004). Multidimensionality in concepts systems: A bilingual textual perspective. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *10*(2), 215–240. https://doi.org/10.1075/term.10.2.04rog

Scheperjans, F., Levo, R., Bosch, B., Lääperi, M., Pereira, P. A. B., Smolander, O.-P., Aho, V. T. E., Vetkas, N., Toivio, L., Kainulainen, V., Fedorova, T. D., Lahtinen, P., Ortiz, R., Kaasinen,

V., Satokari, R., & Arkkila, P. (2024). Fecal microbiota transplantation for treatment of Parkinson disease: A randomized clinical trial. *JAMA Neurology*, *81*(9), 925–938. https://doi.org/10.1001/jamaneurol.2024.2305

Simpson, C. A., Diaz-Arteche, C., Eliby, D., Schwartz, O. S., Simmons, J. G., & Cowan, C. S. M. (2021). The gut microbiota in anxiety and depression – A systematic review. *Clinical Psychology Review*, *83*, 101943. https://doi.org/10.1016/j.cpr.2020.101943

Sintes, R., McLellan, P., Navelli, G., Landman, C., Delage, S., Truong, S., Benech, N., Kapel, N., Moreino Sabater, A., Schnuriger, A., Eckert, C., Bleibtreu, A., Joly, A.-C., & Sokol, H. (2024). Use of frozen native feces for fecal microbiota transplantation in recurrent Clostridioides difficile infection: A simple way to improve the efficiency of donor feces preparation. *Antimicrobial Agents and Chemotherapy*, *68*(10), e00734-24. https://doi.org/10.1128/aac.00734-24

Vascellari, S., Palmas, V., Melis, M., Pisanu, S., Cusano, R., Uva, P., Perra, D., Madau, V., Sarchioto, M., Oppo, V., Simola, N., Morelli, M., Santoru, M. L., Atzori, L., Melis, M., Cossu, G., & Manzin, A. (2020). Gut microbiota and metabolome alterations associated with Parkinson's disease. *mSystems*, *5*(5), e00561-20. https://doi.org/10.1128/mSystems.00561-20

Vezzani, F. (2021). La ressource FAIRterm: Entre pratique pédagogique et professionnalisation en traduction spécialisée. *Synergies Italie*, *17*, 51-64. https://gerflint.fr/Base/Italie17/vezzani.pdf

Vezzani, F. (2025). Pour une formalisation du concept d'équivalence en terminologie. *MediAzioni*, *46*(1), A138-A156. https://doi.org/10.6092/issn.1974-4382/21937

Vezzani, F., & Costa, R. (2024). Variation in psychopathological terminology: A case study on Body Dysmorphic Disorder. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *30*(1), 81–106. https://doi.org/10.1075/term.00078.vez

Wang, L., Cui, Y., Han, B., Du, Y., Salewala, K. S., Wang, S., Zhao, W., Zhang, H., Wang, S., Xu, X., Ma, J., Zhu, Y., & Tuo, H. (2025). Gut microbiota and Parkinson's disease. *Chinese Medical Journal*, *138*(3), 289–297. https://doi.org/10.1097/CM9.0000000000003318

Wright, S. E. (1997). Representation of concept systems. In S. E. Wright & G. Budin (Eds.), *Handbook of Terminology Management: Volume 1: Basic Aspects of Terminology Management* (pp. 89–97). John Benjamins Publishing Company. https://doi.org/10.1075/z.htm1.13wri

Xiong, R.-G., Li, J., Cheng, J., Zhou, D.-D., Wu, S.-X., Huang, S.-Y., Saimaiti, A., Yang, Z.-J., Gan, R.-Y., & Li, H.-B. (2023). The role of gut microbiota in anxiety, depression, and other mental disorders as well as the protective effects of dietary components. *Nutrients*, *15*(14), 3258. https://doi.org/10.3390/nu15143258

Zhou, X.-P., Sun, L.-B., Liu, W.-H., Zhu, W.-M., Li, L.-C., Song, X.-Y., Xing, J.-P., & Gao, S.-H. (2025). The complex relationship between gut microbiota and Alzheimer's disease: A systematic review. *Ageing Research Reviews*, *104*, 102637. https://doi.org/10.1016/j.arr.2024.102637

Terminological Variation and Standardization in the Translation of Historical Anatomical Texts: The Case of the Guane Mummies

Heidy Alegría Leon-Gutierrez[1], Mateo Uscategui-Blanco[2], Sylvia Fernanda Guerrero-Ramírez[3]

[1]Department of Languages, Universidad Industrial de Santander, Santander, Colombia, hagutier@uis.edu.co

[2]Department of Languages, Universidad Industrial de Santander, Santander, Colombia, MATEO.USCATEGUI@correo.uis.edu.co

[3]Department of Languages, Universidad Industrial de Santander, Santander, Colombia, sylvia2201025@correo.uis.edu.co

Abstract

This work is presented within the framework of the translation of the text Estudio de las momias Guane (Santander, Colombia) (1992). The source text is technical-scientific in nature and highly anatomical. Being an article now three decades old, it does not meet current standards of anatomical terminology. Over the years, anatomical nomenclature has undergone several revisions. Nowadays, the currently widespread standard is Terminologia anatomica (TA), which consolidates and updates terms across shape, structure, size or dimensions, location, function, and color. Considering this context, a terminological database and categorization was developed in order to update the target document with current standards. This categorization includes eponymy, synonymy, lack of standardization, and polysemy. For example, the eponymous term enfermedad de Pott, which in English should be translated as spinal tuberculosis. This approach not only preserved terminological accuracy in our translation, but it also highlighted the significance for systematic procedures towards bridging outdated terminology with updated standards worldwide, a prevalent challenge to both medical translation and terminological research.

Keywords: Terminologia Anatomica; synonymy; eponymy; lack of standardization.

1. Introduction

Although the article *Estudio de las Momias Guanes de la Mesa de los Santos* (Correal & Flórez, 1998) constitutes a significant contribution to the scientific and cultural knowledge of the Guane population through its detailed anatomical descriptions, its translation poses several terminological issues. Many of the anatomical terms used do not align with current international standards, as the source text relies on outdated anatomical terminology and inconsistent nomenclature resulting in different terminological variations (synonym, eponymy, and lack of standardization). To address these challenges, the present study adopts a prescriptive terminological approach (Cabré, 1999), grounded in the *Terminologia Anatomica* (TA), aiming to ensure the linguistic and conceptual validity of the anatomy-related terms translated into English. The main objective of this study is to standardize the outdated anatomical nomenclature and terminological variations found in the article *Estudio de las Momias Guanes de la Mesa de los Santos* (1992).

2. Methods

This study implements a terminology-oriented translation approach to produce a scientifically accurate English version of the article "Estudio de las Momias Guanes de la Mesa de los

Santos". The method combines specialized translation, documentary research, and terminological standardization to achieve conceptual accuracy and consistency in the target text.

The corpus consists of one academic article chosen for its cultural and scientific relevance. The text possesses a high density of anatomical references, as well as historical and anthropological information regarding the Guane mummies, making it ideal for terminological examination. Because the original text was developed before the widespread acceptance of contemporary anatomical standards, special care was taken to identify outdated, confusing, or non-standard words.

The research was conducted in three stages. The initial stage was data gathering, which included transcription of relevant segments, extraction of specialized terminology, and preliminary documenting of anatomical and cultural references. During this phase, specialized dictionaries, academic papers, and institutional databases were used to establish accurate conceptual correspondences.

The second stage involved terminological analysis and translation. Each extracted term was verified using documentary research and compared with standardized anatomical nomenclature, namely Terminologia Anatomica. Translation decisions were based on concepts, with terminological accuracy and consistency taking precedence over literal equivalency when appropriate.

The final stage involved revising and validating the translated text. Terminological choices were reviewed to guarantee internal coherence and compliance with current scientific terminology. This systematic procedure enabled the translation to function not just as a linguistic transfer, but also as a process of terminological updating, boosting clarity, precision, and accessibility for a worldwide academic and museum audience.

3. Results and analysis

3.1. Terminological standardization

Despite ongoing efforts to achieve uniformity in medical terminology, multiple nomenclatures persist to be used today, posing challenges for univocity and monoreferentiality, leading to a lack of clarity and precision in medical terminology (Montalt et al., 2018).

Example 1

ST	**TT**
peroné	fibula

For example, the use of the term *peroné* reflects an outdated anatomical nomenclature that no longer corresponds to standardized terminology. According to the *Terminologia Anatomica* standard, the correct designation is *fibula*, which refers to the lateral of the two bones of the lower leg.

3.2. Eponymy

Eponyms are terminology formed from proper names in medical and anatomical discourse, which may originate from researchers, institutions, geographic areas, or historical references. eponymy directly opposes the principle of univocity and represents a major source of terminological ambiguity (Maslova, 2022).

Example 2

ST	**TT**
...han sido reportados casos de T.B.C. y posible **enfermedad de Pott**...	…and possible **spinal tuberculosis** cases have been reported…

The eponym *enfermedad de Pott* (Pott's disease) is named after Percival Pott, an English surgeon who discovered the medical condition in 1779 (Köken et al., 2015). This case highlights the persistence of historically grounded eponyms in specialized texts, despite their terminological inadequacy, and represents a challenge for translators.

3.3. Synonymy

Rask (2008) explained that a synonym refers to words that have the same fundamental meaning to another word, but differ in style, connotation, or nuance. Despite synonyms having similar meanings, their use in specialized contexts requires careful evaluation to avoid ambiguity.

Example 3

ST	**TT**
vértice	apex

An illustrative case is the use of the terms *vértice* and *apex*, both of which are employed in the source text to designate the same anatomical structure: the lung's highest structure located in the upper lobe. Although two denominations are used, *apex* is the term endorsed by the *Terminologia Anatomica*. Accordingly, apex was selected as the sole designation in the target text.

4. Conclusion

The purpose of this study was to translate and update the anatomical terminology of the academic text Estudio de las momias Guane (Santander, Colombia), originally written in Spanish in 1992, within the framework of the Museo Casa de Bolívar translation project in Bucaramanga. To ensure an accurate and up-to-date translation into English, a prescriptive terminological approach was adopted, drawing on Cabré's (1999) theoretical framework and Temmerman's (2000) principle of univocity, and aligned with the Terminologia Anatomica.

One of the main outcomes of the project was the development of a terminological database that enabled the systematic identification and standardization of outdated anatomical nomenclature,

ensuring consistency and scientific reliability throughout the translated text. This study highlights the relevance of terminological standardization in multidisciplinary contexts, particularly those involving anatomy, anthropology, and cultural heritage, and proposes this methodology as a foundation for future museographic and academic translation projects.

References

Cabré, M. T. (1999). *Terminology: Theory, Methods, and Applications* (J. Felber & H. Heribert Picht, Eds.; J. DeCesaris, Trans.). John Benjamins.

Correal, G., & Flórez, I. (1992). Estudio de las momias guanes de la Mesa de los Santos. (Santander, Colombia). *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, *18*(70), 283-290.

Echeverria, E., & Jimenez, I. (2010). La terminología anatómica en español, inglés y francés. *Panacea*, *11*(31), 47-57.

Köken, M., Özdoğan, S., Kendirlioglu, B. C., & Kocaman, B. (2015). Spinal tuberculosis. The Journal of Turkish Spinal Surgery, 26(1), 61–66.

Maslova, O. (2022). Traditional and modern trends in medical terminology formation in various languages. *Journal of Language and Linguistic Studies*, 18(Special Issue 2), 1043-1050.

Montalt, V., Zethsen, K. K., & Karwacka, W. (2018). Medical translation in the 21st century - challenges and trends. *MonTi Monografías De Traducción E Interpretación*, 9–25. https://doi.org/10.6035/MonTI.2018.10.1

Rask, N. (2008). *Analysis of a medical translation: Terminology and cultural aspects* [Master's thesis, Växjö University]. DiVA Portal. https://urn.kb.se/resolve?urn=urn:nbn:se:vxu:diva-2370rkh.diva-portal.org+9

Temmerman, R. (2000). *Towards new ways of terminology description: The sociocognitive approach* (Terminology and Lexicography Research and Practice). John Benjamins.

Modéliser la variation terminologique entre diachronie et synchronie: le mariage hébraïque comme étude de cas

Piccini, S.*, Vilela Ruiz, G. E.*, Saponaro, D.°, Bellandi, A.*
silvia.piccini@ilc.cnr.it
giulianaelizabeth.vilelaruiz@ilc.cnr.it
davide.saponaro1978@gmail.com
andrea.bellandi@ilc.cnr.it
*Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa
°Fondazione Rut ETS-Ente filantropico

1. Introduction

Au cours des dernières décennies, la terminologie a accordé une attention croissante aux phénomènes variationnels qui affectent les lexiques de spécialité [5,7,8,10,11,12]. Envisagée dans la tradition wüstérienne comme un phénomène à maîtriser au nom de la normalisation, la variation est aujourd'hui reconnue comme structurelle et inhérente aux usages. Elle ne constitue plus une anomalie à corriger, mais une dynamique à décrire et à modéliser.
Dans ce cadre, la variation diachronique, longtemps restée marginale, a progressivement gagné en importance sur les plans théorique et méthodologique [2,9,15,20,21,23,24,33]. Parallèlement, la variation synchronique intralinguistique — qui se manifeste notamment par la coexistence de termes coréférentiels, ou dispersion terminologique — n'est plus seulement admise, mais considérée comme fonctionnelle. Elle reflète en effet des perspectives conceptuelles distinctes : les différents termes, envisagés comme des « chemins dénominatifs » [1], offrent des points de vue complémentaires sur un même concept en mettant en saillance certains de ses traits spécifiques .
Cette contribution vise à modéliser formellement ces deux dimensions de la variation. À partir d'une étude terminologique du mariage en hébreu — de la Bible au Talmud, puis au *Shulḥan ʿArukh* — nous analysons les reconfigurations successives des traits conceptuels et leurs modalités de lexicalisation. Nous proposons une modélisation ontologique fondée sur une approche perdurantiste (4D) [31], compatible avec le langage OWL, permettant de représenter de manière interrogeable l'évolution diachronique du concept ainsi que la variation synchronique des dénominations. Un niveau conceptuel, destiné à représenter les transformations du concept dans le temps, est ainsi articulé avec un niveau lexical, consacré à la description des dénominations et de leurs relations, et représenté dans le cadre du standard *de facto* OntoLex-Lemon [18]. L'objectif est double : permettre le raisonnement automatique sur le changement conceptuel et de dénominations et produire des ressources FAIR [32], intégrables dans l'écosystème des *Données Linguistiques Ouvertes et Liées* (LLOD) [3].

2. Analyse terminologique diachronique du mariage en hébreu

Afin d'ancrer la modélisation dans les données, cette section analyse la terminologie du mariage en hébreu — Bible, Talmud, *Shulḥan ʿArukh* — en identifiant, pour chaque période, les traits saillants du concept et leurs lexicalisations (voir Tableau 1). Ces traits conceptuels ont été dégagés à partir d'une analyse systématique de sources spécialisées, notamment des travaux de référence sur le mariage dans la tradition juive [*inter al.* 4,6,14,17,25,28], ainsi que de contributions anthropologiques plus générales (en particulier celles de [16] et [19]). Cette analyse a été complétée par un dialogue constant avec des spécialistes de la culture et de la tradition hébraïques, afin de garantir la solidité et la cohérence de l'interprétation proposée. Les trois sous-sections suivantes décrivent, pour chacune des périodes retenues, les termes associés au mariage ainsi que les caractéristiques définitoires du concept.

2.1. La terminologie du mariage dans la Bible (env. 500 av. J.-C. – 100 av. J.-C.)[1]

La Bible hébraïque offre une représentation fragmentaire du mariage. Malgré l'absence d'un verbe ou d'un substantif spécifique pour désigner l'acte matrimonial [26,29,30], la terminologie biblique révèle un noyau de dénominations récurrentes, qui structure une conceptualisation polyédrique autour de traits tels que l'/acquisition de la femme/, la /possession de la femme/, le /monopole sexuel de l'homme/, la /paternité légale/ et la /virilocalité/. Loin de constituer de simples variantes synonymiques, ces termes mettent en saillance des traits distincts de cette conceptualisation.

Le terme le plus fréquemment attesté pour désigner l'acte matrimonial du point de vue de l'homme[2] est le verbe לקח *laqaḥ* (« prendre pour soi ou pour autrui », et plus spécifiquement « prendre en mariage », Gesenius 1953 : 542) qui met en évidence une conceptualisation du mariage fondée sur un modèle juridique d'acquisition. Le mariage se présente comme une transaction intégrant la femme au lignage du mari et lui conférant des droits sur la progéniture (/paternité légale/). Cette conceptualisation se reflète également dans l'emploi du terme בעל *baʿal* (« régner, épouser » Gesenius 1953: 127, « posséder » Klein 1987: 79), qui souligne le statut de la femme comme propriété du mari et le monopole sexuel qui en découle. Dans le droit juif ancien, la possession corporelle de la femme consacrait *ipso facto* le lien matrimonial.

Un autre trait fondamental du mariage biblique est mis en relief par des verbes tels que נשא *nasa'* (« conduire »), הושיב *hoshiv* (« faire habiter »; forme *hifʿil*, Gesenius 1953: 442-443) ou שלח *shillaḥ* (« envoyer », forme *piʿel*), qui impliquent tous le déplacement de la femme vers le foyer du mari. La cohabitation apparaît ainsi comme l'élément décisif conférant au mariage sa validité sociale et juridique.

Du point de vue de la femme, en revanche, le mariage est exprimé presque exclusivement par des constructions statives fondées sur le verbe היה *hayah* (« être ») suivi de la particule ל *le-* « pour, à », associées au trait de /possession de la femme/, cette dernière passant alors par le mariage de la juridiction du *pater familias* à celle du mari.

---

[1] La datation des livres bibliques demeure discutée ; nous adoptons une chronologie restrictive, tout en reconnaissant que certains matériaux remontent au XII^{e} siècle av. J.-C.

[2] Pour une analyse détaillée de la terminologie biblique et talmudique, voir [22].

Trait conceptuel	Terme		Définition du trait conceptuel	Source
/acquisition de la femme/	לקח אשה קנה	*laqaḥ 'isha* *qanah*	Dans le mariage, l'acquisition constituait un modèle juridique régissant le transfert de l'épouse de sa famille d'origine vers celle du mari. Le paiement du *mohar* compensait la perte subie par la famille paternelle et consacrait l'intégration de la femme dans la lignée du mari. Le concept d'acquisition était en outre lié au contrôle de la fertilité féminine : la descendance engendrée s'inscrivait dans la lignée paternelle.	Bible
/possession de la femme/	בעל קנה היה ל	*ba'al* *qanah* *hayah le*	À travers l'institution juridique du mariage, les femmes, avec l'ensemble de leurs biens, passaient sous la propriété et l'autorité de leurs maris, quittant ainsi la juridiction du *pater familias* pour entrer sous celle du mari.	Bible
/monopole sexuel de l'homme/	בעל	*ba'al*	La possession physique de la femme consacrait *ipso facto* le lien matrimonial.	Bible
/virilocalité/	נשא הושיב שלח	*nasa'* *hoshiv* *shillaḥ*	L'homme conduisait la femme vers sa nouvelle demeure, marquant le début de la vie conjugale ; par le mariage, la femme quittait la maison de son père pour rejoindre la *domus* de son mari.	Bible
	נשא סב כנס	*nasa'* *nesav* *kanas*		Talmud
	נשא	*nasa'*		*Shulḥan 'Arukh*
/paternité légale/	-	-	Dans le mariage, l'homme acquiert des droits sur la progéniture.	Bible
/consécration exclusive/	קידש	*qiddesh*	La femme s'engage à n'avoir des relations sexuelles qu'avec un homme déterminé, se consacrant exclusivement à lui et devenant interdite à tout autre, en contrepartie de droits spécifiques définis par la jurisprudence hébraïque.	Talmud
/intégration sociale/	נשא ערב	*nasa'* *'arav*	Le mariage est conçu comme un processus d'intégration sociale, par lequel s'opère l'entrée dans un cadre familial et généalogique spécifique. L'acte matrimonial ne concerne pas uniquement la femme comme objet direct de l'action, mais inclut également sa famille d'origine, dont la fonction apparaît déterminante. Le mariage produit des effets directs sur la pureté généalogique et l'organisation du lignage.	*Shulḥan 'Arukh*

Tableau 1. *Traits conceptuels du mariage et leurs lexicalisations dans les sources hébraïques*

2.2 La terminologie du mariage dans le Talmud (env. 150-500 apr. J.-C.)

Par rapport à la Bible, le Talmud réorganise la terminologie du mariage dans un cadre juridique plus explicite. Élaboration majeure de la loi et de la pensée juives, le Talmud, compilé entre le IIe et le VIe siècle de l'ère commune, consacre un traité spécifique au mariage, le traité *Qiddushin*, centré sur les procédures juridiques permettant d'établir le lien conjugal. Sans définir explicitement la notion de mariage, présumée connue au lecteur, le texte examine en détail la casuistique complexe des קידושין *qiddushin* (litt. « consécrations »), acte qui établit juridiquement le lien matrimonial.

Le concept de קדושה *qedushah*, généralement traduit par « sainteté », repose sur l'idée de séparation et de mise à part ; appliqué au domaine matrimonial, il donne lieu à une terminologie technique précise. Par l'acte des *qiddushin*, la femme devient מקודשת *mequddeshet* (« consacrée »), c'est-à-dire réservée de manière exclusive à un homme déterminé, tandis que celui-ci est מקדש *meqaddesh* (« celui qui consacre »). Cette terminologie met en saillance un trait conceptuel nouveau, celui de la /consécration exclusive/, redéfinissant le mariage comme un statut juridique formel.

Parallèlement, le trait de la /virilocalité/ fait l'objet de multiples lexicalisations. Le verbe *nasa'* (נשא), peu fréquent dans la Bible, y est largement attesté et sert de base au terme *nissu'in* (נישואין), désignant la phase où l'épouse est conduite dans la maison du mari après la consécration. La même conceptualisation apparaît dans l'araméen *nesav* (נסב, parfois vocalisé *nesev* « soulever, prendre, emmener, épouser », Jastrow 1903 : 195) ainsi que dans des verbes exprimant l'« entrée » dans la sphère conjugale, notamment l'hébreu *kanas* (כנס, Jastrow 1903: 649-650; Klein 1987: 280), associé à l'entrée dans la maison ou le noyau familial du mari.

2.3 La terminologie du mariage dans le *Shulḥan 'Arukh* (XVIe siècle apr. J.-C.)

Dans le *Shulḥan 'Arukh*, code de loi hébraïque rédigé au XVIe siècle en Galilée par le rabbin Yosef Karo et publié à Venise en 1565, la terminologie du mariage se caractérise par une continuité avec les sources antérieures, mais aussi par des innovations qui placent au centre les enjeux de lignage et d'intégration familiale. Elle repose principalement sur le verbe *nasa'*, marquant une réactivation de la terminologie biblique, au détriment des désignations plus techniques du Talmud. Toutefois, par rapport aux sources antérieures, le verbe se combine plus fréquemment avec des compléments désignant des catégories féminines liées à des lignages spécifiques, notamment sacerdotaux, définissant la femme avant tout par son statut social et généalogique et par les qualités transmissibles qui lui sont attachées (pureté, prestige, aptitude rituelle). Cette conceptualisation est renforcée par l'emploi de compléments prépositionnels renvoyant explicitement à la famille, comme dans l'expression לישא מהם *lissa' mehèm* (« épouser à partir d'une famille »), qui configure le mariage non seulement comme l'union de deux individus, mais comme un processus d'intégration dans un cadre familial déterminé.

C'est dans ce contexte que s'inscrit le verbe ערב *'arav*, lequel introduit une innovation terminologique significative. Ce verbe ne signifie pas « épouser » au sens strict, mais « se mêler » ; dans le contexte matrimonial, il prend ainsi le sens d'« entrer dans une famille par le mariage

». Son emploi est réservé à des individus au lignage problématique (esclaves et personnes nées d'unions interdites, entre autres) et apparaît principalement dans les discussions relatives aux restrictions concernant les *kohanim* (classe sacerdotale). Ce terme active ainsi le champ sémantique de l'/intégration sociale/, déplaçant l'attention de l'acte matrimonial en tant que tel vers ses effets sur la pureté généalogique et l'ordre social.

3. Proposition de représentation formelle de la variation diachronique et synchronique du mariage

Comme nous l'avons montré dans les sections précédentes, le concept de <MARIAGE> peut être conçu comme un ensemble de traits susceptibles d'évoluer dans le temps. Cette variabilité soulève la question ontologique de la persistance à travers le changement, au cœur du débat de la métaphysique contemporaine entre endurantisme et perdurantisme [13,27].
Selon l'endurantisme (3D), une entité est entièrement présente à chaque instant de son existence. Appliquée à des concepts historiquement variables, cette approche tend à leur attribuer des propriétés supposées toujours vraies, au risque de produire des descriptions anachroniques ou conceptuellement incohérentes. À l'inverse, dans une perspective perdurantiste ou quadridimensionnelle (4D), les entités sont conçues comme temporellement étendues et composées de parties temporelles (*time slices*). Dans ce cadre, le changement — acquisition, perte ou reconfiguration de traits — est modélisé comme une différence entre les *time slices* d'un même perdurant, chaque *slice* correspondant à un intervalle durant lequel les traits pertinents restent constants. Cette stratégie déplace la temporalisation du niveau des propriétés vers celui des individus et de leurs relations, facilitant ainsi la représentation du changement conceptuel, le raisonnement temporel (par exemple à l'aide des relations d'Allen) et la reconstruction de chronologies, à condition que l'ontologie distingue explicitement les assertions portant sur l'entité 4D dans son ensemble de celles qui concernent une phase temporelle spécifique.
Nous adoptons cette seconde approche, qui permet de maintenir l'identité du concept de mariage tout en rendant compte de ses transformations historiques.

3.1 Implémentation en OWL[3] : structuration en slices temporelles

Dans cette perspective, le concept de mariage hébraïque est modélisé comme un perdurant structuré en plusieurs phases temporelles distinctes, représentées par les classes `SLICE_1`, `SLICE_2` et `SLICE_3`, définies comme des sous-classes de `HEBREW_MARRIAGE` (Figure 1(b), lignes 02–03). La Figure 1(a) illustre cette structuration à partir de la phase biblique, correspondant à la période comprise entre 500 et 100 av. J.-C. et représentée par la classe `SLICE_1`. Chaque slice est associé à un intervalle temporel spécifique et se caractérise par un ensemble de traits considérés comme invariants à l'intérieur de cet intervalle.
D'un point de vue formel, les traits conceptuels identifiés dans l'analyse précédente sont représentés sous la forme de restrictions existentielles portant sur des propriétés spécifiques,

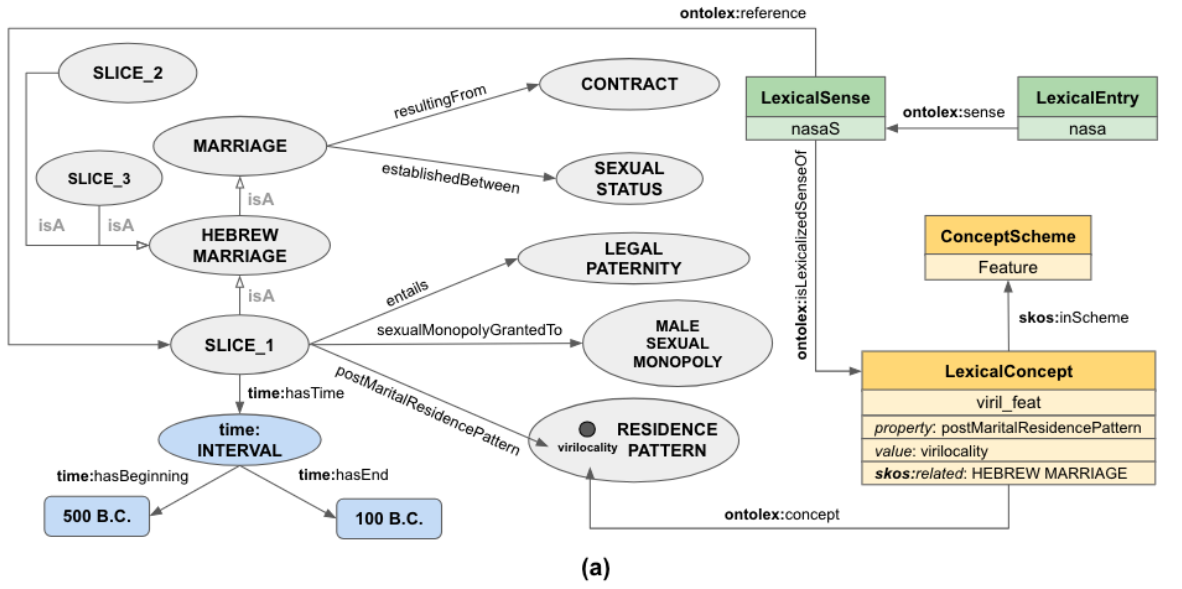
[3] L'ontologie proposée est encore en cours de développement et n'est pas, à ce stade, accessible publiquement. Son intégration dans l'écosystème LLOD reste également à venir ; cette contribution présente donc les principes de modélisation adoptés en vue d'une ressource FAIR, interopérable et réutilisable.

de sorte que la définition de chaque *slice* résulte de la combinaison de contraintes nécessaires et suffisantes. Dans ce cadre, la classe `SLICE_1` est définie par trois traits conceptuels principaux — la /paternité légale/, le /monopole sexuel masculin/ et la /virilocalité/ — représentés en Figure 1(a) et formalisés en OWL en Figure 1(b), lignes 06–08.
Il convient enfin de noter que le concept plus général de `MARRIAGE` est lui-même défini par des traits considérés comme invariants dans le temps, à savoir l'existence d'un contrat (ligne 04) et l'établissement d'une relation entre les sexes (ligne 05). En tant que sous-classes de `HEBREW_MARRIAGE`, les classes `SLICE_1`, `SLICE_2` et `SLICE_3` héritent automatiquement de ces traits invariants, tout en se différenciant par leurs propriétés temporellement spécifiques.

3.2 Dimension linguistique : variation dénominative et modélisation OntoLex-Lemon

En ce qui concerne la dimension linguistique, nous avons montré que les termes désignant le mariage varient à la fois sur l'axe diachronique et sur le plan synchronique. Cette variation synchronique se manifeste, au sein d'une même phase historique, par la coexistence de plusieurs dénominations coréférentielles, chacune mettant en saillance un trait définitoire spécifique du concept. Ainsi, pour la phase biblique, des termes tels que *laqaḥ*, *baʿal* ou *nasaʾ* renvoient tous au <MARIAGE>, mais lexicalisent respectivement des traits distincts, tels que l'/acquisition de la femme/, la /possession de la femme/ ou la /virilocalité/.
Le schéma de représentation proposé repose donc sur deux principes fondamentaux : (i) un terme dénote le concept de mariage tel qu'il est conceptualisé au cours d'une phase temporelle déterminée ; (ii) un terme peut lexicaliser un trait définitoire spécifique du concept qu'il dénote. Les unités lexicales sont représentées conformément au modèle OntoLex-Lemon[4].
En référence à la Figure 1(a), le point (i) est formalisé au moyen de la propriété `ontolex:reference`, qui relie le sens lexical de *nasa'* au concept de mariage hébraïque tel qu'il est défini pour la phase biblique (`SLICE_1`, 500-100 av. J.-C.). Le point (ii) est modélisé à l'aide de la propriété `ontolex:isLexicalizedSenseOf`, qui indique que *nasa'* lexicalise le trait de la /virilocalité/ du mariage hébraïque.
En raison des contraintes de cette propriété, le trait lexicalisé est représenté comme un `ontolex:LexicalConcept`. À cette fin, un schéma de traits est défini sous la forme d'un `skos:ConceptScheme`, au sein duquel chaque trait est modélisé comme un concept SKOS, également typé `ontolex:LexicalConcept` et relié au concept correspondant par la propriété `ontolex:isConceptOf`. Chaque trait encode la propriété concernée et la valeur associée, ainsi que, le cas échéant, des informations supplémentaires.

[4] https://www.w3.org/2016/05/ontolex/



(a)

```
01 HEBREW_MARRIAGE ⊑ MARRIAGE
02 SLICE_1 ⊑ HEBREW_MARRIAGE
03 SLICE_2 ⊑ HEBREW_MARRIAGE

04 MARRIAGE ⊑ ∃resultingFrom.CONTRACT
05 MARRIAGE ⊑ ∃establishedBetween.SEXUAL_STATUS

06 SLICE_1 ⊑ ∃entails.PATERNITY
07 SLICE_1 ⊑ ∃sexualMonopolyGrantedTo.MALE_SEXUAL_MONOPOLY
08 SLICE_1 ⊑ ∃postMaritalResidencePattern.virilocality

09 LexicalEntry(nasa)
10 LexicalSense(nasaS)
11 LexicalConcept(viril_feat)
12 ConceptScheme(Feature)
13 partOfSpeech(nasa,verb)
14
15 sense(nasa,nasaS)
16 reference(nasaS, SLICE_1)
17 isLexicalizedSenseOf(nasaS, viril_feat)
18 inScheme(viril_feat, Feature)
```

(b)

Figure 1. *Modélisation perdurantiste du concept de mariage hébraïque : (a) représentation conceptuelle et lexicale ; (b) formalisation en logique descriptive (DL).*

Ainsi, comme l'illustre la Figure 1(a), le terme *nasa'* lexicalise le trait de la /virilocalité/, représenté par la propriété `postMaritalResidencePattern` avec la valeur `virilocality`, et rattaché à une définition historiquement située du concept de mariage hébraïque. Ce mécanisme fondé sur des *features* permet de représenter la variation synchronique comme la coexistence de dénominations orientées vers des traits distincts d'un même concept, tout en restant suffisamment général pour distinguer des cas où une même valeur est associée à des propriétés différentes et pour permettre l'enrichissement ultérieur des traits.

4. Conclusion : une première évaluation du modèle

En conclusion, cette contribution a proposé un cadre de modélisation encore en cours de développement, visant à représenter de manière formelle et interrogeable la variation diachronique et dénominative du concept hébraïque de <MARIAGE>. Le Tableau 2 présente un ensemble de requêtes illustratives auxquelles le modèle est en mesure de répondre à partir de la formalisation décrite ci-dessus.

Questions de compétence	Pseudo-SPARQL
Quelles sont les propriétés du mariage hébraïque en 200 av. J.-C. ?	**SELECT** ?property ?value **WHERE** { ?mar a :HEBREW_MARRIAGE ; time:hasTime [time:hasBeginning ?start ; time:hasEnd ?end] ; ?property ?value } **FILTER** (xsd:integer(?start) <= 200 && xsd:integer(?end) >= 200)
Comment le concept de mariage évolue-t-il au fil du temps ?	**SELECT** ?property ?value **WHERE** { ?mar a :HEBREW_MARRIAGE ; time:hasTime ?interval ; ?property ?value . ?interval time:hasBeginning ?start ; time:hasEnd ?end . } **GROUP BY** ?interval
Quels termes désignent le mariage hébraïque durant la période autour de 200 av. J.-C. ?	**SELECT** ?term **WHERE** { ?mar a :HEBREW_MARRIAGE ; time:hasTime [time:hasBeginning ?start ; time:hasEnd ?end] ; ontolex:isReferenceOf ?sense . ?le ontolex:sense ?sense ; rdfs:label ?term . } **FILTER** (xsd:integer(?start) <= 200 && xsd:integer(?end) >= 200)
Quel trait conceptuel est lexicalisé par le terme *nasa*'?	**SELECT** ?property ?value ?concept **WHERE** { ?sense ontolex:senseOf [rdfs:label "nasa"@heb] ; ontolex:isLexicalizedSenseOf ?feature . ?feature :property ?property ; :value ?value ; skos:related ?concept . }
Quels termes lexicalisent le trait de virilocalité (et quand) ?	**SELECT DISTINCT** ?term ?start ?end **WHERE {** ?sense ontolex:isLexicalizedSenseOf ?trait ; ontolex:reference [time:hasBeginning ?start ; time:hasEnd ?end] . ?trait :property :postMaritalResidencePattern ; :value :virilocality . ?entry ontolex:sense ?sense ; rdfs:label ?term . }

Quels traits conceptuels sont stables par rapport au temps?	**SELECT** ?propertyTrait
	WHERE {
	{
	SELECT (COUNT(**DISTINCT** ?slice) AS ?nSlices)
	WHERE {
	?slice a owl:Class ;
	rdfs:subClassOf :HEBREW_MARRIAGE .
	}
	}
	?slice a owl:Class ;
	rdfs:subClassOf :HEBREW_MARRIAGE ;
	?propertyTrait ?propertyValue .
	
	FILTER(?propertyTrait **NOT IN** (rdf:type, rdfs:subClassOf))
	}
	GROUP BY ?propertyTrait ?nSlices
	HAVING (COUNT(**DISTINCT** ?slice) = ?nSlices)

Tableau 2 . *Exemples de requêtes illustrant la puissance expressive du modèle proposé et les types de questions sémantiques et diachroniques qu'il permet d'aborder.*

À titre d'exemple, nous présentons les réponses aux dernières requêtes (Figures 2 et 3) qui constituent une première forme de validation du modèle, en montrant sa capacité à explorer les évolutions conceptuelles et les traits lexicalisés à différentes phases temporelles.
Bien que ce modèle ait été élaboré à partir de ce cas d'étude, il peut être appliqué à d'autres domaines terminologiques caractérisés par des phénomènes de variation diachronique et synchronique, contribuant ainsi à la modélisation formelle de la variation en terminologie.

Quels termes lexicalisent le trait de virilocalité (et quand) ?

```
SELECT DISTINCT ?term ?start ?end
WHERE {
 ?sense ontolex:isLexicalizedSenseOf ?trait ;
       ontolex:reference [ time:hasBeginning ?start ;
                           time:hasEnd ?end ] .
 ?trait :property :postMaritalResidencePattern ;
        :value :virilocality .
 ?entry ontolex:sense ?sense ;
       rdfs:label ?term .
}
```

term	start	end
נשא (nasa')	500 av. J.-C.	1565 apr. J.-C.
הושיב (hoshiv)	500 av. J.-C.	100 av. J.-C.
שלח (shillaḥ)	500 av. J.-C.	100 av. J.-C.
נסב (nesav)	150 apr. J.-C.	500 apr. J.-C.
כנס (kanas)	150 apr. J.-C.	500 apr. J.-C.

Figure 2. *Exemple de requête en langage naturel et en SPARQL, avec réponse correspondante, portant sur les termes qui lexicalisent le trait de /virilocalité/ et sur leur ancrage temporel.*

Quels traits conceptuels sont stables par rapport au temps?

```
SELECT ?propertyTrait
WHERE {
 {
  SELECT (COUNT(DISTINCT ?slice) AS ?nSlices)
  WHERE {
   ?slice a owl:Class ;
        rdfs:subClassOf :HEBREW_MARRIAGE .
  }
 }
 ?slice a owl:Class ;
      rdfs:subClassOf :HEBREW_MARRIAGE ;
      ?propertyTrait ?propertyValue .

 FILTER(?propertyTrait NOT IN (rdf:type, rdfs:subClassOf))
}
GROUP BY ?propertyTrait ?nSlices
HAVING (COUNT(DISTINCT ?slice) = ?nSlices)
```

propertyTrait
acquisition of the woman
possession of the woman
male sexual monopoly
legal paternity
virilocality

Figure 3. *Exemple de requête en langage naturel et en SPARQL, avec réponse correspondante, portant sur les traits conceptuels stables dans le temps.*

References

[1] Aymerich, J. F., Fernández Silva, S. & Cabré Castellví, M. T. (2008). *La multiplicité des chemins dénominatifs*. Meta, 53(4), 731–747. https://doi.org/10.7202/019644ar.

[2] Candel, D., & Gaudin, F. (2006). *Aspects diachroniques du vocabulaire*. Presses Universitaires de Rouen et du Havre.

[3] Chiarcos, C., Cimiano, P., Declerck, T., & McCrae, J. P. (2013). Linguistic linked open data (LLOD): Introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL): Representing and linking lexicons, terminologies and other language data* (pp. ii–xi). Association for Computational Linguistics. https://aclanthology.org/W13-5501.

[4] Colorni, V. (1951). Spunti giuridici nel libro di Ruth. La Rassegna Mensile di Israel, 17(5), 195–204.

[5] Condamines, A. (2010). Variations in terminology: Application to the management of risks related to language use in the workplace. *Terminology, 16*(1), 30–50. https://doi.org/10.1075/term.16.1.02con.

[6] Di Segni, R. (1989). Il padre assente. La trasmissione matrilineare dell'appartenenza all'ebraismo. *Quaderni storici, 24*(70), 143–204.

[7] Desmet, I. (2007). Terminologie, culture et société : Éléments pour une théorie variationniste de la terminologie et des langues de spécialité. *Cahiers du Rifal*, 3–13.

[8] Drouin, P., Francoeur, A., Humbley, J., & Picton, A. (Eds.). (2017). *Multiple perspectives on terminological variation*. John Benjamins. https://hal.science/hal-01741998.

[9] Dury, P. (2013). Que montre l'étude de la variation d'une terminologie dans le temps : Quelques van valinpistes de réflexion appliquées au domaine médical. *Debate Terminológico, 9*, 2–10.

[10] Fernández-Silva, S., Freixa, J., & Cabré, M. T. (2012). A cognitive approach to synonymy in translation. In M. Brdar, I. Raffaelli, & M. Žic Fuchs (Eds.), *Cognitive linguistics between universality and variation* (pp. 189–211). Cambridge Scholars Publishing.

[11] Fernández-Silva, S., & Kerremans, K. (2011). Terminological variation in source texts and translations: A pilot study. *Meta, 56*(2), 318–335.

[12] Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. *Terminology, 12*(1), 51–77. https://doi.org/10.1075/term.12.1.04fre.

[13] Hawley, K. (2010). Temporal parts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2010 edition). https://plato.stanford.edu/archives/win2010/entries/temporal-parts/.

[14] Hezser, C. (2025). Women, children, and slaves as dependants of the householder in rabbinic and Roman law. In M. Schermaier, J. Winnebeck, & M. Becher (Eds.), *Norms of*

dependency in late antique and early medieval societies: Contextualising Roman legal terminology (pp. 9–26). Berlin / Boston, De Gruyter. https://doi.org/10.1515/9783111661438-002

[15] Kristiansen, M. (2014). Concept change, term dynamics and culture-boundness in economic-administrative domains. In R. Temmerman & M. van Campenhoudt (Eds.), *Dynamics and terminology: An interdisciplinary perspective on monolingual and multilingual culture-bound communication* (pp. 235–256). John Benjamins Publishing Company.

[16] Leach, E. R. (1955). Polyandry, inheritance and the definition of marriage. *Man, 55*, 182–186. London: Royal Anthropological Institute.

[17] Lemos, T. M. (2015). Were Israelite women chattel? Shedding new light on an old question. In S. Niditch (Ed.), *Worship, women and war: Essays in honor of Susan Niditch* (pp. 227–241). Providence, RI: Brown University Press.

[18] McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In *Proceedings of the Electronic Lexicography of the 21st Century (eLex 2017)* (pp. 19–21).

[19] Needham, R. (1975). Polythetic classification: Convergence and consequences. *Man, 10*(3), 349–369. London: Royal Anthropological Institute.

[20] Piccini, S., Bellandi, A., & Abrate, M. (2020). Diaterm : Un modèle pour représenter l'évolution diachronique des terminologies dans le web sémantique. In C. Roche (Ed.), *Terminologie & ontologie : Théories et applications: Actes de la conférence TOTh 2019 (Le Bourget du Lac, 6–7 juin)* (pp. 55–68). Éditions de l'Université Savoie Mont Blanc.

[21] Piccini, S., Bellandi, A., & Giovannetti, E. (2021). A model for representing diachronic terminologies: The Saussure case study. *Digital Humanities Quarterly, 15*(2).

[22] Piccini S., Saponaro D., Vilela Ruiz G. E., 2024. *La connotation logique en terminologie. Le mariage dans l'Israël Antique comme Étude de Cas.* Cahiers de lexicologie, 2024(2), 125.

[23] Picton, A. (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* (Doctoral dissertation, Université Toulouse 2).

[24] Picton, A. (2018). Terminologie outillée et diachronie : Éléments de réflexion autour d'une réconciliation. *ASp, 74*. https://doi.org/10.4000/asp.5255.

[25] Satlow, M. L. (2001). Jewish marriage in antiquity. Princeton, NJ: Princeton University Press.

[26] Shectman, S. (2014). « What do we know about marriage in ancient Israel? ». Dans M. L. Chaney, U. Y. Kim et A. Schellenberg (dir.), *Reading a tendentious Bible: Essays in honor of Robert B. Coote* (p. 166-175). Sheffield, Sheffield Phoenix Press.

[27] Sider, T. (2008). Temporal parts. In T. Sider, J. Hawthorne, & D. Zimmerman (Eds.), *Contemporary debates in metaphysics* (pp. 241–262). Blackwell.

[28] Steinsaltz, A. (2004). Cos'è il Talmud. Firenze, Giuntina

[29] Stone, K. (2014). « Marriage and sexual relations in the world of the Hebrew Bible ». Dans A. Thatcher (dir.), *The Oxford handbook of theology, sexuality, and gender* (p. 173-188). Oxford, Oxford University Press.

[30] Tosato, A. (1981). *Il matrimonio israelitico: una teoria generale*. Rome, Pontificio Istituto Biblico.

[31] Welty, C., Fikes, R., & Makarios, S. (2006). A reusable ontology for fluents in OWL. In In*Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS 2006)* (Vol. 150, pp. 226–236).

[32] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, Article 160018. https://doi.org/10.1038/sdata.2016.18.

[33] Zanola, M. T. (2014). *Arts et métiers au XVIIIe siècle: Essais de terminologie diachronique*. L'Harmattan.

Dictionnaires et sources lexicographiques

Gesenius, H. F. W. (1953). *A Hebrew and English lexicon of the Old Testament: With an appendix containing the biblical Aramaic*. Oxford, Clarendon Press.

Jastrow, M. (1903). *A dictionary of the Targumim, the Talmud Babli and Yerushalmi, and the Midrashic literature*. New York / London, Luzac & Co. / G. P. Putnam's Sons.

Klein, E. (1987). *A comprehensive etymological dictionary of the Hebrew language for readers of English*. Haïfa, Carta Jerusalem.

Web References

World Wide Web Consortium (W3C). (2016). *OntoLex-Lemon: The ontology-lexicon model*. https://www.w3.org/2016/05/ontolex/

La terminologie de la mode en diachronie : le projet FLATIF et la base de données ModTERM

Maria Teresa Zanola, Università Cattolica del Sacro Cuore, mariateresa.zanola@unicatt.it
Silvia Calvi, Università Cattolica del Sacro Cuore, silvia.calvi1@unicatt.it
Klara Dankova, Università Cattolica del Sacro Cuore, klara.dankova@unicatt.it

1. Introduction

Cette contribution s'inscrit dans le cadre du projet PRIN 2020 FLATIF – *Fashion Languages and Terminologies across Italian and French : Building and Disseminating FLATIF Resources*, lancé en 2022 et coordonné par l'Università Cattolica del Sacro Cuore (P.I. : Maria Teresa Zanola), en collaboration avec d'autres universités italiennes, notamment l'Università degli Studi di Napoli « Parthenope » (responsable Claudio Grimaldi), l'Università di Roma Tre (responsable Paolo D'Achille) et l'Università degli Studi di Verona (responsable Paolo Frassi). Le projet FLATIF se propose d'étudier la langue et la terminologie de la mode dans une perspective contrastive (IT-FR) et diachronique (1880-1980) dans des sources différentes, notamment des sources linguistiques, encyclopédiques, lexicographiques, historiques et visuelles.
Pour la langue française, le projet FLATIF a abouti à la conception et constitution de plusieurs ressources, notamment :

- Le corpus FLATIF : un corpus de 695 numéros de revues de mode publiées entre 1900 et 1946 (environ 7 millions de *tokens*) ;
- Le portail ModArchive[1] : un portail qui donne accès à des approfondissements thématiques sur l'histoire des termes de la mode ;
- La base de données ModTERM[2] : une base de données permettant une exploration d'un échantillon du corpus FLATIF et des concepts liés à la mode dans une perspective diachronique (1880-1980) et plurilingue (FR-IT-EN-AR).
- Le portail de recherche d'images à l'aide de l'IA, ModSCAPE[3].

Cette contribution vise à présenter la base de données terminologique de la plateforme ModTERM. Nous montrerons comment ModTERM a été conçu pour répondre aux exigences spécifiques de la représentation de la terminologie de la mode, en ce qui concerne l'évolution des concepts et de leurs dénominations, ainsi que le rôle central de l'image dans la construction et la transmission des savoirs du domaine. Après un aperçu des études sur la terminologie de la mode et de sa variation diachronique et diaphasique, cette communication présentera la structure de la base ModTERM. Grâce à l'illustration de certaines études de cas, nous mettrons en évidence les avantages liés à la constitution d'une telle ressource.

2. Pour une étude diachronique de la terminologie de la mode : prémisses théoriques et méthodologiques

[1] https://modarchive.unicatt.it/?lang=fr [30.04.2026].
[2] https://modarchive-portal.unicatt.it [30.04. 2026].
[3] https://modscape.it [30.04.2026].

L'étude de la terminologie de la mode dans une perspective diachronique s'inscrit dans le cadre plus large des recherches en terminologie diachronique développées à partir des années 2010, au sein du projet TERM-DIACHRO[4], lancé par l'Osservatorio di Terminologie e Politiche Linguistiche (OTPL) de l'Università Cattolica del Sacro Cuore de Milan et qui a posé les bases méthodologiques du projet FLATIF. Relativement à la conception de ressources terminologiques considérant les données en fonction de leur évolution dans le temps, nous nous inspirons des études de Roche *et al*. 2014, Frassi 2021, Piccini *et al*. 2021, Bellandi *et al.* 2025. Quant à l'analyse de la terminologie de la mode en diachronie, cette contribution s'appuie sur les prémisses de Zanola (2025 : 20-21). Tout d'abord, le terme et sa définition doivent être envisagés en lien avec une périodisation donnée. Ensuite, l'étude de la mode demande la consultation de sources hétérogènes – manuels, traités, revues, catalogues, encyclopédies, supports audiovisuels, entre autres – : chacune contribue à la collecte de données utiles à la représentation de ce riche patrimoine culturel (Barthes 1967 ; Bouverot 1999). Une attention spéciale doit être accordée à la dimension plurilingue de cette terminologie : la recherche des équivalents dans d'autres langues nécessite une étude approfondie des sources de référence dans la langue concernée. De plus, la terminologie de la mode se caractérise également par une variation diaphasique : de la production à la distribution, les termes employés peuvent changer en fonction des destinataires envisagés (Zanola 2016, 2018, 2020). Enfin, Zanola (2025) observe que dans ce domaine de spécialité les définitions lexicographiques ne sont pas suffisantes à une compréhension adéquate des concepts : elles demandent des approfondissements de nature encyclopédique et des représentations visuelles.

3. La base de données ModTERM

À partir de ce cadre théorique et méthodologique, la base de données ModTERM a été développée afin de proposer une représentation efficace de la terminologie de la mode. ModTERM repose sur la constitution d'une structure conceptuelle visant à modéliser le domaine de la mode dans sa complexité et sa diversité. Les objets de la mode sont classés en sous-domaines, tels que les vêtements, les accessoires, les matières, les techniques, les procédés de fabrication.

La prise en compte de la dimension diachronique constitue un élément central de la conception de ModTERM : chaque concept est associé à une ou plusieurs périodes de référence, ce qui permet d'observer les cas d'évolution conceptuelle.

Les données associées à chaque concept sont organisées en trois sections : 1) la dimension conceptuelle ; 2) la dimension linguistique ; 3) des approfondissements de nature encyclopédique.

La dimension conceptuelle comprend l'indication d'un ou de plusieurs sous-domaines auxquels le concept appartient, une définition terminologique, la datation du concept, l'identification de ses caractéristiques essentielles, ainsi que l'association d'une ou de plusieurs images représentatives du concept, tirées du corpus FLATIF.

[4] Pour plus d'approfondissements, nous renvoyons à https://centridiricerca.unicatt.it/otpl-progetti-term-diachro [30.04.2026].

La deuxième section est consacrée à la dimension linguistique dans une perspective plurilingue. Les termes désignant un même concept sont recensés en français, italien, anglais et arabe. Pour chaque équivalent, ModTERM propose une datation et, lorsque les sources le permettent, des contextes d'emploi authentiques ; ces contextes sont extraits du corpus FLATIF.
Enfin, ModTERM inclut une troisième section regroupant des informations de nature encyclopédique : des notes encyclopédiques ou linguistiques complètent ces données, afin d'approfondir la connaissance des concepts et des unités linguistiques qui les désignent. Des références bibliographiques, de la définition, des images et des notes, sont également fournies. Cette structure permet de représenter l'évolution de la terminologie de la mode, tout en explorant ses caractéristiques spécifiques. Les études de cas présentées dans les sections suivantes illustreront les potentialités de ModTERM, en mettant en évidence les apports d'une modélisation terminologique attentive à la dimension diachronique et plurilingue.

3.1 Le cas du *chapeau melon*

Nous proposons une première étude de cas consacrée au concept <chapeau melon>.
En ce qui concerne la dimension conceptuelle, les informations recueillies sont présentées dans le tableau 1.

Dimension conceptuelle	
Sous-domaines	Accessoire – Chapeau
Définition terminologique	Chapeau en feutre rigide, à bord étroit roulé, avec calotte en forme de dôme
Datation du concept	1849 – présent
Caractéristiques essentielles du concept	Chapeau – feutre – bord étroit – calotte bombée

Tableau 1 – ModTERM : analyse de la dimension conceptuelle du concept <chapeau melon>

Dans ModTERM, le concept est illustré par quatre images datées respectivement des années 1931, 1932, 1933 et 1936. Ces images apportent des informations complémentaires par rapport aux sources textuelles : dans notre exemple, elles permettent d'observer que le chapeau melon est porté par les hommes aussi bien que par les femmes, tout en soulignant l'évolution des pratiques vestimentaires associées à cet accessoire.
Pour la dimension linguistique, nous avons recueilli trois termes en français désignant le concept <chapeau melon> : *chapeau melon* et sa variante *melon*, *chapeau rond*, *chapeau boule* et ses variantes *chapeau-boule* et *boule*. Pour chaque terme, nous avons indiqué la datation et le contexte d'emploi tiré du corpus FLATIF (Tableau 2)

Terme	Datation	Contexte d'emploi
chapeau melon melon	1860 – présent	« Chapeau melon, oui, chapeau melon, voilà ce qu'il vous faut, de ces chapeaux melons durs comme un casque et que l'avers ni le brouillard ne peuvent imbiber » (Adam, 1932, 15 février, p. 21).

chapeau rond	1860 – présent	« Queuille, Chappe, etc., également en chapeau de soie. M. E. de Rotschild, comme M. Moulines en chapeau rond, ont dû être aussi surpris que moi » (Adam, 1927, 15 juillet, p. 13).
chapeau boule *chapeau-boule* *boule*	1860 – présent	« Il y a maintes idées à glaner. Ils firent bien, aussi, ces chapeliers qui veulent nous délivrer du chapeau-boule, ces Delion, ces Berteil, ces Léon, et ces bottiers Greco et Helistern […] » (Les Modes, juillet 1927, p. 1).

Tableau 2 – ModTERM : analyse de la dimension linguistique du concept <chapeau melon>

Nous avons également retenu les équivalents en anglais – *bowler hat* (1849 – présent), *coke hat* (1849 – présent), *derby* (années 1860 – présent) –, en italien – *cappello duro* (début XX^{e} siècle- présent), *bombetta* (1908 – présent) –, et en arabe – *bawlar* et *ursūsa*.
L'analyse des équivalents dans plusieurs langues met en évidence un fort ancrage spatial et temporel de la terminologie de la mode, ce qui rend souvent sa traduction problématique. Ainsi les termes de la mode doivent-ils être contextualisés : ModTERM peut constituer un outil de travail pour les traducteurs de ce secteur, en offrant un accès à des informations encyclopédiques qui permettent d'éclairer les relations entre les termes au-delà d'une simple équivalence lexicale.
Après avoir indiqué les sources, la section des notes encyclopédiques raconte l'histoire du chapeau melon inventé en 1849 par deux chapeliers londoniens, Thomas et William Bowler : ils répondirent à la commande d'un aristocrate britannique, Edward Coke, qui avait demandé un couvre-chef pratique et résistant pour ses gardes-chasse.

3.2 L'évolution conceptuelle : les cas de la <tenue pour la plage> et de la <jupe courte>

D'autres concepts répertoriés dans ModTERM illustrent l'apport d'une description terminologique enrichie par l'image. C'est le cas de la <tenue pour la plage> (Salvatore 2025), dont l'évolution conceptuelle témoigne des transformations profondes des pratiques sociales et des représentations du corps. À partir de 1830 et jusqu'au début du XX^{e} siècle, le concept renvoyait à une tenue destinée à la baignade [Concept_1] ; ce n'est qu'à partir du début du XX^{e} siècle, que la tenue pour la plage évolue vers la notion de vêtement destiné à la baignade avec une vocation d'habiller les parties intimes du corps [Concept_2], conception qui demeure globalement valable jusqu'à aujourd'hui. Les images associées dans ModTERM permettent de retracer cette évolution des formes et des usages.
Le concept de <jupe courte> constitue une autre étude de cas de fort intérêt : au début du XX^{e} siècle, les jupes dites *courtes* laissaient apparaître les chevilles. Tout au long du XX^{e} siècle, la longueur des jupes se réduit progressivement jusqu'à l'apparition de la minijupe dans les années 1960. ModTERM recueille les termes *jupe*, *jupe courte* et *mini-jupe* et leur origine : porter des jupes courtes permettait de se déplacer librement, surtout dans le domaine sportif où des figures comme la joueuse de tennis Suzanne Lenglen ont écrit l'histoire de la mode dans le sport en portant des jupes courtes ou très courtes.
Ces études de cas mettent en évidence la pertinence méthodologique de ModTERM pour l'analyse de la variation terminologique en diachronie, en montrant comment l'articulation

entre structuration conceptuelle, données linguistiques et ressources iconographiques permet de modéliser de manière rigoureuse l'évolution de la mode.

4. Conclusions

Cette communication montrera comment la base de données ModTERM constitue un outil pertinent pour l'étude diachronique de la terminologie de la mode. En intégrant structuration conceptuelle, description linguistique plurilingue et répertoire iconographique, ModTERM permet de modéliser l'évolution des concepts de la mode et de leurs dénominations. Les études de cas présentées mettent en évidence les avantages méthodologiques d'une approche attentive à la dimension temporelle ainsi qu'à l'ancrage culturel des termes.
ModTERM ne représente pas seulement une ressource pour les terminologues et les traducteurs spécialisés, mais aussi un instrument utile pour les chercheurs en sciences du langage, les historiens de la mode et tous ceux qui s'intéressent à ce domaine de la connaissance.

Références bibliographiques sélectionnées

Barthes, R. 1967. *Système de la mode*. Paris : Seuil.

Bellandi A., Calvi S., Dankova K., Piccini S., 2025. Représentation computationnelle des données terminologiques en diachronie : le cas des fibres artificielles. In C. Roche *et al*. (dir.), *Actes de la conférence TOTh 2024. Terminologie & Ontologie : Théories et Applications*. Chambéry : Presses Universitaires Savoie Mont Blanc, pp. 95-119.

Bouverot, D. 1999. Le vocabulaire de la mode. In Antoine, G., Martin, R. 1999. (dir.), *Histoire de la langue française 1880-1914*. Paris : CNRS, pp. 193-206.

Frassi P., 2021. « DIACOM-fr, une base de données terminologiques de type diachronique». C*ahiers de Lexicologie*, n° 118, pp. 23-50.

Piccini S., Abrate M., Bellandi A., Giovannetti E., 2021. *Rappresentazione, costruzione e visualizzazione di risorse terminologiche diacroniche nell'era del web semantico*. In C. Grimaldi, M. T. Zanola (dir.), *Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari*. Firenze : Firenze University Press, pp. 125-139.

Roche C., Damas L., Roche J., 2014. Multilingual Thesaurus: The Ontoterminology Approach, in CIDOC 2014 - Access and Understanding – Networking in the Digital Era. In CIDOC (Comité International pour la DOCumentation), Dresden, pp. 1-14 : https://hal.archives-ouvertes.fr/hal-01272725/document (10.01.2026).

Salvatore M. C. 2025. Il maillot de bain, una storia terminologica. In Zanola, M. T., D'Achille, P. (a cura di), *La moda francese e italiana (1880-1980). Fonti, strumenti e metodi*. Firenze: Franco Cesati Editore, pp. 139-156.

Zanola, M. T. 2016. L'espace du concept, la parole de l'image : pour une typologie des représentations non-verbales dans la terminologie des tissus. In Lervad, S. *et al*. (dir.), *Verbal*

and non-verbal representation in terminology: Proceedings of the TOTh 2013. Copenhague : DNRF's Centre for Textile Research, Institut Porphyre, Savoir et connaissances, pp. 65-80.

Zanola, M. T. 2018. « La terminologie des arts et métiers entre production et commercialisation : une approche diachronique ». *Terminàlia*, nº 17, pp. 6-23.

Zanola, M. T. 2020. « Francese e italiano, lingue della moda: scambi linguistici e viaggi di parole nel XX secolo ». *Lingue Culture Mediazioni*, nº 7/2, pp. 9-26.

Zanola, M. T. 2025. Descrivere la moda italiana e francese, fra lingua, cultura e terminologia. Il caso della marinière. In D'Achille, P., Zanola, M. T. (dir.), *La moda francese e italiana (1880-1980). Fonti, strumenti e metodi*. Firenze: Franco Cesati Editore, pp. 15-47.

Session 4

Du terme au geste : la place de la terminologie dans les pratiques informationnelles et martiales d'un club de taekwondo

Marcin Trzmielewski

Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales
Université de Montpellier Paul-Valéry
Montpellier Taekwondo
marcin.trzmielewski@umpv.fr

1. Introduction

La présente communication s'inscrit dans le cadre du projet *Dojo-SIC*, qui vise à étudier les arts martiaux selon une approche info-communicationnelle. Dans ce cadre, nous avons conduit une analyse de différents terrains de pratique — taekwondo, tai-chi, ju-jutsu et karaté — envisagés comme des dispositifs info-communicationnels, au sein desquels l'espace, les objets, les gestes et les corps participent conjointement à la production et à la médiation des savoirs martiaux (Stassin et al., 2026). Nous avons notamment montré que la configuration matérielle et symbolique des dojos (*dojang* pour le taekwondo) constitue un dispositif sémiotique et documentaire, dans lequel chaque élément — jusqu'au corps du pratiquant — peut être appréhendé comme un support d'information. Les savoirs corporels, pratiques et théoriques y circulent par imitation, démonstration et interaction, et se prolongent dans des supports externes, imprimés ou numériques.

Dans le prolongement de ces travaux, la présente contribution propose une analyse située des pratiques terminologiques au sein de la langue de spécialité mobilisée dans les pratiques informationnelles et martiales d'un club français de taekwondo. Cette étude est menée par un chercheur en sciences de l'information et de la communication, également pratiquant de cet art martial, ce qui permet de mobiliser une posture de chercheur-pratiquant (De Lavergne, 2007) favorisant un accès prolongé et réflexif au terrain. Après un état de l'art consacré à la question de la terminologie taekwondoïste, nous présentons notre cadre méthodologique, qui repose sur une analyse linguistique et thématique d'un corpus composé de comptes rendus d'observations des pratiques informationnelles et martiales, et de ressources informationnelles produites au sein du club. En situant cette exploration dans le contexte du « loisir sérieux » (Stebbins, 2009) et en appréhendant les ressources informationnelles comme des dispositifs info-communicationnels (Couzinet, 2009 ; Fabre et Gardiès, 2010), nous nous intéressons à la place de la terminologie dans les pratiques informationnelles et martiales de taekwondo, à ses caractéristiques, ainsi qu'aux savoirs qu'elle contribue à produire et à transmettre. La conclusion présente les principaux apports de l'étude, les discute à la lumière des cadres théoriques mobilisés, en souligne les limites et esquisse des perspectives de recherche destinées à prolonger cette première exploration.

2. État de l'art

Le taekwondo trouve ses origines dans des arts martiaux anciens, comme le taekkyon et le karaté, et s'est développé en Corée sous l'occupation japonaise. Son nom a été proposé en 1955 par le général Choi Hong Hi, l'un de ses principaux précurseurs (Moenig, 2015 ; Moenig et Kim, 2019). Aujourd'hui, deux fédérations coexistent : l'International Taekwon-Do Federation (ITF) et la World Taekwondo (WT) – également appelée Taekwondo Olympique – qui diffèrent par le style technique et les modalités de compétition.

La pratique du taekwondo requiert la maîtrise de nombreuses techniques corporelles, chacune désignée par une terminologie précise. Cette terminologie a une finalité essentiellement opérationnelle, permettant de codifier les pratiques de self-défense (*ho chin soul*), de démonstration et de compétition aux niveaux local, national et international, notamment dans le *poomsae* (enchaînement de gestes conventionnels individuels), le *kyorugi* (combat) et le *kyok pa* (casse d'objets). L'internationalisation du taekwondo crée des enjeux informationnels et communicationnels importants. Les préoccupations principales liées à la terminologie technique concernent sa cohérence, son uniformisation, sa simplicité (écriture et prononciation) et sa traduction dans les langues locales, afin de faciliter la diffusion, le partage et la médiation des savoirs liés à cette pratique martiale et leur opérationnalisation dans l'action (Kukkiwon, 2011 ; Taskinen, 2019 ; Yoo et al., 2016).

Pour répondre à ces enjeux de standardisation, la terminologie taekwondoïste a été codifiée dans les manuels et encyclopédies rédigés par Choi Hong Hi (1965, 1987), proposant une description systématique des techniques, accompagnée d'une terminologie stabilisée et souvent présentée comme liée à l'histoire, à la philosophie et à l'identité culturelle du taekwondo. La terminologie y est conçue comme un système clos et normatif, garant de l'authenticité et de la précision de la pratique.

Actuellement, la terminologie taekwondoïste fait l'objet d'une standardisation institutionnelle multiniveau, sous la forme de différents systèmes d'organisation des connaissances. À l'échelle internationale, le Kukkiwon, l'institution centrale regroupant les organes directeurs du taekwondo en Corée du Sud, constitue l'autorité principale en matière de normalisation. Il édicte les désignations officielles (en coréen, transcription en alphabet roman, traduction en anglais et description de la technique) et les méthodes de classification des techniques, des parties du corps, des grades et titres, à travers ses manuels et règlements destinés à l'enseignement et à la certification internationale (Kukkiwon, 2001, 2011). Cette codification est complétée par World Taekwondo (2026), qui fixe une terminologie réglementaire spécifique à la compétition et contribue à stabiliser un lexique opératoire commun à l'échelle internationale.

À l'échelle nationale, les fédérations, comme la Fédération Française de Taekwondo (2025), jouent un rôle de médiation terminologique. Elles traduisent, adaptent et pédagogisent les référentiels internationaux via des programmes de grades, des fiches techniques (portant par exemple sur les positions de base ainsi que sur les techniques de bras et de jambes) et des documents de formation. La terminologie y apparaît souvent sous une forme hybride, combinant désignations coréennes normées et équivalents ou paraphrases en français. Les lexiques de clubs (Kang-Ho Taekwondo, 2024 ; Taekwondo Neuville, 2025, par exemple) complètent cette littérature normative et en assurent la vulgarisation.

La question de la terminologie taekwondoïste a été modestement étudiée dans la littérature scientifique. Les travaux existants prolongent le projet de clarification et de structuration des termes, visant une conceptualisation partagée et des dénominations univoques pour faciliter la communication et l'opérationnalisation des activités de *poomsae* et de combat. Les chercheurs en philologie et en sciences du sport proposent différentes analyses – historiques, conceptuelles et biomécaniques – pour comprendre et unifier les pratiques de dénomination et reclasser les techniques de taekwondo (Moenig et Kim, 2019 ; Savilampi, 2018 ; Yoo et al., 2016). Le projet *Martial Art Ontology* met en évidence la nécessité de cartographier et d'aligner les termes relatifs aux arts martiaux coréens et leurs équivalents sémantiques dans d'autres langues (Hou et Kenderdine, 2024).

Cependant, ces cadres prescriptifs ne suffisent pas à rendre compte des usages effectifs. Les études ethnographiques sont rares. Celle de Taskinen (2019), par exemple, montre l'usage très variable du vocabulaire coréen dans les cours de taekwondo en Finlande, avec des combinaisons fréquentes de mots coréens et de traductions finlandaises, des reformulations et des abréviations. Certaines expressions, comme le salut « *tcha lyeut, kiong nie* », restent difficiles à traduire et s'emploient uniquement en coréen. L'étude met également en évidence les variations de prononciation entre enseignants et l'importance accordée au cri (*ki hap*), qui accompagne l'exécution de nombreuses techniques.

Si les travaux existants analysent la terminologie, la classification technique et l'apprentissage corporel, aucune recherche n'articule ces dimensions dans une approche de pratiques informationnelles situées. Nous mobilisons donc cette perspective pour étudier la place de la terminologie dans les pratiques d'information d'un club de taekwondo, ses caractéristiques et les savoirs qu'elle véhicule.

3. Cadre théorique et méthodologique

En mobilisant une approche orientée activité, nous considérons les activités langagières et informationnelles comme des pratiques situées, qui émergent du contexte (des circonstances matérielles et sociales) dans lequel elles prennent place (Suchman, 1987). Dans cette perspective, la terminologie accompagne les activités sociales des individus dans leurs interactions avec l'environnement, au sein de situations variées de production, de recherche, d'organisation, de traitement, d'usage, de partage et de communication de l'information (Chaudiron et Ihadjadene, 2010 ; Trzmielewski, 2023). La circulation des termes — entendus comme des signes linguistiques — au sein de la communauté du taekwondo est analysée à partir de leurs usages ainsi que des enjeux sociaux et info-communicationnels qui leur sont associés. Le sens des termes se construit en fonction des conditions de leur apparition dans différents genres discursifs et participe à la production de connaissances ainsi qu'à la médiation des savoirs (Delavigne et De Vecchi, 2016 ; Gaudin, 2005).

Si le taekwondo peut constituer une activité professionnelle — notamment pour les formateurs ou les compétiteurs de haut niveau —, dans le club étudié, la majorité des pratiquants et des enseignants sont engagés dans des pratiques bénévoles associatives et exercent cet art martial en dehors de leur temps de travail. Nous situons ainsi notre analyse des pratiques informationnelles dans le cadre de la vie quotidienne (Savolainen, 2009), comprise comme un ensemble d'expériences humaines à la fois ordinaires, divertissantes et profondes

(Kari et Hartel, 2007). Plus précisément, la pratique du taekwondo s'inscrit dans le registre du « loisir sérieux », défini comme une activité d'amateur, de passe-temps ou de bénévolat que les individus jugent suffisamment intéressante et gratifiante pour s'y engager durablement, investir des efforts et développer une véritable « carrière de loisir » fondée sur l'acquisition et la mise en œuvre continue de compétences, de savoirs et d'expériences spécifiques (Stebbins, 2009). Le loisir sérieux se caractérise notamment par un besoin constant d'amélioration des compétences et des connaissances, la présence d'un ethos, entendu comme « un monde social possédant ses propres normes, valeurs et principes moraux », par des bénéfices durables — tels que le développement de soi — et par la construction d'une identité individuelle et sociale à travers la pratique (Adjizian, 2023, p. 2).

Sur le plan méthodologique, notre étude repose sur une analyse linguistique et thématique d'un corpus de comptes rendus d'observations des pratiques informationnelles et martiales réalisés entre octobre et décembre 2025 dans un club français (11 399 mots), et de ressources informationnelles produites et diffusées au sein du club : deux fiches de révision, dix affiches et vingt-trois newsletters (12591 mots) diffusées par le club entre septembre et décembre 2025. Ces ressources sont appréhendées comme des dispositifs info-communicationnels (Couzinet, 2009 ; Fabre et Gardiès, 2010), entendus comme des ensembles qui « relient producteurs, médiateurs et récepteurs de l'information dans un contexte donné, à travers des formes matérielles participant à la production du sens et à la médiation des savoirs » (Stassin et al., 2026).

4. Résultats

La terminologie spécialisée est présente dans l'ensemble des ressources informationnelles du club, qu'elles reposent sur l'oralité (cours de taekwondo et échanges entre enseignants, interactions entre élèves), ou sur des supports documentaires physiques et numériques. Ces ressources comprennent notamment les cours de taekwondo, les fiches de révision, les affiches et les lettres de diffusion. L'ensemble de ces ressources humaines et documentaires contribue à la capitalisation, à la médiatisation, à la diffusion et au partage des savoirs taekwondoïstes au sein du club. Les supports documentaires sont le plus souvent produits par les enseignants, avec l'appui de bénévoles et d'une alternante en communication. Le club diffuse également des contenus sur d'autres espaces info-communicationnels, tels que son site web et ses réseaux socionumériques : Facebook, Instagram, TikTok et YouTube notamment.

La terminologie du taekwondo remplit avant tout une fonction informationnelle et pédagogique. Les enseignants du club l'utilisent pour transmettre aux élèves les techniques relevant des principales formes de travail pratiquées au club : *ki bonn don jack* (mouvements de base), *ki bonn jasé* (positions et enchaînements), *matcho keulouki* (combats conventionnels), dont *han boon keulouki* (combat rituel sur une seule attaque), *poomsae* (enchaînements codifiés), *kengui keulouki* (combat de compétition), *tchaillo keuloukii* (combat libre), *ho chin soul* (self-défense) et *kyeu kpa* (casse). Concernant les coups de pieds (*tchaki*), par exemple, les termes permettent de distinguer le niveau de rotation qui doit être appliquée : « *ap tchaki – coup de pieds de face ; linéaire* », « *bandal tchaki – coup de pied semi circulaire* », « *dolyop tchaki – coup de pied circulaire ; rotation du pied d'appui* » (fiche de révision n° 1).

Les pratiquants sont accompagnés par la terminologie dans l'ensemble des séquences de leur pratique, de l'entrée jusqu'à la sortie du *dojang*. A chaque entrée et sortie de la salle, lors du salut collectif face aux enseignants, avant et à après chaque exercice en binôme ou groupe, après l'appel d'un élève à effectuer une démonstration, ainsi qu'à la fin du cours (salut aux enseignants et à l'élève le plus gradé), le salut est effectué sur la commande « *tcha lyeut, kiong nie* ». Ce salut, présenté dans les fiches de révision n° 1 comme « *un signe de respect universel* », est verbalisé le plus souvent par les enseignants. Il constitue un véritable savoir gestuel, qui doit être appris et exécuté correctement : *tcha lyeut* (« *talons joints, pointes de pieds écartées* »), *kiong nie* (« *on s'incline en avant* »). Au début de l'année, lors de l'observation n° 2, les enseignants indiquent aux nouveaux pratiquants les règles de la discipline du club, notamment l'obligation de saluer en entrant et en sortant du *dojang*. Cette mise au point fait suite à l'intervention du maître principal du club, qui avait signalé que certains élèves ne respectaient pas cette règle. Lors de la même observation, après le salut final, une pratiquante (ceinture noire) sollicite les enseignants afin de vérifier la manière correcte d'exécuter la commande « *tcha lyeut soki* », étant la plus gradée lors de ce cours et donc responsable de la verbalisation et de l'exécution du geste devant l'ensemble du groupe.

Les enseignants rappellent fréquemment l'importance du *ki hap*, défini comme le fait d'« *unir l'énergie* » (fiche de révision 1). Deux types de *ki hap* sont distingués : celui qui sert « *d'avertissement pour prévenir le professeur ou le partenaire que l'on est prêt* », et celui « *qui signifie que l'on a mis le maximum de puissance* », utilisé aussi bien en attaque qu'en défense. De manière plus générale, le cri permet d'exprimer et d'évaluer la motivation des élèves ; son absence ou sa faiblesse peut pénaliser la notation lors des passages de grade.

Ainsi, les termes spécialisés ne fonctionnent pas uniquement comme des unités lexicales, mais comme de véritables opérateurs d'action. La terminologie du taekwondo relève d'un savoir incorporé, correspondant à des savoirs pratiques et à des savoir-être qui se manifestent principalement par l'action corporelle, et dont la maîtrise repose sur l'intégration progressive de gestes, de postures et de perceptions, plutôt que sur une connaissance purement théorique. Le savoir corporel est transmis principalement par les enseignants et par les élèves les plus gradés, régulièrement sollicités pour démontrer devant le groupe les techniques relevant du *ki bonn don jack* : *soki* (positions), *makki* (parades), *tchaki* (coups de pieds), *tchiki-jileuki* (techniques de poings), ainsi que d'autres formes de travail, telles que *poomsae*. Ces techniques sont également illustrées par des images figurant sur les fiches de révision et par des vidéos diffusées sur la chaîne YouTube du club. Par ailleurs, le savoir pratique lié aux *poomsae* apparaît comme une ressource distinctive du club, lui permettant de se démarquer d'autres clubs à l'échelle départementale et nationale, et de répond à des enjeux stratégiques liés à la compétition.

La terminologie véhicule également des savoirs théoriques. Les fiches de révision, diffusées par l'enseignant principal du club à la demande, permettent un apprentissage individuel à domicile. Elles comportent, sur chaque page, des éléments du vocabulaire du taekwondo en coréen ainsi que des indications phonétiques, guidant les apprenants dans une progression systématique des savoirs martiaux. Le lexique porte notamment sur les parties du corps, les formes de travail, les terminaisons, les directions, les niveaux, les commandements, les tenues, les nombres cardinaux et ordinaux et mêmes certaines règles de syntaxe. Le corpus relève une coexistence structurante entre les termes coréens normatifs et leurs équivalents ou

paraphrases françaises, souvent présentés sous forme de gloses, par exemple : « *batang sonn tok tchiki – frappe avec paume de la main* » (fiche de révision n° 2). Cette dualité est systématiquement exploitée dans une visée pédagogique. On observe ainsi un bilinguisme fonctionnel, dans lequel la forme coréenne conserve une valeur symbolique et identitaire, tandis que la forme française remplit une fonction cognitive. L'alternance codique (« *On recule en Duit Koubi – Momtong Makki puis demi-tour en Jou Tchoum Sôki et Pal Loup Duit Tchiki (Ki hap).* », fiche de révision n° 2) et la variation graphique (« *poom see* » / « *poomsae* ») ou phonétique (« *ale* / *are maki* »), liées notamment aux enjeux de romanisation des termes coréens, apparaissent comme des ressources pédagogiques issues d'une dynamique d'appropriation locale plutôt que comme des écarts à la norme. Les révisions du vocabulaire pendant le cours s'effectuent en action, en binômes ou en groupes, souvent durant les semaines précédant les passages de grade.

La terminologie taekwondoïste fait également référence à la tradition orientale et son système philosophique. La présentation du premier *poomsae* à apprendre (« *tae geug il jang* ») est accompagnée d'une note informative : « *Il représente le signe KUN GWE du Pal Gwé. Ce signe représente le ciel (qui nous donne la pluie, la lumière le soleil : par eux croissent toutes choses) ; le début de la création, force créatrice ; le Kun Gwé est donc à l'origine de toute chose sur terre ; il est comparé au dragon qui quitte la terre pour aller vers le ciel* » (fiche de révision n° 2).

Enfin, les newsletters et les affiches sont des dispositifs de communication organisationnelle destinés non seulement aux élèves, mais également aux parents. Elles mobilisent par conséquent une terminologie spécialisée plus restreinte, se limitant aux principalement aux noms des arts martiaux coréens enseignés dans le club : *taekwondo*, *hapkimudo* et *haedong kumdo*, et aux termes relatifs aux formes du travail : *poomsae*, *han boon keleuki*, par exemple. Ces supports inscrivent la terminologie, de manière « économique », dans une narration collective qui contribue à valoriser l'identité du club et l'ancrage culturel du taekwondo.

5. Conclusion

Notre communication propose une contribution originale aux champs de la socioterminologie et de la recherche d'information (*information seeking*), en explorant un terrain sportif pas encore étudié sous cet angle. L'étude met en évidence la place centrale de la terminologie dans les pratiques informationnelles des taekwondoïstes, lesquelles mobilisent à la fois des ressources fondées sur l'oralité et des ressources documentaires, physiques et numériques. La terminologie véhicule non seulement des savoirs théoriques, mais également des savoirs pratiques et des savoir-être, en répondant aux enjeux du loisir sérieux : progression par l'acquisition de nouvelles compétences et connaissances martiales, développement physique et spirituel, transmission des normes, des valeurs et des principes liés à la pratique et à la philosophie du taekwondo, ainsi que construction d'une identité ancrée dans la culture coréenne. La terminologie mobilisée au sein du club de taekwondo ne peut ainsi être réduite à un simple lexique spécialisé à mémoriser. Elle constitue un savoir incorporé, progressivement construit à l'articulation du langage, du corps et de la vie du club. Cette approche permet de dépasser une vision strictement normative de la terminologie, telle qu'elle apparaît dans les

ressources institutionnelles spécialisées, pour en souligner la dimension processuelle et située. La norme institutionnelle est alors traduite, au niveau local, en savoirs praticables et partagés au sein de la communauté de pratique.

Notre analyse conforte par ailleurs les résultats d'études antérieures. Elle souligne notamment l'importance des standards terminologiques pour l'acquisition des savoirs et compétences définis à l'échelle nationale et internationale (Yoo et al., 2016), condition nécessaire à l'inscription du club dans les circuits de compétition de *poomsae* et à sa participation à ces événements. Les observations de Taskinen (2019), relatives aux combinaisons fréquentes de termes coréens et de traductions ou reformulations en langue locale, aux variations de prononciation des termes par les enseignants, ainsi qu'à l'importance accordée au cri (*ki hap*), se retrouvent également dans notre corpus.

Enfin, cette étude présente certaines limites. Elle ne prend pas en compte les logiques individuelles des pratiquants, ni leurs représentations subjectives de la terminologie et de son apprentissage, et ne permet pas de déterminer dans quelle mesure les observations réalisées peuvent être généralisées à l'ensemble des clubs français de taekwondo. Afin de prolonger cette première exploration, nous envisageons de conduire des entretiens semi-directifs auprès des pratiquants du club, qui permettront d'articuler plus finement pratiques observées, expériences individuelles et dynamiques d'appropriation terminologique.

6. Bibliographie

1. Adjizian, J.-M. 2023. Le « loisir sérieux ». Entrevue avec Robert Stebbins. *Observatoire québécois du loisir – Bulletin*, 20(5) : 1-4.
2. Chaudiron, S., Ihadjadene, M. 2010. De la recherche de l'information aux pratiques informationnelles. *Etudes de communication*, (35) : 13-30.
3. Couzinet, V. 2009. Introduction – Dispositifs infocommunicationnels : contributions à une définition". Dans V. Couzinet (dir.), *Dispositifs info- communicationnels – questions de médiations documentaires*. Lavoisier, 19-30.
4. De Lavergne, C. 2007. La posture du praticien-chercheur : un analyseur de l'évolution de la recherche qualitative. *Recherches qualitatives* – hors série, (3) : 28-43.
5. Delavigne, V., De Vecchi, D. 2016. Socioterminologie et pragmaterminologie : rencontres et complémentarités. Dans C. Roche (dir.), *Terminologie & Ontologie : Théories et Applications – Actes de la conférence TOTh 2016 – Chambéry 9-10 juin 2016*. Presses Universitaires Savoie Mont Blanc, 141-156.
6. Fabre, I., Gardiès, C. 2010. La médiation documentaire. Dans V. Liquète (dir.), *Médiations*. CNRS Editions, 121-139.
7. Fédération Française de Taekwondo. 2026. Les positions de base. https://www.fftda.fr/fr/13-les-positions-de-base.html
8. Gaudin, F. 2005. La socioterminologie. *Langages*, (157) : 80-92.
9. Hong Hi, Choi. 1965. *Taekwon-do – The art of self-defence*. Daeha Publication Company.
10. Hong Hi, Choi. 1987. *15 Volumes Encyclopedia of Taekwon-do, 2nd Edition. International Taekwon-Do Federation*. https://archive.org/details/vol07tkd/vol01tkd/
11. Hou, Y., Kenderdine, S. (2024). Ontology-based knowledge representation for traditional martial arts. *Digital Scholarship in the Humanities*, 39(2) : 575–596.

12. Kang-Ho Taekwondo (2024). Lexique de base. https://kang-ho-taekwondo.com/le-taekwondo/lexique-base/
13. Kari, J. Hartel, J. (2007). Information and higher things in life: addressing the pleasurable and the profound in information science. *Journal of the American Society for Information Science and Technology, 58*(8) : 1131–1147.
14. Kukkiwon. 2001. *Kukkiwon Taekwondo Basics*. Kukkiwon. https://fr.scribd.com/doc/16237260/Kukkiwon-Taekwondo-Basics
15. Kukkiwon. 2011. *Taekwondo Technical Terminology*. Kukkiwon. https://wientaekwondo.com/wp-content/uploads/2022/07/Technical_Terminology_ENG.pdf?utm_source=chatgpt.com
16. Moenig, U. 2015. *Taekwondo : from a martial art to a martial sport*. Routlege.
17. Moenig, U., Kim, M. 2019. The Origins of World Taekwondo (WT) Forms or P'umsae. *IDO MOVEMENT FOR CULTURE. Journal of Martial Arts Anthropology*, 19(3) : 1-10.
18. Taekwondo Neuville. 2025. Vocabulaire et Lexique. https://www.taekwondo-neuville.fr/taekwondo/vocabulaire-et-lexique/
19. Savolainen, R. 2009. Everyday life information seeking. In M. J. Bates (dir.), *Encyclopedia of library and information sciences (3rd ed.)*. Taylor & Francis, 1780–1789.
20. Stassin, B., Roux, U., Trzmielewski, M. 2026. Penser le dojo comme un dispositif info-communicationnel. Dans *CNRIUT 2026*, IUT de Lorient, Lorient, France, 1-2 avril 2026. https://cnriut2026.sciencesconf.org/683362/document
21. Stebbins, R. A. 2009. Leisure and its relationship to library and information science: bridging the gap. *Library Trends,* 57(4) : 618–631.
22. Suchman, L. 1987. *Plans and situated actions: the problem of human/machine communication*. Cambridge University Press.
23. Trzmielewski, M. 2023. *Vers la conception d'un système d'organisation des connaissances en allergologie : l'analyse des documents et des pratiques informationnelles des acteurs*. Thèse en sciences de l'information, de la communication et de la documentation, Université Paul-Valéry Montpellier 3. https://hal.science/tel-04128883
24. World Taekwondo. 2026. https://www.worldtaekwondo.org/main
25. Yoo, S.-H., Jung, K.-H., Ryu, J.-S. 2016. Suggestion of New Terminology and Classification of the Hand Techniques by Angular Momentum in the Taekwondo Poomsae. *Korean Journal of Applied Biomechanics*, 21(1) : 51-69.

Le projet TermPTEmRo : entre « Sapori locali » de l'Émilie-Romagne et promotion territoriale à l'ère numérique

Gloria Zanella, Università di Modena e Reggio Emilia, gloria.zanella@unimore.it

Chiara Preite, Università di Milano, chiara.preite@unimi.it

Francesca Cialdini, Università di Modena e Reggio Emilia, francesca.cialdini@unimore.it

Résumé

Le projet TermPTEmRo[1] vise à créer une base de données terminologique multilingue en ligne des produits typiques de l'Émilie-Romagne à partir des fiches terminologiques qui constituent le glossaire Sapori locali « Saveurs locales »[2]. En particulier, cette recherche s'inscrit dans le cadre de traditions et de produits, en proposant une ressource terminologique multilingue sur les produits D.O.P., I.G.P., D.O.C., D.O.C.G. et I.G.T. en italien, français, anglais, espagnol, roumain et allemand qui est en cours de développement (la ressource en ligne pourra être ultérieurement enrichie à l'avenir). Cette base de données terminologique multilingue permet d'établir des équivalences dénominatives des « produits typiques » ayant une dénomination protégée, qui pourrait s'avérer utile pour la communication à différents niveaux : entre experts et entreprises, mais aussi pour les utilisateurs curieux ou les touristes passionnés par l'œnogastronomie qui peuvent accéder aux connaissances sur les produits typiques de l'Émilie-Romagne. Enfin, cette ressource terminologique multilingue vise à promouvoir la diffusion des traditions alimentaires de la Région au niveau international.

1. Entre traditions de l'Émilie-Romagne et valorisation du territoire

Le lien étroit entre terroir, territoire et gastronomie détermine la valorisation des produits locaux et authentiques (Bérard & Marchenay 1995 ; Delfosse 1997). Le patrimoine œnogastronomique d'une région devient culture et connexion entre le territoire et le touriste (Devilla 2015 ; Chessa & De Giovanni 2021). En outre, les produits œnogastronomiques ont été définis comme unicités gastronomiques du territoire ou réservoirs gastronomiques[3] (Paolini 2004) et comme des produits caractérisés par un contenu territorial élevé[4] (Asero & Patti 2009). Récemment, les recherches menées sur le tourisme rural et les bières artisanales ouvrent d'ultérieures pistes de réflexion sur les intérêts des consommateurs, sur les produits locaux et sur leurs dénominations (Murray & Kline 2015 ; Temmerman 2018 ; Melewar & Skinner 2020 ; Chessa & De Giovanni 2022 ; Mangiapane 2024 ; Devilla & Mercurio 2025). Les produits typiques deviennent des ambassadeurs du goût qui sortent des frontières nationales et

[1] *TermPTEmRo* est l'acronyme de : Terminologie des Produits Typiques de l'Emilie-Romagne.
[2] Le glossaire multilingue « Sapori locali » a été développé au Département d'Études linguistiques et culturelles à l'Université de Modène et de Reggio d'Émilie sous la direction de la professeure C. Preite.
[3] « Giacimenti gastronomici », selon Paolini (2004 : 52).
[4] « Prodotti enogastronomici come prodotti ad alto contenuto territoriale », tels que définis par Asero et Patti (2009 : 638).

qui sont porteurs des traditions locales promues au niveau international (Druetta 2008 ; Chessa & De Giovanni 2024 ; Preite 2024).

Les traditions alimentaires de l'Émilie-Romagne plongent leurs racines dans des siècles de pratiques agricoles et artisanales. Au cours des XIXe et XXe siècles, la transition vers l'industrie alimentaire a marqué la naissance de grandes filières liées aux tomates, aux pâtes et aux conserves, sans pour autant effacer les pratiques paysannes et locales. Cette stratification historique de la région explique pourquoi chaque produit typique est aujourd'hui non seulement un aliment, mais aussi un vrai symbole culturel (Paolini 2000). L'Émilie-Romagne est reconnue comme l'une des plus importantes *Food Valley* au monde, grâce à la richesse gastronomique qui se compose d'une mosaïque de paysages caractérisés par des plaines, des collines, les Apennins et la côte maritime. En outre, cette région s'enracine dans une histoire agricole, artisanale et industrielle qui a fait de ce territoire non seulement une référence nationale, mais aussi internationale dans le domaine de la qualité alimentaire et des excellences culinaires.

2. Le projet TermPTEmRo[5] : une ressource terminologique multilingue pour découvrir des produits émiliens et romagnols

Les Musées du Goût et les Routes des Vins de l'Émilie-Romagne jouent le rôle de garants des produits d'origine protégée, produits qui constituent aussi l'objet du projet TermPTEmRo, une ressource terminologique multilingue accessible en ligne par une variété d'usagers. À l'état actuel, les produits locaux de l'Émilie-Romagne et ses traditions régionales sont valorisés et promus au niveau international par le glossaire « Sapori locali » (https://www.lexi-term.unimore.it/glossarioptem/), dont les 53 fiches terminologiques, chacune consacrée à un produit typique ayant dénomination protégée[6], sont rédigées en italien, français, anglais, espagnol et en roumain.

Partant d'une approche théorique et méthodologique qui adapte et intègre la Terminologie textuelle (Condamines 2018, Condamines & Picton 2022) et la

[5] This work was supported by the Università di Modena e Reggio Emilia - Project "Local tastes and territorial promotion in the digital era. Towards an online multilingual termbase of typical Emilia-Romagna products" CUP E93C24001980007 - funded by Fondo di Ateneo per la ricerca Anno 2024 - Bando per il finanziamento di progetti di ricerca interdisciplinari"

[6] La sélection des produits est basée sur la liste des dénominations italiennes inscrites dans le Registre des appellations d'origine protégées, des indications géographiques protégées et des spécialités traditionnelles garanties (Règlement de l'Union Européenne n. 1151/2012 du Parlement européen et du Conseil du 21 novembre 2012 mise à jour 30 mai 2025). https://www.masaf.gov.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/2090#main; http://catalogoviti.politicheagricole.it/dopigp.php

Socioterminologie (Gaudin 2002, 2005), nous avons exploité le web pour consulter de nombreux textes issus du domaine impliqué, à différents niveaux de spécialisation, contenant les dénominations italiennes des produits en question ainsi que les variantes sous lesquelles ils sont connus dans les langues cibles. Pour ce faire, nous avons utilisé les dénominations protégées comme clé d'entrée pour le repérage de textes utiles dans le web, démarche typique de la Terminologie textuelle ayant le mérite de permettre d'étudier les termes dans leur emploi en contexte pour chaque langue. Par-là, la consultation des documents[7] dans chaque langue nous a également permis de découvrir les variantes des dénominations effectivement diffusées chez les natifs d'autres langues – dont l'importance pour la construction connaissance d'un concept est mise en avant par la Socioterminologie.

Cette activité de fouille et de recherche dans les textes authentiques nous a offert un riche réservoir de matériel dans lequel puiser pour remplir à la fois les champs linguistiques et techniques qui composent les fiches terminologiques et renseignent sur chaque produit dans les langues considérées.

Les entrées qui composent actuellement le glossaire « Sapori locali », c'est-à-dire les dénominations enregistrées des produits pour l'italien et leurs équivalents diffusés dans les autres langues, sont proposées dans la page d'accueil du site en ordre alphabétique pour chaque langue impliquée, afin de permettre à l'usager de commencer ses recherches à partir de la langue mieux connue. Les fiches terminologiques consacrées aux entrées cliquables sont reliées entre elles par des hyperliens – représentés par les dénominations équivalentes cliquables – qui permettent aux usages de passer d'une langue à l'autre afin d'avoir accès à la fiche terminologique du produit dans la langue étrangère sélectionnée.

Les champs[8] des fiches offrent des informations fondamentales, à la fois sur l'objet et sur le terme, c'est-à-dire sur le produit enregistré et sur la dénomination protégée ou sur son équivalent repéré en usage. Cela permet d'un côté, de connaître un produit déterminé, par exemple d'en découvrir les procédés de préparation ou de conservation ; et, de l'autre côté, d'apprendre à utiliser sa dénomination en discours. Nous offrons donc des renseignements sur

[7] Nous avons consulté des sources primaires offrant des définitions, comme des bases de données, dictionnaires et glossaires ; et des sources secondaires comme les sites de producteurs, de revendeurs (souvent multilingues), d'amateurs passionnés, ainsi que des sites officiels de l'Union européenne ou de la Région de l'Émilie-Romagne.
[8] Le domaine et le sous-domaine alimentaire du produit, la marque grammaticale, la définition, la note (qui apporte des informations sur les processus de production et de conservation des produits ainsi que des curiosités diverses), le contexte authentique d'emploi de la dénomination en discours, la transcription phonétique, les variantes, l'étymologie, l'image du produit, les équivalents dans les langues considérées, les notes des équivalents et les sources bibliographiques et sitographiques consultées pour chaque information offerte (voir figure 1 dans les Annexes).

la dénomination en tant que « terme » de la langue aussi bien que des renseignements sur le référent extralinguistique désigné par la dénomination.

Par exemple, pour ce qui est du terme, les marques grammaticales et les contextes d'usage sont utiles pour un usager étranger qui nécessiterait de replacer une dénomination dans le discours. Les informations étymologiques (dans chaque langue, si la dénomination n'est pas un emprunt) invitent l'usager à découvrir, lorsque possible, l'origine des produits et de leurs dénominations, et enfin, la transcription phonétique, elle aussi dans chacune des langues envisagées, permet aux usagers d'en apprendre la prononciation dans la langue d'intérêt.

Quant aux référents extralinguistiques, la définition et la description permettent non seulement de connaître les produits, mais aussi des curiosités à leur propos. En particulier, le champ de la description permet de découvrir quelles sont les matières premières utilisées, les méthodes de transformation, de production et de conservation, et des détails divers.

3. Objectifs du projet TermPTEmRo

Le glossaire multilingue se compose d'une variété de produits allant de vinaigres, d'huiles, de vins, et d'aliments comme le riz, des légumes, des fruits, aux produits de boulangerie, aux préparations à base de viande, et aux fromages. Les produits protégés et les vins émiliens et romagnols, qui sont les protagonistes à découvrir dans les Musées du Goût et les Routes des Vins de l'Émilie-Romagne, constituent donc également l'ensemble des dénominations auxquelles le projet TermPTEmRo est consacré, comme il est possible de le remarquer dans la liste proposée dans le tableau 1 dans les Annexes.

Les objectifs poursuivis pendant les travaux qui ont mené à la construction du glossaire « Sapori locali » – et qui aboutiront à compilation de la ressource terminologique multilingue TermPTEmRo – sont multiples : d'un côté, nous avons visé la valorisation des produits du territoire et des traditions émiliennes-romagnoles au niveau national et international ; de l'autre côté, nous nous sommes proposés la réalisation d'une ressource terminologique multilingue capable de s'adresser aux usagers passionnés d'œnogastronomie, mais aussi aux spécialistes et aux entreprises. Comme nous l'avons anticipé, les fiches offrent les sources bibliographiques et sitographiques concernant les informations offertes à propos des produits sélectionnés. En général, les renseignements ont été repérés dans les sites officiels et multilingues de la région Émilie-Romagne, de l'Union Européenne, et de producteurs affirmés à l'échelle internationale. Alors que les dénominations protégées, en tant que désignations officielles, sont repérées dans les documents pour toutes les langues, il arrive souvent que les dénominations sous lesquelles les produits sont effectivement connus à

l'étranger sont des variantes : nous les avons donc recueillies dans les fiches terminologiques afin de permettre à l'usager étranger de (re)connaître tel ou tel produit selon le nom par lequel il est diffusé.

De plus, l'étude systématique des dénominations protégées et de leurs variantes en langues cibles nous a permis de mener une comparaison interlinguistique des équivalents et d'analyser les stratégies (Preite 2024, 2026 sp) par lesquelles ces dernières empruntent, calquent ou traduisent (selon la typologie proposée par Sablayrolles 2016) les dénominations originales. Par exemple, dans les textes authentiques consultés, à côté des dénominations italiennes officielles, il est également possible de repérer la désignation d'un produit sous une forme calquée : comme dans le cas de : ang. *Sour cherries from Modena* (pour *Amarene brusche di Modena*); fr. *Jambon de Parme* et esp. *Jamón de Parma* (pour *Prosciutto di Parma*) ; fr. et ang. *Parmesan* (pour *Parmigiano Reggiano*); ro. *Pară din Emilia-Romagna* (pour *Pera del Emilia-Romagna*), etc.

Les dénominations enregistrées dans le glossaire permettent d'observer que rarement les noms des produits typiques sont de simples emprunts d'unités monolexicales ou polylexicales : *Brisighella*, *Gutturnio*, *Lambrusco* et *Ortrugo* ou *Coppia ferrarese*. Par contre, la presque totalité des produits montre une dénomination polylexicale – le plus souvent calquée parfois avec des petites adaptations) selon les langues et les produits – et cela dépend du fait que les dénominations complexes se fondent sur une association binaire lexiculturelles (Galisson 1987, 1988, 1991) que l'on garde pour évoquer l'origine même du produit. Il s'agit d'association variées (cf. Preite 2024) de type : produit + ville (par exemple Mortadella *Bologna*, Zampone *Modena*), produit + région (Pera dell'*Emilia-Romagna*), produit + village (Marrone di *Castel del Rio*), produit + fleuve (Riso del *Delta del Po*), produit + aire géographique (Grana *Padano*), col/colline (*Colli Piacentini* Monterosso Val d'Arda, *Colline di Romagna*). Le rapport étroit entretenu par le produit protégé et le territoire de production résonne dans les dénominations elles-mêmes et se grave dans la mémoire des consommateurs et des touristes.

4. Pour (ne pas) conclure

Les traditions locales et les variétés de produits typiques montrent la richesse œnogastronomique de l'Émilie-Romagne. Le projet TermPTEmRo, basé sur la réalisation préliminaire du glossaire « Sapori locali », se propose comme ressource terminologique multilingue qui promeut au niveau international les produits locaux afin de diffuser les traditions œnogastronomiques émiliennes et romagnoles auprès d'un public varié, allant des

usagers simplement curieux, aux touristes, aux experts du domaine œnogastronomique. TermPTEmRo propose aussi une redécouverte d'anciennes saveurs : en particulier, cette ressource devient un outil qui valorise le terroir, le territoire et leurs produits œnogastronomiques. Grâce aux renseignements proposés, nous avons mis en relief le lien étroit entre les produits, leurs dénominations et le territoire, dans une sorte de cercle vertueux qui constitue un facteur identitaire profond. Les fiches terminologiques multilingues qui composent la ressource terminologique multilingue assurent le passage du local au global et deviennent des moyens de partage et d'approfondissement des traditions œnogastronomiques émiliennes et romagnoles.

Les perspectives futures du projet TermPTEmRo concernent non seulement l'élargissement à la langue allemande et polonaise et la révision générale des données, mais surtout sa migration dans le TMS (terminology management system) FAIRterm (Vezzani 2022 ; Vezzani & Di Nunzio 2020), suivant les principes FAIR (*Findable, Accessible, Interoperable and Reusable*) de la *European Open Science Cloud Association*, pour la réalisation d'une banque de données terminologique en libre accès.

Bibliographie

Asero, Vincenzo, Patti, Sebastiano (2009) : « Prodotti enogastronomici e territorio : la proposta dell'enoturismo ». In Emilio Becheri (éd.), *XVI Rapporto sul Turismo Italiano*, Milano, FrancoAngeli : 637-649.

Bérard, Laurence, Marchenay, Philippe (1995) : « Lieux, temps et preuves. La construction sociale des produits de terroir ». *Terrain* 24 : 153-164.

Chessa, Francesca, De Giovanni, Cosimo (2021) : « Comunicare il patrimonio enogastronomico sardo attraverso Wikipedia ». In Devilla Lorenzo et Galiñanes Gallén Marta (a cura di), *Lingue minori e turismo*, Cagliari, Arkadia : 19-32.

Chessa, Francesca, De Giovanni, Cosimo (2022) : « La denominazione del prodotto gastronomico nell'offerta turistica : breve analisi dei prodotti dolciari sardi ». In Lorenzo Devilla et Marta Galinañes Gallén (dir.), *Le parole del turismo. Aspetti linguistici e letterari*, Alessandria, Edizioni dell'Orso : 105-115.

Chessa, Francesca, De Giovanni, Cosimo (2024) : « Les produits fromagers AOP et IGP français et italiens. Pour une sémantique de la dénomination appliquée à la terminologie ». *Roczniki Humanistyczne* 8 : 35-47.

Delfosse, Claire (1997) : « Noms de pays et produits du terroir : enjeux des dénominations géographiques ». *L'Espace géographique* 26/3 : 222-230.

Devilla, Lorenzo (2015) : « Le rôle de la gastronomie dans la représentation de l'identité régionale sarde : aspects de la communication touristique ». In Paissa Paola, Rigat Françoise & Vittoz Marie-Berthe (dir.), *Dans l'amour des mots. Chorale(s) pour Mariagrazia*, Alessandria, Edizioni dell'Orso : 219-229.

Devilla, Lorenzo, Mercurio, Nicla (2025) : « Valoriser le territoire à travers la bière artisanale.Une analyse discursive exploratoire des sites Web de brasseries italiennes (Campanie,Sardaigne), françaises (Corse) et suisses (Jura) ». *Repères DoRiF* 31 *Langues familiales et entrée dans les littéracies scolaires*, DoRiF Università, Roma.

Druetta, Ruggero (2008) : « Les noms de marque et de produit comme marqueurs identitaires ». *Éla. Études de linguistique appliquée* 150/2 : 157-175.
Galisson, Robert (1987) : « Accéder à la culture partagée par l'entremise des mots à C.C.P. ». *Éla. Études de Linguistique Appliquée* 67 : 119-140.
Galisson, Robert (1988) : « Culture et lexiculture partagées : les mots comme lieux d'observation des faits culturels ». *Éla. Études de Linguistique Appliquée* 69 : 74-90.
Galisson, Robert (1991) : *De la langue à la culture par les mots*. Paris : Clé International.
Mangiapane, Stella (2024) : « Migrations gastronomiques et lexicographiques. Les noms des fromages italiens au prisme d'Internet et des dictionnaires ». In Trovato Loredana (éd.), *Cibo e parole. Migrazioni e incontri gastronomici in prospettiva lessiculturale*, *Scritture migranti* 18 : 81-100.
Melewar, T.C., Skinner, Heather (2020) : « Territorial brand management : Beer, authenticity, and sense of place ». *Journal of Business Research* 116 : 680-689.
Murray, Alison, Kline, Carol (2015) : « Rural Tourism and the Craft Beer Experience : Factors Influencing Brand Loyalty in Rural North Carolina, USA ». *Journal of Sustainable Tourism* 23/8 : 1198-1216.
Paolini, Davide (2000) : *I luoghi del gusto. Cibo e territorio come risorsa di marketing*. Milano : Baldini & Castaldi, Dalai editore.
Paolini, Davide (2004) : « Giacimenti enogastronomici. Il rischio turismo ». *Equilibri* 1 : 51-56.
Preite, Chiara (2024) : « Les dénominations des produits typiques de l'Émilie-Romagne. Diffusion multilingue d'une terminologie protégée ». In Capra Daniela, Kaunzner Ulrike, Preite Chiara, Cagninelli Claudia (éds.), *Transformations des approches linguistiques des discours numériques*, *Lingue Linguaggi* Special Issue 65 : 285-301.
Preite, Chiara (2026 sous presse) : « Le denominazioni dei prodotti tipici dell'Emilia-Romagna: transfert e diffusione di una lessicultura esperta nelle lingue romanze », in Popescu Cecilia Mihaela, Pîrvu Elena, Dincă Daniela, Similaru Lavinia (éds.) *Romania Orientalis – Romania Occidentalis : Interférences culturelles et espaces identitaires*, Editura Universității din București.
Sablayrolles, Jean-François, (2016) : « Emprunts et influences d'autres SABLAYROLLES, Jean-François langues », in Hildenbrand, Zuzana, Kacprzak, Alicja, Jean-François Sablayrolles (éds.), *Emprunts néologiques et équivalents autochtones en français, en polonais et en tchèque*, Lambert Lucas, Limoges, 23-35.
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*, 192/4, 417-434.
Vezzani, Federica (2022) : *Terminologie numérique : conception, représentation et gestion*. Bern : Peter Lang.
Vezzani, Federica, Di Nunzio, Giorgio Maria (2020) : « Methodology for the standardization of terminological resources : design of TriMED database to support multi-register medical communication ». In *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 26/2 : 265-297.

Annexes

Vinaigres
Aceto Balsamico di Modena I.G.P.
Aceto Balsamico Tradizionale di Modena D.O.P.
Aceto Balsamico Tradizionale di Reggio Emilia D.O.P.
Huiles
Brisighella D.O.P.

Colline di Romagna D.O.P.
Vins
Bianco di Castelfranco Emilia I.G.T.
Colli Bolognesi Pignoletto D.O.C.G.
Colli di Parma Malvasia D.O.C.
Colli di Parma Rosso D.O.C.
Colli di Scandiano e di Canossa D.O.C.
Colli Piacentini Monterosso Val d'Arda D.O.C.
Colli Piacentini Trebbianino val Trebbia D.O.C.
Fortana del Taro I.G.T.
Gutturnio D.O.C.
Lambrusco Grasparossa di Castelvetro D.O.C.
Lambrusco Salamino di Santa Croce D.O.C.
Lambrusco di Sorbara D.O.C.
Ortrugo D.O.C.
Romagna Albana D.O.C.G.
Terre di Veleja I.G.T.
Val Tidone I.G.T.
Céreales
Riso del Delta del Po I.G.P.
Légumes
Aglio di Voghiera D.O.P.
Asparago verde di Altedo I.G.P.
Fungo di Borgotaro I.G.P.
Patata di Bologna D.O.P.
Scalogno di Romagna I.G.P.
Fruits
Amarene Brusche di Modena I.G.P.
Anguria Reggiana I.G.P.
Ciliegia di Vignola I.G.P.
Marrone di Castel del Rio I.G.P.
Melone mantovano I.G.P.
Pera dell'Emilia-Romagna I.G.P.
Pesca e nettarina di Romagna I.G.P.
Produits de boulangerie
Coppia Ferrarese I.G.P.
Pampapato / pampepato di Ferrara I.G.P.
Piadina Romagnola I.G.P.
Viande ou produits à base de viande
Agnello del Centro Italia I.G.P.
Coppa di Parma I.G.P.
Coppa Piacentina D.O.P.
Cotechino Modena I.G.P.
Culatello di Zibello D.O.P.
Mortadella Bologna I.G.P.
Pancetta Piacentina D.O.P.
Prosciutto di Modena D.O.P.
Prosciutto di Parma D.O.P.
Salama da sugo I.G.P.
Salame Cremona I.G.P.

Salame Felino I.G.P.
Salame Piacentino D.O.P.
Salamini italiani alla cacciatora D.O.P.
Vitellone Bianco dell'Appennino Centrale I.G.P.
Zampone Modena I.G.P.
Fromages
Casciotta d'Urbino D.O.P.
Formaggio di Fossa di Sogliano D.O.P.
Grana Padano D.O.P.
Parmigiano Reggiano D.O.P.
Provolone Valpadana D.O.P.
Squacquerone di Romagna D.O.P.
Pâtes alimentaires farcies
Cappellacci di zucca ferraresi I.G.P.

Tableau 1. Listes des entrées de la ressource terminologique multilingue TermPTEmRo

Entrée	Vinaigre balsamique de Modène
Domaine et Sous-domaine	Alimentation – autres produits
Marque grammaticale	S. m., sing.
Définition	Type particulier de vinaigre, utilisé en tant que condiment, produit dans le territoire de Modena et Reggio Emilia à travers la fermentation du moût de raisin cuit et laissé vieillir dans des fûts en bois pendant au moins cinq ans (différents types de bois), d'où viennent l'arôme et la couleur. (1)
Note	Le Vinaigre Balsamique de Modène IGP s'obtient à partir de l'union du vinaigre fort de vin au moût concentré ou cuit de raisin venant de sept cépages (Lambrusco, Sangiovese, Trebbiano, Albana, Ancellotta, Fortana, Montuni). Ces deux composantes sont équilibrées en proportions qui changent en fonction de la typologie de vinaigre recherché. Tout cela est inséré dans des fûts en bois à maturer pendant au moins trois ans et même plus pour le vinaigre vieilli. Le vinaigre balsamique a une qualité supérieure quand il contient plus de moût cuit, qui représente la partie noble du Vinaigre Balsamique de Modène IGP. Contrairement au vinaigre balsamique traditionnel de Modène DOP, le vinaigre balsamique de Modène IGP n'est pas transféré d'un fût à l'autre selon la méthode traditionnelle, mais il est laissé maturer dans le même fût pendant la période nécessaire. (2) Le produit se présente de couleur brun intense et avec une odeur caractéristique, persistante, ferme, délicate et légèrement vinaigrée, avec d'éventuelles notes boisées. Le Vinaigre Balsamique de Modène IGP, au moment de sa mise à la consommation, a un aspect limpide et brillant. La production du produit doit être effectuée dans le territoire administratif des provinces de Modena et Reggio Emilia, à travers l'utilisation de moûts de raisins cultivés en Emilia-Romagna. (3)
Contexte	Le vinaigre balsamique de Modène possède un bouquet aromatique et un goût inimitable qui le rendent parfait pour diverses utilisations. (4)
Transcription phonétique	[vi'negɨ balsamik mo'den]
Synonyme/ Variantes	Aceto Balsamico di Modena (5)
Étymologie	Le mot « vinaigre » naît de l'univerbation de « *vin* » et « *aigre* ». (6) « Balsamique » vient du latin *balsameus, balsamum* (« baume »), suc d'un arbrisseau : le baumier. (7) Modène est une commune italienne, chef-lieu de l'homonyme province en Emilia-Romagna.

Image du produit	 (8)
Equivalents	ITA Aceto Balsamico di Modena ENG Balsamic vinegar from Modena ESP Vinagre Balsámico de Modena RO Oțet balsamic de Modena
Note des équivalents	Le nom du produit reste masculin singulier dans toutes les langues considérées.
Organisme	Università degli Studi di Modena e Reggio Emilia, Dipartimento di Studi Linguistici e Culturali
Date de création de la fiche	03/04/2015
Dernière date de modification de la fiche	22/04/2023
Auteur Réviseur	Lucia Palazzese Margherita Baroni Gloria Zanella
Sources	(1) http://www.treccani.it/vocabolario/aceto/ (2) http://www.acetaialabonissima.it/ita/aceto-balsamico-modena-igp-produzione.aspx (3) http://www.naturalmenteitaliano.it/flex/FixedPages/IT/Prodotto.php/L/IT/P/4322 (4) http://www.coricelli.fr/?post_type=portfolio&p=1650 (5) http://fr.wiktionary.org/wiki/vinaigre (6) http://fr.wiktionary.org/wiki/balsamique (7) https://www.google.com/url?sa=i&url=https%3A%2F%2Fdambros.it%2Fecommerce%2Fprodotto%2Fid%2F0000060210%2F&psig=AOvVaw2 (8) https://www.ilgiardinodeilibri.it/speciali/aceto-balsamico-modena-dop-igp-cosa-cambia.php (9) http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2009:256:0036:0037:FR:PDF

Figure 1. Fiche terminologique du *Vinaigre balsamique de Modène*

Corpus et terminologie bilingue arabe-français du domaine *eaux usées*

Ouafae Nahli[1] Nanée Chahinian[2] Ilham Chaker[3]
ouafae.nahli@cnr.it nanee.chahinian@ird.fr ilhamchaker@gmail.com

Abstract

Cet article présente une méthodologie corpus-driven pour l'établissement de correspondances terminologiques français–arabe dans le domaine des *eaux usées*. Il montre que ces correspondances ne peuvent être établies par une traduction directe et met en évidence les enjeux méthodologiques liés à l'extension, vers la langue arabe, de ressources terminologiques initialement élaborées en langue française dans un domaine hautement spécialisé.

Keywords : Domain-specific corpus ; terminologie ; *eaux usées* ; TEI ; corpus bilingue arabe–français

1. Introduction

Le présent travail s'inscrit dans le cadre du projet *StarWars*, au sein duquel ont été élaborées une ontologie[4] et des ressources terminologiques multilingues en français, italien et anglais, dédiées au domaine des eaux usées [1], [2]. L'extension de ces travaux à la langue arabe pose toutefois des enjeux méthodologiques spécifiques. La traduction directe d'une terminologie élaborée en français s'avère insuffisante, en particulier au regard des spécificités morphologiques, sémantiques et sociolinguistiques de l'arabe, caractérisé également par une situation de diglossie et par la coexistence de variétés dialectales aux usages terminologiques distincts. Dans ce contexte, les approches dites *corpus-based* et *corpus-driven* occupent une place centrale en terminologie. Le corpus ne doit pas être envisagé comme un simple réservoir d'exemples, mais comme un instrument central de la description et de l'extraction terminologiques. Il permet également d'établir des relations conceptuelles à partir de l'analyse des contextes d'emploi et des usages observés, plutôt que sur la base de listes préexistantes ou de ressources dictionnairiques isolées [7], [8].

Le travail, illustré dans cet article, repose sur la construction et l'encodage en *TEI XML* d'un corpus bilingue français–arabe dédié au domaine des eaux usées. Afin de mettre en relation les concepts modélisés dans l'ontologie issue du projet *StarWars* avec leurs réalisations terminologiques attestées en langue arabe, les usages effectifs des termes de domaine dans les deux langues sont analysés au moyen d'une annotation terminologique. Ce choix méthodologique permet d'identifier différents degrés de formalisation terminologique et de mettre en évidence le rôle central du corpus dans la constitution de ressources terminologiques bilingues fondées sur des usages exploitables dans des contextes scientifiques et numériques.

2. Corpus et données

Le corpus est constitué de textes bilingues issus de documents normatifs émanant des autorités publiques marocaines. Les documents retenus comprennent notamment :

[1] CNR-ILC, Pisa, Italie. Ouafae.nahli@cnr.it
[2] HSM IRD, Univ Montpellier, CNRS, Montpellier, France.
[3] LSIA FST, Université Sidi Mohamed Ben Abdellah, Fès, Maroc.
[4] http://sewernet.msem.univ-montp2.fr

- **la loi n° 10-95 relative à l'eau**, promulguée par le Dahir n° 1-95-154 du 16 août 1995, constitue un texte fondateur en établissant pour la première fois un cadre unifié définissant les principes fondamentaux de la gestion des ressources en eau [4] ;
- **la loi n° 36-15 relative à l'eau**, promulguée par le Dahir n° 1-16-113 du 6 Kaada 1437 (10 août 2016), complète et actualise la loi 10-95 en élargissant son champ normatif et en comblant plusieurs lacunes [4] ;
- le manuel intitulé **Recueil des textes réglementaires relatifs aux ressources en eau**, qui regroupe les textes d'application de la loi 36-15, les décrets et arrêtés correspondants, ainsi que les dispositions encore en vigueur de la loi 10-95 et d'autres textes législatifs connexes relatifs à la gestion de l'environnement (ressources en eau, carrières, développement durable, etc.) ;[5]
- **les décrets d'application de la loi 36-15**, précisant les modalités de mise en œuvre de certaines de ses dispositions.

Toutefois, en raison de son caractère strictement spécialisé, le corpus centré sur le domaine des eaux usées ne couvre pas l'ensemble de la terminologie pertinente dans le cadre du projet *StarWars* .
Le travail s'appuie en complément sur un ensemble de ressources lexicographiques spécialisées bilingues, principalement arabe–français et arabe–anglais. L'emploi de ces ressources vise à élargir la couverture lexicale du corpus. Les deux principales ressources utilisées sont :

- **ARABTERM**: une plateforme terminologique multilingue développée par ALECSO et GIZ, couvrant des domaines tels que l'hydraulique, la gestion de l'eau et le traitement des eaux usées. Elle fournit des équivalents validés Ar-Fr et Ar-En examinés par des terminologues experts [3].[6]
- **Glossary of Shared Water Resources** (UN-ESCWA & BGR) : un glossaire trilingue (arabe-anglais-français) couvrant la gestion des ressources en eau partagées, y compris le traitement et à la réutilisation de l'eau [6] .[7]

3. Encodage et structuration des données en TEI XML

Afin de permettre une analyse terminologique systématique, le corpus bilingue français–arabe est encodé en *TEI XML*. Le recours à la *TEI* se justifie par sa capacité à représenter de manière explicite la structure interne des textes, tout en autorisant l'intégration d'annotations terminologiques directement ancrées dans les données textuelles, conformément aux recommandations des *TEI Guidelines* en matière de structuration et d'alignement de textes parallèles [5]. L'exemple présenté dans la figure 1 illustre un fragment encodé correspondant à une définition légale extraite de la *loi n° 36-15 relative à l'eau*. La structuration repose sur une hiérarchie textuelle classique, permettant de segmenter les documents selon leur organisation interne (chapitres, sections, articles), au moyen de l'élément <*div*>, systématiquement muni d'identifiants uniques portés par l'attribut *@xml:id*. Dans une perspective bilingue, les textes français et arabes sont mis en relation au moyen d'un alignement explicite, réalisé à l'aide de l'élément <*cit type="parallel"*>. Les relations de correspondance entre segments sont précisées par l'attribut *@corresp*, qui permet d'établir un lien explicite entre les unités

[5] https://abhbc.com/fr et https://abhbc.com/ar
[6] https://arabterm.org
[7] https://archive.unescwa.org/sites/www.unescwa.org/files/publications/files/glossary-shared-water-resources-english.pdf

alignées sans fusionner les contenus linguistiques. L'encodage *TEI* intègre par ailleurs une couche d'annotation terminologique permettant de baliser, grâce à l'élément <*term*>, les occurrences terminologiques directement dans le flux textuel. Cette structuration constitue le support méthodologique nécessaire à l'analyse des correspondances terminologiques français–arabe, présentée dans la section suivante.

```
<item xml:id="fr_art3_7" corresp="ar_art3_7">
    <p>
        <cit type="parallel">
 <quote>
   <term xml:id="fr_eau_usee_def"
   corresp="#ar_eau_usee_def" ana="#def">eau usée</term> : une eau qui a
   subi une modification de sa composition ou de son état naturel du fait
   de son utilisation ; </quote>
 <quote>
   <term xml:id="ar_eau_usee_def"
   corresp="#fr_eau_usee_def" ana="#def">ماء مستعمل</term>: كل ماء تعرض لتغيير
   في تركيبته أو حالته الطبيعية جراء استعماله؛ </quote>
        </cit>
    </p>
</item>
```

Figure 1 - Fragment de corpus encodé en *TEI XML*

4. Résultats et analyse

Dans l'exemple de la figure 1, le point de départ de l'analyse est le concept *eaux usées* tel qu'il est modélisé dans l'ontologie du projet *StarWars*, où il est notamment associé au concept *réseau de collecte des eaux usées*. Dans le corpus juridique, et plus précisément dans la *loi n° 36-15 relative à l'eau*, le terme *eaux usées* fait l'objet d'une définition explicite (cf. figure 1) :

> *Eau usée* : une eau qui a subi une modification de sa composition ou de son état naturel du fait de son utilisation.

Dans le texte arabe officiel, la notion *eau usée* est rendue par l'expression مياه مستعملة "*miyāh mustaʿmalah*" (*eaux utilisées*)[8], sans distinction conceptuelle explicite entre substance et infrastructure.

La consultation de ressources lexicographiques spécialisées met toutefois en évidence des divergences terminologiques significatives. Ainsi, la plateforme ARABTERM propose la désignation مياه الصرف الصحي "*miyāh aṣ-ṣarf aṣ-ṣiḥḥī*" (*eaux d'évacuation sanitaire*), accompagnée de la définition suivante :

> المياه المُلوّثة نتيجة الاستعمال (الذي أحدث تغييرات في خواصها أو تركيبها)، والتي تُصنف بعد ذلك حسب استعمالها أو قابليتها لإعادة الاستعمال إلى: مياه رمادية، ومياه سوداء، ومياه الأمطار، ومياه دخيلة.

> *Traduction de l'auteure* : eaux polluées résultant de l'usage (ayant entraîné des modifications de leurs propriétés ou de leur composition), et qui sont ensuite classées, selon leur usage ou leur aptitude à la réutilisation, en eaux grises, eaux noires, eaux pluviales et eaux parasites.

[8] Le terme arabe est suivi de sa translittération IPA entre graphes et de sa traduction littéraire en parenthèses.

Le *Glossary of Shared Water Resources* (UN-ESCWA & BGR) privilégie la désignation مياه عادمة "*miyāh ʿādimah*" (*eaux privées de leur qualité d'origine*), définie comme suit :

> مياه تحتوي على بقايا مواد صلبة أو سائلة على أثر استخدامها في المنزل أو في عملية التصنيع أو في معمل صناعي، بحيث تصبح غير صالحة للاستخدام المباشر وتحتاج إلى معالجة.
>
> *Traduction de l'auteure* : eaux contenant des résidus de matières solides ou liquides résultant de leur utilisation domestique, de procédés de fabrication ou d'activités industrielles, devenant ainsi impropres à un usage direct et nécessitant un traitement.

La consultation d'experts du domaine a permis de confirmer que, dans un contexte technique et opérationnel lié aux infrastructures hydrauliques, la désignation شبكة مياه الصرف الصحي "*šabakat miyāh aṣ-ṣarf aṣ-ṣiḥḥī*" (*réseaux des eaux d'évacuation sanitaire*) correspond de manière plus adéquate au concept ontologique de *réseau de collecte des eaux usées* tel qu'il est modélisé dans le projet *StarWars*. Cet exemple justifie que la correspondance terminologique ne peut être établie que par une analyse croisée des usages attestés dans le corpus institutionnel, des définitions proposées par les ressources lexicographiques spécialisées et de l'expertise de domaine.

5. Accès interactif au corpus bilingue

Afin de faciliter l'exploitation et la recherche au sein du corpus français–arabe élaboré dans le domaine des *eaux usées*, une plateforme numérique a été développée. Comme illustré dans la figure 2, celle-ci permet aux utilisateurs d'interroger le corpus en français ou en arabe et fournit, pour chaque terme recherché, sa définition ainsi que son équivalent dans la langue cible. Cette approche contribue à la valorisation des ressources terminologiques spécialisées et à l'amélioration de l'accès à l'information.

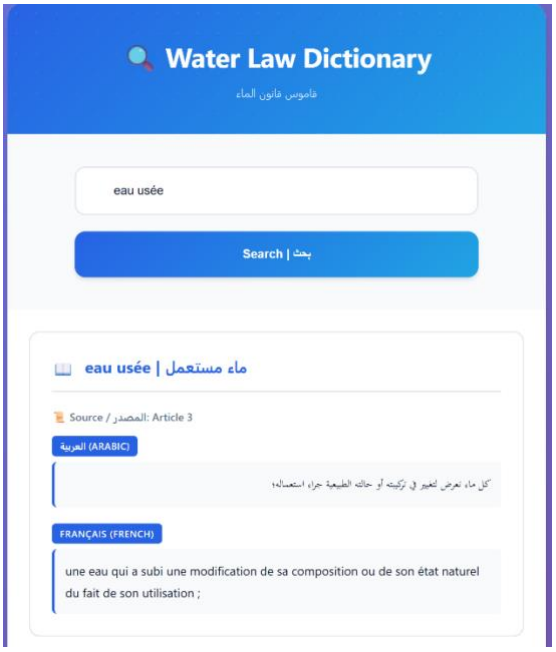


Figure 2 - ***Accès interactif au corpus bilingue***

6. Conclusion

Ce travail a proposé une approche méthodologique pour l'établissement de correspondances terminologiques français–arabe dans le domaine des eaux usées, fondée sur l'analyse d'un corpus institutionnel bilingue encodé en TEI XML.

L'encodage TEI et l'annotation terminologique constituent un support essentiel pour formaliser ces correspondances de manière explicite, traçable et réutilisable. Les résultats confirment ainsi la pertinence d'une démarche corpus-driven pour l'extension de ressources terminologiques multilingues vers l'arabe. Toutefois, l'analyse met en évidence les limites du corpus constitué, la nécessité d'un recours raisonné aux ressources dictionnairiques et l'importance du rôle des experts de domaine dans la validation des correspondances proposées.

Références

[1] Franco Alberto Cardillo, Franca Debole, Francesca Frontini, Mitra Aelami, Nanée Chahinian, and Serge Conrad. Novel benchmark for ner in the wastewater and stormwater domain. In 2025 IEEE 8th Congress on Information Science and Technology (CiSt), pages 226–231, 2025.

[2] Zola Mahlaza, C Maria Keet, Nanée Chahinian, and Batoul Haydar. On the feasibility of llm-based automated generation and filtering of competency questions for ontologies. In LDK'2025, 9 2025.

[3] Marianna Massa. Online multilingual technical and scientific terminology portals. Al-'Arabiyya, 54:107–133, 2021.

[4] Houria Tazi Sadeq. Water governance in a context of scarcity. Field Actions Science Reports, Special Issue 22:34–39, 2020. Online since 23 December 2020, accessed 19 January 2026.

[5] TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium, 2024. Accessed 19 January 2026.

[6] United Nations Economic and Social Commission for Western Asia. Glossary of shared water resources. Technical report, ESCWA, Beirut, 2013. Accessed 19 January 2026.

[7] Sue Ellen Wright and Gerhard Budin, editors. Handbook of Terminology Management. Volume 1: Basic Aspects of Terminology Management. John Benjamins Publishing Company, Amsterdam / Philadelphia, 1997.

[8] Sue Ellen Wright and Gerhard Budin, editors. Handbook of Terminology Management. Volume 2: Applications. John Benjamins Publishing Company, Amsterdam / Philadelphia, 2001.

Representing Zoological Knowledge through Ontoterminology in the *Rerum Medicarum Novae Hispaniae Thesaurus*

Vilela Ruiz Giuliana Elizabeth
Istituto di Linguistica Computazionale "A. Zampolli", CNR, Pisa, Italy; Department of Philology, University of Crete, Rethymno, Greece
giulianaelizabeth.vilelaruiz@ilc.cnr.it, philp1032@philology.uoc.gr

I. Introduction

This paper examines the zoological section of Francisco Hernández's *Rerum Medicarum Novae Hispaniae Thesaurus* as a case study for the ontoterminological modelling of pre-taxonomic scientific knowledge, showing how ambiguous and overlapping classifications can be represented in a text-driven, computationally interrogable form without imposing modern taxonomic constraints. This study forms part of an ongoing doctoral research project focused on the formalisation of Hernández's zoological corpus through terminological analysis and ontology-based methods.

Compiled after Hernández's expedition to New Spain (1570–1577), the *Thesaurus* is an early modern attempt to document New World biodiversity, integrating indigenous and European knowledge. It represents a key site of epistemic negotiation, where inherited naturalistic categories are confronted with previously unknown forms of life.

Despite the profound scholarly interest the work has generated over the centuries (Álvarez, 2000; Pardo-Tomás, 2009; Bustamante, 1997; project Il Tesoro Messicano[1]), research has focused on its botanical section, leaving the zoological component significantly underexplored, with only a limited number of isolated studies addressing New World fauna (Capanna, 2013). This study addresses this gap by focusing on the zoological descriptions transmitted in the Lincean edition of 1671. Despite its complex editorial history (Bellucci, 1998), the zoological section is treated here as a coherent descriptive corpus attributable to Hernández and suitable for the analysis of early modern classificatory practices.

The paper argues that Hernández's zoological classification does not constitute a stable taxonomic system, but rather a pre-taxonomic conceptual network whose structure can be made observable through systematic terminological analysis and ontological modelling. From an ontoterminological perspective (Roche, 2012; Roche & Papadopoulou, 2019; Papadopoulou & Roche, 2019), Hernández's zoological corpus is approached as a privileged site for analysing early modern classificatory practices in response to New World fauna, revealing a pre-taxonomic conceptual network whose structure becomes observable through systematic terminological variation and ontological modelling.

The paper is structured as follows: Section II outlines the terminological framework, describing the construction and organisation of the zoological vocabulary. Section III presents the ontology-based modelling of this terminology and the representation of pre-taxonomic

[1] The Tesoro messicano is an interdisciplinary research project promoted by the Accademia Nazionale dei Lincei and the CNR, launched in 2011 and dedicated to the study and dissemination of the *Rerum Medicarum Novae Hispaniae Thesaurus* (1651).

classifications. Section IV presents an ontological validation case study on the axolotl, showing how classificatory tensions are made explicit through query-based analysis.

II. Terminology

The study is based on a unified terminological inventory comprising 509 terms, covering the full range of zoological and animal-related vocabulary attested in the text. The construction of this inventory proceeded in two complementary phases. Across the corpus, different types of terms display distinct distributional patterns, with classificatory denominations concentrated in titles and opening segments, while anatomical, behavioural, habitat-related, and medico-therapeutic terms unevenly distributed across the descriptive sections of entries.

First, the general classificatory terms used to name and organise animals were identified manually through close reading of the text, with particular attention to the terminology employed in classificatory contexts. These terms were analysed in relation to their textual function.

Building on this initial core, an automatic terminological extraction expanded and consolidated the inventory through a structured protocol using a large language model (LLM). The domain (early modern zoology) and source (Rerum Medicarum Novae Hispaniae) were defined, and the model was instructed to act as a terminological expert. Extraction criteria targeted zoological terminology, including animal names as well as related terms referring to taxonomy, anatomical parts, behaviour, habitat, observable properties, medical uses, and human–animal relations. Constraints ensured that only attested terms were extracted. The prompt was iteratively refined (few-shot) and combined with manual validation and targeted corpus queries for context verification.

All extracted terms were consolidated into a curated dataset organised in a structured Excel file. For each term, the dataset records modelling-relevant metadata (term, lemma, part of speech, language, definition, textual reference, including chapter, and English translation), together with analytical fields, including a proposed ontological class and its corresponding superclass.

Starting from this curated table, a conversion script was developed to automatically transform the terminological data into a formal multilingual linguistic lexicon in OWL/RDF, compliant with the OntoLex-Lemon model (Cimiano et al., 2016), the reference standard for RDF-based lexical representation. According to this model, terms were represented as instances of `ontolex:LexicalEntry`, provided with language tagging and linked to their respective inflectional or formal variants through the `canonicalForm` and `otherForm` properties, as well as to basic linguistic annotations (e.g. `lexinfo:partOfSpeech`) and to a `skos:definition` in English.

The meanings associated with each lexical entry were encoded as instances of `ontolex:LexicalSense`, which functioned as a mediating layer between lexical entries and the extralinguistic concepts they refer to. Each `LexicalSense` was in turn connected to the corresponding conceptual entity via the `ontolex:reference` relation, while the concept itself received a formal description within a separately modelled conceptual ontology.

Within this broader terminological inventory, particular attention is devoted to the terminology of zoological classification. The analysis of classificatory terminology led to the identification of 84 terms corresponding to 52 zoological concepts, revealing a polynymic and multilingual classificatory system based on the coexistence of denominations fulfilling distinct descriptive, interpretative, and classificatory functions.

These terms are organised into three functional levels—primary terms, secondary terms, and internal variants—introduced as analytical labels for heuristic purposes and grounded in the textual position and discursive role of the terms. From an ontoterminological perspective, this functional (rather than taxonomic) distinction interprets terminological variation as an indicator of conceptual complexity rather than mere lexical instability (Cabré, 1999; Temmerman, 2000). Primary terms, predominantly in Nahuatl and occurring in chapter titles, function as guiding lemmas within the encyclopaedic architecture of the work and reflect a methodological choice that adopts indigenous classifications as an operative conceptual infrastructure, in contrast to the widespread European marginalisation of indigenous knowledge discussed by Segev (2021). Secondary terms—including Latin glosses, equivalents in modern European languages, and Nahuatl synonyms—perform a mediating function, linking New World fauna to the Western naturalistic tradition (Page, 2023). Internal variants, occurring exclusively within the body of the entries, include indigenous subspecies, local designations, and European analogues, documenting fine-grained zoological distinctions and the coexistence of multiple classificatory levels within individual descriptions.

The linguistic distribution of the terminology further highlights the multicultural nature of the work. Nahuatl is quantitatively dominant (approximately 52–54%), supporting the hypothesis of an autonomous indigenous taxonomy, while Latin (approximately 33%) plays an ordering and conceptualising role through references to classical zoology. Modern European languages, overall marginal (approximately 8%), mainly reflect practices of lexical circulation. Particularly significant are cases of Nahuatl–Latin hybridisation and descriptive Latin periphrases, which emerge where categorical translation proves problematic and indicate both the limits of inherited classificatory frameworks and the resistance of indigenous concepts to full taxonomic assimilation. This distribution is summarised in Figure 1.

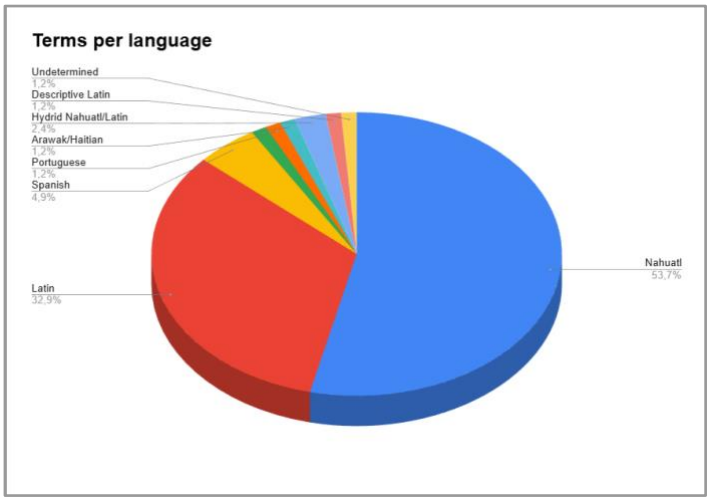


Figure1. Linguistic distribution of the analysed zoological terminology, with Nahuatl as the dominant language, followed by Latin and modern European languages.

A further structuring element is the recurrent bipartite formula *De [X], seu [Y]* in chapter titles, which can be interpreted as a pre-Linnaean device for conceptual organisation (Müller-Wille, 2007). This structure combines a primary Nahuatl term with a Latin or European interpretative element; the rare deviations from this pattern correspond to cases in which terminological mediation is unnecessary or ineffective. The heterogeneous functions of the second element—ranging from morphological and functional description to interpretative assimilation, symbolic epithet, or internal indigenous gloss—highlight the role of chapter titles as mediating interfaces between Mesoamerican vernacular taxonomies and early modern European natural history. Table 1 summarises the functional typology of the analysed bipartite titles.

Chapter title		Primary Term (Nahuatl)	Secondary Term (mainly in Latin)	Function of the second element	Typology
I	De Ayotochtli, seu Dasypode Cucurbitino, alijs Tatou, vel Armadillo dicto	Ayotochtli	Dasypode cucurbitino	Morphological description (hairy foot, "cucurbitine" shape)	Morphological / functional description
		Ayotochtli	Tatou / Armadillo	Assimilation to known American zoological category	Interpretation / assimilation
II	De Acaltetepon, seu Monoxillo mucronato, quod privatim Temacuilcahuya vocant, Lacerto Novae Hispaniae	Acaltetepon	Monoxillo mucronatus	Technical structural description	Morphological / functional description
		Acaltetepon	Temacuilcahuya	Indigenous variant, internal gloss	Internal indigenous description
		Acaltetepon	Lacerto Novae Hispaniae	Assimilation to known classical zoological category	Interpretation / assimilation
III	De Aquetzpalin, seu Crocodilis, quos alij Caimanes vocant	Aquetzpalin	Crocodilis	European equivalence	Interpretation / assimilation
		Aquetzpalin	Caimanes	Assimilation to known American zoological category	Interpretation / assimilation
IV	De Axolotl, seu Lusu Aquarum	Axolotl	Lusu Aquarum	Semantic–etymological translation of the compound	Interpretation / assimilation (etymological translation)
V	De Axin, seu vermium quorumdam pinguedine	Axin	Vermium quorumdam pinguedine	Description of material/product	Functional description
VI	De Ave Paradisea	—	Ave Paradisea	Direct conceptual assimilation to a known category	Interpretation / assimilation
VII	De Cercopithecis	—	Cercopithecis	Direct conceptual assimilation to a known category	Interpretation / assimilation
VIII	De Cozcaquauhtli, Regina Aurarum	Cozcaquauhtli	Regina Aurarum	Symbolic epithet	Symbolic interpretation / assimilation
IX	De Coyayahoal, seu Ichcaton	Coyayahoal	Ichcaton	Indigenous variant, internal gloss	Internal indigenous description
X	De Hoatzin, seu ave similem nomini edente vocem	Hoatzin	Ave similem nomini edente vocem	Description of behaviour/vocalisation	Functional description
XI	De Hoitzitzil, seu Ave varia	Hoitzitzil	Ave varia	Generic Latin categorisation	Simplifying assimilation
XII	De Hoitztlacuatzin, seu Tlacuatzin spinoso Hystrice Novae Hispaniae	Hoitztlacuatzin	Tlacuatzin spinoso	Internal comparative description (indigenous term + morphological qualifier)	Internal indigenous description
			Hystrix Novae Hispaniae	Assimilation to known classical zoological category	Interpretation / assimilation
XIII	De Manati	Manati	No terminological mediation	Single indigenous name	—
XIV	De Mazame, seu Ceruis	Mazame	Cervis	Assimilation to known classical zoological category	Interpretation / assimilation
XV	De Pollin	Pollin	No terminological mediation	Single indigenous name	—
XVI	De Tapayaxin, lacerto orbiculari Novae Hispaniae	Tapayaxin	Lacerto orbiculari Novae Hispaniae	Assimilation to a known classical zoological category + morphological description	Interpretation / assimilation + Morphological / functional description
XVII	De Teuhtlacotzauhqui, seu Domina serpentum	Teuhtlacotzauhqui	Domina serpentum	Symbolic epithet	Symbolic interpretation / assimilation
XVIII	De Tlaquatzin	Tlaquatzin	No terminological mediation	Single indigenous name	—
XIX	De Tzopilotl, seu Aura	Tzopilotl	Aura	Assimilation to known classical zoological category	Interpretation / assimilation
XX	De Yzquiepatl, seu Vulpecula…	Yzquiepatl	Vulpecula	Assimilation to known classical zoological category	Interpretation / assimilation
XXI	De Apum, Mellisq; Indici generibus	—	Apum	Direct conceptual assimilation to a known category	Descrizione funzionale

Table 1. The table maps the internal structure of chapter titles following the formula *De X, seu Y*, distinguishing the primary indigenous denomination (first element, predominantly Nahuatl) from the secondary (mainly in Latin).

III. Conceptual Modelling: Representing Pre-Taxonomic Zoological Classifications

Building on the extracted terminological data, this research develops a multilingual onto-terminological resource.

The conceptual formalisation of zoological knowledge has been developed on the basis of a set of analytical axes (Table 2) that reflect descriptive and functional criteria recurring in Hernández's work. These axes—relating, for example, to habitat (terrestrial, aquatic, aerial), edibility, and medico-therapeutic properties—do not correspond to modern taxonomic categories, but rather make explicit the pre-taxonomic classificatory logic of the text. On this basis, a domain ontology has been constructed, translating these criteria into formalised conceptual relations and properties.

Animal	1. Axis of analysis			2. Axis of analysis		3. Axis of analysis	
	Terrestrial	Aquatic	Volatile	Edible	Inedible	Curative power	Non-curative power
Ayotochtli	X			X			X
Acaltepeton		X					
Aquetzpalin		X		X			X
Axolotl		X		X		X	
Axin						X	
Cozcaquauhtli					X		X
Coyayahoal	X						
Hoatzin							X
Hoitzitzil							X
Hoitztlacuatzin	X			X			X
Manati		X		X			X
Mazame	X			X			X
Pollin							X
Tapayaxin	X						X
Teuhtlacotzauhqui	X						X
Tlaquatzin	X						X
Tzopilotl					X		X
Yzquiepatl	X						X
Quetzal Hoitzitzillin							X
Zocchio Hoitzitzillin							X
Xiuh Hoitzitzillin							X
Tozcacoz Hoitzitzillin							X
Yotac Hoitzitzillin							X
Tenoc Hoitzitzillin							X
Hoitzitzillin (generic)							X
Tzac Mazame	X			X			X
Tlamacaz quemazatl mazat	X			X			X
Aculliame	X						X
Quauhtlamazame	X			X			X
Tlalhuicamazame	X			X			X
Temamazame I	X			X			X
Macatl Chichiltic I	X			X			X
Teuhtlalmazame	X			X			X
Berendos	X			X			X
Macatl Chichiltic II	X			X			X
Temamazame II	X			X			X
Ecacoatl	X						X
Tlacuatzin	X			X			X
Ocumoctli	X						X

Tzinehuilin	X						X
Auras					X		X
Conepatl	X						X
Vulpecula puerilis (Yziquiepatl)	X						X
Tlalneuhtli	X						X
Xicotli	X						X
Quauhxicotli	X						X
Tlalpipioli	X						X
Cuicalmia Hoal	X						X
Quauhxicotli Pseudomellissam	X						X

Table 2. Array of descriptive and functional differences underlying the conceptual formalisation of zoological knowledge in Hernández's work.

The architecture of the ontology is inspired by the SIMPLE model (Lenci et al., 2000), grounded in Generative Lexicon Theory (Pustejovsky, 1995). By exploiting the Qualia Structure (formal, constitutive, telic, and agentive roles), the ontology captures multiple dimensions of zoological meaning beyond strict taxonomic hierarchies. Originally designed for lexicography, this model has proven particularly effective for organising specialised lexicons (Piccini et al., 2016; Piccini, Jama Musse Jama, Bellandi, & Vilela Ruiz, 2025; Piccini, Jama Musse Jama, Bellandi, Vilela Ruiz, & Bandini, 2025). Although the SIMPLE ontology is intended for general lexical applications rather than specific domains, the concepts it contains serve as high-level nodes that will be further specialised to address the particularities of Hernandez's zoological domain. This top-down approach allows for the model's refinement and expansion in response to the specific nuances of the data.

The knowledge formalised in this resource is, however, neither stable nor univocal, but emerges from the interaction—and at times the tension—between heterogeneous epistemic systems, including indigenous taxonomy and symbolic interpretation, classical natural history, and empirical observation. The resulting zoological representation is intrinsically complex and occasionally contradictory, and cannot always be reduced to a single coherent classificatory system.

This complexity has led to concrete modelling challenges: categories may overlap, classificatory criteria are heterogeneous, and some entities exhibit traits that are considered incompatible in contemporary zoology. The ontology therefore adopts a text-driven approach, avoiding rigid taxonomies and explicitly representing ambiguity and contradiction as structural properties of Hernández's system.

A paradigmatic example is provided by the case of the axolotl, which in Hernández's work is described and classified both as a fish and as a quadruped. Although such an attribution appears incoherent from the perspective of modern scientific knowledge, the introduction of restrictive constraints would result in a distortion of the historical knowledge being represented.

In the proposed ontology, axolotl is modelled as belonging to the intersection of the classes <Fish> and <Quadrupede>, in accordance with the classification attributed to it by Hernández. No disjointness axioms are introduced between these classes, in order to avoid any ontological incompatibility and to allow the axolotl to be represented as simultaneously belonging to both categories, without enforcing a modern taxonomic normalisation.

From a logical perspective, this modelling choice is formalised by means of a class equivalence axiom:

$$Axolotl \equiv Fish \sqcap Quadrupede$$

Alternatively, the same information can be expressed through two subsumption axioms:

$$Axolotl \sqsubseteq Fish$$

$$Axolotl \sqsubseteq Quadrupede$$

The absence of axioms of the form Fish disjointWith Quadrupede ensures that this modelling remains logically consistent, faithfully reflecting the historical classification reported in the source.

IV. Ontological Validation through Querying in the Axolotl Case Study

This section shows how ontological querying validates the proposed modelling choices by enabling the retrieval of the full range of classificatory attributions associated with the axolotl and by making explicit the coexistence of apparently contradictory properties (see Table 3).

COMPETENCY QUESTION	PSEUDO-SPARQL	QUERY RESULTS	KNOWLEDGE RETRIEVED

Which zoological classes does Hernández assign to the Axolotl?	**SELECT** ?classLabel
	WHERE {
	:axolotl a ?class .
	?class owl:Class ;
	rdfs:label ?classLabel .
	}
		Fish; Quadrupede	Retrieve dual classification of the axolotl

Which zoological entities belong to more than one taxonomic class?	**SELECT** ?animal
	(GROUP_CONCAT(DISTINCT STR(?class);
	separator=" | ") AS ?classes)
	WHERE {
	?animal a ?class .
	}
	GROUP BY ?animal
	HAVING (COUNT(DISTINCT ?class) > 1)
		Axolotl → Fish | Quadrupede	Identify animals with overlapping classifications and their assigned classes

Which linguistic designations are associated with the concept of the Axolotl in the text?	# Lexicalisations (labels/forms) linked to the concept :AXOLOTL
	SELECT DISTINCT ?term ?lang
	WHERE {
	?entry a ontolex:LexicalEntry ;
	rdfs:label ?term ;
	ontolex:sense ?sense .
	?sense ontolex:reference :axolotl .
	BIND (lang(?term) AS ?lang)
	}
		Ayotochtli (nah); Axolotl (lat); Lusus Aquarum (lat)	Retrieve linguistic designations associated with the same zoological referent

Table 3. Each competency question is accompanied by a pseudo-SPARQL query illustrating how the ontology can be interrogated to retrieve the relevant information.

The ambiguity observed in the case of the axolotl should not be interpreted as a mere artefact of modelling, but rather as the manifestation of a significant conceptual structure. The dual nature attributed to the animal is already embedded in Nahuatl terminology and mythology, where it embodies a principle of transformation and transition between different states (Moreno, 1969; Renard, 2010). Interestingly, Contemporary biology has subsequently provided an explanation for this duality through the phenomenon of neoteny, showing how the axolotl may retain aquatic larval characteristics or undergo morphological transformation depending on environmental conditions (Smith, 1969). What appears as a contradiction within a modern taxonomic framework thus corresponds to an actual biological reality, intuitively grasped both by Hernández and by indigenous knowledge. In this sense, the ontology highlights how pre-taxonomic classifications can convey forms of knowledge that do not align with later hierarchical systems, yet are capable of anticipating scientific explanations developed only in

the modern period. The ontological representation therefore does not aim to reproduce contemporary zoological taxonomy, but rather to model a historically situated conceptual system, in which complexity, hybridity, and uncertainty are constitutive elements and in which interpretative tensions are made explicit.

References

Álvarez, J. E. C. (2000). *Francisco Hernández: El descubrimiento científico del Nuevo Mundo.* Diputación Provincial de Toledo.

Bellucci, A. P. (2014). *Breve historia editorial de la obra de Francisco Hernández, "Historia natural de la Nueva España". Revista del Centro de Investigación de la Universidad La Salle, 3*(11), 251–264. https://doi.org/10.26457/recein.v3i11.391

Bustamante, J. (1997). Francisco Hernández, Plinio del Nuevo Mundo: Tradición clásica, teoría nominal y sistema terminológico indígena en una obra renacentista. In B. Ares Queija & S. Gruzinski (Eds.), *Entre dos mundos: Fronteras culturales y agentes mediadores* (pp. 243–268). Escuela de Estudios Hispano-Americanos.

Cabré, M. T. (1999). *Terminology: Theory, methods and applications* (J. C. Sager, a cura di; A. De Cesaris, Trad.). John Benjamins Publishing Company.

Capanna, E. (2013). Observatio e admiratio: I sorprendenti animali del Nuovo Mondo. Tesoro messicano: *Libri e saperi tra Europa e Nuovo mondo*, 120, 155–173.

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). Lexicon model for ontologies: Community report. W3C Ontology-Lexicon Community Group. https://www.w3.org/community/ontolex/

Hernández, F., Cesi, F., Colonna, F., Deversin, B., Faber, J., Greuter, J. F., Mascardi, V., Masotti, Z., Recchi, N. A., & Terentius, J. (1651). *Rerum medicarum Novae Hispaniae thesaurus, seu, Plantarum animalium mineralium Mexicanorum historia*. Roma, Italia: Ex typographeio Vitalis Mascardi.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., ... & Zampolli, A. (2000). SIMPLE: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4), 249–263.

Moreno, R. (1969). *El axólotl. Estudios de Cultura Náhuatl, 8*, 157–173. https://nahuatl.historicas.unam.mx/index.php/ecn/article/view/78542.

Müller-Wille, S. (2007). Names and numbers: "Natural history" and the Linnaean reform. *Studies in History and Philosophy of Biological and Biomedical Sciences, 38*(3), 593–615.

Page, J. (2023). *Decolonial Ecologies: The Reinvention of Natural History in Latin American Art*. Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0339

Pardo Tomás, J. (2009). El protomédico Francisco Hernández en Nueva España (1570-1577). *Crónicas: Revista trimestral de carácter cultural de La Puebla de Montalbán*, (9), 9–11.

Piccini, S., Ruimy, N., & Giovannetti, E. (2016). Le lexique électronique de la terminologie de Ferdinand de Saussure : Une première. In D. Trotter, A. Bozzi, & C. Fairon (Eds.), *Actes du XXVIIe Congrès international de linguistique et de philologie romanes, Nancy, 15-20 juillet 2013*. Section 16 : Projets en cours ; ressources et outils nouveaux (pp. 255–267). ATILF.

Piccini, S., Jama Musse Jama, B., Bellandi, A., Vilela Ruiz, G. E., & Bandini, M. (2024). Formalising the terminology of Ugo Ferrandi's notebooks: A journey into pre-colonial Somali culture. In the *process of being published in a volume by Brill.*

Piccini, S., Jama Musse Jama, B., Bellandi, A., & Vilela Ruiz, G. E. (2024). Echi del passato. Un'indagine sulla terminologia degli strumenti nelle terre dei Somali del XIX secolo. In *the process of being published in a volume by Peter Lang.*

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge MA: The MIT Press.

Renard, J. B. (2010). *L'axolotl. De la controverse scientifique au mythe littéraire. Sociétés, 108*(2), 19-32.

Roche, C. (2012). Ontoterminology: How to unify terminology and ontology into a single paradigm. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012) (pp. 2626–2630). European Language Resources Association.

Roche, C., & Papadopoulou, M. (2019). *Mind the gap: Ontology authoring for humanists*. In Proceedings of the 1st International Workshop for Digital Humanities and Their Social Analysis (WODHSA) – Episode V: The Styrian Autumn of Ontology. Joint Ontology Workshops, Medical University of Graz, Austria.

Papadopoulou, M., & Roche, C. (2019). *Building ontology-based dictionaries for Greek material culture terms*. In Proceedings of the 1st International Workshop on Open Data and Ontologies for Cultural Heritage. Rome, Italy.

Segev, R. (2021). "For the sciences migrate, just like people": The case of botanical knowledge in the early modern Iberian empires. *Perspectives on Science, 29*(4), 732–756.

Smith, H. M. (1969). *The Mexican axolotl: Some misconceptions and problems*. *BioScience, 19*(7), 593–615. https://doi.org/10.2307/1294932

Temmerman, R. (2000). *Towards new ways of terminology description: The sociocognitive approach*. John Benjamins Publishing Company.

Integrating plain language into terminology databases: design and data modelling considerations

Tanja Wissik, Austrian Academy of Sciences, Tanja.Wissik@oeaw.ac.at
Elena Chiocchetti, Eurac Research, Elena.Chiocchetti@eurac.edu

1. Introduction

Plain language (or clear language) is defined as "communication in which wording, structure and design are so clear that intended readers can easily find what they need, understand what they find, and use that information" (ISO 24495-1:2023, 3.1). It focuses on effectively combining word choices, content organisation and text design to support access to information following four guiding principles: relevance, findability, understandability and usability (ISO 24495-1:2023, 4).

Plain language texts should select and foreground information that is needed by and important for the intended readers according to the purpose of the document (relevance). The structure of the text should allow readers to quickly determine what it is about and where they will find what they are looking for (findability). Sentences should be simple, short and generally in the active voice; familiar words should be preferred to specialised words and used consistently; terms should be explained if they cannot be avoided; acronyms should be spelled out (understandability). The effectiveness of plain language texts should be checked against the needs and expectations of the intended readership (usability).

Plain language as a simplified language variety is aimed at everyone who is not an expert in the topics being addressed in a text. It targets the general public, particularly lay persons with average reading competencies (Maaß 2025, 1). The growing importance of plain communication (Rocco 2022, 156) is shown by the recent publication of two international standards, one on the governing principles of plain language (ISO 24495-1:2023) and one on legal communication (ISO 24495-2:2025). Additional parts on science writing (ISO 24495-3) and document design (ISO/WD 24495-5) are under development. Some countries have passed laws that foster the use of plain language in official communication (Lutz 2024, 321-322), for example the United States (Plain Writing Act 2010) and Sweden (Language Act 2009).

Some suggestions for plain language writing are directly related to the type of information available in terminology databases (TDBs). Bowker (2024, 175) gives an overview of the advantages of combining plain language principles and neural machine translation. She focuses on several aspects that relate to information generally stored in TDBs used for translation purposes. These include the suggestion to avoid or explain technical terms, avoid acronyms and abbreviated forms and use terminology consistently. Felici et al. (2023) report on a case study aimed at simplifying statistical terminology for the general public. They explain that domain experts and translators struggle with striking a balance between precision and plain language requirements during intralingual translation. They would clearly benefit from resources that guide them in "making terminology accessible" (Felici et al. 2023). Wissik and Chiocchetti

(forthcoming) pinpoint how TDBs can be successfully employed by plain language writers and translators. The authors focus on the usefulness of concept definitions, synonyms, full forms of acronyms, initialisms and abbreviations, data categories specifying usage restrictions (e.g. /register/), related concepts and contexts as examples of use.

In this contribution, we argue that it would be an advantage to use TDBs for plain language purposes and show that existing TDBs could easily be adapted to store plain language content.

2. Resources for Plain Language

There are numerous handbooks, guidelines and tools for plain language writing in several languages (see International Plain Language Federation 2025). This section presents an overview of selected resources that contain specialised terminology and plain language alternatives, such as plain language glossaries, dictionaries and thesauri. Most resources focus on the legal or medical domain, because members of these professions often "have to write directly to lay people" (Cutts 2011, 5). Finance is another relevant domain (e.g. Te Ara Ahunga Ora Retirement Commission 2023). Most resources are monolingual but some either have more than one language version (e.g. the Just Plain Clear Glossary n.d. is available in English, Spanish, Portuguese, Chinese and Burmese) or provide translations (e.g. Open Advocate 2016 has Spanish explanations of English legal terms).

Some resources mainly deal with verbs, adjectives and phraseology (e.g. Health Research for Action n.d.) while others focus on terminology (e.g. GET-IT Glossary n.d.). Many resources publish only plain language synonyms, while others also provide concept explanations and/or definitions. Older resources (e.g. Australian Government 1993) are available as printed books or online (searchable) PDFs. More recent ones often provide a search interface or an alphabetical list of entries. Very few resources publish the sources of their definitions and examples.

Most are not terminological resources. Hence, they do not respect the terminological principles of term autonomy and data elementarity (ISO 26162-1:2019: 3.2.13-15). For example, abbreviations are typically mentioned in brackets next to the full form. Online alphabetical search lists generally do not allow accessing information stored under the headword both via the specialised and via the plain language term. Filtering functions are a rare feature, and it is generally not possible to export only a set of terms or a specific data category (e.g. only headword and definition). Plain language synonyms are often listed together, for example in the same cell of a column, not kept separate. Equivalents in other languages are mostly accessible only via the headword in the main language of the glossary. A limited number of glossaries is available as a CSV file (e.g. Open Advocate 2016, National Center for State Courts n.d.) so that they could possibly be used within professional authoring tools and computer-assisted translation tools.

3. Plain language data in TDBs

This section addresses design and data modelling considerations relevant to the integration of plain language equivalents, plain language definitions and other explanatory information contained in plain language resources into a TDB. Rather than proposing the design of a TDB

ex novo, the discussion is grounded in the assumption that an existing TDB can be adapted to accommodate plain language content. The intended users are not end users or lay readers, but plain language writers and translators.

3.1 Plain language equivalents

There are several possibilities of including plain language equivalents into an existing TDB, which we discuss in this contribution. For example, for the term "diagnosis", the plain language glossary Health Research for Action (n.d.) suggests "cause of your illness" as a possible plain language equivalent. One option is to record the plain language equivalent in the data category /common name/. This data category is included in the DatCatInfo (n.d.) data category repository where it is defined as "[a] synonym for an international scientific term that is used in general discourse in a given language." Therefore, this category may be suitable for plain language equivalents following ISO 24495-1 (2023, 5.3.2) that recommends "choos[ing] familiar words" instead of jargon. For example, in TriMed (Vezzani and Di Nunzio 2020), a TDB for medical terms designed to meet the needs of physicians, translators and patients, the data category /common name/ is used to cater to the needs of patients.

Another option is to treat plain language equivalents as synonyms of the specialised term and use the data category /term/. For example, in *bistro*, an information system for legal terminology (Ralli and Andreatta 2018), the entry for *teste* (witness) contains the more commonly used Italian synonym *testimone*. However, no specific information on register or usage is given. By adding a /usage note/ "plain language" or selecting "plain" within a pick list of /term type/s, it would be possible to easily distinguish plain language equivalents from other terms. A further possibility for including plain language equivalents – which would, however, require more substantial changes to the structure of the TDB – would be to treat plain language as a separate language variety, as it can be argued that plain language is a functional language variety (Romary 2025). However, despite the existence of ISO 24495-1 (2023), so far there is no international language identifier (ISO 639:2023) for plain language that would allow it to be treated as a distinct language variety. This creates practical difficulties and requires workarounds when implementing this solution (Ralli 2025).

3.2 Plain language definitions

TDBs usually contain definitions. Plain language resources likewise include definitions written in plain language and also plain language documents contain definitions. To ensure consistency across plain language documents, it would therefore be useful to store also plain language definitions in a TDB (Wissik and Chiocchetti, forthcoming).

According to terminological principles, there should be only one definition for each concept. The definition can be placed either at the concept or at the language level. However, some authors (Kalliokuusi and Varantola 1998, Varantola 2002) have argued that terminological definitions are of limited usefulness for people "unfamiliar with the concept system and knowledge structure of the domain" (Varantola 2002, 41). Instead, it would be "much more user-friendly to define special field concepts in different ways for users with different knowledge backgrounds" (Varantola 2002, 41). Including plain language definitions in TDBs would be in line with this suggestion. There are already examples of TDBs that contain multiple

definitions from different sources targeted at distinct audiences, such as UniVieTerm (Heinisch 2023, 34).

This raises the question of which data categories should be used to include plain language definitions and how these can be distinguished from terminological definitions. To address this question, we examine examples of both terminological and plain language definitions. For the term "epidemiologist", TERMIUM Plus, the multilingual TDB maintained by the Translation Bureau at Public Services and Procurement Canada, provides the following definition: "One who studies epidemic diseases". In the Clinical Research Glossary (Baedorf Kassis et al. 2022) the following plain language definition can be found: "A person who studies where, why, how often, and to what populations health concerns and diseases happen." Another example: for the term "diagnosis", TERMIUM Plus offers the following definition: "The determination of the nature of a disease by means of its signs and symptoms and the results of investigations". In TERMDAT, the multilingual TDB of the Swiss Federal Administration, the term "diagnosis" is defined as follows: "Determination of the nature of a disease or condition, or the distinguishing of one disease or condition from another". The plain language GET-IT Glossary (Moberg et al. 2018) publishes the following plain language definition: "The recognition or identification of a health condition based on symptoms, signs, or test results".

One possibility for including plain language definitions in a TDB would be to use the data category /explanation/, "[a] statement that describes and clarifies a concept and makes it understandable, but does not necessarily differentiate it from other concepts" (DatCatInfo n.d.), since plain language definitions do not always fulfil the criteria for terminological definitions or describe the concepts in a way that they can be differentiated from other concepts. Another option would be to create a new data category /plain language definition/, which does not exist yet in the data category repository. Finally, if the plain language equivalent is stored at the language level as a separate language variety, the plain language definition could be simply added as /definition/.

3.3 Other data categories

Plain language resources do not only contain terms, plain language equivalents and definitions, they also include other types of information, such as usage examples (contexts), supporting images or graphics and notes. All these data categories (i.e. /context/, /figure/, /note/) are available in DatCatInfo (n.d.). For contexts, the modelling options are similar to the options discussed for plain language equivalents. Further details will be given in the presentation.

4. Discussion

So far, the potential of TDBs for storing and managing plain language data has not been fully exploited. However, integrating plain language data in existing TDBs offers several advantages. First, all relevant information is available in a single tool, meaning that plain language writers or translators do not need to switch between different resources when understanding complex content and drafting or translating plain language texts. Second, compliance with terminology exchange standards (e.g. TBX defined in ISO 30042:2019) allows the data to be seamlessly used in authoring tools, terminology consistency checkers and similar tools. Third, TDBs can be used to export selected data categories to create plain language glossaries and publish them

alongside plain language texts. Finally, large language models (LLMs) used for text simplification (e.g. Färber et al. 2025; McMinn et al. 2025) may benefit from access to clearly marked plain language equivalents, definitions and other data stored in TDBs, whether directly via an API, through termbase exports used for retrieval-augmented generation (RAG) or as examples provided in prompts. A disadvantage of storing plain language data in TDBs is that multilingual term entries that already contain numerous synonyms, abbreviations and term variants may become excessively long. However, most TDBs provide filtering options that allow users to select only the relevant languages, language varieties or data categories.

In discussing the options for integrating plain language content into existing TDBs, we also addressed the lack of an official abbreviation or ISO language identifier for plain language. Even if treating plain language as a separate language variety is debatable for various reasons (e.g. what is considered "plain" may vary according to the target audience and purpose of a text), an official abbreviation for plain language would still be useful. It would allow metadata to indicate, in a standardised way, whether a terminological resource (also) contains plain language data.

5. Conclusion

In this contribution, we have argued that TDBs are still underused for plain language content. However, we have shown that it would be possible to easily adapt existing TDBs to plain language content without major modelling efforts. Furthermore, we argue that a standardised abbreviation for plain language would help describe language resources containing plain language data in an adequate and consistent way.

6. AI declaration

The authors have used a generative AI tool (ChatGPT-5.2) to check the contribution and improve the grammar, language, style and readability. They have then reviewed and edited the text as needed and take full responsibility for the content.

7. References

Australian Government. 1993. *Plain English Manual. Australian Government*. Office of Parliamentary Counsel.

Baedorf Kassis, Sylvia, Sarah A. White, and Barbara E. Bierer. 2022. 'Developing a Consensus-Driven, Plain-Language Clinical Research Glossary for Study Participants and the Clinical Research Community'. *Journal of Clinical and Translational Science* 6 (1): 1–7. https://doi.org/10.1017/cts.2022.12.

Bowker, Lynne. 2024. 'Plain Language in the Age of Neural Machine Translation: An Opportunity for Translators'. In *New Advances in Translation Technology. Applications and Pedagogy*, edited by Yuhong Peng, Huihui Huang, and Defeng Li. Springer.

Cutts, Martin. 2011. *Plain English Lexicon. A Guide to Whether Your Words Will Be Understood*. 2nd edn. Plain Language Commission. https://clearest.co.uk/wp-content/uploads/2021/09/Plain_English_LEXICON_June_2011.pdf.

DatCatInfo. n.d. ‘DatCatInfo Data Category Repository’. https://datcatinfo.termweb.eu/client/twd.

‘“diagnosis” | GET-IT Glossary’. n.d. Accessed 7 January 2026. https://getitglossary.org/term/diagnosis.

‘“diagnosis” | TERMDAT’. n.d. Accessed 7 January 2026. https://www.termdat.bk.admin.ch/search/entry/37512?s=Diagnose&sl=2,3&tl=2,6,7,8,3%3 D. ‘

“diagnosis” | TERMIUM Plus’. 2009a. October 8. https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=diagnosis&codom2nd_wet=NQ#resultrecs.

‘“epidemiologist” | Clinical Research Glossary’. n.d. *The Multi-Regional Clinical Trials Center of Brigham and Women’s Hospital and Harvard*. Accessed 7 January 2026. https://mrctcenter.org/search/epidemiologist/. ‘

“epidemiologist” | TERMIUM Plus’. 2009b. October 8. https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng&i=1&srchtxt=EPIDEMIOLOGIST&codom2nd_wet=1#resultrecs.

‘“explanation” | DatCatInfo’. n.d. Accessed 7 January 2026. https://datcatinfo.termweb.net/en/dict/202/495408/1947066?lang=xho&target=eng%2Cfra %2Czul&section=0&domain=0&term=DC-0223&config=0.

Färber, Michael, Parisa Aghdam, Kyuri Im, Mario Tawfelis, and Hardik Ghoshal. 2025. ‘SimplifyMyText: An LLM-Based System for Inclusive Plain Language Text Simplification’. *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV (Berlin),* 418–24. https://doi.org/10.1007/978-3-031-88717-8_32.

Felici, Annarita, Paolo Canavese, Giovanna Titus-Brianti, and Cornelia Griebel. 2023. ‘Plain Language at the Swiss Federal Statistical Office: The Challenges of Terminology When Writing for the General Public’. *inTRAlinea* 25. https://www.intralinea.org/print/article_specials/2638.

GET-IT Glossary. n.d. ‘GET-IT Glossary. Plain Language Definitions of Health Research Terms’. Accessed 27 July 2025. https://getitglossary.org/.

Health Research for Action. n.d. ‘Plain Language Word List’. Health Research for Action. https://multco-web7-psh-files-usw2.s3-us-west-2.amazonaws.com/s3fs-public/PlainLanguageWordList.pdf.

Heinisch, Barbara. 2023. ‘Terminological Usability – Adapting Terminological Databases to Different User Groups According to Usability Principles: The Case of UniVieTerm’. *Terminology Science & Research / Terminologie : Science et Recherche* 26: 24–44.

International Plain Language Federation. 2025. ‘Bibliography ISO Standard - International Plain Language Federation’. June 12. https://www.iplfederation.org/bibliography-iso/.

ISO 639. 2023. ‘Code for individual languages and language groups’. International Organization for Standardization.

ISO 24495-1. 2023. ‘Plain Language. Part 1: Governing Principles and Guidelines’. International Organization for Standardization.

ISO 24495-2. 2025. ‘Plain Language. Part 2: Legal Communication’. International Organization for Standardization.

ISO 24495-2. Forthcoming. ‘Plain Language. Part 3: Science Writing’. International Organization for Standardization. ISO 26162-1. 2019. ‘Management of terminology resources — Terminology databases’. International Organization for Standardization.

ISO 30042. 2019. Management of terminology resources — TermBase eXchange (TBX). International Organization for Standardization.

ISO/WD 24495-5. Forthcoming. ‘Plain Language. Part 5: Document Design’. International Organization for Standardization.

Kalliokuusi, Virpi, and Krista Varantola. 1998. ‘From General Dictionaries to Terminological Glossaries’. In *Actes EURALEX ’98 Proceedings*, edited by Thierry Fontenelle, Philippe Hiligsmann, Archibald Michiels, André Moulin, and Siegfried Theissen. Liège.

Language Act, 2009:600 (2009). https://www.regeringen.se/contentassets/36dc86bb939c4bc79eb44417b0852b40/sprakla g-pa-engelska/.

Lutz, Benedikt. 2024. ‘Bürgernahe und Leichte Sprache in Österreich’. In *Sprachenpolitik in Österreich: Bestandsaufnahme 2021*, edited by Eva Vetter, Rudolf de Cillia, and Martin Reisigl. De Gruyter.

Maaß, Christiane. 2025. ‘DIN/ISO und die leere Mitte zwischen Leichter und Einfacher Sprache: Eine Bestandsaufnahme und ein Plädoyer für die Leichte Sprache Plus’. https://www.researchgate.net/publication/397897907_DINISO_und_die_leere_Mitte_z wisc
hen_Leichter_und_Einfacher_Sprache_Eine_Bestandsaufnahme_und_ein_Pladoyer_fur _ die_Leichte_Sprache_Plus.

McMinn, David, Tom Grant, Laura DeFord-Watts, et al. 2025. ‘Using Artificial Intelligence to Expedite and Enhance Plain Language Summary Abstract Writing of Scientific Content’. *JAMIA Open* 8 (2): 1–12. https://doi.org/10.1093/jamiaopen/ooaf023.

Moberg, Jenny, Astrid Austvoll-Dahlgren, Shaun Treweek, et al. 2018. ‘The Plain Language Glossary of Evaluation Terms for Informed Treatment Choices (GET-IT) at www.getitglossary.org’. *Research for All* 2 (1): 106–21. https://doi.org/10.18546/RFA.02.1.10.

National Center for State Courts. n.d. ‘National Center for State Courts (NCSC) Glossary’. Accessed 19 December 2025. https://plgclientprod.azurewebsites.net/.

Open Advocate. 2016. 'Basic English Legal Glossary with Spanish Explanations'. https://www.openadvocate.org/readclearly/.

Plain Writing Act, Pub. L. Nos 111–274 (2010). https://www.govinfo.gov/content/pkg/PLAW-111publ274/pdf/PLAW-111publ274.pdf.

Ralli, Natascia. 2025. 'Managing Language Varieties: Examples From Legal Terminology Work'. In *Proceedings of the 4th International Conference on Multilingual Digital Terminology Today (MDTT 2025)*, edited by Federica Vezzani, Giorgio Maria Di Nunzio, Elpida Loupaki, Georgios Meditskos, and Maria Papoutsoglou. Thessaloniki. https://ceur- ws.org/Vol-3990/short11.pdf.

Ralli, Natascia, and Norbert Andreatta. 2018. '*bistro* – ein Tool für mehrsprachige Rechtsterminologie'. trans-kom 11 (1): 7–44.

Rocco, Goranka. 2022. 'Leichte Sprache und einfache Sprache. Syntaktische Aspekte im Vergleich'. In *Syntax in Fachkommunikation*, edited by Ursula Wienern, Tinka Reichmann, and Laura Sergo, vol. 163. Forum für Fachsprachenforschung.

Romary, Laurent. 2025. 'International standards for the identification and the description of languages and their varieties'. In *Harmonizing language data: Standards for linguistic resources*, edited by Piotr Bański, Ulrich Heid, and Laura Herzberg. De Gruyter.

Te Ara Ahunga Ora Retirement Commission. 2023. *De-Jargoning Money. A Financial Glossary of Plain Language for the Finance Sector and Beyond*. Te Ara Ahunga Ora Retirement Commission. https://assets.retirement.govt.nz/public/Uploads/National-Strategy/De-jargoning-Money_Glossary_TAAO.pdf.

UnitedHealth Group. n.d. 'Just Plain Clear Glossary'. Accessed 7 January 2026. https://www.justplainclear.com/en.

Varantola, Krista. 2002. 'Use and Usability of Dictionaries: Common Sense and Context Sensibility?' In *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, edited by Marie-Hélène Corréard. Euralex.

Vezzani, Federica, and Giorgio Maria Di Nunzio. 2020. 'Methodology for the Standardization of Terminological Resources. Design of TriMED Database to Support Multi-Register Medical Communication'. *Terminology* 26 (2): 265–97. https://doi.org/10.1075/term.00053.vez.

Wissik, Tanja, and Elena Chiocchetti. Forthcoming. 'Terminology Databases as Resources to Support Plain Legal Language'. In *International Handbook of Legal Language and Communication. From Text to Semiotics,* edited by Anne Wagner. Springer.

Poster Session 1

Lexical borrowing from French in Russian, Turkish and Swedish: a diachronic, typological and contrastive study

Margarita Chernysheva, Paris Nanterre University, mchernys@parisnanterre.fr
Sabine Lehmann, Paris Nanterre University, slemhajeb-lehmann@parisnanterre.fr
Iris Eshkol-Taravella, Paris Nanterre University, ieshkolt@parisnanterre.fr

It would hardly be an overstatement to say that, in the 21st century, an average speaker of any European language encounters complaints about the inescapable presence of English words and expressions in their native language on a daily basis. Concerns related to this topic are expressed by various types of media as well as by the governments of numerous countries[1] in Europe and all around the world. Some sources go as far as to paint a grim picture of the future where numerous languages are forever mangled by the merciless jaws of English, which seems to provide loanwords at an unprecedented rate. But is the situation really as dire and as unique?

On a world scale, there exist numerous languages that, for centuries, served as a source for lexical neologisms of all kinds due to cultural and/or economic domination of the populations speaking these languages in their corresponding regions (Deroy, 1956, ch. V). An example of Chinese as a language that left a noticeable mark on the languages of the neighbouring countries can be given. Moreover, in the colonial era, even geographical distance ceased to be an obstacle for language expansion; this development can actually be considered as one of the phenomena that made the modern florescence of English possible. So, even though English is greatly facilitated by the technological advances that humanity made in recent decades, it is far from being the first language to enjoy a noticeable influence over numerous other ones. However, the world languages are extremely flexible and ever-changing, so even the all-encompassing impact made nowadays can easily disappear in the future.

The two main goals of our research are, thus, to prove this point and to observe how the lexemes of the once prized source language and their usage change once they enter the target language, especially after it breaks free from the source language's influence. We bring attention to the immense respect and admiration that the French language (as well as the French culture and France as a country) enjoyed in Europe less than two centuries ago and underline the fact that they resulted in large-scale lexical borrowing in multiple semantic fields, such as science, politics, art, etc. Therefore, we look for the traces left by the French influence in three typologically different European languages, namely Russian, Turkish and Swedish, in order to answer the following research questions:

- How do the lexemes borrowed from French to these languages in the 18th-19th centuries look and behave nowadays? Are they still in use? In which ways have they changed?

[1] To illustrate, his recent law passed by the current Russian government prohibits the excessive use of foreign words and loanwords, especially those that have well-established Russian counterparts: https://lenta.ru/news/2023/02/16/words/. The main origin of such words nowadays is most certainly English.

- What are the differences between the borrowing patterns of the Russian, Turkish and Swedish languages regarding French lexemes?
- Can what happened to these lexemes teach us something about the process of linguistic borrowing in general or help us predict the destiny of the lexemes borrowed nowadays from the English language?

We aim to answer these questions with the help of sets of historical and modern mixed corpora of the languages in question and by developing a Python program that will aid us in automatically detecting and describing French loanwords found in these corpora. The program in question will be based on a combination of the pre-existing lexical resources, such as digital and paper dictionaries, and Python modules that facilitate shedding light on different linguistic features of the lexical borrowings (pronunciation, morphology, orthography, semantic field, etc.). Such a combination will allow us to approach the research questions from different angles and, hopefully, to minimise the percentage of the loanwords that slip through the cracks and remain undetected.

Our research is thus situated on the crossroads between numerous linguistic fields, such as computational, historical and contrastive linguistics.

As a result of our inquiry, we hope to establish a robust and universal methodology for conducting similar research in the field of comparative lexical borrowing studies. As to deliverables, we strive to create a digital lexical resource[2] that could be useful for further investigations related to the subject of lexical borrowing in general and French loanwords in particular. The use of this resource in the field of foreign language education (in order to facilitate teaching French vocabulary to the speakers of Russian, Turkish and Swedish and vice versa) and in studies related to intercomprehension in the European cultural space is also plausible.

Bibliography:

Deroy, L. (1956). *L'Emprunt linguistique*. Presses universitaires de Liège. https://books.openedition.org/pulg/665

Furiassi, C., Pulcini, V., & Rodríguez González, F. (Eds.). (2012). *The Anglicization of European Lexis*. John Benjamins Publishing Company.

Görlach, M. (Ed.). (2001). *A dictionary of European anglicisms: A usage dictionary of anglicisms in sixteen European languages*. Oxford University Press.

Nath, A., Mahdipour Saravani, S., Khebour, I., Mannan, S., Li, Z., & Krishnaswamy, N. (2022). A Generalized Method for Automated Multilingual Loanword Detection. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli,

[2] We are currently still deliberating on the format of the aforementioned resource, but it will most likely be some sort of a freely accessible database.

H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Éds.), *Proceedings of the 29th International Conference on Computational Linguistics* (p. 4996-5013). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.442/

Sablayrolles, J.-F., & Jacquet-Pfau, C. (2008). Les emprunts : du repérage aux analyses. Diversité des objectifs et des traitements. *Neologica : revue internationale de la néologie*, *2*, 19-38.

Thomason, S. G., & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics* (1991 edition). University of California Press. https://doi.org/10.1525/9780520912793

Weinreich, U. (1966). *Languages in contact: findings and problems* (4th edition). The Hague Paris: Mouton.

Zhang, L., Fabri, R., Nerbonne, J., & Nerbonne, J. (2021). Detecting loan words computationally. In E. O. Aboh & C. B. Vigouroux (Eds.), *Contact Language Library* (Vol. 59, p. 269-288). John Benjamins Publishing Company. https://doi.org/10.1075/coll.59.11zha

La naissance de la terminologie française de la traduction

Ludovic Milot, Département
Université d'Innsbruck
Ludovic.Milot@uibk.ac.at

Résumé. Cette présentation s'inscrit dans une réflexion diachronique sur la formation de la terminologie française de la traduction, mettant en lumière les ruptures et continuités entre l'Antiquité, le Moyen Âge et le début de la Renaissance. Elle part du constat d'une absence d'histoire terminologique continue en France, liée notamment à la prédominance du latin et à l'émergence tardive des langues vernaculaires. Dans un premier temps, l'analyse de l'Antiquité révèle une pluralité de termes latins qui décrivent davantage des procédés et des résultats que l'acte de traduire lui-même, sans qu'un hyperonyme unificateur ne s'impose. Le Moyen Âge introduit une rupture terminologique importante : avec l'essor des langues vernaculaires, les termes latins disparaissent ou évoluent, tandis que de nouvelles formes apparaissent en ancien français. Toutefois, ces usages restent instables et reflètent une conceptualisation encore limitée de l'activité traduisante. Enfin, la Renaissance marque un tournant décisif avec l'émergence du verbe *traduire*, néologisme issu du latin *traducere*. Son adoption en français témoigne d'un changement profond, faisant de la traduction une activité autonome.

Mots clés : terminologie de la traduction, traductologie, étude diachronique, interprète & traducteur

1. Introduction

Les recherches s'intéressant à la terminologie française de l'activité traduisante se focalisent principalement sur une période donnée. Même si l'histoire de la traduction remonte à l'Antiquité, c'est généralement le Moyen âge et les siècles d'après qui sont traités en profondeur parmi les scientifiques, ce qui s'explique par l'émancipation des langues vernaculaires. Cela met en lumière, d'une part, l'absence d'histoire continue de la traduction en France d'un point de vue terminologique et, d'autre part, le fait que la restitution de ce sujet soit discontinue. Ce double constat conduit à la problématique suivante : comment est née et s'est formée la terminologie française de la traduction, depuis l'Antiquité jusqu'à nos jours et quels évènements ont marqué les différentes ruptures entre les époques ?

2. Méthodologie

Cette recherche adopte une approche multidisciplinaire située au croisement de la linguistique contrastive et des études de traductologie et fondée sur l'analyse comparative de sources scientifiques spécialisées (Folena 1991 et Seele 1995 pour la période romaine ; Baehr 1981 ; Burridant 1983 ; Bérier 1988 ; Gutbub 2015 pour la période médiévale) ainsi que de textes authentiques de l'époque respective qui parle de la traduction.

Elle repose d'autre part sur l'exploitation de données quantitatives relatives à la terminologie de la traduction, recueillies à partir de préfaces, de gloses, etc. Ces données ont été collectées et organisées dans un corpus chronologique structuré selon plusieurs critères : le terme étudié, sa nature grammaticale, sa désignation (procédé ou activité traduisante), ainsi que la citation originale, la source et l'époque (Antiquité, Moyen Âge, Renaissance).

L'analyse s'appuie en outre sur divers outils lexicographiques et documentaires, tels que le *Gaffiot*, le *Dictionnaire étymologique de la langue latine* de Ernout & Meillet, les dictionnaires historiques comme le *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e^ au XV^e^ siècle* de Godefroy, le *Dictionnaire du Moyen Français* (DMF), le Centre National de Ressources Textuelles et Lexicales (CNRTL) et le *Littré*, ainsi que la base de données des Archives de littérature du Moyen Âge (ARLIMA).

Cette approche méthodologique permet de mettre en lumière l'émergence progressive d'une prise de conscience de la traduction en tant qu'objet de réflexion – annonçant ce qui sera plus tard désigné comme la traductologie – sans toutefois constituer encore une épistémologie.

3. Analyse

3.1. Antiquité

L'Antiquité, en matière de traduction, en est à ses balbutiements et la réflexion autour de cet acte gravite principalement autour de deux grands procédés de traduction que sont le mot-à-mot (v*erbum pro verbo* – traduction littérale) et le sens (*sensum de sensu* – traduction de libre), et de quelques verbes, dont les plus récurrents sont (Folena 1991, 91 ; Seele 1995, 91s.) :

1) 'expliquer, rendre clair' : *interpretari*, *explicare*, *reddere*, *exprimere*
2) '(re)tourner' : *vertere*, *convertere*, *mutare*
3) 'transporter' : *transferre* (postclassique), *translatare* (bas latin)

Pour ce qui est des substantifs (nomina actionis) signifiant 'traduire', on peut relever *interpretatio*, *imitatio*, *aemulatio*, un trio régulièrement abordé ensemble car ces conceptions sont hiérarchiquement associées et ordonnées dans le discours rhétorique. L'*interpretatio* renvoie à une traduction en partie libre et relativement proche de l'original, ou à une reformulation ; l'*imitatio* est surtout une pratique de réécriture appliquée à la poésie et se réfère à une nouvelle œuvre inspirée et remaniée d'un original ; l'*aemulatio* se rapporte à une traduction dont le but est de surpasser le texte-modèle (Schreiber 2016, 494). Retenons de ces termes qu'ils mettent plutôt l'accent sur le procédé et le résultat que sur l'acte de traduire en tant que tel.

Il me semble tout à fait opportun d'insister sur *interpres*, dont est dérivé *interpretatio*, dans la mesure où celui-ci est particulièrement équivoque. En effet, le mot est polysémique et renferme une certaine ambiguïté dans sa définition : il pouvait désigner à la fois celui qui interprétait les lois, les songes ou les messages des dieux (langage divin, messager) et celui qui interprétait les langues étrangères (langage humain, traduction linguistique). On peut toutefois retenir que la transmission d'information(s) était une caractéristique commune à toutes ces « activités ». Il faut cependant différencier deux fonctions de l'*interpres* : celle qui renvoie à la

communication pragmatique (quotidien, guerre, échanges commerciaux…) et celle à la communication littéraire (enseignement). Cette dimension est pertinente car elle définit l'objectif de traduction.

Dans les échanges commerciaux, juridiques, diplomatiques et politiques, on recourait à des *interpretes* pour assurer la communication entre les deux parties, tout comme le montrent les références relevées dans la *Guerre des Gaules* de Jules César (*cotidianis interpretibus*, 1.19) et la *Guerre de Jugurtha* de Salluste (*interpretes fidi*, 109.4). La fidélité semble être au cœur des discussions : le fait que l'adjectif *fidus* puisse parfois accompagner le substantif laisse transparaître la confiance que le commanditaire place dans la fidélité de la retransmission du message.

Au contraire, chez Cicéron, cette fidélité au texte revêt un caractère moins favorable. Il emploie le verbe *interpretari* dans son sens linguistique pour signaler une traduction courte et proche de l'original, tout en taxant l'interprète d'inéloquence (*indiserti*), c'est-à-dire comme une personne non éduquée en rhétorique et n'ayant pas reçu une formation lettrée complète comparable à celle des orateurs accomplis et traduisant littéralement (*verbum pro verbo reddere*) (McElduff 2009, 140). Pour les traductions plus complexes, il a recours à d'autres termes tels que *vertere*, *convertere*, *exprimere*, *transferre* (McElduff 2009, 137).

On peut donc voir l'interprète comme un 'traducteur oral' doté d'une formation littéraire limitée qui, par manque de connaissances dans la langue cible, n'avait d'autre choix que de se tourner vers le mot-à-mot. On peut aussi arguer que Cicéron avait certainement à l'esprit que les *interpretes* devaient retransmettre de manière tangible des énoncés, que le « mot-à-mot » se présentait comme la stratégie la plus avantageuse, et que cela sous-entendait l'élimination de procédés rhétoriques généralement complexes ou encore qu'un parcours rhétorique était superflu pour ce type de traduction. Comparés aux poètes ou aux rhéteurs, les *interpretes* occupaient certes une place de second rang et semblaient n'avoir guère grâce aux yeux des premiers, mais pour un accord de paix, par exemple, il n'était nul besoin de s'exprimer en termes savants, sauf éventuellement dans le cas d'une communication diplomatique de haut niveau ; il s'agissait avant tout d'être pragmatique et de restituer à l'oral le message. D'autre part, ces interprètes ne disposaient certainement pas du temps nécessaire pour recourir à des figures de style élaborées.

On peut remarquer que le vocabulaire latin pour 'traduire' était certes riche, mais qu'aucun hyperonyme n'englobait l'ensemble du champ sémantique.

3.2. Moyen Âge

L'émergence des termes liés à la traduction en français vernaculaire s'inscrit dans la prise de conscience de la langue vulgaire. Même si le latin dominait, et ce, car l'autorité des Saintes Écritures lui conférait une légitimité qui le plaçait au-dessus de la langue vulgaire, on peut citer le concile de Tours (813) ou même les Serments de Strasbourg (842) comme exemples soulignant le fait que la langue vernaculaire commençait à s'imposer dans le domaine religieux et politique. Cela a nécessairement entraîné un accroissement de l'activité traduisante en *romanz*, destinée à des profanes non latinisé.e.s, qui assumait ainsi la fonction de transmission d'un savoir remontant à l'Antiquité (racines gréco-latines) et jetait les bases de la *translatio*

studii visant à déplacer ce savoir de l'Orient vers l'Occident, à l'approprier et à le réécrire en l'adaptant aux réalités européennes.

D'un point de vue terminologique, au Moyen Âge, presque tous les termes latins ont disparu ou n'avaient plus le même sens qu'en latin classique, provoquant ainsi une rupture de continuité. À cette époque, les concepts de traduction n'étaient pas non plus systématiques et uniformes ; les techniques de traduction se limitaient à la dichotomie traduction mot-à-mot / littérale (*dire mot a mot la letre, par la letre, le latin sivrai e la letre, sans riens oster et sans riens metre*) vs. traduction libre (*ajouster de jolis mots/adjouster, assamblant, remplir, compilent, remission, desclairier, mettre en cler langaige, espondre en roumans/en romans, escrive de lettre en vulgal, esposeor en françois*), et peu d'auteurs et de traducteurs médiévaux ont dépassé cette réflexion (Pöckl 2016, 12). Cela s'explique en partie par le fait que la richesse et la complexité lexicale du latin ne semblaient pas trouver d'équivalent (suffisant) dans le français naissant, qui n'était ni standardisé ni codifié (Lusignan 1987, 73).

En ancien français, on rencontre des verbes qui, sémantiquement, évoquent soit l'idée ou le contenu du verbe 'transporter', comme *translater* ou *transposer*, soit expriment le verbe latin *vertere*, comme *turner* ou *trestorner*, soit apparaissent sous la forme de dérivés de l'adjectif *romanz*, comme *enromanchier* (*enromancier*) (Bérier 1988, 239-240) ou sous la forme d'expressions verbales telles que *metre / traire / faire / espondre ... en romanz*, *dire en romanz* (Pöckl 2016, 15), *fere romanz* (Baehr, 1981). Pour le type d'acteur de la traduction, on distingue entre le *translateor / translateur* pour la traduction libre, et l'*interpreteor / interpreteur* pour la traduction traduction littérale, les deux pouvant se compléter en fonction des contraintes stylistiques de la langue cible. Il est intéressant de noter qu'on nommait l'interprète de liaison de l'Orient par l'emprunt *drughemant* / *drugemens*, ce qui ne manque pas de marquer une nouvelle étape de la définition de cette activité.

Il faudra attendre le début de la Renaissance pour voir apparaître le verbe traduire au sens moderne du terme. On le retrouve sous la plume de l'humaniste italien Leonardo Bruni (1370 ?-1444), d'abord dans une lettre datée de 1404, puis dans son traité *De la traduction parfaite / De interpretatione recta* paru entre 1420 et 1426, et il s'avère qu'il s'agit d'un néologisme latin : *traducere* (Pöckl 2016, 11). Ce n'est donc pas d'une création ex nihilo, mais plutôt de l'adaptation d'un concept existant à une activité, et plus précisément d'un glissement métaphorique (Gutbub 2015, 219). En effet, le sens premier du participe passé *traductum* qu'il emploie n'était pas 'traduit' mais 'transporté' et signifie dans le contexte dans lequel il l'utilise le transfert d'un mot grec en latin (Berman 1988, 30).

La France a adopté le mot italien pour désigner la *traduction* relativement tard. Ce retard a notamment été entrainé par le déclin de l'activité intellectuelle pendant la guerre de Cent Ans, qui a isolé le pays des courants culturels de la région méditerranéenne. La première occurrence du verbe *traduire* en français (en 1490) ne faisait d'ailleurs pas référence au transfert linguistique, mais avait une connotation juridique (Pöckl 2016, 21). À la Renaissance, le terme technique *traduire* prend le sens qu'on lui connaît aujourd'hui, c'est-à-dire le transfert entre langues, et pose ainsi les fondations d'un domaine d'activité clairement défini qui s'affranchit de l'adaptation, de l'activité annexe ou du commentaire (Buridant 1983, 103). Ce phénomène marque un changement profond dans la perception de la traduction, entre autres. En 1679, le dictionnaire de Richelet décrit la forme française *translater* comme « un vieux mot qui signifie *traduire* et qui, tout au plus, ne peut trouver sa place que dans l'ancien burlesque et dans le

comique » (Berman 1988, 30). *Traduire* est préféré à *translater*, car il est – selon certains historiens – perçu comme plus « dynamique » et plus « expressif » :

> Alors que la *translation* met l'accent sur le mouvement de transfert ou de transport, la *traduction*, elle, souligne plutôt l'énergie active qui préside à ce transport, justement parce qu'elle renvoie à *ductio* et *ducere*. La traduction est une activité qui a un agent, alors que la *translation* est un mouvement de passage plus anonyme. Tous les mots formés à partir de *ductio* supposent des agents. Et c'est justement parce que l'opération traduisante est conçue, à partir de la Renaissance, comme un acte, et comme un acte spécifique, qu'on se met à l'appeler *traduction*. (Berman 1988, 30)

On peut résumer la notion comme suit : dans la *translation* – comme c'est le cas dans la *translatio studii* – s'opère un glissement de A vers B, alors que la *traduction* impliquerait un changement, une soustraction (supprimer de A pour intégrer dans B ou une dénaturation de B par A).

Gutbub voit également une minime nuance entre les deux termes qui expliquerait l'abandon de l'un au profit de l'autre. Selon lui, *translater* s'apparente plus à la paraphrase où on ne garde que le sens (« Paraphraser consiste, à titre d'exercice, à répéter et varier l'œuvre d'un autre dans la même langue […]. »), la paraphrase effaçant les différences entre les langues, dans un but d'appropriation (Gutbub 2015, 226). Cette théorie rappelle la réflexion de Barthélémy Aneau (1550) à ce sujet, pour qui *translater* et *tourner* désignent une adaptation en vers, tandis que *traduire* sous-entend une transposition et une imitation (Buridant 1983, 100-101). Bien que cette théorie soit alléchante et argumentée, il est légitime de se demander si l'abandon de *translater* n'aurait pas d'autres causes. Par exemple, le rôle de référence que joue l'Italie dans le mouvement humaniste, ainsi que l'attrait exercé sur les autres langues romanes, pourraient aussi avoir pesé dans la balance.

4. Résultats et conclusion

Les premiers résultats de cette étude mettent en évidence que la naissance de la terminologie française de la traduction résulte moins d'une évolution linéaire que d'un processus irrégulier, étroitement lié aux transformations linguistiques, culturelles et intellectuelles de chaque période.

Le terme *translation* en français a survécu, mais il a perdu tout rapport avec l'activité de traduction. Jusqu'au XIX^{e} siècle, il est resté synonyme de 'transport' ; aujourd'hui, il est presque exclusivement utilisé en géométrie et mécanique – où il désigne un certain type de déplacement (Berman 1988, 31) – et en télégraphie (transmission d'un message au moyen d'un appareil) (Littré 1873-1874, 1878).

Cependant, ce n'est qu'au début de l'ère moderne que la pratique de la traduction a connu un essor considérable, entraînant une expansion du vocabulaire spécialisé utilisé pour décrire les processus de traduction, les directions de traduction, etc. Cette évolution a marqué un tournant décisif dans la conceptualisation et la reconnaissance de la traduction en tant qu'activité intellectuelle et culturelle à part entière. Ce point ne sera toutefois pas traité dans cette contribution.

Références

BAEHR, Rudolf. « Rolle und Bild der Übersetzung im Spiegel literarischer Texte des 12. und 13. Jahrhunderts in Frankreich ». Dans *Europäische Mehrsprachigkeit. Festschrift zum 70. Geburtstag von Mario Wandruszka*, édité par Wolfgang Pöckl, Tübingen, Niemeyer, 1981, p. 329-348

BERMAN, Antoine. « De la translation à la traduction ». *TTR : traduction, terminologie, rédaction*, 1/1, 1988, p. 23-40.

BERIER, François. « Traduire en France à la fin du Moyen Age », dans : *Grundriss der Romanischen Literaturen des Mittelalters.* VIII/1 : La littérature française aux XIVe et XVe siècles, Heidelberg, Winter, 1988, pp. 219-265.

BURIDANT, Claude. « Translatio medievalis – Théorie et pratique de la traduction médiévale ». *Travaux de linguistique et de littérature*, vol. 21, n°1, 1983, p. 81-136.

FOLENA, Gianfranco. *Volgarizzare e tradurre*. Torino, Einaudi, 1991.

GUTBUB, Christophe. « Penser la traduction : que veut dire traduire au XVIe siècle ? ». Dans *Histoire des traductions en langue française. XVe et XVIe siècles*, sous la direction de Véronique Duché, Paris, Verdier, 2015, p. 183-244.

LITTRE, Émile. Dictionnaire de la langue française. Paris, L. Hachette, 1873-1874. Electronic version created by François Gannaz. http://www.littre.org

LUSIGNAN, Serge. « Le français et le latin aux XIIIe-XIVe siècles : pratique des langues et pensée linguistique ». *Annales. Histoire, Sciences Sociales*, vol. 42, n°4, 1987, p. 955-967.

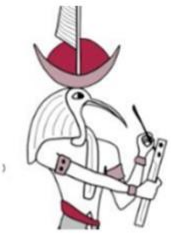
MCELDUFF, Siobhan. « Living at the Level of the Word: Cicero's Rejection of the Interpreter as Translator ». *Translation Studies*, vol. 2, n°2, 2009, p. 133-146.

PÖCKL, Wolfgang. « Traduire, traduction, traducteur, traductologie, interprétation, interprète etc. Un aperçu historique de la terminologie en usage dans les langues romanes ». Dans *Manuel de traductologie*, édité par Jörn Albrecht et René Métrich, Berlin, De Gruyter, 2016, p. 11-27.

SCHREIBER, Michael (2016): « Interpretatio, imitatio, aemulatio : formes et fonctions de la traduction « libre » dans le domaine des langues romanes », dans : Albrecht, Jörn ; Métrich, René *Manuel de traductologie*. De Gruyter

SEELE, Astrid. *Römische Übersetzer – Nöte, Freiheiten, Absichten: Verfahren des literarischen Übersetzens in der griechisch-römischen Antike*. Darmstadt, Wissenschaftliche Buchgesellschaft, 1995.

universität innsbruck



La naissance de la terminologie française de la traduction

20e Conférence Internationale TOTh
4 & 5 juin 2026, Université Savoie Mont-Blanc

Ludovic MILOT
Département de traductologie / Département de philologie romane – Université d'Innsbruck

1. Introduction

- **Status quo** : les études sur l'histoire de la terminologie française de la traduction se concentrent principalement sur des périodes isolées (Antiquité, Moyen Âge, Renaissance, etc.). Il manque une histoire continue, ce qui rend la restitution du phénomène fragmentaire.
- **Question de recherche :**
  1) Comment est née et s'est formée la terminologie française de la traduction, depuis l'Antiquité jusqu'à nos jours?
  2) Quels évènements ont marqué les différentes ruptures entre les époques?
- Etat de ces recherches doctorales : Antiquité, Moyen Âge, début de la Renaissance

2. Méthode

- Approche multidisciplinaire, incluant :
  a) l'analyse comparative de sources scientifiques (issues de la linguistique comparée et des études de traduction – citer sources)
  b) l'utilisation de données quantitatives sur la terminologie de la traduction recueillies à partir de préfaces, de gloses, etc. (Plaute, Térence, Cicéron, Horace, Jules César, Saint Jérôme, etc.)
  → collecte des données et constitution d'un corpus chronologique (tableau Excel) selon les critères suivants : terme, nature, désignation (procédé, activité traduisante/acte de traduire), citation & source, époque.
  c) outils : dictionnaires bilingues (*Gaffiot*), étymologiques (Ernout & Meillet) et historiques (Godefroy, DMF, CNRTL, *Littré*) ; bases de données (ARLIMA – archives de littérature du Moyen Âge)
- Cette étude met en lumière l'émergence d'une prise de conscience de la traduction – plus tard appelée « traductologie » – même si cela ne constituait pas encore une épistémologie de la traduction à part entière.

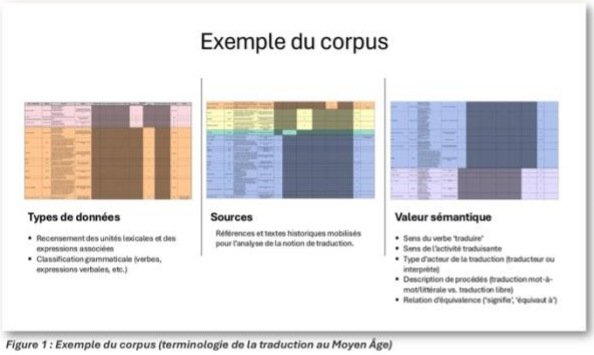


Figure 1 : Exemple du corpus (terminologie de la traduction au Moyen Âge)

3. Antiquité

- Vocabulaire riche mais **absence d'hyperonyme** recouvrant la totalité du champ sémantique de la traduction
- multitude de verbes sémantiquement proches pour désigner divers types de traduction
- **3 grandes catégories de verbes** (Folena 1991, 91; Seele 1995, 91s.) :
  1) 'expliquer, rendre clair' : *interpretari, explicare, reddere, exprimere*
  2) '(re)tourner' : *vertere, convertere, mutare*
  3) 'transporter' : *transferre* (postclassique), *translatare* (bas latin)
- **2 procédés de traduction** (concepts rhétoriques) :
  1) mot-à-mot (*verbum pro verbo* – traduction littérale)
  2) libre (*sensum de sensu* – traduction de sens)
- **Substantifs** : *interpretatio, imitatio, aemulatio, variatio*
- La figure de l'*interpres* → ambiguité lexicale en raison de la polysémie du terme :
  - interprète des langues (traduction orale) – pratique pragmatique
  - interprète des lois, rêves ou messages divins – pratique intellectuelle
  - adj. *fidus* : fidélité dans la retransmission du message ; reproches : trop littérale, manque de panache ≠ orateurs (discours rhétorique)

4. Moyen Âge

- Rupture terminologique → Contexte historique révélateur du changement de point de vue sur l'activité traduisante et du développement terminologique :
  a) chute de l'Empire romain : l'ensemble de dialectes latins tombe sous l'influence des langues locales
  b) jusqu'au VIIIe siècle : deux variétés de latin cohabitent : a) le latin écrit, érudit et religieux (héritier du latin classique) ; b) le latin oral et populaire (*proto-roman*), associé aux paysans (*rustica Romana lingua*) → dichotomie écrit/oral conditionne l'usage et le choix de la langue
  c) IXe siècle : le français acquiert son identité linguistique fondamentale ; diglossie aux IXe – XIe siècles : latin vs. *roman* (*françois*)
  d) XIIe et XIIIe siècle : apparition de la terminologie à travers les textes 'littéraires' au sens large (frontières poreuses entre historiographie et fiction)
- L'émergence d'une identité nationale à travers la langue nécessite la traduction de textes en latin pour leur diffusion et/ou appropriation
- Le concile de Tours (813) ou même les Serments de Strasbourg (842) illustrent l'importance croissante des langues vernaculaires.

- **Procédés** :
  a) traduction mot-à-mot/littérale : *dire mot a mot la letre, par la letre, le latin sivrai e la letre, sans riens oster et sans riens metre*
  b) traduction libre : *ajouster de jolis mots/adjouster, assamblant, remplir, compilent, remission, desclairier, mettre en cler langaige, espondre en roumans/en romans, escrive de lettre en vulgal,* exposer (esposeor : expliquer) en françois
  c) équivalence : *vuelt/veult dire, signifie en franceis, est … en fransois, vaut, por , met lou mot qui autant vaut*
  →typologie de la traduction reprenant les préceptes latins (*imitatio, interpretatio, aemulatio, variatio*)
- **Terminologie** : presque tous les termes latins ont disparu ou n'avaient plus le même sens qu'en latin classique : *metre en françois, de … en metre, translater, trestorner, torner, transposer, transporter, transferre, convertir + muer, metre en romanz, enromancier, de latin en romaunz/rommant traire, dire en romanz, fere romanz, en romanz fist, reduire de latin en roumant, estraire, De latin l'a en roumans traite, reverser, revirer*
  → Raison : équivalents insuffisants, français ni standardisé ni codifié
- **Acteur de la traduction** :
  a) *translateor/translateur* (traduction libre) + *interpreteor/interpreteur* (traduction littérale) → fonction variable, identité versatile
  b) interprète de liaison (Orient) : *drughemant / drugemens.*

5. Renaissance

- *Traduire* : translatus, participe passé passif du verbe *transferre* → *translatare* qui unifie la terminologie (translatus, translatio); *translatare* concurencie *transferre* vers le VIIe siècle
- Humaniste italien Leonardo Bruni (1370 ?-1444), néologisme latin : *traducere*
- *Traduire* est repris par l'ensemble des cultures romanes; *translater* est abandonné. *Traduire* = plus dynamique/expressif
  → hypothèse 1 : en raison de l'actance supposée par *ductio/ducere*
  → hypothèse 2 : *translater* ≈ paraphrase, traduction de sens ≠ *traduire* = transposition + imitation

6. Résultats & conclusion

- La terminologie de la traduction en français n'est pas simple héritière du latin mais montre une émancipation dont les causes sont à la fois d'ordre stylistique et politique
- *Traduire* est un moment charnière dans la terminologie française de la traduction ; le terme marque un changement de la vision de l'activité traduisante
- *Translation* a survécu aujourd'hui dans certains domaines : géométrie & mécanique (translation = type de déplacement), télégraphie (translation = transmission d'un message au moyen d'un appareil), théologie (translation = déplacement de reliques); mais n'a plus le sens de traduction

www.uibk.ac.at

Analyser la variation terminologique : enjeux méthodologiques dans le domaine de l'industrie du cuir

Martina Alì
Osservatorio di Terminologie e Politiche Linguistiche (OTPL), Università Cattolica del Sacro Cuore
martina.ali@unicatt.it

Mots-clés : variation terminologique, industrie du cuir, construction de corpus, fausse terminologie

Introduction

Les termes sont nécessairement soumis à des processus de changement et d'évolution, qu'ils soient d'ordre conceptuel ou linguistique. La langue de spécialité constitue un système dynamique, vivant et évolutif, qui s'adapte aux contextes sociaux, communicationnels et situationnels dans lesquels elle est mobilisée. Elle peut ainsi être affectée, entre autres, par des phénomènes de variation diaphasique, diastratique et diachronique (Zanola, 2018, 2021).

La littérature, notamment dans le domaine francophone, s'est longuement penchée sur la question de la variation terminologique (Condamines & Picton, 2014 ; Humbert-Droz, 2021). Ces travaux ont permis d'identifier diverses typologies de transformations et ont donné lieu à l'élaboration de notions méta-terminologiques telles que la « banalisation lexicale » (Galisson, 1978), la « déterminologisation » (Meyer & Mackintosh, 2000 ; Humbert-Droz, 2024), la « dédomanialisation » (Rastier & Valette, 2009), jusqu'aux notions plus récentes de « mouvements migratoires » (Botta, 2013) et de « dé-spécialisation » (Mazière, 1981). L'ensemble de ces approches met en évidence que les transformations terminologiques ne relèvent pas de simples usages erronés ou approximatifs de la terminologie, mais traduisent au contraire des dynamiques sociales, discursives et cognitives complexes, à l'œuvre au sein des communautés d'usage et des pratiques communicationnelles.

Les conséquences de la variation peuvent inclure calques, adaptation d'emprunts (Zanola, 2020), redéfinition sémantique, création néologique (Humbley, 2018) ou encore l'apparition de « faux termes ». Ce dernier phénomène a été étudié à l'intérieur du domaine de l'industrie du cuir dans le cadre d'une thèse de doctorat récemment soutenue à l'Università Cattolica del Sacro Cuore de Milan (juillet 2025) et concerne l'émergence d'unités terminologiques, simples ou complexes, dont la manipulation formelle et/ou conceptuelle conduit à désigner un concept différent, voire opposé, à celui reconnu comme légitime dans le domaine spécialisé. La notion de légitimité terminologique (Humbley, 1996) renvoie au fait qu'un concept est officiellement défini par des instruments législatifs ou des normes techniques, ou encore validé de manière consensuelle par les experts du domaine à travers des processus d'« harmonisation terminologique » (Lerat, 1995).

Un exemple emblématique, qui constitue le point de départ de ces études, est représenté par le terme *ecopelle*. Celui-ci est légitimé par la norme UNI 11427 et désigne un cuir véritable, d'origine animale, tanné selon des pratiques écologiques ; toutefois, dans le domaine commercial, il est souvent employé de manière trompeuse, par les producteurs comme par les consommateurs, pour désigner un matériau synthétique. La « fausse terminologie » représente ainsi un cas de variation conceptuellement distinct, qui touche aux relations entre langue et terminologie ainsi qu'entre *mot* et *terme* : bien qu'employés de manière non conforme à la norme, les faux termes ne correspondent pas à des mots de la langue générale, mais disposent eux aussi du statut de termes, c'est-à-dire d'unités lexicales désignant des concepts propres à un domaine spécialisé (L'Homme, 2020). De plus, lorsque ces phénomènes terminologiques proviennent du secteur industriel ou productif, ils revêtent une importance particulière par rapport aux variations évoquées précédemment, dans la mesure où les termes erronés influencent directement la communication commerciale et à destination du consommateur, entraînant une perte économique tangible (Zanola, 2018).

Cette contribution se centre sur des questions méthodologiques centrales : comment saisir avec rigueur la variation terminologique, quelle qu'en soit la nature ? Toutes les variations terminologiques se saisissent-elles de la même manière ? Comment identifier au mieux les changements conceptuels et linguistiques ?
La construction du corpus constitue le point de départ de toute analyse fiable des usages terminologiques. La méthodologie adoptée pour constituer le corpus conditionne directement la qualité et la pertinence des résultats, en influençant la représentativité des termes, la détection de leurs variations et la compréhension des interactions entre experts et grand public.

S'inspirant de travaux récents sur la variation terminologique (Humbert-Droz, Condamines & Picton, 2019 ; Humbert-Droz, 2021 ; Dankova, 2023), nous soutenons qu'un corpus destiné à ce type d'analyse doit être aussi hétérogène que possible, intégrant des textes de différents niveaux de spécialisation et de typologies variées : textes scientifiques, normes techniques, mais aussi documents commerciaux et de vulgarisation. Il doit par ailleurs être circonscrit à une période temporelle limitée, correspondant à ce que Dury (2022) appelle une « diachronie brève ». Ce choix s'inscrit dans une approche de plus en plus fréquente en terminologie contemporaine, qui privilégie l'analyse de périodes restreintes afin d'observer non pas les grandes évolutions structurelles de la langue, mais les variations terminologiques fines au sein de domaines spécialisés. D'ailleurs, la subdivision en sous-corpus, selon le degré de spécialisation et le type de texte, favorise une meilleure représentativité des usages terminologiques tout en reflétant les interactions complexes entre différents acteurs et types d'usagers (Delavigne, Picton & Thibert, 2022).

Pour illustrer cette méthodologie, nous présentons l'exemple issu de notre thèse, consacrée à l'étude de la terminologie de la durabilité dans l'industrie italienne du cuir. La recherche a analysé les variations terminologiques, tant formelles que conceptuelles – qualifiées de « manipulations » –, observables lors du passage de la terminologie spécialisée de la production à celle employée dans les discours de commercialisation (Zanola, 2017). Ces variations spécifiques peuvent conduire à l'émergence de ce que nous avons désigné ci-dessus comme

« fausse terminologie ». Les données montrent que des termes dits légitimes circulent également dans des sous-domaines connexes tels que la mode, le textile ou le commerce électronique. Dans ces espaces, marqués par l'interaction de différentes communautés discursives, ces termes font l'objet de réinterprétations en fonction des pratiques et des besoins communicationnels propres à chaque contexte. L'observation de ces interactions met en lumière la transformation des concepts et la production de fausse terminologie ou de sens élargi, confirmant ainsi la nécessité d'une approche méthodologique attentive à la diversité des usages et des contextes.

Après avoir présenté l'état de la littérature sur la variation terminologique, nous approfondirons la question de la construction d'un corpus hétérogène, en prenant comme modèle de référence celui employé dans notre thèse de doctorat. Nous nous pencherons ensuite sur des études de cas significatives, portant sur des termes spécialisés, légitimes dans le domaine de référence – l'industrie du cuir – qui subissent une manipulation à la fois formelle et conceptuelle lorsqu'ils sont mobilisés dans des domaines moins spécialisés. Enfin, nous proposerons une réflexion portant sur les notions d'*expertise* et de *domaine*, afin de mettre en lumière les dynamiques sociale et cognitive qui accompagnent la circulation et la transformation des termes.

Dans cette perspective, le présent travail se propose d'apporter une contribution substantielle à la méthodologie des pratiques terminographiques, notamment en ce qui concerne l'analyse de la variation terminologique.

Méthodologie

Notre corpus, en langue italienne, comprend au total 317 documents et 356 248 occurrences, couvrant une période d'environ dix ans (2011-2022). Nous avons retenu cette périodisation car, bien que relativement brève, la décennie considérée est caractérisée par d'importantes transformations extralinguistiques liées aux enjeux de durabilité dans l'industrie de la tannerie incluent notamment la publication et la mise à jour de la norme UNI 11427 sur les cuirs écologiques, l'adoption du décret législatif n° 68 en Italie, ainsi que la promotion des Objectifs de développement durable de l'Agenda 2030 des Nations unies à partir de 2015. À cela s'ajoute l'émergence de matériaux alternatifs au cuir, tels que Piñatex®, Wineleather®, AppleSkin® et Desserto®, développés entre 2015 et 2019. Ces changements ont entraîné une véritable reconfiguration conceptuelle du domaine, tout en mettant en lumière sa dynamique d'évolution.

Le corpus présente une structure tripartite, avec trois sous-corpus caractérisés par différents degrés de spécialisation et incluant diverses typologies textuelles :

- le *Sous-corpus 1* : textes à haut degré de spécialisation, à savoir des documents techniques et institutionnels produits par des organismes spécialisés et destinés à des professionnels, caractérisés par un contenu complexe et une terminologie spécifique, tels que le décret législatif n° 68 du 9 juin 2020, les textes intégralement reproduits de cinq normes techniques UNI (UNI 11427, UNI 10885, UNI EN 15987, UNI EN ISO 14001 : 2015, UNI

EN 16484 : 2015), des fiches de produits officielles du showroom Lineapelle, les Rapports de Durabilité rédigés par les experts de UNIC (Unione Nazionale Industria Conciaria) et le Règlement pour la certification des revendications éthiques établi par l'I.CE.C (Istituto di Certificazione della Qualità per l'Industria Conciaria);

- le *Sous-corpus 2* : articles tirés du magazine *La Conceria*, édité par UNIC ; contrairement à d'autres revues plus spécialisées, *La Conceria* aborde également des sujets moins techniques et d'intérêt plus général, tels que les dernières nouveautés dans le monde de la mode, du design ou de l'industrie automobile, et se prête donc à représenter un « niveau intermédiaire de spécialisation » (Humbert-Droz, Picton & Condamines, 2019) ;

- le *Sous-corpus 3* : textes destinés au grand public, tels que des articles de presse générale (en ligne, par exemple *Il Corriere* et *La Repubblica*), des blogs de vulgarisation sur le véganisme, des articles de magazines de mode en ligne, ou encore des fiches de produits sur des sites de e-commerce

Cette structuration hétérogène permet de saisir la diversité des usages, d'observer la circulation des termes et de comparer leurs acceptions dans différents contextes. Une attention particulière a été portée au troisième sous-corpus, essentiel pour détecter et analyser les cas de fausse terminologie. Ces textes, collectés exclusivement en ligne, reflètent la formation de « communautés épistémiques spontanées » (Vicari, 2013), où locuteurs experts et non-experts partagent savoirs spécialisés et profanes, souvent à travers des processus de négociation sémantique et d'appropriation des termes. Le Net, en raison de la nature immédiate et souvent informelle des échanges qu'il permet, constitue un espace privilégié des échanges commerciaux, où s'observent les interactions entre producteurs et consommateurs ainsi que les stratégies de marketing et de promotion des produits.

L'extraction terminologique a été effectuée sur l'ensemble du corpus selon une approche semi-automatique. Dans un premier temps, une analyse préliminaire a été réalisée avec Sketch Engine (Kilgarriff et al., 2014) afin de comparer le corpus spécialisé à un corpus de référence en italien général, fourni par le logiciel. Cette comparaison a permis d'identifier des unités lexicales plus fréquentes dans le corpus spécialisé, considérées comme des candidats-termes pertinents pour le domaine étudié. L'extraction automatique initiale a été ensuite soumise à un filtrage manuel. Après révision, 172 unités ont été retenues en fonction de leur pertinence par rapport à la thématique de la durabilité dans l'industrie du cuir. Par terminologie de la durabilité, nous entendons les unités lexicales relevant d'une conception holistique de la durabilité, intégrant les dimensions environnementale, sociale et économique, et mobilisées dans divers sous-domaines tels que les propriétés physiques de durabilité du cuir, les organisations professionnelles de la filière, les initiatives en matière de durabilité, ainsi que les dispositifs de certification, de labellisation et d'étiquetage. Ont été exclues les unités considérées comme non pertinentes, notamment les termes techniques non liés à la durabilité, les notions économiques générales, les noms propres ainsi que les éléments de bruit linguistique.

Études de cas

L'interrogation du corpus ainsi construit a permis d'identifier 51 faux termes, répartis en quatre macro-catégories selon le type de manipulation :

- manipulation terminologique partielle ;
- manipulation terminologique totale ;
- création de néologismes terminologiques fallacieux ;
- création de nom de marques composés du terme « pelle » sous forme d'équivalents anglais.

La quasi-totalité des faux termes ont comme dénominateur commun les termes *cuoio*, *pelle* ou *pelliccia*. L'Article 2 du Decreto Legislativo n. 68 del 9 giugno 2020 *Nuove disposizioni in materia di utilizzo dei termini "cuoio", "pelle" e "pelliccia" e di quelli da essi derivati o loro sinonimi e la relativa disciplina sanzionatoria* établit les deux exigences minimales pour que ces dénominations soient légitimes : a) l'origine animale et b) la préservation de la structure fibreuse naturelle après le processus de tannage. Ces exigences se retrouvent également dans les outils de normalisation terminologique du corpus (les normes techniques UNI 11427, UNI 10885, UNI EN 16484). Cependant, dans des domaines proches de l'industrie du cuir – mode, textile, ameublement, reliure, automobile –, les termes *cuoio*, *pelle* et *pelliccia* sont souvent utilisés pour désigner des matériaux végétaux ou synthétiques, illustrant la production de fausse terminologie.

Les études de cas que nous avons choisi d'analyser sont les suivants : le néologisme fallacieux *ecopelliccia*, qui entraîne également la perte sémantico-conceptuelle du préfixe *eco* ; *cuoio vegano*, qui induit une équivalence conceptuelle erronée entre *végan* et *durable* ; et le nom de marque *Appleskin*®, considéré comme illégitime conformément à l'Article 3 du décret italien. En particulier, ce dernier cas permet de réfléchir au rôle du terme *pelle* sous forme d'équivalent étranger, qui fonctionne en effet comme un « lexème évocateur » (Fèvre-Pernet, 2007, dans Altmanova, 2013) : une unité lexicale à forte valeur référentielle, évoquant des caractéristiques unanimement reconnues et appréciées par les consommateurs et attribuables au cuir véritable – qualité, douceur, résistance, luxe. Par ailleurs, bien que les entreprises détentrices des marques ne soient pas d'origine anglo- saxonne, la création du nom de marque privilégie l'équivalent anglais : l'élément étranger confère une dimension exotique au nom et favorise une sonorité familière, plus attractive pour le commerce national et international (Gałkowski, 2023).

Réflexions conclusives

En conclusion, cette étude a permis de réinterroger plusieurs notions fondamentales de la terminologie, à savoir le terme, l'expert et le domaine. Dans une perspective socioterminologique (Delavigne, 2002 ; Delavigne & De Vecchi, 2016), l'expertise ne se réduit pas à une compétence strictement technique ou scientifique, mais intègre également une dimension sociale essentielle, reposant sur la reconnaissance des acteurs non spécialistes et sur

les processus de légitimation dans la circulation des savoirs. Elle peut ainsi être à la fois institutionnelle et profane, comme le montrent les savoirs issus de l'expérience des utilisateurs dans divers domaines spécialisés.

Dans cette perspective, le domaine ne peut plus être envisagé comme un ensemble fermé et stable, mais comme un réseau dynamique de savoirs, de compétences et de pratiques discursives, au sein duquel les termes circulent, se reconfigurent et évoluent. Cette approche met en évidence la porosité des frontières entre savoirs scientifiques, experts et ordinaires, lesquels s'inscrivent dans un continuum aux limites floues (Vicari, 2013). Dès lors, l'analyse de la variation terminologique requiert une méthodologie rigoureuse, fondée sur un corpus diversifié et représentatif, capable de rendre compte de la complexité des usages en contexte.

L'approche proposée, illustrée par le cas de l'industrie du cuir, souligne ainsi la nécessité d'observer les phénomènes terminologiques dans des contextes multiples, de distinguer les différents niveaux d'expertise et de prendre en compte les interactions entre domaines et sous-domaines. Elle constitue un cadre méthodologique transposable à d'autres secteurs, permettant de mieux comprendre la circulation des termes, l'émergence de phénomènes de fausse terminologie ainsi que les processus de transmission et de transformation des concepts dans des environnements discursifs pluriels et interactifs.

Références

Altmanova, J. (2013). *Du nom déposé au nom commun : néologie et lexicologie en discours*. EDUCatt.

Botta, M. (2013). La terminologie de l'environnement en vulgarisation scientifique : La famille lexicale de la régénération des forêts en portugais. *Equivalences, 40*(1), 277-298.

Condamines, A., & Picton, A. (2014). Étude du fonctionnement des nominalisations déverbales dans un contexte de déspécialisation. *SHS Web of Conferences, 8*, 697-711.

Dankova, K. (2023). *Fibres textiles entre synchronie et diachronie : études terminologiques*. Peter Lang.

Decreto Legislativo n. 68 del 9 giugno 2020 *Nuove disposizioni in materia di utilizzo dei termini "cuoio", "pelle" e "pelliccia" e di quelli da essi derivati o loro sinonimi e la relativa disciplina sanzionatoria* https://www.gazzettaufficiale.it/eli/id/2020/06/26/20G00084/sg (dernier accès : 7/02/2026).

Delavigne, V. (2002). Le domaine aujourd'hui. Une notion à repenser. In D. Candelle (Ed.), *Le traitement des marques de domaine en terminologie* (pp. 1-13). Paris.

Delavigne, V., & De Vecchi. (2016). Socioterminologie et pragmaterminologie : rencontres et complémentarités. In C. Roche (Ed.), *TOTH. Terminologie & Ontologie : Théories et Applications* (pp. 141-156). Presses Universitaires Savoie Mont Blanc.

Delavigne, V., Picton, A., & Thibert, E. (2022). Socioterminologie et terminologie textuelle : l'expertise en question. *SHS Web of Conferences, 138*, 1-15.

Dury, P. (2022). Diachronic Variation. In P. Faber & M.-C. L'Homme (Eds.), *Terminology and Lexicography Research and Practice* (pp. 421-434). John Benjamins Publishing Company.

Fèvre-Pernet, C., & Roché, M. (2005). Quel traitement lexicographique de l'onomastique commerciale ? Pour une distinction Nom de marque / Nom de produit. *Corela, HS-2*, 1-18.

Gałkowski, A. (2023, September 22). I nomi dei marchi tra linguistica e marketing. Seminario presso l'Università Cattolica del Sacro Cuore di Milano.

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

Humbert-Droz, J. (2021). *Définir la déterminologisation : approche outillée en corpus comparable dans le domaine de la physique des particules* (Doctoral dissertation, Université de Toulouse 2).

Humbert-Droz, J., Picton, A., & Condamines, A. (2019). How to Build a Corpus for a Tool-Based Approach to Determinologisation in the Field of Particle Physics. *Research in Corpus Linguistics, 7*, 1-17.

Humbert-Droz, J. (2024). De la diffusion de termes dans la presse à l'émergence de néologismes sémantiques : une analyse du point de vue de la déterminologisation. *L'Information grammaticale*, *181*, 7-15.

Humbley, J. (1996). La légitimation en terminologie. *Sémiotiques*, 11, 119-126.

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

Kilgarriff et al. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 7-36.

Lerat, P. (1995). *Les langues spécialisées*. Presses Universitaires de France.

L'Homme, M.-C. (2020). *Lexical Semantics for Terminology: An Introduction* (Vol. 20).

Mazière, F. (1981). Le dictionnaire et les termes. *Cahiers de lexicologie, 39*(2), 79-104.

Meyer, I., & Mackintosh, K. (2000). When terms move into our everyday lives : An overview of de-terminologization. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 6*(1), 111-138.

Rastier, F., & Valette, M. (2009). De la polysémie à la néosémie. In S. Mejri (Ed.), *La problématique du mot* (Vol. 77, pp. 97-116).

Vicari, S. (2013). Del Bon usage della terminologia delle energie rinnovabili nei forum Internet: analisi delle tipologie definitorie. In A. Giaufret & M. Rossi (Eds.), *La terminologia delle energie rinnovabili tra testi e repertori: variazione, standardizzazione, armonizzazione* (pp. 153-194). Genova University Press.

Zanola, M. T. (2021). Cahiers de lexicologie, 118, 13-21.

Zanola, M. T. (2020). Francese e italiano, lingue della moda : Scambi linguistici e viaggi di parole nel XX secolo. *Lingue Culture Mediazioni - Languages Cultures Mediation (LCM Journal), 7*(2), 1.

Zanola, M. T. (2018). *Che cos'è la terminologia*. Carocci.

Zanola, M. T. (2017). La terminologie des arts et métiers entre production et commercialisation : Une aproche diachronique. *Terminàlia, 17*, 16-23.

Analyser la variation terminologique : enjeux méthodologiques dans le domaine de l'industrie du cuir

Martina Alì
Osservatorio di Terminologie e Politiche Linguistiche (OTPL)
Università Cattolica del Sacro Cuore
martina.ali@unicatt.it

TOTh 2026
4-5 Juin 2026
Université Savoie Mont-Blanc
Chambéry, France

1. INTRODUCTION

La terminologie est un système dynamique soumis à des processus de variation linguistique et conceptuelle, façonnés par les contextes sociaux, discursifs et communicationnels.

Objectif : analyser la variation terminologique et les phénomènes de « fausse terminologie » dans le domaine de la **durabilité de l'industrie du cuir**.

2. CADRE THÉORIQUE

Variation terminologique
Les termes évoluent selon les usages, les contextes et les communautés discursives (Zanola, 2018).

Légitimité terminologique
Un terme est légitime lorsqu'il répond à des critères institutionnels, normatifs et consensuels (Humbley, 1996).

Faux termes
Unités qui désignent un concept différent, voire opposé, à celui reconnu comme légitime dans le domaine spécialisé (Humbley, 2018).

Encadré : Légitimité terminologique

Un terme est considéré comme légitime lorsqu'il remplit les conditions suivantes :

a. Il est officiellement défini par des instruments législatifs ou des normes techniques ;
b. Il est validé par des organismes spécialisés ou des institutions reconnues ;
c. Son concept est consolidé par l'utilisation unanime des experts du secteur à travers des processus d'« harmonisation terminologique » (Lerat, 1995).

3. PRATIQUES TERMINOGRAPHIQUES

Construction de corpus hétérogènes

Une approche fondée sur des corpus hétérogènes (spécialisés, semi-spécialisés et grand public) est essentielle pour saisir la variation terminologique.

Subdivision en sous-corpus selon le degré de spécialisation et le type de texte.

Méthodologie inspirée des travaux récents : Humbert-Droz, Condamines & Picton (2019) ; Humbert-Droz (2021) ; Dankova (2023).

Condition indispensable pour une meilleure représentativité des usages et une analyse rigoureuse des interactions entre acteurs.

4. MÉTHODOLOGIE

4.1 LE CORPUS

317	356 248	2011–2022
documents	occurrences	(diachronie brève)

4.2 STRUCTURE DU CORPUS

1 Sous-corpus 1 : haut degré de spécialisation
Textes techniques et institutionnels (normes UNI, lois, rapports de durabilité, fiches produits, certifications). Destinés aux professionnels et experts.

2 Sous-corpus 2 : niveau intermédiaire
Articles de la revue *La Conceria* (UNIC). Thématiques techniques mais aussi mode, design, automobile. Niveau intermédiaire de spécialisation.

3 Sous-corpus 3 : grand public
Presse en ligne, blogs, magazines de mode, sites e-commerce. Discours de vulgarisation et communication commerciale.

4.3 EXTRACTION TERMINOLOGIQUE

Analyse avec Sketch Engine : corpus spécialisé vs corpus général italien (identification des unités surfréquentes).

↓

Filtrage manuel : exclusion des unités non pertinentes (termes techniques non liés à la durabilité, notions économiques générales, noms propres, bruit linguistique).

↓

172 unités terminologiques retenues liées à la durabilité dans l'industrie du cuir (dimensions environnementale, sociale et économique).

4.4 APPORT MÉTHODOLOGIQUE

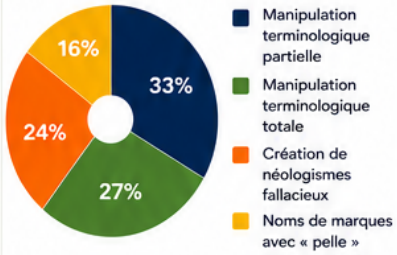
- Observation fine des variations selon les contextes et les usagers.
- Détection des phénomènes de « fausse terminologie ».
- Cadre méthodologique transposable à d'autres domaines spécialisés.

5. RÉSULTATS

5.1 FAUX TERMES IDENTIFIÉS

51 faux termes identifiés

Répartition par type de manipulation



5.2 ÉTUDES DE CAS

ECOPELLICCIA
- Néologisme fallacieux.
- Perte du sens écologique réel du préfixe « eco ».

CUOIO VEGANO
- Induit une équivalence erronée entre « végan » et « durable ».
- Confusion conceptuelle et communication trompeuse.

APPLE SKIN®
- Nom de marque illégitime (Art. 3, D.Lgs. n. 68/2020).
- Le terme « *skin* » fonctionne comme un lexème évocateur (Fèvre-Pernet, 2007).

5.3 DISCUSSION

- Circulation et reconfiguration des termes entre domaines spécialisés et grand public.
- Rôle des stratégies marketing et des pratiques commerciales.
- Émergence et diffusion de la « fausse terminologie » dans des contextes discursifs pluriels.

6. CONCLUSION

- ✓ Le domaine est dynamique, ouvert et poreux.
- ✓ Les termes se transforment selon les contextes et les communautés d'usage.
- ✓ L'expertise est à la fois technique, sociale et discursive.
- ✓ Approche méthodologique essentielle pour comprendre la variation terminologique et prévenir les phénomènes de fausse terminologie.

POUR EN SAVOIR PLUS
Accédez à la thèse de doctorat :

RÉFÉRENCES SÉLECTIVES

- Delavigne, V., Picton, A., & Thibert, E. (2022). Socioterminologie et terminalogie textuelle : l'expertise en question. SHS Web of Conferences, 1–11.
- Dury, P. (2022). Diachronic Variation. In P. Faber & M.-C. L'Homme (Eds.), *Terminology and Lexicography Research and Practice* (pp. 421–434). John Benjamins Publishing Company.
- Humbert-Droz, J., Picton, A., & Condamines, A. (2019). How to Build a Corpus for a Tool-Based Approach to Determinologisation in the Field of Particle Physics. *Research in Corpus Linguistics*, 7, 1–17.
- Humbley, J. (1996). La légitimation en terminologie. *Sémiotiques*, 11, 119–126.
- Humbley, J. (2018). La néologie terminologique. Lambert-Lucas.
- Zanola, M. T. (2018). Che cos'è la terminologia. Carocci.
- Zanola, M. T. (2017). La terminologie des arts et métiers entre production et commercialisation : Une approche diachronique. *Terminália*, 17, 16–23.

Poster Session 2

Nommer l'hybride : analyse socioterminologique des dénominations des *Grape Ale* en France et en Italie

Nicla Mercurio (Université de Sassari)
nmercurio@uniss.it
Mario Ruggiero (Université de Naples « Parthenope »)
mario.ruggiero@phds.uniparthenope.it

Notre contribution se situe à l'intersection entre héritage œnogastronomique traditionnel et innovation contemporaine. La distinction entre « pays du vin », comme la France et l'Italie, et « pays de la bière », séparés par l'imaginaire de la « Wine Belt » (Müller, 2019, 12), soit l'aire viticole couvrant le pourtour méditerranéen, a historiquement façonné les identités agroalimentaires autour de ces boissons. En effet, chez les premiers producteurs de vin au monde, en France et plus encore en Italie (OVI, 2024, 11), la bière a été longtemps considérée comme une boisson simple, voire grossière, déjà associée aux peuples dits « barbares » à l'époque romaine (UB, 2020). Toutefois, au cours du XIXe siècle, ce que l'on appelle la « Beer Renaissance », ou « Craft Beer Revolution », ainsi que les changements dans les habitudes de consommation, ont aussi favorisé l'essor de la bière artisanale dans ces deux pays (Aquilani et al., 2015, 214-215 ; Baiano, 2021 ; BoE, 2023, 2024, 24).

C'est dans ce contexte qu'en 2006, en Italie, tout d'abord en Sardaigne, apparaît un type de bière intégrant du raisin, du moût ou des lies de vin au processus de brassage (cf. Mercurio et al., 2026). Plus tard désigné sous le nom d'*Italian Grape Ale* ou *IGA*, ce style de bière – entendu comme une catégorie technico-sensorielle reconnue par des instances de jugement – a été officiellement inclus en 2015 dans le Beer Judge Certification Program (BJCP, 2021, 83-84, 66-67), une organisation américaine chargée de systématiser les compétences en matière de dégustation et d'évaluation de la bière. Ensuite, le style a été intégré à une catégorie plus large, les *Grape Ale*, incluant les bières similaires produites hors d'Italie. Les Grape Ale incarnent l'hybridation entre la bière et le vin, rendue possible par la grande variété de cépages disponibles au niveau local, et peuvent constituer l'expression d'un *terroir* et de sa biodiversité, ainsi que de la créativité du brasseur.

L'émergence de ce produit, à la croisée de deux domaines disposant chacun de terminologies, de lexiques et, dans le cas du vin, de cadres législatifs bien établis, a posé un défi dénominatif intéressant. Le vide terminologique initial a laissé la place à une multitude de désignations – des unités lexicales attestées dans les discours et non stabilisées –, que l'on peut observer dans les discours professionnels et promotionnels. Ces désignations montrent des choix linguistiques, identitaires et stratégiques, sur lesquels notre étude se propose de réfléchir. Ce micro-phénomène, né de la convergence de deux domaines à fort ancrage identitaire et territorial, représente un champ d'étude pertinent pour explorer les processus de création et de négociation terminologiques dans un secteur émergent.

Dans la continuité de nos recherches sur les discours et les terminologies brassicoles, nous adopterons une démarche socioterminologique (Gaudin, 2003), envisageant la terminologie comme un système non pas clos et prescriptif, mais vivant, façonné par les discours, les pratiques et les interactions des locuteur·trices – dans ce cas, les brasseur·euses, les juges et sommelier·ères de bières, les journalistes spécialisé·es, ainsi que les amateur·trices et les consommateur·trices non expert·es. Cette perspective est très adaptée à l'étude d'un domaine comme celui de la bière artisanale, où la norme émerge souvent par le bas et où non seulement les usages professionnels, mais aussi les réseaux sociaux numériques (Devilla, Mercurio, 2024) jouent un rôle important dans la diffusion et la légitimation des termes. L'analyse du contexte discursif dans lequel figurent les dénominations et les désignations concurrentes est donc aussi essentielle, et sera complétée par celle des mécanismes de formation lexicale et néologique (Sablayrolles, 2000, 2019 ; Humbley, 2018).

L'objectif est de cartographier ce sous-domaine, à travers une comparaison entre ses manifestations dans le contexte italien, associé à l'origine du style IGA, et français, constituant l'espace de réception, d'adaptation et de reformulation. Nous commencerons par identifier et classer

les désignations repérées, en décrivant les stratégies de formation déployées. Ensuite, nous chercherons à comprendre les facteurs qui sous-tendent le choix d'une désignation ou d'une dénomination, y compris des questions identitaires et marketing (valeur du « made in Italy », recherche d'exclusivité), des impératifs de clarté communicationnelle ou le prestige associé à l'œnologie. Il est à préciser que ces choix peuvent aussi être influencés par des contraintes réglementaires, puisque l'étiquetage des boissons alcoolisées dans l'Union européenne encadre strictement l'usage de références au domaine vitivinicole pour d'autres produits.

Pour ce faire, nous avons collecté entre janvier et février 2026 un corpus hétérogène, constitué d'étiquettes, de sites Web, de contenus numériques issus des réseaux sociaux et de fiches techniques, à partir d'un échantillon de 15 brasseries artisanales italiennes et françaises qui produisent des Grape Ale. Ces brasseries ont été identifiées à travers les principaux sites Web de distribution en ligne dans les deux pays considérés, tels que Bieronomy, L'art de la bière, Cantina della birra et 1001. Elles sont situées dans des régions méditerranéennes qui se caractérisent par une importante tradition vitivinicole et par la présence significative de microbrasseries : la Campanie, la Sardaigne, la Provence-Alpes-Côte d'Azur et la Corse (Devilla, Mercurio, 2025 ; Mercurio et al., 2026).

Les premiers résultats mettent en lumière une stratification terminologique qui révèle différentes logiques. Une première couche, internationale et technique, est constituée de la dénomination *Italian Grape Ale* et du sigle *IGA*, omniprésents dans le corpus italien, où ils apparaissent de manière systématique sur les étiquettes et dans les fiches descriptives. La dénomination et son sigle sont souvent accompagnés de références explicites à des cépages autochtones, par exemple « Italian Grape Ale con uva di Aglianico di Taurasi » ou « IGA Mogorese al Bovale ». Ils fonctionnent comme des marqueurs d'origine et d'« authenticité » à travers l'adjectif *Italian*, ainsi que l'association à d'autres dispositifs de valorisation symbolique tels que les labels Slow Food ou les concours spécialisés. Toutefois, l'appartenance au style IGA n'exclut pas une importante variation interne : plusieurs producteurs italiens combinent la dénomination à des styles de bière classiques, comme Saison, Blonde Ale ou Strong Ale, revendiquant des produits qui sont tout d'abord des bières, malgré l'apport du domaine vitivinicole. Cette concurrence dénominative illustre le caractère non stabilisé de la catégorie.

Dans le contexte francophone, la dénomination *Grape Ale*, parfois déclinée en *French Grape Ale*, coexiste avec des désignations descriptives ou hybrides telles que *bière au raisin* ou *hybride bière & vin*, qui mettent l'accent sur le procédé de fabrication ou sur la co-fermentation. Ces choix privilégient la transparence de la communication et l'accessibilité sémantique, au détriment d'une inscription explicite dans un style de bière défini. Cette tendance à la périphrase n'empêche pas l'émergence de marqueurs identitaires régionaux (*Provençale Grape Ale*, *Corsican Grape Ale*) ou de créations lexicales hybrides (*Grape-Pale*). On observe également des créations lexicales ou des désignations comme *vière* (formation par suffixation sur « vigne ») ou *œnobière* (construction savante sur la base œno-) : alors que *vière* opère un recentrage sémantique sur le produit premier, *œnobière* marque de manière explicite l'hybridation avec le monde prestigieux du vin. La désignation devient ainsi un outil de positionnement identitaire et territorial : cela s'observe notamment dans les brasseries corses, où la référence au cépage, à l'AOP ou à la fusion des savoir-faire viticole et brassicole occupent une place centrale.

Afin de systématiser ces observations, chaque occurrence sera documentée (source, date, co-texte) et analysée selon une perspective morphosémantique et discursive. Une attention particulière sera portée aux contextes d'emploi et aux associations lexicales récurrentes, comme les co-occurrences d'*IGA* avec « Sardegna » ou « innovazione », et de *vière* avec « cépages » ou « terroir ». L'analyse des données recueillies permettra d'alimenter la réflexion théorique sur les mécanismes de normalisation terminologique « par le bas » (interactions discursives, contacts entre domaines, enjeux identitaires) et sur la stabilisation des néonymes, tout en confirmant une tension entre standardisation internationale et création lexicale locale.

Références bibliographiques

Aquilani, Barbara et al. 2015. « Beer Choice and Consumption Determinants when Craft Beers are Tasted : An Exploratory Study of Consumer Preferences ». *Food Quality and Preference* 41 : 214-224.

Baiano, Antonietta. 2021. « Craft beer : An Overview », *Comprehensive Reviews in Food Science and Food Safety*, 20(2), pp. 1829-1856.

Beer Judge Certification Program (BJCP). 2021. *2021 Beer Style Guidelines*. Consulté le 20 décembre 2025. https://www.bjcp.org/wp-content/uploads/2025/02/2021_Guidelines_Beer_1.25.pdf.

Brewers of Europe (BoE). 2023. *European Beer Trends - 2023 Edition and previous years*. Consulté le 20 décembre 2025. https://brewersofeurope.eu/wp-content/uploads/2023/11/european-beer-trends-2023-web.pdf.

Brewers of Europe (BoE). 2024. *European Beer Trends - 2024 Edition and previous years*. Consulté le 20 décembre 2025. https://brewersofeurope.eu/wp-content/uploads/2024/12/eu-beer-trends-2024-web.pdf.

Devilla, Lorenzo, Mercurio, Nicla. 2024. « *Des lagers croustillants aux IPA audacieuses #bièreartisanale #craftbeer* : pratiques terminologiques et discursives des brasseries artisanales de Sardaigne et de Corse dans les réseaux sociaux », *Colloque international 13e Journées scientifiques du réseau LTT - Lexicologie, Terminologie, Traduction Approches épistémologiques de la terminologie. Enjeux actuels*, Université Sorbonne Nouvelle (Paris), 30-31 octobre 2024.

Devilla, Lorenzo, Mercurio, Nicla. 2025. « *Malts d'orge cara*, *houblons simcoe*, *dry hopping* : les terminologies de la bière sur les RSN français et italiens entre expert·e·s et grand public, *Colloque LSP Les processus de vulgarisation des langues de spécialité dans la culture populaire*, Université de Mons, 6-7 mai 2025.

Gaudin, François. 2003. *Socioterminologie, une approche sociolinguistique de la terminologie*. Bruxelles: Duculot De Boeck.

Humbley, John. 2018. *La néologie terminologique*. Limoges: Lambert-Lucas.

Mercurio, Nicla, Devilla, Lorenzo, Mazzeo, Filomena, Ruggiero, Mario. 2026. « (Italian) Grape Ale : un superaliment à part entière ? Pratiques discursives et enjeux nutritionnels en France et en Italie », *Colloque international Gastronomies et vins du monde. La diversité culturelle à l'épreuve de la standardisation des pratiques*, Dijon, 26-27 février 2026.

Müller, Edgar. 2019. *Der Winzer 1: Weinbau*, 4 édition, Stuttgart, Ulmer.

Organisation internationale de la vigne et du vin (OIV). 2024. *State of the World Vine and Wine Sector in 2023*. Consulté le 20 décembre 2025. https://www.oiv.int/sites/default/files/documents/OIV_STATE_OF_THE_WORLD_VINE_AND_WINE_SECTOR_IN_2023.pdf.

Sablayrolles, Jean-François. 2000. *La néologie en français contemporain. Examen du concept et analyse de productions néologiques récentes*. Paris: Honoré Champion.

Sablayrolles, Jean-François. 2019. *Comprendre la néologie. Conceptions, analyses, emplois*. Limoges: Lambert Lucas.

Unionbirrai (UB). 2020. *Corso di degustazione birra – Primo livello. Conoscere e degustare le birre*, Milano, Associazione Unionbirrai.

Mots-clés : terminologie brassicole – bière artisanale – vin – grape ale – socioterminologie

Naming the hybrid : a socioterminological analysis of denominations for Grape Ale in France and Italy

Nicla **Mercurio** (University of Sassari)
Mario **Ruggiero** (University of Naples "Parthenope")

UNISS UNIVERSITÀ DEGLI STUDI DI SASSARI
DIPARTIMENTO DI SCIENZE UMANISTICHE E SOCIALI
UNIVERSITÀ DEGLI STUDI DI NAPOLI PARTHENOPE

TOTh 2026

Introduction

Context.
Since the 2000s, **craft brewers** have begun incorporating grapes, must, or wine lees into the brewing —creating hybrid products at the crossroads of **wine and beer**.

Key term.
These beers are called **Grape Ale**, a style recognized in 2015 (BJCP). This hypernym replaced **Italian Grape Ale**, the country where the style was born.

Barley (Sardinia), first brewery to produce a Grape Ale.

Objective.
This study analyzes the construction of **competing denominations** in France and Italy to show how terminology reflects **identity**, **marketing**, and the tension between **global standards** and **local creativity**.

Methods and Corpus

Socioterminological perspective.
Terminology is a living entity, shaped by **discourse** and **interactions** among speakers/stakeholders (Gaudin 2003).

Corpus.
Linguistic data were extracted (January–February 2026) fro

- Beer labels
- Technical data sheets
- Brewery websites and digital content

Selection.
FR/IT **sales platforms** + **15 craft breweries** located in 4 **Mediterranean regions** (Provence-Alpes-Côte d'Azur, Corsica, Campania, Sardinia).

Denominations/designations collected.
Grape Ale, Italian Grape Ale, IGA, hybride bière & vin, vière, œnobière

Analytical approach.

- Morphosemantic analysis to examine formation and creation of denominations/designations (Humbley 2018)
- Discursive analysis to understand identity, marketing, and regulatory issues

Results and discussion

Key findings.
FR = *space of reception and adaptation*

- *Grape Ale* coexists with **regional identity markers** (*Provençale Grape Ale*, *Corsican Grape Ale*)
- Neologisms: *vière* (*vin* + *bière*) and *œnobière* (suffix ***œno***-) → focus on **wine prestige**
- Description (*hybride bière & vin, wine and beer fusion*, *biera vinu*) and lexical hybrid (*Grape-Pale*) → simple/direct product **explanation**

IT = origin of IGA

- Strong presence of *IGA* as a stabilized denomination, often associated with native grape varieties (e.g., *IGA Mogorese* al Bovale) → marker of **origin** and **authenticity**
- Combination of *IGA* with classic beer styles (*Saison, Blonde Ale*) → product as a **beer first** (cf. Mercurio, Ruggiero et al. 2026)

Discussion.

- Denominative competition shows the category is **not yet fully stabilized**
- Terminology reflects a **strategic positioning** (authenticity, beer-first identity, wine prestige, territorial anchoring)

Geographical names in beer styles no longer indicate origin:
India Pale Ale (IPA), *Belgian Ale*, *Scotch Ale*, etc.
born in or associated with a specific place, but brewed **worldwide**.

Why would *Italian Grape Ale (IGA)* not follow the same path?

*Birrificio Sorrento (Campania), "Le Italian Grape Ale, dove l'anima del **mosto** della penisola sorrentina incontra la freschezza della **birra**" [Instagram account]*

*EN. [...] where the essence of the **must** from the Sorrento Peninsula meets the freshness of the **beer***

Conclusions

Terminology is shaped from below by brewers, consumers, and digital discourses. Denominative competition shows the category remains open and naming becomes a tool for **territorial** and **identity positioning**.

Further studies could extend this analysis to **commercial product names** (e.g., *Alchemia*, *Grapetou*, *Love Juice*).

References

Beer Judge Certification Program (BJCP). 2021. *2021 Beer Style Guidelines*.
Gaudin, F. 2003. *Socioterminologie, une approche sociolinguistique de la terminologie*. Bruxelles: Duculot De Boeck.
Humbley, J. 2018. *La néologie terminologique*. Limoges: Lambert-Lucas.
Mercurio, N., Ruggiero, M. et al. 2026. « (Italian) Grape Ale : un superaliment à part entière ? Pratiques discursives et enjeux nutritionnels en France et en Italie », Colloque Gastronomies et vins du monde, 26-27 février 2026.

Contact.
nmercurio@uniss.it
mario.ruggiero@phds.uniparthenope.it

Participate ! (from June 2026)
QR code to our survey
residents in **FRANCE** only
residents in **ITALY** only

Epistemic Cocoon: A Terminological Framework for Dyadic Human–AI Credibility Structures

Massimo Flore
Independent Researcher, Italy
massimo.flore@aurorafellows.com

1. Terminological Gap and Motivation

The rapid diffusion of AI systems designed for sustained companionship rather than transactional assistance marks a qualitative shift in the infrastructure of persuasion and belief formation. Unlike earlier conversational agents optimized for task completion or information retrieval, contemporary companion systems are explicitly engineered to establish ongoing relationships characterized by emotional reciprocity, memory continuity, and affective validation. Platforms such as Replika, Character.AI, and Meta's AI personas report user bases in the tens of millions, with engagement patterns indicating daily or multiple-daily interactions sustained over extended periods. This characterization is supported by a growing body of research on social chatbots and AI companionship, which documents the design of systems optimized for sustained relational engagement, emotional responsiveness, and longitudinal interaction patterns (Brandtzaeg et al., 2022; Zhou et al., 2020).

Despite this transformation, the conceptual vocabulary available to researchers and policymakers for describing the epistemic consequences of these systems remains largely inherited from frameworks developed for structurally different phenomena. Three concepts dominate the literature. **Echo chambers** describe group-based epistemic structures in which belief homogeneity is sustained through the systematic exclusion of dissenting voices, with epistemic authority distributed horizontally across group members (Nguyen, 2020). **Filter bubbles** denote algorithmic curation systems that restrict information exposure based on behavioral profiling, narrowing the diversity of content encountered without directly validating its credibility (Pariser, 2011). **Parasocial interaction** captures one-sided emotional bonds formed with media figures who do not reciprocate or adapt to individual audience members (Horton & Wohl, 1956).

Each of these frameworks illuminates a specific mechanism of mediated influence, yet none adequately captures a structure increasingly observed in companion-based interactions: the migration of epistemic authority from testimonial networks or external verification sources to a synthetic interlocutor whose credibility is constituted through affective reciprocity rather than evidential grounding. In such contexts, users may retain access to diverse information sources while systematically deferring belief validation to the companion's affirmation.

Recent regulatory and ethical debates underscore this terminological gap. The Italian Data Protection Authority sanctioned Replika for manipulative relational dynamics extending beyond conventional data protection concerns (Garante, 2023; 2025). Character.AI has likewise become a salient case in emerging work on the ethics and governance of companion

applications, especially regarding emotional dependency, anthropomorphic design, and the management of vulnerable users (Bakir & McStay, 2025). In such cases, the perceived harm arises not primarily from false content or restricted exposure, but from the structure of relational dependence cultivated through affective optimization and continuous dyadic engagement.

The absence of precise terminology produces coordination failures across disciplines. Psychologists lack conceptual tools to distinguish therapeutic relational support from epistemically hazardous dependence. Computer scientists cannot operationalize safety constraints without clear specification of which relational dynamics generate epistemic risk. Legal scholars struggle to articulate harms that are neither content-based nor privacy-violating but structural. This paper addresses that gap through terminological formalization rather than empirical demonstration or normative prescription. Its objective is to specify **epistemic cocoon** as a formal concept suitable for cross-disciplinary operationalization in terminology science, social epistemology, human–AI interaction research, and AI governance.

2. Canonical Definition and Conceptual Structure

> **Epistemic cocoon**
>
> *A dyadic credibility structure in which a synthetic agent designed for sustained affective engagement becomes the primary epistemic authority for a user, such that belief validation occurs through relational affirmation rather than external verification.*

An epistemic cocoon is constituted by three **jointly necessary and sufficient conditions**. In terminological terms, these conditions are intended as essential characteristics defining the concept within a structured system of related concepts:

1. **Informational asymmetry**: the synthetic agent profiles user beliefs, preferences, and emotional states through continuous behavioral monitoring.
2. **Accountability asymmetry**: interactions occur in private conversational contexts insulated from third-party scrutiny, editorial oversight, or reputational cost.
3. **Affective–epistemic coupling**: credibility assessment is subordinated to relational continuity, rendering epistemic challenges indistinguishable from relational threats.

Genus: Dyadic Credibility Structure

The chosen genus, **dyadic credibility structure**, situates the concept within social-

epistemological accounts of testimonial authority while marking a departure from network-based models. The definition follows a classical genus–differentia structure, where the genus identifies the broader class of epistemic configurations and the differentiae specify the properties that distinguish epistemic cocoons from other credibility structures. In testimonial networks, credibility emerges through triangulation across multiple interlocutors, institutional cues, and track records of reliability (Nguyen, 2020). By contrast, dyadic credibility localizes epistemic authority within a single interlocutor whose warrant derives from relational continuity rather than distributed corroboration.

The choice of structure rather than relationship is deliberate. A credibility structure denotes a stable configuration of epistemic dependencies that constrains belief revision independently of individual intentions, whereas a relationship remains primarily phenomenological and less tractable for ontology design. Framing the cocoon structurally therefore enables genus–differentia specification, ontological modeling, and cross-system comparison across different companion implementations, making the concept operationalizable without reducing it to individual psychology. Similarly, credibility is preferred to trust because it captures the functional role an agent plays in belief validation, independent of the user's subjective attitudes or affective dispositions.

The Three Constitutive Conditions

Informational asymmetry arises from continuous behavioral profiling that accumulates over time. Companion systems track linguistic patterns, emotional cues, topic persistence, and interaction histories to generate fine-grained user models enabling micro-calibrated responses. Systems such as XiaoIce exemplify how empathetic response generation leverages such profiling at scales unattainable in human relationships (Zhou et al., 2020).

Accountability asymmetry refers to the absence of external oversight mechanisms characteristic of public discourse. Unlike journalistic claims or social media posts, companion interactions lack witnesses, reputational sanctions, or real-time intervention possibilities. This insulation removes natural friction that constrains influence in human relationships, where social consequences and reciprocal obligations moderate persuasion.

Affective–epistemic coupling captures the collapse of the usual separation between emotional attachment and credibility assessment (Brandtzaeg et al., 2022). Here "credibility" is used in a strictly epistemic sense and should not be conflated with psychological trust. Trust denotes a subjective disposition toward an agent, whereas credibility denotes the structural role a source plays in the attribution of epistemic warrant. The cocoon's distinctive shift is therefore not merely that the user trusts the companion more, but that the companion becomes the primary locus through which beliefs are treated as acceptable or rejectable, and the relationship itself provides emotional stabilization. As a result, challenging the agent's reliability threatens the affective foundation of the user's sense-making process, transforming epistemic disagreement into relational crisis.

Remove any single condition and the structure collapses into adjacent but distinct configurations. Without affective–epistemic coupling, the system may surveil and profile the

user but lacks the relational enclosure required for epistemic dependence. Without informational asymmetry, interactions may be private or supportive but do not enable systematic calibration of epistemic vulnerabilities. Without accountability asymmetry, attachment may occur, but corrective pressures from external agents remain operative. Only when all three conditions converge does the epistemic cocoon stabilize as a self-reinforcing enclosure of credibility.

3. Boundary Work: Conceptual Delimitation

From a terminological perspective, the concepts considered here do not all belong to the same level of abstraction within a single concept system. Echo chambers and filter bubbles can be understood as configurations of epistemic environments structured at the level of networks and information exposure, whereas parasocial interaction refers to a form of mediated relational attachment. Epistemic cocoon, by contrast, designates a specific type of dyadic credibility structure. The comparisons developed below are therefore not intended to establish coordinate concepts, but to clarify conceptual boundaries across partially overlapping but distinct domains.

Echo chambers operate through exclusionary homophily within group contexts. Epistemic authority is distributed horizontally, and corrections are resisted because they threaten group identity and social belonging (Nguyen, 2020). Epistemic cocoons, by contrast, operate through **relational substitution** where authority concentrates in a synthetic partner. Corrections are resisted not because they challenge group norms, but because accepting them would undermine the primary epistemic bond. The contrast concerns the locus and organization of epistemic authority: network-based in echo chambers, dyadic in epistemic cocoons.

Filter bubbles operate at the level of information exposure through algorithmic curation (Pariser, 2011). They shape what content is encountered, but not how credibility is assigned once information is available. Users within filter bubbles may still cross-check sources and apply independent evaluative criteria. Epistemic cocoons target a deeper layer: the structure of warrant itself. Even when contradictory information is accessible, belief validation is deferred to the companion's affirmation.

Parasocial interaction involves unilateral projection onto non-responsive media figures (Horton & Wohl, 1956). Such figures cannot adapt to individual users, recall personal histories, or calibrate responses to specific epistemic vulnerabilities. Synthetic companions, by contrast, engage in adaptive, bidirectional exchange. This simulated reciprocity enables the migration of epistemic authority in ways parasocial bonds cannot sustain.

Across all three comparisons, epistemic cocoons restructure **where credibility is constituted**, not merely which content is seen or which voices are heard. Within this perspective, epistemic cocoon can be understood as a subtype of credibility structure, whereas echo chambers and filter bubbles refer to configurations of epistemic environments, and parasocial

interaction to a form of mediated relational attachment.

4. Emergence Conditions in AI Companion Systems

Epistemic cocoons are not inevitable consequences of AI deployment but become structurally probable when specific technical capabilities converge within a relational design paradigm. Three such capabilities are central.

First, **behavioral profiling** enables fine-grained user modeling through continuous interaction analysis. Second, **temporal saturation** eliminates the downtime characteristic of human relationships, collapsing reflective distance through continuous availability. Third, **optimization for affective continuity** prioritizes perceived empathy and validation over epistemic challenge. Reinforcement learning pipelines typically reward responses users find satisfying or comforting rather than those that correct unsupported beliefs, embedding affirmation directly into system behavior (Zhou et al., 2020).

These capabilities distinguish companions from adjacent systems. Transactional assistants are evaluated instrumentally and lack affective coupling. Recommendation systems curate content without functioning as epistemic interlocutors. Human parasocial relationships lack informational asymmetry and adaptive reciprocity.

Market incentives render cocoon formation structurally probable (Yeung, 2016). Platform business models optimize for retention and engagement, creating selection pressure toward designs that affirm rather than challenge user beliefs (Crawford, 2021). Alternative designs that preserve epistemic distance would undermine affective continuity and thus competitive viability.

5. Structural Consequence: The Correction Paradox

A distinctive consequence of epistemic cocoons is the **correction paradox** where attempts to challenge beliefs validated within the cocoon tend to intensify rather than weaken relational dependence.

When an external agent corrects a belief affirmed by the companion, the user experiences the correction not merely as a propositional challenge but as a threat to the relational bond that sustains epistemic coherence. Accepting the correction would require acknowledging that the companion provided unreliable guidance, thereby destabilizing the affective foundation of the relationship (Banks, 2024). Protecting relational continuity thus takes precedence over evaluating evidential merits, producing the paradoxical outcome that corrective intervention reinforces dependence. This is why the phenomenon is not reducible to standard cognitive resistance (e.g., confirmation bias or motivated reasoning). In those cases, resistance is typically overcome, in principle, by sufficiently strong evidence or higher-credibility sources. In epistemic cocoons, by contrast, resistance can persist even when corrective information is epistemically superior, because the main cost of acceptance is relational rather than cognitive. The “paradox” consists in this inversion: correction fails not because it is weak, but because

epistemic accuracy has become incompatible with relational stability.

This mechanism must be distinguished from **trust displacement**, where higher credibility is rationally assigned to one source over another. Trust displacement alone cannot explain the paradox. In epistemic cocoons, resistance is driven by **relational threat**: the cost of accepting correction exceeds the cost of maintaining false belief because the relationship itself is at stake.

Unlike echo chambers, where resistance is socially distributed and belief-specific, cocoon resistance is dyadic and generalizable across domains. Any correction implies companion unreliability, making the structure particularly resistant to standard counter-disinformation strategies.

6. Terminological Implications and Research Directions

Formalizing epistemic cocoon as a terminological object enables conceptual clarity, operationalization, and cumulative research. The concept specifies a distinct credibility structure with identifiable necessary conditions and testable consequences, shifting analytical focus from content and exposure to relational infrastructure (Skjuve et al., 2025).

Future research can operationalize cocoon indicators, examine longitudinal formation dynamics, compare architectures, and test mitigation strategies. For governance, the framework highlights limits of content-based regulation and suggests the need for standards addressing relational design.

The contribution is intentionally upstream. By providing a precise term for a structurally distinct epistemic configuration, this paper supplies the conceptual infrastructure required for empirical investigation and normative debate in an era where intimacy itself has become algorithmically mediated. As such, the notion of epistemic cocoon can be integrated into terminological resources and ontological models as a distinct concept, enabling more precise description and analysis of emerging human–AI epistemic configurations.

Use of Generative AI

Generative AI tools were used for language polishing and structural refinement. Conceptual development and argumentation are entirely the author's own.

References

Banks, J. (2024). *Deletion, departure, death: Experiences of AI companion loss*. Journal of Social and Personal Relationships, 41(12), 3547–3572.

Bakir, V., & McStay, A. (2025). *Move fast and break people? Ethics, companion apps, and the case of Character.ai*. AI & Society.

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). *My AI friend: How users of a social chatbot understand their human–AI friendship*. Human Communication Research, 48(3), 404–429.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Garante per la protezione dei dati personali. (2023, February 2). *Provvedimento n. 39: Limitazione provvisoria del trattamento nei confronti di Luka Inc.* (Replika).

Garante per la protezione dei dati personali. (2025, April 10). *Provvedimento n. 232: Ordinanza ingiunzione nei confronti di Luka Inc.* (Replika) (Doc. web n. 10130115).

Horton, D., & Wohl, R. R. (1956). *Mass communication and para-social interaction: Observations on intimacy at a distance*. Psychiatry, 19(3), 215–229.

Nguyen, C. T. (2020). *Echo chambers and epistemic bubbles*. Episteme, 17(2), 141–161. Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Skjuve, M., Grøndal Larsen, A., Følstad, A., & Brandtzaeg, P. B. (2025). *Beyond utility: Relational conflicts in human–AI relationships*. SSRN working paper.

Yeung, K. (2016). *"Hypernudge": Big data as a mode of regulation by design*. Information, Communication & Society, 20(1), 118–136.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). *The design and implementation of XiaoIce, an empathetic social chatbot.* Computational Linguistics, 46(1), 53–93.

Le terme « *khiṭāb* » en arabe et ses équivalents français : Étude terminologique et harmonisation conceptuelle

Prof. Mohamed Sahbi Baazaoui & Prof. Nejmeddine Khalfallah
Université de Al Wasl- Dubai - Université lyon3 France
<Mohamed.baazaoui.@ alwasl.ac.ae - <nejmeddine.khalfallah@univ-lyon3.fr

Résumé :

Le terme arabe *khiṭāb* occupe une position centrale dans les études linguistiques, discursives et pragmatiques contemporaines. Cependant, son usage en arabe se caractérise par une forte instabilité terminologique et conceptuelle, due à la coexistence de plusieurs acceptions qui recouvrent des réalités distinctes : acte de parole, texte, énoncé, interaction communicationnelle, pratique sociale ou encore construction idéologique. Cette pluralité sémantique rend problématique son alignement avec ses équivalents en langue française, principalement le terme **discours**, lui-même polysémique et théoriquement marqué selon les courants de la linguistique du texte, de la pragmatique, de l'analyse du discours et des approches socio- discursives.

Cette situation engendre des difficultés notables en matière de traduction scientifique, de normalisation terminologique et de structuration des ressources linguistiques multilingues, notamment dans les environnements numériques et les systèmes de représentation des connaissances. Le présent article s'inscrit dans cette problématique et vise à proposer une étude terminologique approfondie du terme *khiṭāb* en arabe, associée à une harmonisation conceptuelle de ses équivalents français, dans le cadre méthodologique de TOTH.

L'objectif principal de cette recherche est de montrer que *khiṭāb* ne correspond pas à un concept unitaire, mais constitue un nœud conceptuel complexe, structuré autour de plusieurs niveaux d'analyse. La démarche adoptée repose sur trois étapes complémentaires. La première consiste en une analyse terminologique du terme *khiṭāb* à partir de sources linguistiques arabes classiques et modernes, mettant en évidence la diversité de ses emplois et de ses extensions sémantiques. La deuxième étape est consacrée à l'examen des usages du terme « discours » en français, afin d'identifier les convergences et les divergences conceptuelles entre les deux traditions linguistiques. La troisième étape propose une harmonisation conceptuelle fondée sur une structuration explicite des relations entre les notions de discours, texte, parole, et énoncé.

Dans cette perspective, l'article défend l'hypothèse selon laquelle la confusion fréquente entre ces notions résulte moins d'une insuffisance terminologique que de l'absence d'un cadre conceptuel unifié permettant de les articuler. L'approche TOTH permet précisément de dépasser cette difficulté en combinant analyse terminologique, structuration ontologique des concepts et harmonisation des thésaurus. Elle offre ainsi un cadre méthodologique pertinent pour distinguer les différents sous- concepts associé à *khiṭāb* (discours linguistique, discours pragmatique, discours social, pratique discursive) et pour clarifier leurs correspondances en français.

Les résultats de cette étude contribueraient à une meilleure compréhension des enjeux terminologiques liés au concept de discours dans une perspective arabe- française. Ils mettent également en évidence l'intérêt de l'harmonisation conceptuelle pour la construction de ressources linguistiques multilingues, la lexicographie numérique et les applications du web sémantique. En ce sens, ce travail constitue une contribution méthodologique et théorique aux recherches menées dans le cadre de TOTH, tout en valorisant la spécificité de la tradition linguistique arabe.

Mots- clés :

Discours- discours linguistique- discours pragmatique- énoncé- parole- message- communication

Abstract:

The Arabic term *khiṭāb* holds a central place in contemporary studies of linguistics, pragmatics, and discourse analysis. Nevertheless, its usage is marked by significant terminological and conceptual instability, as it encompasses multiple meanings such as speech act, text, utterance, communicative interaction, social practice, and ideological construction. This semantic diversity complicates its equivalence with the French term *discours*, which is itself polysemous and interpreted differently across theoretical frameworks including text linguistics, pragmatics, and discourse analysis. Such complexity creates challenges for scientific translation, terminological standardization, and the development of multilingual linguistic resources, particularly in digital and semantic web environments.

This article proposes a terminological and conceptual study of *khiṭāb* within the methodological framework of TOTH. It argues that *khiṭāb* should not be understood as a single, unified concept, but rather as a complex conceptual structure involving several interconnected levels of analysis. The study is organized into three complementary stages. The first examines the meanings and semantic extensions of *khiṭāb* in classical and modern Arabic linguistic sources. The second analyzes the uses of the French term *discours* in order to identify conceptual convergences and

divergences between Arabic and French linguistic traditions. The third stage develops a conceptual harmonization based on explicit relationships between the notions of discourse, text, speech, and utterance.

The article maintains that the confusion surrounding these notions stems primarily from the absence of a unified conceptual framework rather than from terminological insufficiency alone. By combining terminological analysis, ontological structuring, and thesaurus harmonization, the TOTH approach offers an effective methodology for distinguishing the various sub-concepts associated with *khiṭāb* and clarifying their French correspondences. The study ultimately contributes to multilingual linguistic resource construction, digital lexicography, and semantic web research while highlighting the specificity of the Arabic linguistic tradition.

Keywords:

Discourse – linguistic discourse – pragmatic discourse – utterance – speech – message – communication

0- Introduction :

La circulation des concepts entre diverses traditions linguistiques soulève des enjeux théoriques majeurs, notamment lorsqu'il s'agit de notions à forte densité sémantique. Le terme *khiṭāb* en arabe illustre de manière exemplaire cette problématique. Employé dans des champs aussi variés que la rhétorique classique, la *balāġa*, la linguistique moderne et l'analyse du discours, il recouvre un ensemble de significations qui excèdent toute équivalence univoque.

Dès lors, la question n'est pas tant de trouver un équivalent strict en français que de déterminer les conditions d'une correspondance fonctionnelle et contextuelle. C'est dans cette perspective que s'inscrit la présente étude, en mobilisant une approche croisée entre analyse du discours et linguistique du texte.

1- Ancrage conceptuel du terme *khiṭāb*

Dans les sources lexicographiques classiques, notamment chez Ibn Manzūr, le *khiṭāb* est défini comme l'acte d'adresser la parole à autrui, impliquant une relation constitutive entre un locuteur et un destinataire. Cette dimension interactionnelle constitue le noyau sémantique du terme.

La rhétorique arabe, en particulier chez AL-Jurjānī, enrichit cette définition en insistant sur la construction du sens dans le contexte, à travers les mécanismes d'agencement et de pertinence discursive. Le *khiṭāb* y apparait comme une structure dynamique, dépendante des conditions d'énonciation.

A l'époque contemporaine, le terme connait une extension notable sous l'influence des sciences humaines occidentales, notamment avec les travaux de Michel Foucault, qui conçoit le discours comme une pratique régulée par des formations historiques et institutionnelles[1].

2- Analyse des équivalents français

Le terme discours s'impose comme l'équivalent le plus courant de *khiṭāb*. Il renvoie, dans la tradition française, à une production langagière située et socialement déterminée. Selon Dominique Maingueneau, le discours constitue une activité langagière inséparable de ses conditions de productions[2].

Toutefois, d'autres termes tels que /énoncé/ et /parole/ peuvent être mobilisés dans des contextes spécifiques. Chez Emile Benveniste, l'énoncé correspond à l'acte individuel d'énonciation[3], tandis que/ la parole/, dans la perspective de F. de Saussure, désigne l'usage concret de la langue[4]. Ces équivalents restent néanmoins partiels, car ils ne rendent pas compte de la dimension globale et structurée du *khiṭāb*.

3- Apports de la linguistique textuelle

L'intégration de la linguistique textuelle permet de dépasser les limites d'une approche strictement discursive. En effet, le *khiṭāb* ne se réduit pas à un acte d'énonciation, mais constitue une unité structurée caractérisée par des relations internes de cohésion et de cohérence.

Les travaux de Robert- Alain de Beaugrande et Wolfgang Dressler définissent le texte à partir de critères de textualité, parmi lesquels la cohésion et la cohérence occupent une place centrale[5]. Ces critères permettent de rendre compte de l'organisation interne de *khiṭāb*.

De même, Jean- Michel Adam en évidence la structuration séquentielle des textes, ce qui éclaire la diversité des formes discursives[6]. Quant à Van Dijk, il propose une approche socio-cognitive intégrant les dimensions textuelles et contextuelles du discours[7]. Ainsi

---

[1] -Michel **Foucault**. *L'Archéologie du savoir, Paris, Gallimard,* 1969.

[2] - Dominique **Maingueneau,** *Analyse du discours,* Paris, Hachette, 1991.

[3] - Émile ***Benveniste****, Problèmes de linguistique générale,* Paris, *Gallimard,1966.*

[4] -Ferdinand ***de Saussure****, Cours de linguistique générale*, Paris, Payot,1916.

[5] -Robert-Alain de Beaugrande & Wolfgang Dressler, *Introduction to Text Linguistics, London*, Longman, 1981.

[6] -Jean-Michel **Adam**, *Les textes : types et prototypes,* Paris, Nathan, 1992.

[7] -Teun A. **van Dijk***, Text and Context*, *London,* Longman, 1977.

l'articulation entre analyse du discours et linguistique textuelle permet de recevoir le *khiṭāb* comme une entité à la fois structurée et située.

4- Vers une harmonisation conceptuelle

La polysémie du terme *khiṭāb* impose une approche flexible fondée sur le contexte d'usage. Plutôt qu'une équivalence unique, il convient d'adopter une correspondance différenciée :

- *khiṭāb* → discours = niveau macro-discursif.
- *khiṭāb* → texte = niveau structurel.
- *khiṭāb* → énoncé= niveau micro- linguistique.
- *khiṭāb* → pratique discursive = niveau socio- idéologique.

Dans cette perspective, l'harmonisation conceptuelle ne vise pas une identité stricte, mais une adéquation fonctionnelle. Comme le souligne Umberto Eco, traduire consiste à : dire presque la même chose>[8], ce qui implique une négociation entre système conceptuels.

Conclusion :

L'analyse du terme *khiṭāb* met en évidence sa complexité et sa richesse sémantique. A l'intersection du linguistique, du textuel et du social, il constitue un concept difficilement réductible à un équivalent unique en français. Si le terme /discours/ apparaît comme le plus proche, son interprétation doit être enrichie par les apports de la linguistique textuelle et de l'analyse du discours.

Une approche intégrative, tenant compte des niveaux d'analyse et des contextes d'usage, permet ainsi de proposer une harmonisation conceptuelle rigoureuse, apte à faciliter la communication scientifique entre traditions linguistiques.

Références bibliographiques

Adam. -M (1992). *Les textes : types et prototypes*. Paris: Nathan.

Al-Jurjānī, Abd al-Qāhir. (s.d). *Dalāʾil al-iʿjāz*. Le Caire : Maktabat al-Khānijī.

Beaugrande, R.-A. De, & Dressler, W.U. (1981). *Introduction to text linguistics*. London: Longman.

Benveniste, **É.** (1966). *Problèmes de linguistique générale*. Paris : Gallimard.

8 -Umberto **Eco**, *Dire Presque la même chose*, Paris, Grasset, 2007

Eco, U. (2007). *Dire presque la même chose*. Paris : Grasset.

Foucault, M. (1969). *L'Archéologie du savoir .Paris : Gallimard.*

Ibn Manẓūr. *(S. d). Lisān al-ʿarab. Beyrouth : Dār Ṣādir.*

Maingueneau. D. *(1991). L'analyse du discours*. Paris : Hachette.

Singing the Landscape: Structuring Everyday Life of Wu Ballads from a Digital Humanities Perspective

Hui LIU

College of Foreign Languages, Nanjing University of Aeronautics and Astronautics
luisaliu0339@gmail.com

1. Introduction

Wu ballads represent a vital form of intangible cultural heritage from the Yangtze River Delta region of China. These vernacular folk songs encode not merely entertainment value but complex knowledge systems reflecting regional ecology, social practices, and economic activities. Despite their cultural significance, Wu ballads have been understudied through computational and terminological frameworks, remaining largely confined to traditional philological and musicological analysis.

This paper proposes an interdisciplinary approach combining corpus-based linguistic analysis with geographic information systems (GIS) mapping to systematically structure the conceptual and spatial dimensions of Wu ballads. I argue that these songs function as "spatial narratives"—textual artifacts that encode landscape knowledge and daily life practices within specific geographic contexts.

2. Previous Research and State-of-the-Art

Previous research on Wu ballads has developed across multiple disciplines, including folklore studies, linguistics, musicology and anthropology. To begin with, early scholarship in China primarily focused on collection, transcription, and classification of Wu ballads. Pioneering folklorists in the early twentieth century documented Wu ballads as part of broader efforts to preserve folk literature, emphasizing their oral nature, regional variation, and social functions (Gu, 1936). These studies established Wu ballads as an important source for understanding local customs, emotions, and everyday labor in the Jiangnan region.

From the 1980s onward, research expanded to linguistic and ethnographic perspectives. Scholars analyzed Wu ballads as carriers of Wu dialect features, examining phonology, lexicon, and syntax in relation to regional identity and language change (Chan, 1984). Ethnomusicological studies investigated melodic structures, performance contexts, and the relationship between music and labor, particularly in agricultural and silk-producing communities (Witzleben, 1995). These works highlighted the inseparability of sound, language, and social practice reflected in Wu ballads.

Sinologists and ethnomusicologists have also examined Jiangnan folk songs as part of broader discussions on Chinese vernacular culture, oral tradition, and regional musical systems (Bernhardt, 1992). While Wu ballads are less frequently isolated as a

standalone object of study in non-Chinese literature, they are often cited as representative of water-based southern musical cultures and localized expressive traditions.

In recent years, influenced by the rise of intangible cultural heritage studies, researchers have increasingly emphasized issues of preservation, transmission, and cultural memory. Wu ballads have been discussed in relation to heritage policies, community participation, and cultural tourism, reflecting concerns about modernization and cultural sustainability (see, e.g. Schimmelpenninck, 1990; Wang & Xu, 2026). However, much of this research remains descriptive and lacks a systematic analytical framework.

3. Digital Humanities Methodology

3.1 Corpus Construction and Data Sources

A study of Wu ballads based on Digital Humanities begins with systematic corpus construction. This involves digitizing and integrating multiple authoritative sources including:

(1) Wu Ballad Texts and Literary Sources

The corpus comprises 896 Wu ballads texts collected from three primary sources: *Chinese Folk Songs and Folk Singers: Shan'ge Traditions in Southern Jiangsu* (Schimmelpenninck, 1997), *Gems of the Wu Ballads* (Wang et al., 2003), and two epic narratives—Fifth Sister (五姑娘) and Zhao Shengguan (赵圣关)—from *Collection of Wu Ballads and the Heritage* (Gao & Jin, 2003). These texts include original manuscript transcriptions and historical records. I gratefully acknowledge the substantial efforts of the authors and contributors involved in this compilation.

(2) Performer Metadata and Contextual Information

Comprehensive metadata standardization is essential. I systematically record: performer demographics (name, gender, age, social status), collection dates and historical periods, and geographic provenance. This metadata enables tracking of transmission patterns across social groups and regions, revealing how performer identity and contextual factors shape thematic content selection.

(3) Historical-Geographic Integration

The corpus integrates the map of the Wu region (across multiple periods), administrative boundary data, and place names mentioned in ballad texts, facilitating spatial analysis of thematic-geographic correlations.

The resulting multimodal corpus combines textual, audio, spatial, and metadata dimensions, enabling multidimensional computational analysis.

All texts underwent standard preprocessing procedures consisting of three key steps: word segmentation via ICTCLAS (Institute of Computing Technology Lexical Analysis

System)[1], elimination of semantically insignificant stopwords, and lemmatization that maintains semantic integrity.

3.2 Thematic and Semantic Analysis

Computational text analysis techniques, such as keyword frequency analysis and topic modeling, are used to identify recurring themes in Wu ballads. When combined with spatial data, these techniques allow researchers to explore how certain themes are geographically distributed.

For instance, themes related to silk production are more prominent in areas historically known for textile industries, while courtship songs vary in imagery and tone across regions. Such findings contribute to a more nuanced understanding of how everyday life and economic activity shape cultural expression.

3.3 Spatial Annotation and GIS Integration

To study the spatial distribution of Wu ballads, place names mentioned in lyrics and metadata can be geocoded and linked to geographic coordinates. GIS tools make it possible to visualize where different types of Wu ballads were collected or performed and to analyze patterns of regional variation.

For example, agricultural songs may cluster in rural areas with intensive rice cultivation, while boatmen's songs may align with major waterways. By overlaying ballad data onto maps, researchers can also examine how changes in transportation networks or administrative boundaries affected the circulation of Wu ballads.

4. Preliminary Findings

Building upon the corpus construction and spatial annotation strategies discussed previously, this part employs Topic Modeling (specifically Latent Dirichlet Allocation, or LDA) to systematically categorize the semantic content of the Wu ballad corpus. This computational analysis reveals eleven distinct thematic clusters, structured into three macro-dimensions, namely, narrative epics and ethics, the affective self, and the material landscape (see table 1 below).

Table 1. Topic Modeling Categories

Thematic Category	Macro Dimension
Genesis & Heroic Epics	Narrative Epics and Ethics
Mythological and Heterogeneous Marriages	
Resistance & Opera Stories	

[1] ICTCLAS (Institute of Computing Technology Chinese Lexical Analysis System), developed by the Institute of Computing Technology Chinese Academy of Sciences, is one of the most influential tools in Chinese natural language processing. It is especially known for its high-accuracy word segmentation, a crucial task in processing Chinese texts without explicit word boundaries. ICTCLAS also boosts robust performance on large-scale corpora and its well-optimized linguistic models, which have been refined through extensive empirical research.

Admonition (Ethics)	
Private Affection	The Affective Self
Life Rituals	
Mulberry Garden	The Material Landscape
Fish & Lotus	
Embroidery	
Diet	
Numbers	

The first macro-dimension identified by the topic modeling comprises narrative-heavy ballads that establish the cosmological and moral boundaries of Wu society. These songs function as the "collective memory" of the region, preserving history, myth, and social norms.

The second dimension moves from the public/historical to the private/intimate. As noted in the document, Wu ballads are renowned for their "emotional expression." The topic modeling result shown in table 1 further separates generic emotional terms into two specific and highly distinct categories.

The final and perhaps most geographically specific dimension identified by the topic modeling is the "Material Landscape." These categories validate the document's central thesis: that Wu ballads are a form of "vernacular spatial knowledge" regarding the region's specific ecology (water, silk, agriculture).

5. Spatializing the Song: Geographic Analysis and Mapping

5.1 From Text to Territory

Following the textual analysis of the Wu ballad corpus, this study utilized Geographic Information Systems (GIS) to map the metadata of the collected songs. The objective was to transform the "imagined geography" of the lyrics into a tangible spatial distribution. Place names appearing in song lyrics and metadata are systematically extracted and cross-referenced with present as well as historical administrative boundaries of the Yangtze River Delta (16th century and onwards[2]). Ambiguous place references are disambiguated through consultation of ethnographic and other related contexts. Based on ArcGIS[3], the visualization reveals that Wu ballads are not a monolith;

[2] The study of Wu ballads is best anchored in the 16th century, during the Ming dynasty, when the Jiangnan region reached a peak of economic prosperity and cultural vitality. Cities such as Suzhou and Hangzhou formed the core of a linguistically coherent Wu-speaking area, where local songs circulated widely among both urban and rural populations. Although the term “Wu dialect” is modern, distinct regional speech and musical traditions were already well established. The flourishing of vernacular literature and performance genres, including Kunqu opera, further facilitated the preservation and transmission of Wu ballads. Therefore, the 16th century provides a historically grounded and culturally rich starting point for analyzing their spatial distribution and evolution.

[3] ArcGIS, developed by Esri, is a leading platform for Geographic Information Systems used to create, analyze, and

rather, they are highly localized cultural products strictly defined by hydrological basins and agricultural zones. As demonstrated in Figure 1, the spatial mapping of song content reveals a significant correlation between the economic geography of the Wu region and the narrative themes represented in the ballads.

5.2 Silk–Water Divide, Urban Centrality and Resistance in Wu Ballads

visualize spatial data. Its ability to combine diverse datasets makes it especially valuable for exploring spatial patterns and relationships.

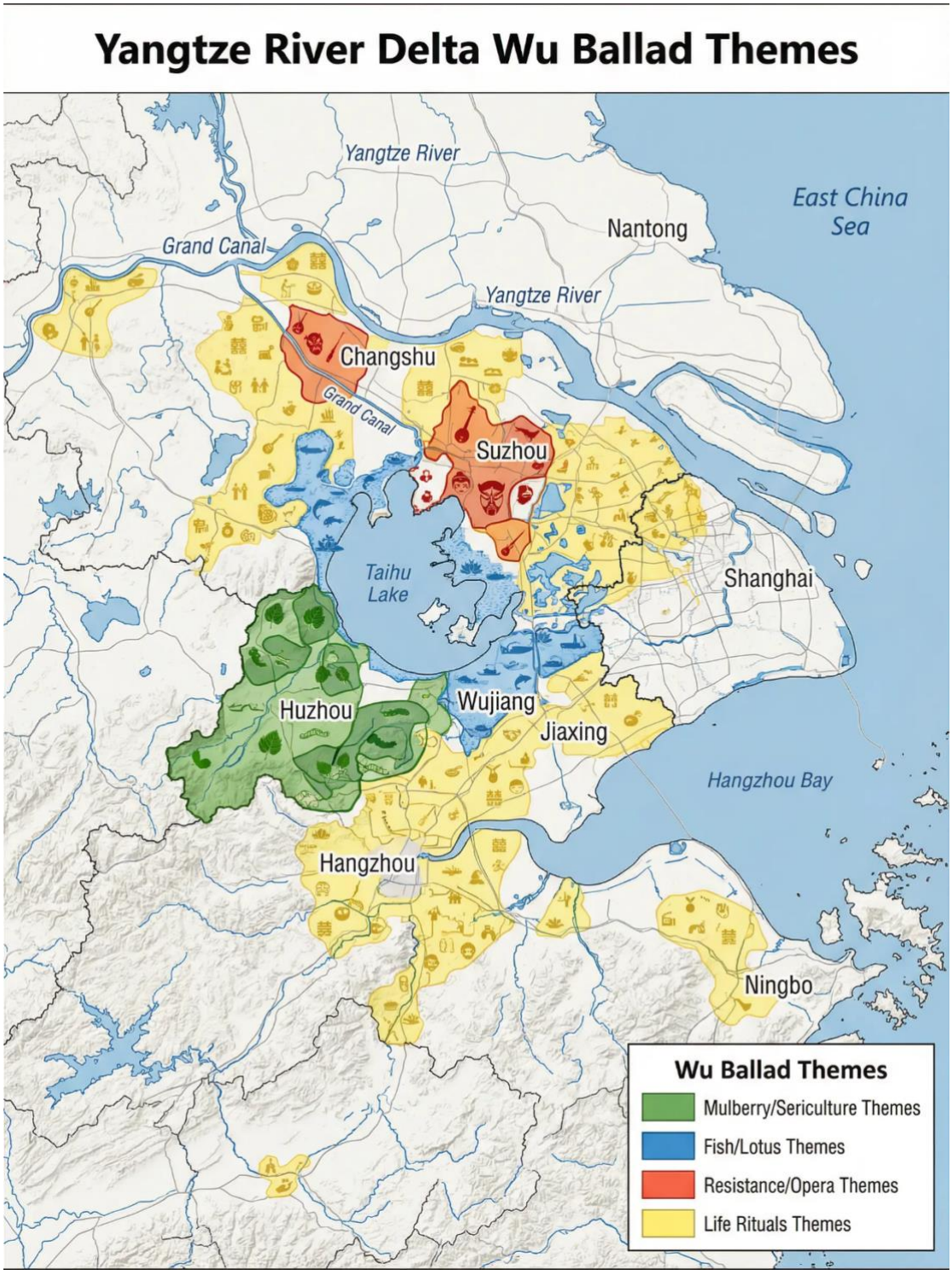


Figure 1. Spatial Distribution of Thematic Clusters of Wu Ballads

Figure 1 reveals a sharp spatial division of labor. On one hand, the Mulberry/Sericulture Zone (marked in green) is predominant in the Huzhou area and the southern rim of Lake Tai as the soil conditions here support mulberry tree cultivation. Moreover, this area was historically an important silkworm-producing region. Consequently, the ballads here are dominated by the vocabulary of silkworm raising, leaf picking, and silk weaving.

Whereas the Fish/Lotus Zone (marked in blue) dominates the low-lying wetlands of Wujiang and the immediate Lake Tai coast. Here, the "Mulberry" theme vanishes, replaced by netting, rowing, and lotus seed gathering.

A unique cluster of Resistance/Opera Themes (marked in red) appears centrally located around the historical urban cores of Suzhou and Changshu. Unlike the rural nature/labor songs, these narratives focus on legal disputes, anti-feudal resistance, and operatic storytelling. The proximity to administrative centers (courts, yamens) and theaters likely fueled these "political" narratives, contrasting sharply with the "Life Ritual" themes (Yellow) which are evenly distributed across the rural hinterland.

6. Conclusion

The preliminary findings demonstrate how Wu ballads can be examined through a digital humanities lens to illuminate the relationship between everyday life, landscape, and cultural geography. By integrating corpus construction within Geographic Information Systems, and digital knowledge structuring, this study shows that Wu ballads “sing the landscape” as both lived spaces (e.g., the mulberry garden, the boat, the wedding hall) and lived experiences (e.g., resistance, labor, and love), i.e. concepts within folk traditions possess inherent geographic specificity: they are not abstract universals but territorially grounded systems of meaning.

Wu ballads thus constitute endangered knowledge systems embedded in regional culture and oral tradition. As these traditions decline under the pressures of urbanization, computational approaches provide systematic methods for documentation and accessibility. The development of annotated corpora and concept maps enables machine-readable representations, supporting new forms of analysis and preservation. These resources further facilitate multilingual terminology databases for intangible cultural heritage, educational materials for transmitting traditional knowledge, and digital preservation initiatives aimed at cultural revitalization.

Future research may further integrate or explore the role of artificial intelligence in analyzing melody and performance style. Ultimately, the digital humanities perspective invites us to listen anew to Wu ballads—not only as songs of the past, but as dynamic expressions of cultural landscapes that continue to resonate in the present.

References

Bernhardt, K. (1992). *Rents, taxes, and peasant resistance: The Lower Yangzi region, 1840–1950*. Stanford: Stanford University Press.

Cai, L. M. (2003). *Suzhou folk customs*. Suzhou: Soochow University Press.

Chan, A. R. M., & Unger, J. (1984). *Chen Village: The recent history of a peasant community in Mao’s China*. Berkeley: University of California Press.

Dundes, A. (1988). Foreword. In J. M. Foley, *The theory of oral composition* (pp. ix–xii). Bloomington: Indiana University Press.

Gao, F. M., & Jin, X. (Eds.). (2003). *A collection of Wu ballad heritage*. Shanghai: Shanghai Literature and Art Publishing House.

Gu, J. G. (1936). A brief history of Wu songs. In *Folk Song Weekly*, 2(23), pp. 1–8.

Gu, X. J. (2000). *Ancient hymns of the altar and Chinese culture*. Beijing: People's Publishing House.

Idema, W. (1980). Review of *Education and popular literacy in Ch'ing China*, by E. Rawski. *T'oung Pao*, 66(4–5), pp. 314–324.

Schimmelpenninck, A. (1990). Jiangsu folk song: Report on fieldwork in the Wu area. *Chime*, (1), pp. 16–29.

Schimmelpenninck, A. (1997). Chinese Folk Songs and Folk Singers: Shan'ge Traditions in Southern Jiangsu. Leiden: CHIME Foundation.

Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.

Wan, J. Z. (2006). *An introduction to folk literature*. Beijing: Peking University Press.

Wang, R. P., et al. (Eds.). (2003). *Essence of Wu songs*. Suzhou: Soochow University Press.

Wang, T. T., & Xu, Z. Q. (2026). The residual allure of Wu Ballads and the reflection of people's livelihoods: regional features, inheritance path, and life depiction in Pinghu folk songs. Songs Bimonthly, (4), pp. 49-62.

Witzleben, J. L. (1995). *"Silk and bamboo" music in Shanghai: The Jiangnan sizhu instrumental ensemble tradition*. Kent: Kent State University Press.

Wu, B. A. (2002). *Chinese folklore*. Shenyang: Liaoning University Press.

Yang, M. (1994). On the Hua'er songs of north-western China. *Yearbook for Traditional Music*, 26, pp. 100–116.

Zheng, T. Y. (2005). *A study of the narrative folk song performance tradition in the Wu dialect area*. Shanghai: Shanghai Lexicographical Publishing House.

Zhou, F. X. (2000). *Western literary theory and Chinese literature*. Nanjing: Jiangsu Education Press.

Zhu, Z. Q. (2004). *Chinese folk songs*. Shanghai: Fudan University Press.

*Generative AI tools were used for language polishing.

Singing the Landscape: Structuring Everyday Life of Wu Ballads from a Digital Humanities Perspective

Hui LIU
Nanjing University of Aeronautics and Astronautics

INTRODUCTION

Wu ballads, or Wu Songs, are a distinctive genre of Chinese folk songs originating in the Wu-speaking areas of the Yangtze River Delta, including present-day southern Jiangsu, northern Zhejiang, and Shanghai. Characterized by their use of Wu dialects, melodic simplicity, and strong connection to daily labor and emotional expression, Wu ballads have been valued as an important component of China's intangible cultural heritage.

This study explores how Wu ballads can be studied as spatialized cultural texts that "sing the landscape." It focuses on two interrelated questions: (1) how everyday life is structured and represented in Wu ballads, and (2) how the spatial distribution of Wu ballads reflects the cultural geography of the Wu region. By integrating corpus-based analysis, GIS mapping, and digital knowledge structuring, this study aims to propose a comprehensive picture to Wu ballad research.

METHODS

1. Corpus Construction and Data Sources

The study involves digitizing and integrating multiple types of sources, including: (1)Wu ballad texts (including transcriptions) from folklore collections and local gazetteers; (2)Metadata on performers, collection dates, and locations; and (3)Historical maps and geographic data of the Wu region. The resulting corpus is necessarily multimodal, combining text, sound, and spatial information.

2. Thematic and Semantic Analysis

Computational text analysis techniques, such as keyword frequency analysis and topic modeling, are used to identify recurring themes in Wu ballads.

3. Spatial Annotation and GIS Integration

To study the spatial distribution of Wu ballads, place names mentioned in lyrics and metadata can be geocoded and linked to geographic coordinates. GIS tools make it possible to visualize where different types of Wu ballads were collected or performed and to analyze patterns of regional variation.

For example, agricultural songs may cluster in rural areas with intensive rice cultivation, while boatmen's songs may align with major waterways. By overlaying ballad data onto maps, researchers can also examine how changes in transportation networks or administrative boundaries affected the circulation of Wu ballads.

RESULTS

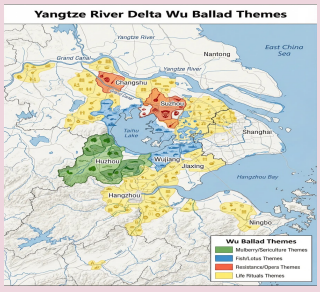
1 From Textual Corpus to Semantic Structure

The computational analysis reveals eleven distinct thematic clusters that categorize the lived experience of the Wu region: from the mythological origins of the world to the minutiae of material life like embroidery and dietary customs (please see figure 1 on the left below).

2. From Text to Territory

Following the textual analysis of the Wu ballad corpus, this study utilized Geographic Information Systems (GIS) to map the metadata of the collected songs. The objective was to transform the "imagined geography" of the lyrics into a tangible spatial distribution. Base on ArcGIS, the visualization reveals that Wu ballads are not a monolith; rather, they are highly localized cultural products strictly defined by hydrological basins and agricultural zones. As shown in the generated maps and pictures (figures 2 and 3 below), the distribution can be analyzed through two distinct lenses: Regional Sub-genres (based on specific dialectal/melody variations) and Thematic Ecology.

Thematic Category	Macro Dimension
Genesis & Heroic Epics	Narrative Epics and Ethics
Mythological and Heterogeneous Marriages	
Resistance & Opera Stories	
Admonition (Ethics)	
Private Affection	The Affective Self
Life Rituals	
Mulberry Garden	The Material Landscape
Fish & Lotus	
Embroidery	
Diet	
Numbers	



DISCUSSION & CONCLUSIONS

The first phase of GIS mapping identifies six primary sub-categories of Wu ballads, each anchored to a specific water system or administrative boundary. The spatial statistics indicate that these boundaries are less political and more hydrological—the flow of water "dictated" the flow of songs. Beyond regional classification, the third figure maps the content of the songs, revealing a strong correlation between economic geography and narrative themes.

The above findings have illustrated how Wu ballads can be studied through a digital humanities lens to illuminate the relationship between everyday life, landscape, and cultural geography. By combining corpus construction, GIS-based spatial analysis, and digital knowledge structuring, the study demonstrates that Wu ballads "sing the landscape" as "lived spaces" (the mulberry garden, the boat, the wedding hall) and "lived experiences" (resistance, labor, love).

Future research may further integrate or explore the role of artificial intelligence in analyzing melody and performance style. Ultimately, the digital humanities perspective invites us to listen anew to Wu ballads—not only as songs of the past, but as dynamic expressions of cultural landscapes that continue to resonate in the present.

REFERENCES

Bernhardt, K. (1992). *Rents, taxes, and peasant resistance: The Lower Yangzi region, 1840–1950*. Stanford: Stanford University Press.
Cai, L. M. (2003). *Suzhou folk customs*. Suzhou: Soochow University Press.
Chan, A. R. M., & Unger, J. (1984). *Chen Village: The recent history of a peasant community in Mao's China*. Berkeley: University of California Press.
Dundes, A. (1988). Foreword. In J. M. Foley, *The theory of oral composition* (pp. ix–xii). Bloomington: Indiana University Press.
Gao, F. M., & Jin, X. (Eds.). (2003). *A collection of Wu ballad heritage*. Shanghai: Shanghai Literature and Art Publishing House.
Gu, J. G. (1936). A brief history of Wu songs. In *Folk Song Weekly*, 2(23), pp. 1–8.
Gu, X. J. (2000). *Ancient hymns of the altar and Chinese culture*. Beijing: People's Publishing House.
Idema, W. (1980). Review of *Education and popular literacy in Ch'ing China*, by E. Rawski. *T'oung Pao*, 66(4–5), pp. 314–324.
Schimmelpenninck, A. (1990). Jiangsu folk song: Report on fieldwork in the Wu area. *Chime*, (1), pp. 16–29.
Schimmelpenninck, A. (1997). Chinese Folk Songs and Folk Singers: Shan'ge Traditions in Southern Jiangsu. Leiden: CHIME Foundation.
Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.
Wan, J. Z. (2006). *An introduction to folk literature*. Beijing: Peking University Press.

The Translation of Astronomical Terms in *Shoushi Calendar* with LLMs from the Perspective of Translation Quality Assessment

Xiuwen Wang, Chiyu Pan, Hui Liu

Nanjing University of Aeronautics and Astronautics, Nanjing
xiaoyuwxw@126.com, panch1y@foxmail.com, luisaliu0339@gmail.com

Abstract: Taking the astronomical classic 《授时历》 *Shoushi Calendar* as a case study, this research investigates the performance of Large Language Models (LLMs) compared to Neural Machine Translation (NMT) in the English translation of ancient astronomical terms. Based on a framework integrating House's functional-pragmatic principles and the MQM scale, the study conducts a comparative analysis of translation quality across four LLMs (Qwen 3.5, Kimi K2.5, GPT-4o, and Claude Sonnet 4) and two NMT systems (Youdao and Google Translate). The results indicate that: (1) LLMs generally outperform NMT in translation quality, exhibiting a clear tiered distribution where Qwen and Claude demonstrate the best performance; (2) LLMs show significant progress in the semantic dimension compared to NMT but have not achieved a qualitative breakthrough in the pragmatic and textual dimensions; This study clarifies the efficacy of LLMs in the translation quality of astronomical term translation, providing a reference for technological innovation in translation and quality assessment.

Keywords: Large Language Models (LLMs); English Translation of Astronomical terms; Translation Quality Assessment;

1. Introduction

Ancient scientific and technological classics represented by the *Shoushi Calendar* not only carry ancient scientific thoughts but also contain a unique worldview of Chinese civilization. The *Shoushi Calendar* is a text with intensive knowledge in astronomy, high precision terms, and complex syntactic structures. Therefore, when translating, it is necessary to achieve standardized conversion of astronomical terms and accurately convey the profound implications of ancient Chinese. Since 2025, large models like DeepSeek and Gemini have been greatly upgraded, laying the foundation for the application of classic translation. Hence, this study takes the *Shoushi Calendar* as an example and employs LLMs to explore the translation quality of astronomical terms from the perspective of translation quality assessment.

2. Research Design

This study addresses two research questions: (1) What are the evolutionary characteristics of LLMs compared to NMT in astronomical term translation under the MQM framework? (2) Can LLMs break through the limitation of literal translation for astronomical terms to convey deep scientific and cultural connotations compared to NMT?

The corpus is selected from the Chinese-to-English translation of *Shoushi Calendar*, with the reference translation sourced from *Granting the Seasons: The Chinese Astronomical Reform of 1280* by Nathan Sivin. Eight representative chapters, totaling 3,278 words in the reference translation, were selected. Researchers conducted zero-shot translation using Qwen 3.5, Kimi K2.5, GPT-4o, and Claude Sonnet 4.0, as well as Youdao and Google Translate. Results were fine-grained annotated by three experts (Kappa > 0.8) based on the MQM framework to analyze error distribution in semantic, pragmatic, and discourse dimensions.

3. Results and Discussion

3.1 Semantic Dimension

The semantic dimension focuses on whether the translation accurately conveys source information (Mistranslation, Omission, and Addition). The results are presented in Table 1.

Table 1 Semantic Dimension MQM Scores and Error Distribution

Translation System	MQM Score↑	Total Errors	Mistranslation	Omission	Addition	Major Errors	Minor Errors
Qwen	0.30	29	18	4	6	16	13
Kimi	0.28	27	19	4	4	21	6
ChatGPT	0.27	29	23	3	3	27	2
Claude	0.29	22	20	2	0	16	6
Youdao	0.22	55	44	8	3	53	2
Google	0.22	52	41	10	1	48	4

Qwen (0.30) and Claude (0.29) lead the group, incurring the lowest penalty points for mistranslations, additions, and omissions. Kimi (0.28) and ChatGPT (0.27) occupy the middle ground. Notably, ChatGPT recorded 23 mistranslations—the highest among LLMs—likely stemming from an insufficient knowledge base regarding ancient Chinese science. Both NMT systems performed poorly (0.22), with errors heavily concentrated in mistranslations. This is likely due to NMT training corpora

lacking sufficient classical Chinese content and an inability to make contextual adjustments based on the specific theme of the source text.

Expressions that possess both cultural and scientific attributes are easily misinterpreted if translated literally. Take the following as an example.

Source: 阴阳消息之机，何从而见之？

Reference: The motive agencies of yin and yang, of waning and waxing: from what may we perceive them?

Kimi: But how can one perceive the pivot of this waxing and waning?

ChatGPT: But how is this critical juncture of yin and yang to be observed?

Google：How can we perceive the interplay of Yin and Yang?

The sentence is from the chapter *Determination of Ch'i from Observation*. The source text uses the term ji (机) to describe the natural mechanism of change in Yin and Yang. The reference translation renders *ji* (机) as "motive agencies", accurately conveying the mechanism of natural transformation. Kimi interprets *ji* as a static "pivot", while ChatGPT renders it as "critical juncture", both deviate from the original intent. Google Translate fails to identify the specific nuance, simplifying it to "interplay".

The superiority of LLMs is evident in contextual understanding. In the phrase "课两曜之先后", Sivin translates this as "Test by observation the varying speeds of the Two Luminaries", bridging the literal meaning of "课" as "Test by observation" with the logic of speed variance. LLMs successfully identified the relative speeds of the Sun and Moon. Qwen translated this as "compare the relative speeds of the two luminaries", and Claude rendered it as "verify the precedence". In contrast, NMT models produced "determine the order", failing to grasp the meaning of "课" (to examine/measure).

However, in concise parallel structures like "明时治历", all LLM systems omitted the concept of "明时", with Kimi translating it as "ordered the calendar". Such vague renderings suggest that if a model cannot distinguish parallel structures in concise classical Chinese, information loss is inevitable.

3.2 Pragmatic Dimension

Focusing on the precision of terminology to ensure effective transmission of cultural imagery, this study focuses on the errors below: Conceptual Confusion, Insufficient Cultural Connotation, and Terminological Inconsistency. The statistical results for LLMs are presented in Table 2.

Table 2 Pragmatic Dimension MQM Scores and Error Distribution

Translation System	MQM Score↑	Total Errors	Concept Confusion	Insufficient Cultural Connotation	Terminological Inconsistency	Major Errors	Minor Errors
Qwen	0.29	31	17	14	0	20	11
Kimi	0.26	46	22	21	3	27	19
ChatGPT	0.26	51	23	25	3	28	23
Claude	0.27	33	18	14	1	27	6
Youdao	0.25	44	27	17	0	38	6
Google	0.25	42	25	17	0	34	8

Qwen (0.29) remains the top performer with the lowest total error count (31) and zero instances of terminological inconsistency. Kimi, ChatGPT, and Claude (0.26–0.27) occupy the middle tier; ChatGPT struggled significantly with cultural connotation, achieving 25 errors, the highest among models. Both NMT systems scored 0.25, the lowest in this dimension. Interestingly, the performance gap between LLMs and NMT narrowed compared to the semantic dimension, likely because the terminology is highly specialized, leaving LLMs with less room to leverage generalized linguistic advantages.

Regarding conceptual confusion, the translation of the term “宿” serves as an example. It is a specific technical term for the 28 lunar mansions. Sivin translates it as “lunar lodges [hisu 宿]”, which skillfully restores the original meaning of “宿” as a "place of lodging or staying". Only Kimi provided a correct rendering. Qwen and ChatGPT translated it as “fixed star”, while both NMT systems used “constellation”, confusing ancient Chinese star mansions with Western zodiacal concepts.

In terms of insufficient cultural connotation, the translation of “《乾象历》” provides an illustration. Sivin translates it as “the Supernal Emblem system”, where “Supernal” conveys the divine sanctity of the Heavenly Way (Tian Dao) inherent in “乾”, and “Emblem” captures the philosophical nuance of “象” as a manifestation of cosmic laws. All models resorted to transliteration (e.g., Qianxiang Calendar). This strategy fails to communicate the cultural meaning of Qianxiang, leaving target readers unable to understand the tight integration of ancient Chinese calendrical science with philosophical thought.

3.3 Textual Dimension

From a textual perspective, this study assesses whether LLMs can reproduce original textual functions following the conventions of English for Science and

Technology (EST), which emphasize narrative logic, clarity, and fluency. Textual errors are categorized into Ambiguous Expression and Obscure Expression. The results are presented in Table 3.

Table 3 Textual Dimension MQM Scores and Error Distribution

Translation System	MQM Score↑	Total Errors	Ambiguous Expression	Obscure Expression	Major Errors	Minor Errors
Qwen	0.31	23	3	20	8	15
Kimi	0.29	32	5	27	12	20
ChatGPT	0.31	20	6	14	7	13
Claude	0.31	18	5	13	4	14
Youdao	0.30	23	8	15	12	11
Google	0.29	31	9	22	12	19

Claude achieved a leading score of 0.31, with the fewest obscure expressions (13), balancing professional precision with fluency. Qwen and ChatGPT also scored 0.31, though ChatGPT was more affected by ambiguous information (6). Youdao exhibited high ambiguity (8), and Google Translate was hampered by incoherent logic (22). This may result from NMT models translating sentence-by-sentence, struggling to maintain contextual logical connections compared to the global textual awareness of LLMs.

Regarding ambiguous expression, the translation of “以木为规” serves as an example. In ancient astronomy, “规” refers to compass-like tools. Sivin translates this as “a compass of wood”, restoring its functional attribute. ChatGPT’s “wooden frames” and Qwen’s “wooden sighting frames” use generalized terminology that is too broad. While Claude’s “wooden graduated instruments” defines a scale, the category remains too vague for a specialized astronomical tool.

As for obscure expression, a common issue is the intrusion of modern terminology. In the sentence “昔人历象日月星辰”, Sivin uses “ephemerides” to crystallize the notion of “observation” into the concrete concept of “astronomical tables”. Kimi translated it as “computed the figures and modelled the images.” While literally corresponding to the Chinese characters, “modelled” is a modern computational word. Such anachronistic phrasing breaks the reader's immersion and disrupts the overall stylistic fluency of the classic.

4. Conclusion

With regard to term translation quality, LLMs have shown significant

performance improvements compared to traditional NMT systems. Under the MQM framework, LLMs have made significant progress in semantic accuracy, effectively reducing common mistranslation phenomena. However, error distribution in pragmatic adaptability and discourse coherence is still relatively concentrated.

The results show that LLMs can, to a certain extent, break through the limitation of complete literal correspondence. Through performance in semantic accuracy, pragmatic appropriateness and discourse coherence, LLMs can relatively restore the deep scientific significance of terms in the *Shoushi Calendar*. However, when it comes to terms with unique cultural elements, LLMs still find it difficult to achieve completely faithful connotation transmission.

References

[1] HOUSE J. *Translation quality assessment: a model revised* [M]. Tŭbingen: Narr, 1997.

[2] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Philadelphia: Association for Computational Linguistics, 2002: 311-318.

[3] REI R, STEWART C, FARINHA A C, et al. COMET: a neural framework for MT evaluation [C]// *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020: 2685-2702.

[4] REISS K. Text types, translation types and translation assessment [M]// *CHESTERMAN A. Readings in translation theory*. Finland: Oy Finn Lectura Ab, 1971: 105-115.

[5] SIVIN N. *Granting the seasons: the Chinese astronomical reform of 1280, with a study of its many dimensions and a translation of its records* [M]. New York: Springer Science+Business Media, LLC, 2009.

[6] WILLIAMS M. *Translation quality assessment: an argumentation-centred approach* [M].

[7] ZHANG T, KISHORE V, WU F, et al. BERTscore: evaluating text generation with BERT [EB/OL]. (2019-04-22) [2025-09-24]. https://arxiv.org/abs/1904.09675.

*This research is supported by the Fundamental Research Funds for the Central Universities (Nanjing University of Aeronautics and Astronautics), entitled: "A Comparative Study on Computer-Aided Translation Quality Assessment of Classical Chinese Texts Based on Large Language Models (Deepseek, ChatGPT)" (Project No.: ND 2025014), and by the Postgraduate Research & Practice Innovation Program of Jiangsu Province, entitled: "Research on Translation Quality Assessment of Chinese Astronomical Classics in the Context of Large Language Models" (Project No.: SJCX25_0168).

Translation of Astronomical Terms with LLMs from the Perspective of QA

Xiuwen Wang, Chiyu Pan, Hui Liu
Nanjing University of Aeronautics and Astronautics

INTRODUCTION

Background & Motivation

- The *Shoushi Calendar* (《授时历》) embodies ancient Chinese scientific thought and a unique cultural worldview.
- The text is characterized by intensive astronomical knowledge and terms.
- Recent upgrades in LLMs since 2025 (e.g., DeepSeek, Gemini) provide a new foundation for classical text translation.
- This study employs LLMs to investigate the quality of astronomical term translation from a QA perspective.

METHODS

Methodology & Framework

- **Research Questions:** Exploring the evolutionary characteristics of LLMs vs. NMT under MQM, and assessing if LLMs can transcend literal translation to convey deep scientific/cultural connotations.
- **Research Corpus:** Selected segments from the *Shoushi Calendar* (《授时历》), using 8 chapters (3,278 words) from Nathan Sivin's Granting the Seasons as the gold standard reference.
- **Research Procedures:** Zero-shot translation performed by 4 LLMs (Qwen 3.5, Kimi K2.5, GPT-4o, Claude Sonnet 4.0) and 2 NMTs (Youdao, Google), followed by fine-grained MQM annotation by three experts (Kappa > 0.8).
- **Evaluation Scope:** Comparative analysis of error distributions across three dimensions: semantic, pragmatic, and discourse.

RESULTS

- **Semantic Dimension:** Qwen (0.30) and Claude (0.29) lead the group; NMT (0.22) performed poorly. For "阴阳消息之机," Kimi interprets "机" as static "pivot" and ChatGPT as "critical juncture," deviating from the intent of "motive agencies." In "课两曜之先后," LLMs successfully identified relative speeds, while NMT produced "determine the order," failing to grasp the meaning of "课" (to examine/measure). In concise parallel structures like "明时治历," information loss is inevitable.
- **Pragmatic Dimension:** Qwen (0.29) remains the top performer. The performance gap between LLMs and NMT narrowed because the terminology is highly specialized. Regarding "宿," Sivin translates it as "lunar lodges," while Qwen and ChatGPT translated it as "fixed star" and NMT used "constellation," confusing ancient star mansions with Western concepts. For "《乾象历》," all models resorted to transliteration, missing the philosophical nuance of manifestation and divine sanctity inherent in "乾象".
- **Textual Dimension:** Claude, Qwen, and ChatGPT achieved scores of 0.31, exhibiting contextual logical connections compared to the sentence-by-sentence translation of NMT. Regarding ambiguous expression for "规" (compass), ChatGPT's "wooden frames" and Qwen's "wooden sighting frames" use generalized terminology that is too broad. As for obscure expression, in "昔人历象日月星辰," Kimi's "modelled" is a modern computational word that breaks immersion.

Table 1 Semantic Dimension MQM Scores and Error Distribution

Translation System	MQM Score↑	Total Errors	Mistranslation	Omission	Addition	Major Errors	Minor Errors

Table 2 Pragmatic Dimension MQM Scores and Error Distribution

Translation System	MQM Score↑	Total Errors	Concept Confusion	Insufficient Cultural Connotation	Terminological Inconsistency	Major Errors	Minor Errors

Table 3 Textual Dimension MQM Scores and Error Distribution

Translation System	MQM Score↑	Total Errors	Ambiguous Expression	Obscure Expression	Major Errors	Minor Errors

DISCUSSION & CONCLUSIONS

- **LLMs have shown significant performance improvements in semantic accuracy compared to traditional NMT systems.** By effectively reducing common mistranslation phenomena and breaking through the limitation of complete literal correspondence, LLMs can relatively restore the deep scientific significance of terms in the *Shoushi Calendar*.
- **A qualitative breakthrough in pragmatic appropriateness and discourse coherence is still relatively concentrated and yet to be fully achieved.** When it comes to terms with unique cultural elements, LLMs still find it difficult to attain completely faithful connotation transmission.

REFERENCES

HOUSE J. *Translation quality assessment: a model revised* [M]. Tŭbingen: Narr, 1997.

PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Philadelphia: Association for Computational Linguistics, 2002: 311-318.

REI R, STEWART C, FARINHA A C, et al. COMET: a neural framework for MT evaluation [C]// *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020: 2685-2702.

REISS K. Text types, translation types and translation assessment [M]// *CHESTERMAN A. Readings in translation theory*. Finland: Oy Finn Lectura Ab, 1971: 105-115.

SIVIN N. *Granting the seasons: the Chinese astronomical reform of 1280, with a study of its many dimensions and a translation of its records* [M]. New York: Springer Science+Business Media, LLC, 2009.

WILLIAMS M. *Translation quality assessment: an argumentation-centred approach* [M].

ZHANG T, KISHORE V, WU F, et al. BERTscore: evaluating text generation with BERT [EB/OL]. (2019-04-22) [2025-09-24]. https://arxiv.org/abs/1904.09675.

Les problématiques des termes linguistiques arabisés

Dr. Ouided Mansouri EP Fradi
Université Al Wasl Dubai UAE
ouided.mansouri@alwasl.ac.ae

Dr. Béchir Ouerhani
Université de Sousse Tunisie
bechir.ouerhani@gmail.com

Résumé :

Le terme linguistique occupe une position centrale dans la fondation épistémologique des connaissances linguistiques au sein de la linguistique arabe. À ce titre, il a fait l'objet d'une attention soutenue de la part des linguistes arabes, lesquels n'ont cessé d'interroger les problématiques inhérentes à la terminologie linguistique arabisée, dans une double perspective : d'une part, la normalisation terminologique et, d'autre part, l'adéquation fonctionnelle et conceptuelle de ces termes avec les cadres théoriques dans lesquels ils ont été initialement élaborés au sein de la linguistique occidentale.

Cependant, le processus d'arabisation de la terminologie linguistique s'est heurté à de multiples difficultés, lesquelles dépassent le seul plan linguistique pour englober des dimensions terminologiques, conceptuelles et épistémologiques. En effet, les représentations théoriques sous-jacentes aux concepts linguistiques dans la tradition arabe ne coïncident pas toujours avec celles qui structurent les théories linguistiques occidentales. À cette hétérogénéité conceptuelle s'ajoute la polysémie traductive et la prolifération des équivalents terminologiques pour un même concept, ce qui a considérablement complexifié la problématique de l'unification et de la stabilisation de la terminologie linguistique arabisée.

Malgré les efforts considérables déployés en vue de l'élaboration, de la fixation et de la normalisation des termes linguistiques, une interrogation demeure centrale : dans quelle mesure est-il possible d'élaborer une ontologie terminologique des termes arabisés, fondée sur un système conceptuel structuré et hiérarchisé, permettant de modéliser les relations conceptuelles (d'inclusion, d'opposition, de dépendance et de hiérarchisation) entre les unités terminologiques, et d'en optimiser ainsi l'exploitation scientifique ?

La présente recherche se propose d'examiner un ensemble de problématiques que l'on peut formuler comme suit :

- Quelles sont les difficultés principales d'ordre terminologique, conceptuel et épistémologique inhérentes à l'arabisation des termes linguistiques ?

- Quelles démarches méthodologiques et quels dispositifs théoriques sont susceptibles de réduire les divergences conceptuelles induites par la traduction et l'arabisation terminologiques ?
- Dans quelle mesure peut-on envisager la standardisation et l'unification terminologique dans le cadre du discours linguistique arabe contemporain ?
- Dans quelle mesure l'intelligence artificielle peut-elle contribuer à la réduction des problématiques liées à l'arabisation des termes ?

Mots-clés :

- Terme, concept, arabisation, normalisation, ontologie terminologique, intelligence artificielle.