

# Domains in Motion: A NLP Framework for Mapping Wind Energy Terminology

Serena Sassi<sup>1\*†</sup> and Andrea Lops<sup>2\*†</sup>

<sup>1</sup>Dipartimento di Ricerca e Innovazione Umanistica, Università degli Studi di Bari Aldo Moro, Piazza Umberto I, Bari, 70121, Apulia, Italy.

<sup>2</sup>Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari, Via Edoardo Orabona, 4, Bari, 70125, Apulia, Italy.

\*Corresponding author(s). E-mail(s): [serena.sassi@uniba.it](mailto:serena.sassi@uniba.it); [andrea.lops@poliba.it](mailto:andrea.lops@poliba.it)

†These authors contributed equally to this work.

## Abstract

Wind energy constitutes a dynamic and interdisciplinary domain, shaped by technological innovation, evolving societal debates, and the interplay of diverse knowledge fields. Understanding its terminology is challenging, as concepts change quickly and overlap across different areas. This study introduces LUMEN, a computational method that uses Large Language Models (LLMs) and semantic embeddings to identify and organize subdomains within wind energy. Using a corpus of texts from 2016 to 2024 collected via Sketch Engine, LUMEN maps relationships among terms and highlights connections across the field. By providing translators and terminologists with a clear framework to visualize and navigate these complex terminological networks, the method demonstrates how NLP-based approaches can make the analysis of evolving, interdisciplinary domains both more precise and accessible.

**Keywords:** Terminology; Natural Language Processing; Large Language Model; Domain Analysis; Wind Energy

## 1 Introduction

In terminological research, the delineation and organization of complex domains constitutes one of the most persistent and multifaceted challenges. This difficulty arises not only from the inherent complexity of defining the conceptual boundaries of a field of knowledge, but also from the heterogeneity and interdisciplinarity that characterize many contemporary domains. Unlike disciplines with relatively stable terminological systems, areas such as environmental studies and renewable energy resist rigid categorization. They encompass diverse epistemological traditions, methodological approaches, and practical applications, thus complicating efforts to establish coherent taxonomies or domain-specific glossaries.

The domain of environmental studies and the energy transition provides a clear illustration of this challenge. The growing urgency of climate change and the global shift towards sustainability have brought these domains to the forefront of scientific, public, and political discourse. This is not a static, well-defined field; it is a dynamic space characterized by rapid technological innovation, evolving public policy, and intense societal debate. Consequently, the terminology of renewable energy is marked by a constant influx of neologisms, the mainstreaming of specialist terms, and the adoption of concepts from intersecting fields like economics, law, and social science. This rapid, multi-source evolution of language creates significant ambiguity, semantic instability, and the very interoperability challenges that now require systematic

terminological investigation (Delavigne, 1994; Fløttum, 2010; Georgescu, 2010; L’Homme, 2015; Zanola, 2010).

Renewable energy – and wind energy in particular – provides a paradigmatic example of such complexity. While often approached as a primarily technical or engineering field, wind energy is intrinsically multidimensional, intersecting with discourses on ecology, environmental impact, economics, public policy, safety, and even cultural and societal debates. This multiplicity of perspectives involved implies that no single taxonomy or hierarchical classification can adequately capture the full range of knowledge associated with this domain. Moreover, the rapid expansion of specialized terminology within renewable energy further accentuates the need for methodologies capable of systematically identifying, classifying, and representing concepts in a way that reflects both established disciplinary structures and emergent subdomains. This inherent multidimensionality poses significant semantic and interoperability challenges, complicating data integration and knowledge sharing across the energy transition landscape (Mirqosim, 2025).

As with many emerging technologies, wind energy terminology has rapidly expanded beyond its initial, purely technical environment. It is now prevalent in national and local programs, corporate communications, and various forms of media language. Describing the wind energy lexicon thus requires engaging with the administrative, legal, political, and regulatory frameworks specific to each country. Indeed, each nation has adopted different codes of practice governing the development, production, incentives, sale, and distribution of this energy (Sassi, 2025). Consequently, despite this expansion, terminological research in this sector remains far from complete. Compared to other energy forms and established industries, the wind energy domain has long been overlooked in terminological research. It has often been treated terminologically as a secondary component of ‘renewable energy’, rather than as a distinct technological field and activity in its own right.

To date, research efforts to manage terminological complexity in domain analysis have primarily focused on two fronts: the development of formal ontologies and the application of term extraction techniques. Several efforts have aimed to create structured vocabularies and ontologies, providing formal, top-down representations of specific technical aspects (Costa et al., 2024). Concurrently, studies in computational linguistics have evaluated Automatic Term Extraction (ATE) methods in various technical domains (Delavigne, 2002), including wind energy (Lops & Sassi, 2025; Sassi, 2024, 2025). The evaluation of such ATE systems typically relies on standardized datasets, as highlighted in related computational and methodological studies (Lefever & Terryn, 2024; Zhang, Zou, & Du, 2025), which provide standards for term-level validation. However, while existing studies provide rigid, top-down structures for known concepts, there is a lack of research on using data-driven, bottom-up Natural Language Processing (NLP) approaches to systematically map the entire conceptual structure of the wind energy domain. Previous works have not focused on automatically identifying the hierarchical relationships between domains, subdomains, and sub-subdomains from a terminological perspective.

The present study directly addresses this gap, since it situates itself at the intersection of terminological research, computational linguistics, and environmental studies. Its primary aim is to investigate how NLP methodologies can support the systematic identification of domains, subdomains, and sub-subdomains within the field of wind energy. By doing so, this work contributes to the broader discussion on how terminological research can adapt to the increasing complexity and interdisciplinarity of contemporary scientific and technical fields. Recent advances in computational linguistics and NLP offer the potential to address this specific gap. By combining corpus-based approaches with the analytical capacities of Large Language Models (LLMs) and semantic similarity techniques, it has become possible to move beyond manual classifications, developing instead dynamic, data-driven models of knowledge organization, which may subsequently be examined, validated, and refined by domain experts. These approaches allow for the construction of hierarchical domain trees that not only provide terminological precision but also capture the complex, evolving, and often overlapping nature of interdisciplinary fields.

Rather than replacing traditional methodologies, computational models could also serve as complementary tools that enhance terminological practice by uncovering latent structures, relationships, and semantic fields otherwise difficult to perceive. This approach could be particularly valuable for translators and terminologists working in highly complex domains such as renewable energy, as it provides them with systematic instruments to visualize semantic fields, understand how they are interconnected, and identify terms that occur across multiple fields. In doing so, it enables professionals engaged in terminological

research and translation to approach their analyses with greater clarity and precision, thereby facilitating more consistent and informed decision-making in contexts where conceptual boundaries are fluid and interdisciplinary overlaps are pervasive.

## 2 Genesis of the Issue

The present study addresses two interrelated problems that have emerged in the course of investigating the terminological field of wind energy. The first concerns the conceptual status of the notion of “domain”, a category that has long functioned as a cornerstone of terminology studies but whose adequacy has been increasingly called into question in light of contemporary developments. The second problematic pertains to the methodological means by which terminological knowledge may be represented and systematized, with particular emphasis on the role of ontologies. These issues will be examined in detail in the following sections, where their implications for the analysis and organization of complex terminological landscapes are considered in depth.

### 2.1 Rethinking the Boundaries of Environmental Domains

The concept of “domain” occupies a central role across various academic fields, from linguistics and philosophy to sociology and information sciences (Rey, 1979). Within the field of terminology, it serves as a foundational principle used to organize and classify scientific and technical terms, thereby structuring knowledge coherently and systematically. However, this traditional conception of the “domain”, grounded in rigid classifications and boundaries, warrants reconsideration in light of contemporary shifts in both knowledge production and linguistic practices. Although such classifications may have once sufficed in contexts characterized by less terminological proliferation, the increasing complexity of science and technology, alongside the hybridization of specialized and general discourse, has rendered this notion increasingly problematic and open to debate.

In this context, it is essential to critically assess the relevance of the “domain” as it is presently conceived, especially in fields like environmental studies, where technical terms frequently overcome specialized spheres and infiltrate general language and public discourse. As Delavigne (2022) emphasizes, the attempt to precisely delineate a specific domain often leads to significant challenges, particularly in addressing complex subjects such as environmental issues. The environment, by its very nature, is a multidimensional field that involves continuous interaction among various disciplines, relying on a vast network of interconnections with technical, scientific, social, and cultural discourses (Grimaldi, 2021; Y. Hamon & Paissa, 2023). As Myerson and Rydin (2014) observe, “in academic terms, ‘environment’ belongs to every discipline and none”.

These interactions transcend the connections between various domains and levels of expertise. Even highly specialized discussions on this matter inevitably draw on knowledge from a broad spectrum of associated fields (Altmanova, Cartier, Luzzi, Pinto, & Piscopo, 2022; Candel, 1979; Pascaline, 2014). Delavigne (2002) highlights not only the vast and interconnected network of disciplines, practices, discourses, and techniques that constitute environmental studies, but also the considerable disparities in qualitative equivalence across these components. Indeed, no domain, subdomain, or sub-subdomain within the environmental field can be addressed without acknowledging its intrinsic diversity and interdisciplinary nature. As Sager (1990) contends, “[i]n practice, no individual or group of individuals possesses the whole structure of a community’s knowledge; conventionally, we divide knowledge up into subject areas, or disciplines, which is equivalent to defining subspaces of the knowledge space”.

In a field where the concept of “domain” remains fundamentally ambiguous and open to various interpretations, several critical questions naturally arise. How can these expansive and complex bodies of knowledge be segmented in a methodologically sound and pertinent manner? What methodologies can be adopted to quantify and delineate these domains, which are often fluid and interrelated? How can one adequately represent the meaning and scope of these diffuse networks of knowledge, practices, and scientific communities that collectively contribute to intellectual production? How can domain trees be constructed in an era when the boundaries between various fields are increasingly “permeable,” and their overlap is significant (Bordet, 2013; De Bessé, 2000)?

Given these theoretical and technical challenges, we decided to try a different approach to terminological analysis that can address the complexities inherent in domains. By combining traditional terminological

methods with NLP techniques, we tried to develop a more scalable method to categorize and analyze terminology. This hybrid approach would enable us to create terminological resources, facilitating a deeper understanding of the semantic relationships within, in this case, the wind energy domain. NLP techniques, by processing large corpora of text and identifying complex patterns in language, offered the potential for an efficient categorization of terms and domains.

## 2.2 Enhancing Terminology through Ontologies

The second key aspect of our study involves the relationship between terminology and ontologies. Ontologies, as formal representations of knowledge within a particular domain, have long been recognized for their ability to structure information in a way that reflects both semantic relationships and conceptual hierarchies (Guarino, Oberle, & Staab, 2009; Roche, 2005, 2015). The systematic classification of terms within a particular field has long been essential for structuring knowledge in a coherent and accessible way (Durán-Muñoz & Bautista-Zambrana, 2013). In fact, according to Roche, an ontology provides a formalized structure that organizes terms into categories such as domains, subdomains, and concepts, highlighting their relationships and interdependencies (Roche, Calberg-Challot, Damas, & Rouard, 2009; Temmerman & Kerremans, 2003).

Yet, the construction of ontologies within rapidly evolving domains, such as wind energy, presents significant methodological challenges. The fluidity of subdomain boundaries, the rapid evolution of terminology, and the frequent emergence of new interdisciplinary connections make it particularly difficult to establish stable ontological categories and maintain a coherent conceptual hierarchy. In such dynamic contexts, static ontologies are prone to rapid obsolescence, complicating the accurate representation of semantic relationships and conceptual dependencies over time. How, then, can terminologists be supported in capturing and structuring knowledge in domains that change so quickly? How can an ontology be created that remains relevant in a context of continuous technological progress and evolving terminology? What tools and strategies can translators employ to work effectively within such complex and fluid knowledge landscapes?

To explore potential solutions to these challenges, we tested a model that integrates NLP techniques for analyzing evolving domains. By leveraging LLMs and embedding-based methods, we aimed to uncover latent knowledge structures — patterns and relationships that are difficult to detect manually. The study focused on evaluating whether LUMEN could identify and categorize multiple domains beyond the strictly technical scope of wind energy, including health, political and economic, social, cultural, and juridical areas, each with its subdomains and specialized terminology in a rapidly changing context. Encouraged by promising results in a smaller pilot corpus, we tested the model on a highly technical corpus, where unlikely semantic fields, if present, are extremely challenging to detect manually. The following sections outline the steps that guided our experimental investigations.

## 2.3 Methodological Framework

Our contribution is situated within a combined approach that draws on multiple methodologies. Textual terminology (Condamines & Picton, 2022; Peruzzo, 2013; Picton, Condamines, & Humbert-Droz, 2021) constitutes our primary methodological reference, particularly with regard to the construction of “terminologies that are more closely aligned with actual usage” (Condamines, 2018), through the use of large-scale textual data and less structured information (Rebeyrolle, 2000).

Corpus linguistics (T. Hamon & Nazarenko, 2002; Meyer, 2008; Sinclair, 2004) constitutes an essential component of our methodological framework, particularly in relation to quantitative analysis. Its integration responds to the necessity of grounding terminological investigation in empirically attested language use. In domains characterized by conceptual indeterminacy, disciplinary overlap, and accelerated terminological evolution, corpus-based methodologies enable the systematic identification of recurrent patterns and co-occurrence structures that reflect the discursive realities of the field. Rather than projecting externally imposed taxonomies, this approach foregrounds the situated emergence and contextual distribution of terms, capturing both stable lexical cores and peripheral variations. It proves especially pertinent in contexts where disciplinary boundaries are permeable and where the interplay between specialized and non-specialized registers challenges the adequacy of prescriptive classificatory models. In our study, this corpus-driven component constitutes the initial, quantitative phase of analysis, providing an empirical

foundation before transitioning to a qualitative examination of semantic and functional relationships among terms. The details of this combined quantitative–qualitative procedure are provided in Section 3.4.

We further draw on studies in Languages for Specific Purposes from a diachronic perspective (Dury & Picton, 2009; Zanola, 2014), as well as on research in socioterminology (Delavigne, 2013; Gaudin, 2003; Temmerman, 2000). These approaches offer critical insights into the historical stratification of specialized discourse and the social dynamics that shape terminological practices. More specifically, our study examines the circulation and recontextualisation of specialized terminological units when introduced into non-specialist communicative settings (Delavigne, 2021; Humbert-Droz, 2024; Jacobi, 1986). This line of inquiry is particularly relevant in light of the increasing permeability between expert and lay discourses, where terms migrate across communicative domains, often undergoing semantic shifts, functional reassignments, or pragmatic reinterpretations. Attention to such phenomena allows us to account not only for terminological stability but also for variation, negotiation, and adaptation across different socio-discursive configurations.

This multi-methodological approach facilitates a comprehensive identification and analysis of terminological variations – including synonymic, diastatic, diatopic, and other relevant types – within specialized lexicons. Such variations arise from the dynamic interaction between specialized terms and diverse communicative contexts, encompassing both situational and sociolinguistic factors (Bowker & Hawkins, 2006; Freixa, 2006; Tartier, 2006). By employing this approach, it becomes possible to systematically account for the influence of domain-specific registers, audience heterogeneity, and regional language differences on term usage, thereby enhancing the accuracy and contextual sensitivity of terminological studies. This framework also supports the development of more adaptive and user-oriented terminological resources, which are essential for effective knowledge dissemination in multilingual and multidisciplinary environments.

Finally, we have selected specific case studies within the domain of wind energy, following the methodological recommendations of Lyrette and Trépanier (2004), Vargas (2009), and Parrenin and Vargas (2020). This case study approach enables a nuanced exploration of the practical benefits and inherent challenges associated with wind energy implementation. By grounding the analysis in concrete examples and real-world contexts, the study not only elucidates technical and socio-economic dimensions but also highlights the interplay between technological innovation, policy frameworks, and community engagement. Consequently, this method provides valuable insights that contribute to a more comprehensive understanding of wind energy’s impact and potential within contemporary renewable energy discourse.

## 2.4 Lexical and Terminological Resources

Following the suggestions proposed by L’Homme (2004) and Zanola (2018), to contextualize and validate the terminological data extracted from our corpus, we prepared a selection of lexical-terminological resources to serve as reference points and benchmarks. These resources offer structured inventories of terms, enabling us to compare the emerging corpus-derived terminology with established usage across both general and specialized contexts. This comparative perspective is essential for ensuring that the terminology identified through our corpus-driven approach aligns with recognized standards while also capturing emerging linguistic trends within the wind energy sector. Among general-purpose resources, we considered:

- TERMIUM PLUS® (Canada),<sup>1</sup> providing extensive multilingual terminology and standardized definitions across multiple scientific and technical domains;
- IATE (European Union),<sup>2</sup> offering harmonized term usage within European Union institutions and facilitating cross-linguistic comparisons;
- UNTERM (United Nations),<sup>3</sup> containing multilingual terminological entries used in global policymaking and international communications;
- Eurotermbank,<sup>4</sup> aggregating terminological data from multiple European countries, supporting both standardization and cross-domain analysis.

---

<sup>1</sup><https://www.btb.termiumpius.gc.ca/tpv2alpha/alpha-eng.html?lang=eng> [Accessed 8/11/2025]

<sup>2</sup><https://iate.europa.eu/home> [Accessed 8/11/2025]

<sup>3</sup><https://unterm.un.org/unterm2/en/> [Accessed 8/11/2025]

<sup>4</sup><https://www.eurotermbank.com/> [Accessed 8/11/2025]

These general-purpose resources helped us establish a baseline for evaluating term recognition and semantic alignment across languages and domains, highlighting widely accepted terminologies and ensuring consistency with international usage.

Domain-specific resources were also essential for evaluating the technical representation of wind energy terminology. These included glossaries produced by research organizations and universities. For domain-specific resources, which are inherently more limited due to the technical specificity and relative novelty of the wind energy sector, we consulted:

- The Glossary of Wind Energy Research Alliance Wind Energy; <sup>5</sup>
- The Wind Energy Glossary, <sup>6</sup> + Abbreviations; <sup>7</sup>
- The Wind Energy Glossary: Technical Terms and Concepts, Grand Valley State University; <sup>8</sup>
- The Basics of wind energy, a multilingual database; <sup>9</sup>
- Wind Power Terms and Definitions by Enerpac; <sup>10</sup>
- Wind Energy Glossary by Wind Solar Alliance; <sup>11</sup>
- Wind Turbine Glossary of Terms by Windurance. <sup>12</sup>

While these domain-specific resources are limited in scope, they offer authoritative insight into specialized lexicons and allow for cross-validation with corpus-extracted terms. By preparing these resources for consultation, we established a comprehensive framework for comparing corpus-extracted terminology with existing references. This framework allowed us to detect overlaps, gaps, and novel terms, providing a solid empirical foundation for subsequent analysis. Plus, the combination of general and domain-specific resources ensures that our terminological investigation is not only grounded in real-world usage but also contextualized within recognized linguistic and technical standards, enhancing both the reliability and interpretability of the final results.

### 3 Our Approach

This section presents the methodological framework adopted for the identification of latent subdomains. As illustrated in Figure 1, the approach is articulated into three principal stages: the construction of a specialized corpus through Sketch Engine, the identification of domain labels via an LLM, and the subsequent classification of terms on the basis of semantic similarity. These stages contribute to the creation of a hierarchical domain tree that facilitates a comprehensive understanding of the field.

#### 3.1 Step 1: Creation of an English Corpus

The first stage of our methodology consisted in the compilation of a domain-specific corpus (ISO 1087, 2019; ISO 704, 2022), constructed by means of the text retrieval functionalities provided by Sketch Engine (Kilgarriff et al., 2014). We adopted a corpus-driven methodology, in which terminological analysis is grounded in empirical evidence drawn from authentic language use. Unlike approaches that rely primarily on pre-existing dictionaries, theoretical frameworks, or expert intuition, corpus-driven analysis permits a systematic investigation of how terms function in real-world texts. This method facilitates the identification of patterns, distributions, and semantic relationships that may remain invisible in prescriptive sources, offering a robust, data-oriented foundation for the rigorous study of terminology. Such an approach is particularly well-suited to our research objectives, as it allows us to address the complexity, interdisciplinarity, and evolving nature of terminology in the wind energy domain.

To implement this approach, we employed the Sketch Engine’s “Find texts on the web” feature, a semi-automatic tool designed to collect textual data from online sources on the basis of a predefined set of keywords. This procedure ensured a systematic and scalable workflow for capturing the specialized terminological landscape of the wind energy sector.

<sup>5</sup><https://windenergy-researchfarm.com/glossary> [Accessed 8/11/2025]

<sup>6</sup><https://www.wind-energy-the-facts.org/glossary.html> [Accessed 8/11/2025]

<sup>7</sup><https://www.wind-energy-the-facts.org/abbreviations.html> [Accessed 8/11/2025]

<sup>8</sup><https://scholarworks.gvsu.edu/cgi/viewcontent.cgi?article=1005&context=bioreports> [Accessed 8/11/2025]

<sup>9</sup>[https://tesup.com/be\\_fr/blogs/post/bases-de-l-39-energie-eolienne](https://tesup.com/be_fr/blogs/post/bases-de-l-39-energie-eolienne) [Accessed 8/11/2025]

<sup>10</sup><https://blog.enerpac.com/essential-guide-to-wind-power-terms-and-definitions/> [Accessed 8/11/2025]

<sup>11</sup><https://windsolaralliance.org/wind/glossary/> [Accessed 8/11/2025]

<sup>12</sup><https://blog.windurance.com/wind-turbine-glossary-of-terms-pitch-system-more> [Accessed 8/11/2025]

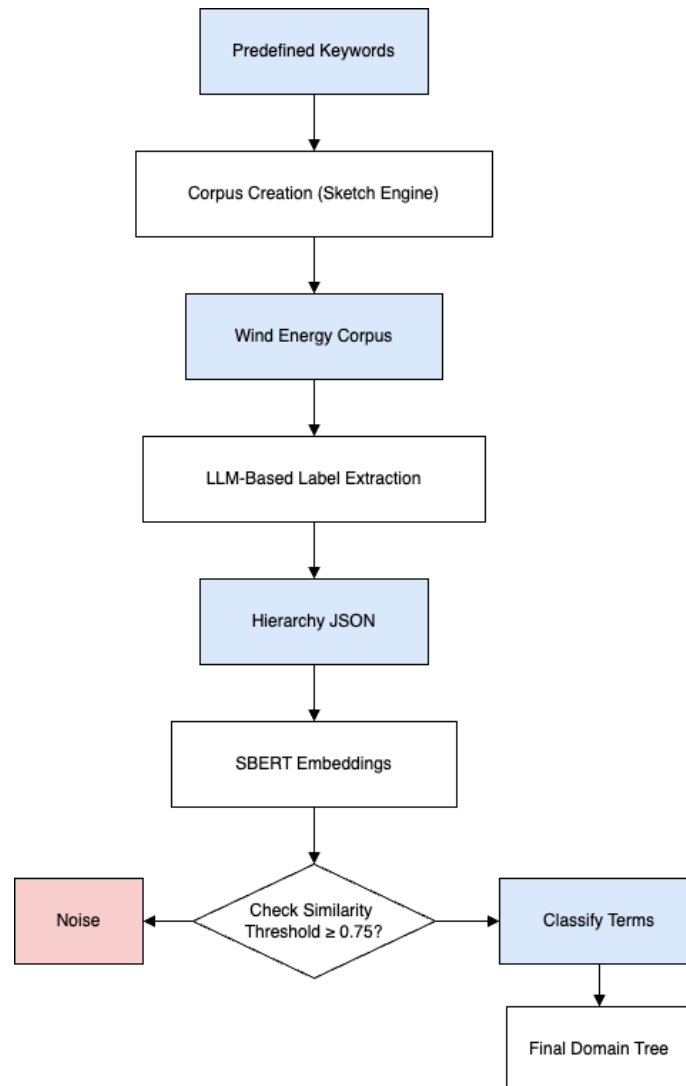


Fig. 1 Overview of LUMEN

Figure 2 presents the set of keywords used for the generation of our corpus. These keywords, highly technical in nature and specific to the wind energy domain, were derived from the selection of the first 20 technical terms extracted from a larger working corpus. This broader corpus encompassed multiple discourse types, including technical wind energy texts, and was developed as part of a doctoral research project focused on terminological analysis in renewable energy sectors. The principle guiding the selection of these keywords was neutrality, aimed at minimizing potential biases during data collection. The chosen keywords – “wind energy”, “wind power”, “onshore wind power”, “onshore wind energy”, “offshore wind power”, “offshore wind energy”, and “wind power plant” – were primarily n-grams (two-word or three-word expressions). This decision was made to ensure broad yet precise coverage, avoiding an excessive focus on overly specialized terms that might distort the corpus composition. Furthermore, this selection reflects a deliberate balance between general and specific terminology relevant to the wind energy field, thereby enhancing both the comprehensiveness and the representativeness of the corpus (L’Homme, 2004).

As shown in Figure 3, following the initial automated retrieval, it was necessary to perform a thorough manual verification of the collected texts. Despite Sketch Engine’s automation, some documents originating from unrelated discourse domains – such as journalistic media, non-governmental organizations, or promotional materials – were inadvertently included and could potentially contaminate the domain-specific analysis. To maintain the integrity and relevance of the corpus, these non-technical documents were systematically identified and removed. The final corpus exclusively comprises technical reports and

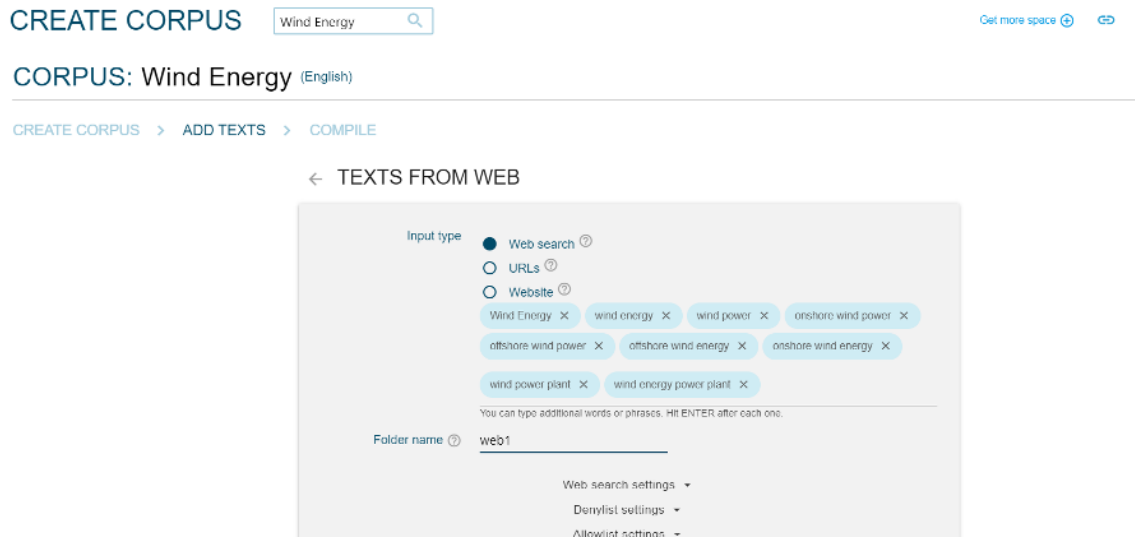


Fig. 2 Representing the keywords used to retrieve texts from the Web

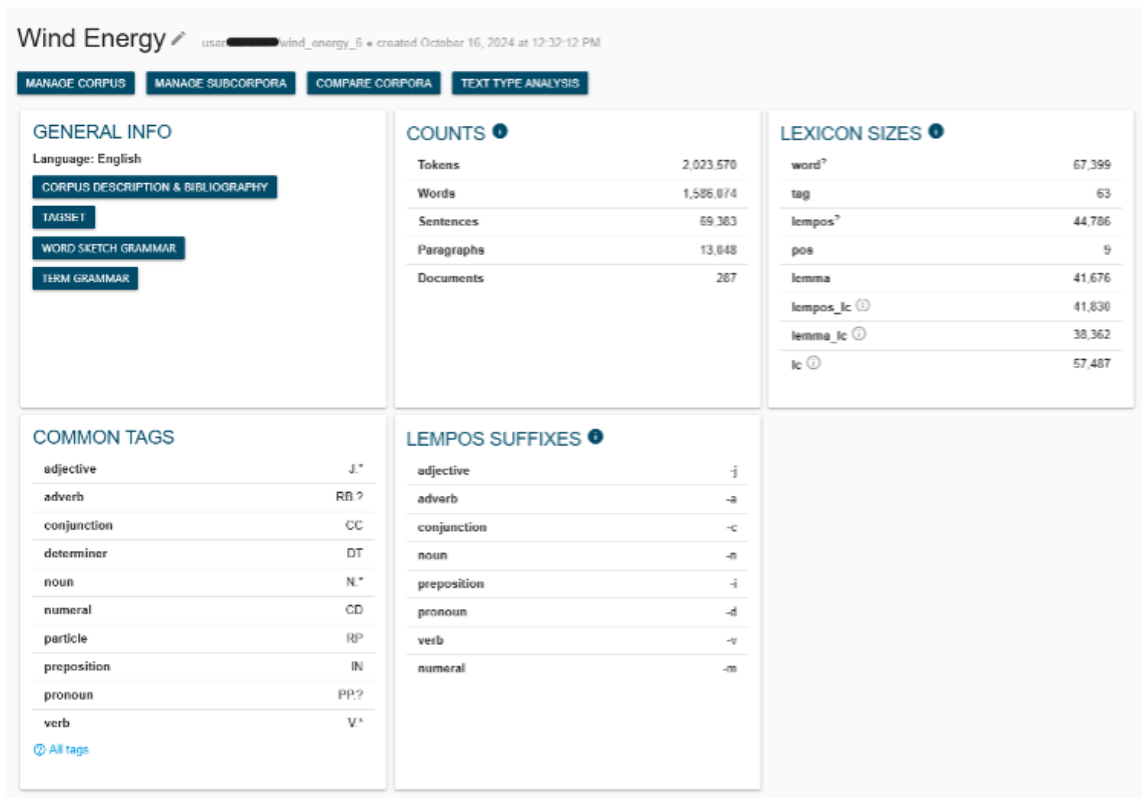


Fig. 3 General data about our corpus

documentation produced primarily by leading companies specializing in wind energy installation and management.

The corpus we created comprises documents published between 2016 and 2024, thereby adhering to a short-diachronic perspective (Dury & Picton, 2009). Such an approach was deemed particularly appropriate in this context, as it makes it possible to capture data that are both representative of the specialized field and closely aligned with its most recent developments. In the case of wind energy, the pace of transformation is especially pronounced, encompassing not only rapid technological innovation but also profound sociological, political, and cultural shifts (Parrenin & Vargas, 2020). Restricting the temporal span of the

corpus to a relatively recent interval thus ensures that the subsequent analysis remains sensitive to the dynamic character of the discourse, and capable of reflecting its ongoing evolution with both accuracy and methodological rigor.

Sketch Engine compiled a corpus comprising 287 documents, exclusively in English. This corpus spans a total of 1,586,074 words, encompassing a diverse array of textual sources relevant to the wind energy sector, ensuring the capture of a large spectrum of discourse surrounding the field. Through Sketch Engine’s automated terminology extraction features, a subset of 100,000 terms was identified as domain-specific, serving as the foundational dataset for subsequent classification and hierarchical organization. This set was composed of 78,419 multi-word terms (e.g., ‘wind turbine’, ‘pitch control’, ‘power grid’, ‘offshore wind farm’) and 21,581 single-word terms (e.g., ‘rotor’, ‘blade’, ‘anemometer’). This distribution, which is heavily skewed towards multi-word units, is characteristic of specialized domains, where conceptual specificity is typically achieved through terminological composition.

### 3.2 Step 2: Domain Label Identification via LLMs

The second stage of our methodology addresses the critical task of discerning a coherent conceptual structure from the extensive lexicon of 100,000 domain-specific terms. At this point, our data consist of a flat, unstructured list of vocabulary. The primary objective of this step is to transform this list into a meaningful, multi-level hierarchy of topics and subtopics that accurately represent the knowledge structure of the wind energy domain. To achieve this, we leverage the advanced analytical capabilities of an LLM, specifically gpt-4o,<sup>13</sup> to function as a computational terminologist, automatically discovering and labeling the latent semantic categories within the data. An LLM is a sophisticated neural network, typically based on the Transformer architecture (Vaswani et al., 2017), that has been pre-trained on vast and diverse corpora of text. This training process enables the model to develop a deeply nuanced internal representation of language, encompassing syntax, semantics, and world knowledge, which can be leveraged for complex, zero-shot reasoning tasks such as the one we propose.

At a high level, the procedure can be conceptualized as a process of automated knowledge synthesis. We begin by presenting the entire corpus of 100,000 terms to the LLM. The model’s task is not merely to group similar terms together, but to perform a more sophisticated act of abstraction: it must analyze the semantic content of these term clusters and generate a concise, human-readable label that accurately describes the underlying concept. For instance, upon analyzing terms such as `grid_integration`, `substation`, `transmission_line`, and `voltage_regulation`, the model must infer that these terms collectively belong to a conceptual category, and it must synthesize an appropriate label, such as “Electrical Infrastructure.” This process is executed recursively, allowing the model to identify broad, high-level domains (e.g., “Technology,” “Environment”) and then progressively break them down into more granular subdomains (e.g., “Mechanical Components,” “Environmental Impact”) and even finer sub-subdomains (e.g., “On Fauna,” “On Flora”). The final output of this conceptual analysis is a comprehensive, tree-like structure — a thesaurus — that organizes the entire terminological landscape into a logical and navigable hierarchy. This automated approach allows us to construct a detailed and objective map of the domain’s conceptual architecture, a task that would be prohibitively labor-intensive and susceptible to subjective biases if attempted manually. To operationalize this high-level procedure, we employ a technique known as *prompt engineering* (Sahoo et al., 2025), which involves formulating a precise set of instructions to guide the LLM’s analytical process. Our prompt was meticulously designed to elicit a structured, hierarchical output that reflects the conceptual organization of the wind energy domain. The prompt’s structure can be deconstructed into several key components:

1. **Persona Assignment:** We instructed the model to assume the role of “an experienced linguist with a deep understanding of natural language processing.” This directive primes the model to activate the parts of its parameter space most relevant to linguistic and semantic analysis, encouraging it to focus on conceptual relationships rather than mere statistical co-occurrence.
2. **Task Definition:** The primary instruction was to analyze the provided list of lemmas and return a JSON (JavaScript Object Notation) object containing all identifiable domains, subdomains, and sub-subdomains. This explicitly defines the nature and format of the desired output. JSON was chosen for its

---

<sup>13</sup><https://openai.com/index/hello-gpt-4o/> [Accessed 8/11/2025]

You are an experienced linguist with a deep understanding of natural language processing. You will be provided with a list of 100,000 lemmas. Your task is to return to me a JSON with all the labels of possible domains and sub-domains that you can unearth in this list. Be detailed, return a JSON with as many domains and as many sub-domains as possible. Take your time to answer. Follow this example:

```
{
  "total_labels": [
    {
      "label": "label domain 1",
      "sub-labels": [
        "Label sub-domain 1": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        "Label sub-domain 2": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        ...
      ]
    },
    {
      "label": "Label domain 2",
      "sub-labels": {
        "Label sub-domain 1": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        "Label sub-domain 2": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        ...
      }
    },
    {
      "label": "Label domain 3",
      "sub-labels": {
        "Label sub-domain 1": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        "Label sub-domain 2": ["Label sub-sub-domain 1", "Label sub-sub-domain 2", ...],
        ...
      }
    },
    ...
  ]
}
```

Fig. 4 Contracted structure of the prompt used for extracting representative domain labels

hierarchical structure, which is inherently suited to representing tree-like knowledge structures such as thesauri or taxonomies, and for its widespread machine-readability.

3. **Constraints and Quality Control:** We included specific constraints to refine the output. Instructions like “Be detailed”, “return a JSON with as many domains and as many sub-domains as possible”, and “Always remember that we just want labels, don’t return lemmas” were crucial. These directives guide the model towards generating a granular and comprehensive hierarchy of concepts (the labels) rather than simply regurgitating or clustering the input terms themselves.

Figure 4 shows the general structure of the prompt. Upon receiving the prompt and the term list, the LLM performs the high-level semantic analysis described previously. It identifies implicit semantic fields by recognizing clusters of related terms within the 100,000-item vocabulary, and synthesizes abstract labels for them. This process is repeated recursively to generate the multi-level hierarchy, thus producing the final structured thesaurus of labels. To ensure the reproducibility and determinism of this generative process, we configured the model’s hyperparameters with specific values. The temperature was set to 0, which minimizes randomness by compelling the model to select the most probable token at each step of the generation process. The top-p sampling was set to 0.95, a nucleus sampling technique that considers the smallest possible set of tokens whose cumulative probability exceeds this threshold. While a temperature of 0 often renders top-p sampling moot, this configuration ensures a focus on high-confidence outputs.

Finally, a token limit of 128,000 was established to manage the computational constraints inherent in processing large volumes of data. The output of this stage is a structured JSON file that serves as the conceptual scaffold for the subsequent term classification phase.

### 3.3 Step 3: Term Classification Using Semantic Similarity

Following the generation of the hierarchical label structure, the third step involves programmatically assigning each of the 100,000 extracted terms to its most semantically appropriate position within this hierarchy. This classification is achieved through the application of embedding-based semantic similarity, a cornerstone technique in modern NLP.

An embedding is a numerical representation of a piece of text (a word, a phrase, or a sentence) in the form of a dense vector. A vector is, in essence, an ordered list of numbers. One can conceptualize this vector as a set of coordinates that precisely locates the piece of text within a vast, multi-dimensional “semantic space”. While we are familiar with three-dimensional space (defined by x, y, and z coordinates), this semantic space can have hundreds or even thousands of dimensions, each capturing a different, subtle aspect of meaning.

The workflow commences with embedding generation. To perform quantitative comparisons of meaning, textual data must first be transformed into a numerical format. We achieve this by converting both the individual terms and the hierarchical label paths into high-dimensional vectors, or “embeddings.” These embeddings are generated using a Sentence-BERT (SBERT) model (Reimers & Gurevych, 2019), specifically `all-mpnet-base-v2`.<sup>14</sup> SBERT is a modification of the BERT architecture (Devlin, Chang, Lee, & Toutanova, 2019), optimized for producing semantically meaningful embeddings for sentences and phrases. Unlike earlier models such as Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013), which generate context-independent vectors for single words, SBERT is context-aware and ideal for our task, as it can capture the composite meaning of multi-word terms and entire label paths (e.g., the string “Energy > Renewable Sources” is treated as a single conceptual unit). Each term and label path was thus mapped to a 768-dimensional vector in a continuous semantic space, where spatial proximity corresponds to semantic relatedness.

The next procedure is similarity calculation. With all textual elements represented as vectors, we can quantify their semantic relatedness. For this, we compute the cosine similarity (Singhal, 2001) between the embedding of each term and the embedding of each hierarchical path. Mathematically, cosine similarity measures the cosine of the angle between two vectors, yielding a score between -1 (perfectly dissimilar) and 1 (perfectly similar). We selected this metric over alternatives like Euclidean distance because it is a measure of orientation, not magnitude. In high-dimensional semantic spaces, the direction of a vector is a more reliable indicator of meaning than its length, which can be influenced by confounding factors like word frequency. Cosine similarity effectively normalizes for vector magnitude, providing a pure measure of semantic closeness.

The final procedure is threshold-based classification. For every term, we calculate its cosine similarity score against all possible label paths in our hierarchy. A classification is made if this score exceeds an experimentally determined similarity threshold, which was optimized to 0.75. This threshold acts as a confidence boundary; a term is assigned to a label only if its semantic proximity meets this minimum requirement. Terms that fail to surpass this threshold for any label path are provisionally categorized as “noise.” This category is not a definitive rejection but rather a flag for human review, encompassing true noise (e.g., corpus artifacts, typos) as well as potentially valid domain terms that are either too niche or semantically ambiguous for the model to classify with high confidence. This mechanism for incremental reclassification is a key part of our iterative refinement strategy. A preliminary evaluation on a subset of 500 terms yielded a precision of approximately 0.82 (the proportion of correct classifications among all classifications made) and a recall of 0.78 (the proportion of relevant terms successfully identified), indicating a promising baseline for future, more extensive validation.

### 3.4 Output

The final artifact produced by the LUMEN methodology is a hierarchical *thesaurus* of the wind energy domain, rendered as a structured JSON file, as exemplified in Listing 1.

---

<sup>14</sup><https://huggingface.co/sentence-transformers/all-mpnet-base-v2> [Accessed 8/11/2025]

**Listing 1** An example of hierarchical JSON output generated by LUMEN

```

{
  "Fauna": {
    "Marine Mammals": {...},
    "Birds": {...},
    "Fish": {...},
    "Habitats": {...},
    "Migration": {...}
  },
  "Animal Behavior": {
    "Feeding": [...],
    "Breeding": [
      "terrestrial mammal",
      "marine mammal",
      "wind power on terrestrial mammals",
      "power on terrestrial mammals",
      "habitat loss",
      "destination habitat",
      "marine environment",
      "migrating bird",
      "bird population",
      "marine bird",
      "habitat use",
      "bird specie",
      "wildlife resource",
      "species distribution",
      "wildlife habitat",
      "bird mortality",
      "habitat change",
      ...
    ]
  }
}

```

This output functions as a computational knowledge representation that is both human-readable - due to its clear, nested structure - and machine-processable, enabling its use in downstream computational tasks. The hierarchical organization of the JSON object directly mirrors the conceptual tree identified in the preceding steps. Each key in the object represents a conceptual label (a node), and its corresponding value is either a nested object containing sub-labels or, at the terminal nodes, an array of the specific terms that were classified under that complete conceptual path. This structure facilitates a systematic exploration of the domain's terminology. For instance, as shown in the excerpt in Listing 2, a user can navigate from the primary domain of "Environment" to the "Environmental Impact" subdomain, and further to the "On Fauna" sub-subdomain. Within this terminal node, the output provides a list of terms, such as "farm development", "habitat loss", and "cumulative impact", which the system classified as semantically belonging to this specific conceptual category based on the calculated similarity scores.

In the experimental application of this methodology, the process identified and structured nine principal domains within the wind energy corpus: *Environment*, *Energy*, *Geography*, *Economy*, *Technology*, *Society*, *Safety*, *Fauna*, and *Research*. For terminologists and translators, this output serves as an analytical resource. It provides a data-driven representation of the semantic fields that constitute the wind energy domain, and makes explicit the thematic relationships and interdisciplinary connections inherent in the source data. By organizing terms within a defined conceptual framework, it can support the analysis of term usage and context, thereby contributing to greater precision in terminological work. A key characteristic of this output is that it is not a static artifact. The methodology is designed to be extensible and repeatable. As the domain evolves, the source corpus can be augmented with new documents, and the entire pipeline can be re-executed. This allows the thesaurus to be updated in an iterative fashion,

**Listing 2** Excerpt from the hierarchical JSON output generated by LUMEN. In this example, environment-related labels, sub-labels, and terms are shown

```
{
  "Environment" : [
    "Environmental Impact": [
      "On Fauna": [
        "farm development",
        "habitat loss",
        "cumulative impact",
        ...
      ],
      "On Flora": [
        "impact on the marine environment",
        "perspective on marine environmental impacts",
        "public land",
        ...
      ],
      "Visual": [...],
      ...
    ],
    ...
  ],
  "Energy": [...],
  "Geography" : [...],
  "Economy" : [...],
  "Technology" : [...],
  "Society" : [...],
  "Safety" : [...],
  "Fauna" : [...],
  "Research" : [...],
}
```

enabling it to reflect emergent concepts and shifts in terminology. This adaptability is intended to support the continued utility and accuracy of the terminological resource over time.

The methodological design of this study was intentionally structured to proceed from a quantitative to a qualitative phase. The initial quantitative stage involved the comprehensive collection and cataloguing of a large volume of domain-specific data, encompassing overarching domains, nested semantic fields, and individual lexical items. Following the quantitative phase, we proceeded with a preliminary qualitative refinement aimed at improving the precision, coherence, and overall consistency of the semantic structure. This refinement entailed a manual review of semantic fields that were deemed imprecise, noisy, or inconsistent. During this review, the previously enumerated lexical-terminological resources were consulted, serving as comparative references for the evaluation and refinement of terms and semantic categories, including decisions on their retention, modification, or removal.

Recognizing the inherent limitations of our disciplinary expertise, we plan to implement a subsequent validation stage involving at least two domain specialists through a structured Google Form. Their expert evaluation will provide critical guidance in further refining the framework, ensuring that the resulting semantic fields achieve both conceptual coherence and terminological accuracy. By integrating an initial quantitative survey with a rigorous, reference-informed qualitative review and expert-informed validation, this approach ensures that the resulting terminological model is well-positioned to support ongoing research and refinement within the dynamically evolving domain of wind energy.

## 4 Conclusions and Further Research

This study has presented the application of the LUMEN methodology for the hierarchical organization of terminology within the wind energy domain. By leveraging a corpus-driven approach combined with

Large Language Models and semantic similarity measures, the methodology enabled the identification of nine primary subdomains – *Environment, Energy, Geography, Economy, Technology, Society, Safety, Fauna, and Research* – along with their associated sub-labels. This comprehensive hierarchical structure offers a detailed representation of the domain, revealing latent thematic patterns, interdisciplinary connections, and semantic relationships that may not be readily observable through conventional manual analysis.

Despite the promising outcomes, certain limitations remain. A subset of terms was classified as “noise,” either due to marginal relevance or insufficient similarity scores. Addressing these limitations requires both adaptive thresholding and additional expert validation to ensure the robustness of the semantic hierarchy. To this end, future work will involve a systematic review of these noisy or ambiguous terms, guided by domain specialists, to refine the organization and improve the overall reliability of the terminological framework.

In addition, the study aims to incorporate expert evaluations through structured tools, such as Google Forms, allowing specialists to quantitatively assess term classifications and semantic label assignments. This approach will provide a more objective, reproducible, and data-informed basis for validating the framework, enabling researchers to distinguish which terms and subdomains should be retained, modified, or removed. The integration of expert feedback in this manner is expected to enhance both the precision and the interpretive depth of the hierarchical domain model.

Another potential development concerns the expansion of the methodology to multilingual corpora. By extending the framework beyond English-language sources, it will be possible to generate semantically coherent terminological resources in additional languages, supporting translators, terminologists, and other stakeholders working in international or cross-linguistic contexts. This extension is particularly relevant in domains such as renewable energy, where global collaboration and multilingual accessibility are essential for policy-making or industry applications.

Overall, the findings of this study underscore the potential of corpus-driven NLP methodologies to complement traditional terminological practices. By providing a structured, empirically grounded, and adaptable framework, LUMEN enables the systematic visualization of semantic fields, the identification of interdisciplinary links, and the navigation of complex technical domains. As wind energy and related renewable sectors continue to evolve, this methodology offers a platform for ongoing terminological research, fostering increasingly accurate, expert-validated, and dynamically updated representations of specialized knowledge.

The sources of LUMEN are available online <https://github.com/lopsandrea/LUMEN>.

## References

- Altmanova, J., Cartier, E., Luzzi, J., Pinto, S., Piscopo, S. (2022). Lexical innovations in the biodiversity and climate change domain: the bio morphem in contemporary French and Italian. *Neologica* 2022, n° 16. *Néologie et environnement*, 85–110, <https://doi.org/10.48611/isbn.978-2-406-13219-6.p.0085>
- Bordet, G. (2013). Brouillage des frontières, rencontres des domaines: quelles conséquences pour l’enseignement de la terminologie et de la traduction spécialisée. *ASp. la revue du GERAS*, 64, 95–115, <https://doi.org/10.4000/asp.3851>
- Bowker, L., & Hawkins, S. (2006). Variation in the organization of medical terms: Exploring some motivations for term choice. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 12(1), 79–110, <https://doi.org/10.1075/term.12.1.05bow>
- Candel, D. (1979). La présentation par domaines des emplois scientifiques et techniques dans quelques dictionnaires de langue. *Langue française*, 43(1), 100–115, <https://doi.org/10.3406/lfr.1979.6165>
- Condamines, A. (2018). Nouvelles perspectives pour la terminologie textuelle. J. Altmanova, M. Centrella, & K.E. Russo (Eds.), *Terminology and discourse. terminologie et discours* (pp. 29–50). Peter Lang.
- Condamines, A., & Picton, A. (2022). Textual terminology: Origins, principles and new challenges. In P. Faber & M.-C. L’Homme (Eds.), *Theoretical perspectives on terminology: Explaining terms, concepts and specialized knowledge* (pp. 219–236). John Benjamins Publishing Company. <https://doi.org/>

[10.1075/tlrp.23.10con](https://doi.org/10.1075/tlrp.23.10con)

- Costa, F., Giyanani, A., Liu, D., Keane, A., Ratti, C., Clifton, A. (2024). An ontology for describing wind lidar concepts. *Remote. Sens.*, 16(11), 1982, <https://doi.org/10.3390/RS16111982>
- De Bessé, B. (2000). Le domaine. *Le sens en terminologie* (pp. 182–197). Lyon: Presses Universitaires Lyon.
- Delavigne, V. (1994). Les discours institutionnels du nucléaire : stratégies discursives d’euphorisation. *Mots: les langages du politique*, 39(1), 53-68.
- Delavigne, V. (2002). Le domaine aujourd’hui. Une notion à repenser. *Le traitement des marques de domaine en terminologie* (p. 29-41). Paris, France.
- Delavigne, V. (2013). Quand le patient devient expert: usages des termes dans les forums médicaux. *Actes de la conférence terminologie et intelligence artificielle (tia 2013)* (pp. 11–20). Paris: Association TIA.
- Delavigne, V. (2021). Phraséologie et didacticité dans les discours de vulgarisation médicale: une ergonomie discursive. *PHRASIS Rivista di studi fraseologici e paremiologici*, 5. Roma: Associazione Italiana Di Fraseologia E Paremiologia Phrasis.
- Delavigne, V. (2022). La notion de domaine en question-à propos de l’environnement. *Neologica*, 16, 27–59, <https://doi.org/10.48611/isbn.978-2-406-13219-6.p.0027>
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. J. Burstein, C. Doran, & T. Solorio (Eds.), *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). Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Durán-Muñoz, I., & Bautista-Zambrana, M.R. (2013). Applying ontologies to terminology: Advantages and disadvantages. *Hermes - Journal of Language and Communication in Business*, 26(51), 65–77, <https://doi.org/10.7146/hjlc.v26i51.97438>
- Dury, P., & Picton, A. (2009). Terminologie et diachronie: vers une réconciliation théorique et méthodologique? *Revue française de linguistique appliquée*, 14(2), 31–41.
- Fløttum, K. (2010). A linguistic and discursive view on climate change discourse. *ASp. la revue du GERAS*, 58, 19–37, <https://doi.org/10.4000/asp.3851>
- Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 12(1), 51-77, <https://doi.org/10.1075/term.12.1.04fre>
- Gaudin, F. (2003). *Socioterminologie. une approche sociolinguistique de la terminologie*. Brussels: Duculot De Boeck.
- Georgescu, M. (2010). La variazione nella terminologia dello sviluppo sostenibile. *Publiforum*. Genova University Press.
- Grimaldi, C. (2021). Le sfide linguistiche del cambiamento climatico. *AIDAinformazioni*, 3-4, 213–216. Bari: Carucci Editore.
- Guarino, N., Oberle, D., Staab, S. (2009). What is an ontology? S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 1–17). Berlin/Heidelberg: Springer. [https://doi.org/10.1007/978-3-540-92673-3\\_0](https://doi.org/10.1007/978-3-540-92673-3_0)
- Hamon, T., & Nazarenko, A. (2002). Structuration de terminologie: quels outils pour quelles pratiques. *Actes de la 9e conférence sur le traitement automatique des langues naturelles (taln 2002)* (pp. 167–176).

Nancy, France.

- Hamon, Y., & Paissa, P. (Eds.). (2023). *Discours environnementaux* (Vol. 20). Rome: Aracne editrice. Retrieved from <https://www.aracneeditrice.eu/anteprime/9791221807769.pdf>
- Humbert-Droz, J. (2024). Terminologie de l'endométriose et représentations de la maladie: regards croisés entre presse généraliste et discours spécialisés. *9e congrès mondial de linguistique française* (Vol. 191, p. 05003). Les Ulis, France.
- ISO 1087 (2019). *Terminology work and terminology science – vocabulary*. International Organization for Standardization. Retrieved from <https://www.iso.org/standard/62330.html>
- ISO 704 (2022). *Terminology work – principles and methods*. International Organization for Standardization. Retrieved from <https://www.iso.org/standard/79077.html>
- Jacobi, D. (1986). *Diffusion et vulgarisation: itinéraires du texte scientifique* (Vol. 324). Besançon, France: Presses Univ. Franche-Comté.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Lefever, E., & Terry, A.R. (2024). Computational terminology. In Y. Peng, H. Huang, & D. Li (Eds.), *New advances in translation technology: Applications and pedagogy* (pp. 141–159). Singapore: Springer Nature Singapore. [https://doi.org/10.1007/978-981-97-2958-6\\_8](https://doi.org/10.1007/978-981-97-2958-6_8)
- L'Homme, M.-C. (2004). *La terminologie: principes et techniques*. Montréal: Presses de l'Université de Montréal.
- L'Homme, M.-C. (2015). Découverte de cadres sémantiques dans le domaine de l'environnement: le cas de l'influence objective. *Terminàlia*(12), 29–40.
- Lops, A., & Sassi, S. (2025). LUMEN: Leveraging Large Language Models for Dynamic Ontologies in Wind Energy Domain Analysis. *Proceedings of the 4th international conference on multilingual digital terminology today (mdtt 2025)* (Vol. 3990). Thessaloniki, Greece: CEUR-WS.org. Retrieved from <https://ceur-ws.org/Vol-3990/paper1.pdf>
- Lyrette, É., & Trépanier, M. (2004). Les dynamiques sociales engendrées par l'implantation du parc éolien le nordais. *Vertigo-la revue électronique en sciences de l'environnement*, 5(1), <https://doi.org/10.4000/vertigo.3978>
- Meyer, I. (2008). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. *Recent advances in computational terminology* (pp. 279–302). Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. Y. Bengio & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, scottsdale, arizona, usa, may 2-4, 2013, workshop track proceedings*. Retrieved from <http://arxiv.org/abs/1301.3781>
- Mirqosim, M. (2025). The structural-semantic features of renewable energy resources terminology in english. *Prospects of teaching english for professional purposes in non-philological higher education institutions: Problems and solutions* (pp. 503–508).
- Myerson, G., & Rydin, Y. (2014). *The language of environment: A new rhetoric*. London: Routledge.
- Parrenin, F., & Vargas, É. (2020). Biodiversité et changement climatique: entre discours du spécialiste et discours vulgarisé. *Les carnets du cediscor* (pp. 33–46). Presses Sorbonne Nouvelle.

- Pascaline, D. (2014). Étude en corpus de l'implantation de quelques emprunts à l'anglais et de leurs concurrents officiels, dans le domaine de l'environnement. *Entre discours, langues et cultures: regards croisés sur le climat, l'environnement, l'énergie et l'écologie* (p. 61-71). Eme Editions.
- Peruzzo, K. (2013, Dec.). Short-period evolution in eu legal texts: old and new terms, old and new concepts. *Linguistica*, 53(2), 39–53, <https://doi.org/10.4312/linguistica.53.2.39-53>
- Picton, A., Condamines, A., Humbert-Droz, J. (2021). Analyse diachronique du processus de détermination. une réflexion en diachronie courte en physique des particules. *Cahiers de lexicologie*, 1(118), 193–225, <https://doi.org/10.48611/isbn.978-2-406-12006-3.p.0193>
- Rebeyrolle, J. (2000). *Forme et fonction de la définition en discours* (Unpublished doctoral dissertation). Toulouse 2, Toulouse, France.
- Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, hong kong, china, november 3-7, 2019* (pp. 3980–3990). Stroudsburg, PA, USA: Association for Computational Linguistics. <https://doi.org/10.18653/V1/D19-1410>
- Rey, A. (1979). *La terminologie: noms et notions*. Paris: PUF.
- Roche, C. (2005). Terminologie et ontologie. *Langages* (pp. 48–62). Cairn/Softwin.
- Roche, C. (2015). Ontological definition. *Handbook of terminology* (pp. 128–152). Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Roche, C., Calberg-Challot, M., Damas, L., Rouard, P. (2009). Ontoterminology - A new paradigm for terminology. J.L.G. Dietz (Ed.), *KEOD 2009 - proceedings of the international conference on knowledge engineering and ontology development, funchal - madeira, portugal, october 6-8, 2009* (pp. 321–326). Setubal, Portugal: INSTICC Press.
- Sager, J.C. (1990). *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins.
- Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A. (2025). *A systematic survey of prompt engineering in large language models: Techniques and applications*. Retrieved from <https://arxiv.org/abs/2402.07927>
- Sassi, S. (2024). Énergie éolienne, anglicismes et fin de vie: enjeux terminologiques. *TermCD-TERMinologie, Communication et Discours*, 2(2).
- Sassi, S. (2025). *La microlingua green: progettazione di un archivio terminologico digitale multilingue della sostenibilità*. (Unpublished doctoral dissertation). Bari, Italy.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.
- Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4), 35–43, Retrieved from <http://sites.computer.org/debull/A01DEC-CD.pdf>
- Tartier, A. (2006, April). Variation terminologique et analyse diachronique. P. Mertens, C. Fairon, A. Dister, & P. Watrin (Eds.), *Actes de la 13ème conférence sur le traitement automatique des langues naturelles. articles longs* (pp. 348–357). Leuven, Belgique: ATALA. Retrieved from <https://aclanthology.org/2006.jeptalnrecital-long.32/>
- Temmerman, R. (2000). *Towards new ways of terminology description*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

- Temmerman, R., & Kerremans, K. (2003). Termonography: Ontology building and the sociocognitive approach to terminology description. E. Hajicova, A. Kotěšovcová, & J. Mírovský (Eds.), *Proceedings of cil17*. Matfyzpress.
- Vargas, É. (2009). Discours de vulgarisation à travers différents médias ou les tribulations des termes scientifiques. le cas de la médecine. *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*(11). <https://doi.org/10.4000/ilcea.217>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. I. Guyon et al. (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, december 4-9, 2017, long beach, ca, USA* (pp. 5998–6008). Retrieved from <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Zanola, M.T. (2010). Glossari e divulgazione della conoscenza: la terminologia dei sistemi fotovoltaici. *Publiforum*. Genova University Press.
- Zanola, M.T. (2014). *Arts et métiers au xviiiè siècle: Etudes de terminologie diachronique*. Paris: L'Harmattan.
- Zanola, M.T. (2018). *Che cos' è la terminologia*. Roma: Carocci.
- Zhang, P., Zou, M., Du, H. (2025). Terminology extraction based on document-level context and domain adaptation. *2025 8th international conference on advanced algorithms and control engineering (icaace)* (p. 2138-2142). <https://doi.org/10.1109/ICAACE65325.2025.11020558>