

# Ontology-Guided Domain Entity Recognition in Environmental Texts: Evaluating Syntax-Driven and LLM Approaches Using BabelNet and GEMET

Elisa Chierchiello<sup>1,†</sup>, Patricia Chiril<sup>2\*,†</sup>, Cristina Bosco<sup>1</sup> and Adriana Pagano<sup>3,1</sup>

<sup>1</sup>Università degli Studi di Torino

<sup>2</sup>University of Chicago

<sup>3</sup>Universidade Federal de Minas Gerais

## Abstract

This paper investigates the identification and ontological classification of domain-specific entities to enable large-scale analysis of environmental discourse. While general-purpose Named Entity Recognition (NER) systems reliably detect standard categories such as persons, organizations, and locations, specialized domains like environmental communication require the recognition of additional, domain-relevant entities. These entities, often realized as common nouns, represent abstract, evolving concepts that are highly dependent on context and vary across languages. To address this challenge, we compare two pipelines for identifying domain-specific environmental entities in a bilingual corpus of WWF Living Planet Reports: (i) a traditional NLP pipeline that extracts noun phrases using dependency syntax parsing and matches them to BabelNet and GEMET, and (ii) a Large Language Model (LLM)-based pipeline that uses prompt-based instructions to both extract noun phrases and generate corresponding ontology matches. We evaluate the coverage of each approach and analyze the most frequent mapped entities to identify key environmental concepts emphasized in WWF discourse. To further assess the capabilities of LLMs in ontology-based annotation, we also prompted the LLM to generate GEMET-style definitions for phrases not found in the ontology. Our findings contribute practical insights for developing robust, ontology-enriched methods for environmental discourse analysis and knowledge extraction. Though tested on environmental texts, the framework can generalize to other domains via suitable ontologies and extraction rules.

## Keywords

environmental discourse analysis, domain entity recognition, dependency syntax, Large Language Models (LLMs), ontology mapping, BabelNet, GEMET, cross-linguistic annotation

## 1. Introduction

Named Entity Recognition (NER) has become an established task in Natural Language Processing (NLP), reliably identifying standard entity types such as persons, organizations, and locations [1, 2]. In specialized domains, general-purpose NER systems perform well when it comes to detecting these conventional entity types. However, many domain-specific applications require a different focus: the identification of domain entities, i.e. conceptually salient terms that often take the form of common nouns and refer to abstract, evolving phenomena central to the domain. In environmental communication, for example, entities such as *climate change*, *deforestation*, or *ecosystem services* play a key role in discourse but fall outside the typical scope of standard NER systems. This calls for adapted approaches capable of capturing and classifying these domain-relevant entities

in context [3]. This challenge is even more pronounced in multilingual contexts, where consistency in detecting and aligning domain-specific entities is crucial for comparative studies. A closely related but very challenging aspect is the continual adaptation to new terminology and the integration with terminology from related specialized domains — both of which are especially relevant for environmental discourse.

Environmental discourse illustrates this complexity very clearly. Named entities such as organizations (e.g., *WWF*), locations (e.g., *Amazon rainforest*), or events (e.g., *COP28*) tend to maintain lexical stability across languages. In contrast, many core environmental concepts, such as *biodiversity loss*, *carbon offsetting*, or *nature positive*, are common noun phrases [4] that are often paraphrased, technically rephrased, or culturally adapted in translation, making them harder to detect reliably.

To address this, researchers have developed rule-based NLP pipelines that integrate syntactic parsing with domain ontology mapping, providing transparent and precise extraction of candidate domain terms [5, 6]. More recently, advances brought by Large Language Models (LLMs) have enabled new approaches to domain entity recognition. Pre-trained LLMs like BERT and domain-adapted extensions show good performance to detect

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 – 26, 2025, Cagliari, Italy


\*Corresponding author.

<sup>†</sup>These authors contributed equally.

✉ elisa.chierchiello@unito.it (E. Chierchiello);

pchiril@uchicago.edu (P. Chiril); cristina.bosco@unito.it

(C. Bosco); apagano@ufmg.br (A. Pagano)

 © 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

domain mentions [3, 7, 8]. Hybrid architectures, such as the Ontology-Attention Layer, demonstrate that coupling LLMs with explicit ontology guidance further improves accuracy in specialized contexts [9].

Despite this progress, to the best of our knowledge, no study has systematically compared rule-based NLP pipelines and LLM-based methods for domain entity recognition in multilingual environmental discourse. Our study addresses this gap by focusing exclusively on domain-specific entities, with conventional named entities such as persons, organizations, and locations to be examined in future work, and implementing two distinct pipelines for their identification and classification. By implementing and comparing two pipelines on a bilingual corpus of WWF Living Planet Report Executive Summaries (2014–2024). The first pipeline uses dependency parsing to extract noun phrases and matches them to BabelNet and GEMET using string-based similarity. The second pipeline uses a prompt-based LLM to both detect noun phrases and suggest ontology matches directly. We then assess the coverage of each approach and qualitatively examine the most frequent shared mapped entities to highlight the core concepts that characterize WWF environmental discourse. To address these aims, this study investigates the following research questions:

- How much coverage do a dependency syntax-based pipeline and a prompt-based LLM pipeline each achieve when extracting and mapping noun phrases to BabelNet and GEMET in a bilingual corpus of WWF Living Planet Reports?
- Which environmental concepts emerge as the most frequently mapped entities and what do these frequent concepts reveal about thematic emphases in WWF environmental discourse?

In order to answer these questions, we assess the two pipelines in terms of coverage — that is, the number and proportion of extracted noun phrases that can be mapped to domain concepts in BabelNet and GEMET — and then examine the most frequently mapped entities to highlight key environmental concepts emphasized in WWF discourse.

Finally, to explore how LLMs can contribute to expanding domain ontologies, we prompted the LLM to generate GEMET-style definitions for entities that could not be matched in GEMET.

By comparing these two pipelines, we highlight practical considerations for building ontology-based workflows for semantic search and discourse analysis in environmental texts. While applied here to environmental texts, the general approach can be tested in other domains using suitable ontologies and tailored extraction strategies.

## 2. Related work

Generic NER has been extensively studied as a foundational NLP task, with systems reliably detecting persons, organizations, and locations [1, 2]. However, as Marrero et al. [4] and Zhang et al. [3] observe, such systems perform poorly when applied to specialized domains because domain-specific concepts are often expressed through common noun phrases rather than proper names, and thus lack the distinctive lexical or orthographic cues that standard NER methods exploit.

To address these limitations, domain-specific NER has been pursued to handle technical and abstract terminology in specialized texts. In the biomedical field, for instance, Zhang et al. [10] review biomedical entity recognition as an example of domain-focused extraction, highlighting the essential role of ontologies for semantic precision. In geosciences, Villacorta Chambi et al. [11] pursue NER improvement through the use of specialized geological schemas.

Ontology-based approaches to domain-specific entity recognition have been widely explored. Garcia-Silva et al. [5] proposed an ontology-based pipeline for environmental data that uses dependency parsing to identify candidate terms and maps them to structured environmental ontologies. Zhou and El-Gohary [6] developed a syntax-driven framework to extract provisions from environmental regulations and link them to a compliance ontology, demonstrating high precision for domain-specific phrases. Wei et al. [9] integrated an ontology-attention mechanism within BERT to improve medical entity recognition, while Dai et al. [12] emphasized the combination of entity recognition and ontology linking to build domain-specific knowledge graphs.

More recently, LLMs have emerged as powerful tools for general and domain-specific entity recognition. These LLMs can complement ontology-based systems by providing contextual understanding for domain terms that lack consistent surface forms.

Cross-lingual and multilingual methods support consistent domain entity alignment across languages. Navigli and Ponzetto [13] presented BabelNet, a multilingual lexical network used for semantic linking. GEMET [14] serves as a domain-focused environmental thesaurus, while Ryu et al. [15] and Zhao et al. [16] show how such resources help maintain terminological coherence in translation and cross-lingual NLP.

A disciplinary field that directly benefits from precise domain entity recognition supported by environmental thesauri and ontologies is environmental discourse analysis. Dryzek [17] and Doyle [18] examine how language shapes environmental policy debates and public narratives. Nerlich and Koteyko [19] explore competing frames in climate change discourse, while recent computational studies by Jørgensen et al. [20] and Chen et

al. [21] apply NLP and machine learning to large-scale climate communication data.

The aforementioned studies demonstrate the benefits of combining NLP methods with structured ontologies for domain-specific entity recognition across multiple domains. However, comparative studies on these methods in the context of multilingual environmental discourse remain limited. This work builds on these foundations to advance ontology-enriched environmental text analysis.

### 3. Methodology

This section introduces the corpus and outlines the two methodological pipelines, which combine noun phrase extraction with ontology mapping.

#### 3.1. Corpus

The corpus used in this study is the English and Italian subcorpus of the TreEn corpus [22], which compiles environmental discourse from the 2014 to 2024 editions of the WWF Living Planet Report.<sup>1</sup> WWF typically publishes a suite of documents tailored to different audiences, including a full report (a comprehensive publication containing detailed data, methodology, case studies, visualizations, and policy analysis) and an executive summary, which distills the key findings and recommendations for policymakers and stakeholders. It is important to note that for the Italian subcorpus, we were only able to locate full reports for the 2022 and 2024 editions, while for the other years under analysis, only the executive summaries were available. As such, to ensure comparability, the English subcorpus is based on the same type of document (i.e., full reports for 2022 and 2024, and executive summaries for the remaining years).

Both the English and the Italian texts were manually cleaned to retain only the plain text, with all non-textual content — such as images, captions, infographics, footnotes, and bibliographic references — systematically removed to support syntactic and semantic annotation. For each English and Italian edition of the WWF Living Planet Report published between 2014 and 2024, we computed the number of sentences, words, and lemmas using a custom Python pipeline built with pandas and language-specific spaCy models ({en, it}\_core\_web\_sm). Table 1 presents the resulting counts across reporting years.

<sup>1</sup>The timeframe (2014–2024) reflects the period for which we were able to retrieve the Italian editions of the report, starting from the earliest available publication up to the most recent: <https://www.wwf.it/cosa-facciamo/pubblicazioni/living-planet-report/>

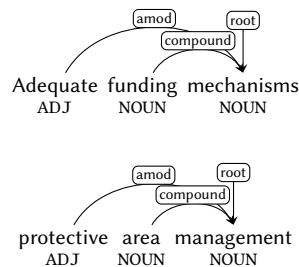
#### 3.2. Noun Phrase Extraction

Drawing on the assumption that most entities are grammatically realized as noun phrases, we applied two different methods to extract noun phrases from the corpus. As described in the following sections, the first method is rule-based and the second is LLM-based.

**Rule-based Noun Phrase Extraction.** We performed rule-based noun phrase extraction relying on annotations following the Universal Dependencies (UD) guidelines.<sup>2</sup> Sentences were annotated morphologically and syntactically using a neural state-of-the-art dependency parser [23] using the language models english-gum-ud-2.15 and italian-isdt-ud-2.15. For each sentence in CoNLL-U format, we identified head tokens tagged as NOUN or PROP and expanded them by recursively including adjectival modifiers (amod), compounds (compound), and nominal modifiers (nmod). The extraction algorithm builds each noun phrase starting from the head and prepending modifiers according to their dependency links. The lemma column in each CoNLL-U representation was used in order to reduce lexical variation and support downstream concept mapping. For instance, from the sample sentence:

- (1) *Adequate funding mechanisms are needed if protective area management is to be effective.*

the extracted noun phrases are: *Adequate funding mechanisms* and *protective area management*, which according to the UD guidelines have the same internal structure and are represented as shown in Figure 1:



**Figure 1:** Dependency syntax annotation for sample noun phrases.

**LLM-based Noun Phrase Extraction.** Our second method of noun phrase extraction employed GPT-o3.<sup>3</sup>

<sup>2</sup><https://universaldependencies.org/guidelines.html>

<sup>3</sup><https://platform.openai.com/docs/models/o3>

WWF report	Language	sentences	tokens	unique words	lemmas	avg. sentence length
2014	English	263	4,795	1,268	999	18
	Italian	261	5,759	1,541	1,140	20
2016	English	371	6,973	1,727	1,361	18.6
	Italian	371	8,308	2,033	1,506	21
2018	English	253	5,203	1,372	1,101	20.2
	Italian	237	5,855	1,663	1,268	20.7
2020	English	378	6,786	1,777	1,444	17.6
	Italian	377	7,948	2,102	1,604	18.9
2022	English	852	19,531	3,309	2,545	22.8
	Italian	853	22,153	3,963	2,892	23.2
2024	English	1,042	23,462	3,163	2,346	22.5
	Italian	1,048	26,976	3,988	2,743	25.7

**Table 1**

Basic statistics of the the English and Italian corpora across six reporting years (2014–2024).

Specific prompts were iteratively developed for each language under analysis (i.e., Italian and English), with instructions highlighting syntactic constraints, lemmatization, and complete modifier preservation in order to ensure consistency with the rule-based noun phrase extraction method. Figure 4 (see Appendix A) presents the English prompt used for this task.

### 3.3. Ontology Mapping

#### 3.3.1. Ontology String Matching

Following the extraction of candidate noun phrases, we performed concept-level mapping using two distinct ontologies: GEMET and BabelNet. The objective was to link each phrase to a unique identifier representing an environmentally relevant concept within a structured semantic resource.

Our multilingual setting required different strategies for the two resources. For GEMET, which is primarily designed around English entries and offers more limited multilingual coverage, we relied on aligned sentence pairs in English and Italian to propagate annotations. Specifically, we used an alignment file where each English sentence was paired with its Italian equivalent. Once GEMET concepts were identified in the English sentence, we transferred them to the Italian version whenever the same noun phrase (or a direct translation) was present. This allowed us to enrich the Italian portion of the corpus even when direct GEMET matches were not available in Italian. To support this transfer, we first checked whether the same noun phrase annotated in English occurred verbatim in the aligned Italian sentence. If no exact match was found, we used automatic translation to bridge the gap between the two languages. Specifically, we translated the English noun phrase into Italian using Google

Translate,<sup>4</sup> and then applied basic normalization (e.g., lowercasing, removal of diacritics) before comparing it to the set of Italian noun phrases extracted from the aligned sentence using the same syntactic rules. If a match was found, the corresponding GEMET concept was propagated to the Italian sentence. For example, in the English sentence:

- (2) “*Around the world, many languages are used to communicate science.*”

the noun phrases *science* and *world* were mapped to GEMET concepts. Their Italian equivalents, *scienza* and *mondo*, appeared among the extracted noun phrases in the aligned Italian sentence “*In tutto il mondo si usano molte lingue per comunicare la scienza*”. In this way, we could propagate the annotations to the Italian side, even though the GEMET concept is originally linked to the English noun phrase.

In contrast, BabelNet provides multilingual support by design. Therefore, we queried noun phrases directly in both English and Italian, allowing us to retrieve language-specific senses without relying on sentence alignment. This approach enabled broader coverage and avoided the need for cross-lingual projection.

**GEMET.** We queried GEMET via its public REST API.<sup>5</sup> For each noun phrase, we attempted an exact string match using the `getConceptsMatchingKeyword` endpoint. To maximize recall, we also applied fallback strategies by decomposing multiword expressions and querying each component token separately (e.g., *climate vulnerability* → *climate*, *vulnerability*). Concept URIs (Uniform Resource Identifier) returned from GEMET were

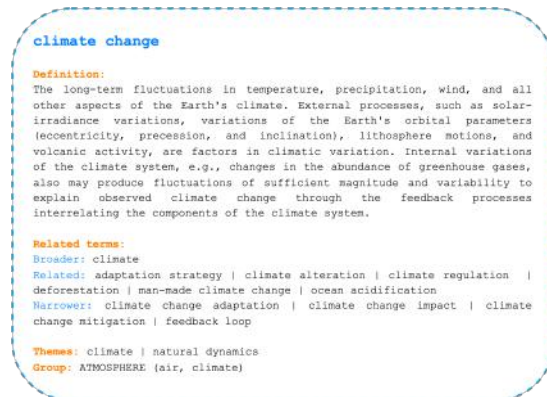
<sup>4</sup><https://cloud.google.com/translate/docs/reference/rest>

<sup>5</sup><https://www.eionet.europa.eu/gemet/en/webservices/>



stored along with the original phrase to support later semantic grouping and analysis. In addition, we retrieved the semantic group associated with each concept using the `getAllConceptRelatives` endpoint (with relation group), allowing us to categorize entities into high-level thematic domains (e.g., BIOSPHERE, SOCIETY, WASTES).

Figure 2 shows the GEMET entry for *climate change*<sup>6</sup> and all the fields we extract: concept URI, label, definition, related terms, and group. These fields were stored to facilitate both downstream semantic analysis and explainability of the mappings.



**Figure 2:** GEMET entry showing extracted fields: definition, related terms, themes, and group.

**BabelNet.** In parallel, we integrated mappings from BabelNet. We accessed BabelNet via its `getSenses` and `getSynsetIds` endpoints,<sup>7</sup> querying each noun phrase both in Italian and in English. This bilingual querying strategy was adopted to maximize coverage and mitigate cases where a concept might be present only in one of the two languages. Unlike GEMET, BabelNet returns disambiguated senses associated with synset identifiers. We retained only senses with part-of-speech NOUN and applied a filtering step to discard irrelevant or ambiguous senses based on glosses and semantic domains. In multiword expressions that failed to return a direct match, we again decomposed the phrase into component tokens and aggregated partial matches when available. As an example, consider the following sentence:

- (3) *Scientists suggest that we have transitioned from the Holocene into a new geological epoch, calling it the 'Anthropocene'.*

<sup>6</sup><http://www.eionet.europa.eu/gemet/concept/1471>

<sup>7</sup><https://babelnet.org/guide>

The extracted noun phrases are: *scientist*, *Holocene*, *new geological epoch*, *Anthropocene*. Figure 3 shows the BabelNet entry for *Anthropocene*, highlighting the fields we extract, namely, definition, categories, relations, synonyms, and semantically related terms.



**Figure 3:** BabelNet entry showing some of the extracted fields.

No GEMET concept was found matching any of these noun phrases, highlighting the wider lexical and multilingual coverage of BabelNet. This example also demonstrates the complementary nature of the two resources: GEMET provides high precision within the environmental domain, while BabelNet ensures broader recall across a wider conceptual space.

This dual mapping strategy enabled both domain-specific grounding (via GEMET) and broader lexical disambiguation (via BabelNet), intensifying the robustness of concept alignment across heterogeneous texts.

### 3.3.2. LLM-based Ontology Mapping

Our second method for performing the concept-level mapping of the extracted noun phrases relies on GPT-o3 through prompts with expected output. Upon extraction, a manual analysis of the concept-level mapping process (cf. Table 2) revealed that several (multi-word) noun phrases were not found in either GEMET or BabelNet. For example, in the following sentence:

- (4) *We need nature positive by 2030 – which, in simple terms, means more nature by the end of this decade than at its start.*

*nature positive* is one of the noun phrases both syntactically and semantically relevant to environmental discourse. While the phrase is made up of a NOUN and ADJ, it is used as a NOUN and has a distinct meaning in current environmental discourse. Both the rule- and LLM-based methods correctly identified *nature positive* as a noun phrase, and it was successfully matched to a corresponding concept in BabelNet. The concept, however, is notably absent in GEMET. To address such coverage gap due to GEMET's limitations and explore the potential of using LLMs for ontology-based annotation, we used GPT-o3 to generate GEMET-style annotations for unmatched phrases in the first output. The prompt used for this task is shown in Figure 4 (see Appendix A).

### 3.4. Thematic analysis

To conduct a diachronic analysis of concepts mapped with GEMET and BabelNet across the 2014–2024 corpus, we identified concepts that appear in all WWF report editions and analysed their frequency per report to gather insights into the evolution of environmental discourse.

## 4. Results

Following the extraction of candidate noun phrases and their subsequent concept-label annotation using GEMET and BabelNet, Table 2 and Table 3 (see Appendix A) present the coverage of noun phrases by the rule- and LLM-based methods across the English and Italian corpora, as well as the number of phrases matched either fully or partially in the two ontologies. A full (exact) match refers to cases in which the entire noun phrase (e.g., *vertebrate species*) was found in the ontology, while a partial match refers to instances in which only a component (substring) of the phrase (e.g., *vertebrate* or *species*) was found. For example, in GEMET, while *vertebrate species* was not found, both *vertebrate* and *species* were matched individually, resulting in a partial match.

Across the 2014–2024 WWF reports, the LLM-based method extracts a number of unique noun phrases comparable to the rule-based method. However, for the 2022 Italian edition, the LLM extracts substantially more phrases. This difference appears to result from the way nested structures were handled: the LLM returned entire noun phrases with several nested noun phrases. This pattern is especially evident in Italian, where nested noun and prepositional phrases are common. For instance, extracted spans such as “*negoziato internazionale della convenzione quadro delle Nazioni Unite sul cambiamento climatico e della convenzione sulla diversità biologica*”<sup>8</sup> or

<sup>8</sup>Original in English: “*international negotiations under the United Nations Framework Convention on Climate Change and the Convention on Biological Diversity*”.

“*accesso a quantità senza precedente di dato da sensore su satellite smartphone e dispositivo*”<sup>9</sup> illustrate the model's tendency to extract the full extent of some noun phrases which have several nested ones.

In terms of coverage, the Italian noun phrases extracted using GPT-o3 show a notable increase in GEMET exact matches, nearly doubling the coverage compared to the rule-based approach. Partial matches also increase across both ontologies, indicating a broader semantic reach. Interestingly, exact matches to BabelNet decline sharply for noun phrases extracted using the LLM-based method after 2016, even when partial BabelNet coverage increases. As noted above, GPT-o3 tends to extract longer and more contextually rich noun phrases that partially align with BabelNet entries. For instance, in the sentence:

- (5) *At the Rio+20 conference in 2012, the world's governments affirmed their commitment to an “economically, socially and environmentally sustainable future for our planet and for present and future generations”.*

one of the noun phrases extracted by GPT-o3 is *economically socially environmentally sustainable future*. While conceptually accurate, this phrase does not match any exact entry in BabelNet, whereas a shorter variant such as *sustainable future* does. This seems to suggest that GPT-o3 extracts entire noun phrases including all modifiers (*economically socially environmentally*), when ontology entries typically include noun phrases made up of classifiers and a few epithets, in this case, *sustainable*.

Regarding the capabilities of LLMs in ontology-based annotation, our manual analysis of the quality of the definitions generated by GPT-o3 for phrases not found in GEMET provided relevant insights into the potential for using LLMs for scaling semantic resources.

For instance, going back to example (4), the term *nature positive*, while absent from GEMET, is present in BabelNet, which defines it as “*outcomes which are net positive for biodiversity, directly and measurably increasing in the health, abundance, diversity and resilience of species, ecosystems and processes*”. GPT-o3, on the other hand, generates the following definition: “*a future state in which nature—biodiversity, ecosystem services and natural capital—is restored and enhanced relative to its current condition*”. While both definitions are valid, the LLM-generated one captures more accurately the forward-looking, goal-oriented nature of the *nature positive* concept. Unlike BabelNet's definition, which frames the concept mainly as a set of measurable biodiversity outcomes, the LLM definition presents it as a “*future state*” in which nature is restored and enhanced. This distinction is significant, as BabelNet treats the concept as a

<sup>9</sup>Original in English: “*access to unprecedented amounts of data from sensors on satellites, smartphones and in situ devices*”.

result, while the LLM version treats it as a trajectory/vision, which aligns more closely with how the term is currently used in WWF discourse (e.g., WWF defines nature positive as a goal to “halt and reverse nature loss by 2030”).<sup>10</sup> This suggests a promising direction for scaling these semantic resources, with domain-relevant entities extracted from domain-specific literature. However, as highlighted above, given the nuanced conceptual distinctions, expert validation remains crucial in order to ensure accuracy and to account for the subtle semantic distinctions that such models may overlook.

**Temporal analysis of concept dynamics.** Our diachronic analysis of the concepts mapped via GEMET and BabelNet across the six-year corpus (2014–2024) yielded the following results.

For GEMET, we identified 59 English concepts that appeared consistently in all years, including domain-specific terms such as *climate change*, *biodiversity*, *ecosystem*, and *habitat loss*, reflecting the controlled and environmentally focused nature of the thesaurus. A parallel analysis of the Italian portion revealed a partially overlapping core set, with terms such as *ambiente* (environment), *specie* (species), and *risorsa* (resource) persistently appearing.

In contrast, BabelNet yielded a smaller set of consistently recurring concepts, such as *biodiversity*, *consumption*, and *development*, but also revealed a much broader and more dynamic tail of emerging concepts (i.e., less frequent terms that vary widely across documents and capture context-dependent discourse). Notably, BabelNet annotations surfaced many general-purpose or discourse-driven terms (e.g., *ambition*, *alarm*, *goal*, *confidence limit*), often reflecting the rhetorical framing of environmental narratives in the source texts.

We also tracked *emerging* and *declining* concepts across both resources. For GEMET, emergent concepts since 2018 include *soil biodiversity*, *plastic*, *ocean acidification*, and *urbanisation*, many of which correspond to increasingly salient ecological issues. Conversely, concepts such as *ammonia*, *energy consumption*, and *ozone* peaked before 2018 and gradually disappeared, suggesting shifting topical focus in environmental discourse. Similar trends were found in BabelNet, where contemporary discourse introduced terms like *sdg*,<sup>11</sup> *carbon sequestration*, and *digital storytelling*, while older narrative anchors like *anthropocene*, *habitat loss*, and even *ocean* saw relative decline. A detailed overview of the five most frequent concepts per year, derived from both GEMET and BabelNet annotations, is provided in Tables 4 and 5 (see Appendix A).

## 5. Discussion

Our findings shed light on both the strengths and limitations of rule-based and LLM-based pipelines for ontology-oriented entity annotation in environmental discourse, aligning with insights from previous work on domain-specific NLP [3, 4, 5, 6].

First, in terms of extraction, the LLM-based approach demonstrated coverage comparable to the rule-based method in line with recent research highlighting LLMs’ strong performance for entity detection [7, 8]. However, the LLM’s tendency to generate longer, contextually rich noun phrases — particularly in Italian, where nesting is frequent — resulted in both higher phrase counts and a greater proportion of partial matches. This confirms observations by Marrero et al. [4] that domain-relevant concepts often appear as complex, nested noun phrases that challenge standard NER boundaries.

Second, our results show that while GEMET provides reliable coverage for core environmental concepts, consistent with its controlled and domain-focused design, BabelNet offers a wider conceptual coverage. This aligns with prior findings that general-purpose lexical networks like BabelNet can capture more entities, but at the same time can include discourse or general entities not so relevant to characterize a domain [13, 15].

Third, the quality of LLM-generated definitions for unmapped phrases suggests potential for semi-automated ontology enrichment. For instance, for the concept *nature positive*, the LLM produced a forward-looking definition more aligned with current environmental discourse framing than the existing BabelNet entry. This supports recent arguments for integrating LLMs into domain ontology extension workflows [9], but also highlights the importance of expert validation, given possible subtleties in sense distinctions.

Finally, our diachronic analysis, though conducted on a very low scale, showed interesting aspects about how sustainability narratives evolve rhetorically, in line with work by Dryzek [17] and Nerlich and Koteyko [19] on shifting environmental frames.

Taken together, our results demonstrate that combining rule-based and LLM-based pipelines may provide complementary strengths for environmental concept annotation: the rule-based method ensures syntactic precision and consistent granularity, while the LLM broadens semantic reach and can supply draft definitions for novel or evolving terms. However, consistent ontology coverage remains an issue, as a substantial proportion of relevant phrases were not found in either resource, underscoring the need for ongoing ontology expansion and domain adaptation, as stressed in recent surveys [10, 24].

Future work should explore refining LLM prompts to better constrain phrase boundaries, integrating syntactic cues during generation, and developing semi-

<sup>10</sup>[https://www.f.panda.org/nature\\_positive/](https://www.f.panda.org/nature_positive/)

<sup>11</sup>Acronym for Sustainable Development Goals.

automatic curation workflows to incorporate validated LLM-generated definitions into existing ontologies. This is a promising path for scaling high-quality, domain-adapted semantic annotation in support of environmental discourse analysis.

## 6. Conclusion and Future Work

In this study, we presented a pipeline for extracting and semantically annotating noun phrases in multilingual environmental texts using both GEMET and BabelNet ontological frameworks. The two resources were used in complementary ways: GEMET provided structured domain-specific knowledge, while BabelNet contributed broader lexical coverage and multilingual flexibility. Through a combination of ontology matching, fallback decomposition strategies, and cross-lingual projection, we achieved wide and meaningful semantic enrichment across languages. Looking ahead, the approach we propose could also support the ongoing evolution of domain ontologies themselves. For instance, GEMET is periodically updated with new concepts and definitions.<sup>12</sup> Automatically extracting candidate terms and associating them with existing or missing concepts, especially through LLM-based suggestion and contextual generalization, might provide curators looking to add to the thesaurus with insightful information.

Several directions can be pursued for the future development of this work. For instance, alternative approaches to named entity propagation — such as alignment-based techniques [25, 26] — can be tested, and additional inventories for entities and concepts can be explored, such as [27].

Finally, it is important to note that our study focused on the task as performed by LLMs. In future work, we will compare these results with human annotations provided by domain experts in order to examine whether more or different entities are extracted from the texts. This comparison will help determine whether more fine-grained analyses are necessary (e.g., to resolve partial matches involving nested entities or syntactically complex modifier structures). Moreover, incorporating expert judgment will allow us to account for diverse disciplinary perspectives (e.g., biology, ecology, chemistry, physics, geography) on environmental issues.

## Acknowledgments

This research is supported by the University of Turin and the TreEn project team. Special thanks to the collaborators who contributed to the treebank annotation.

The work of Elisa Chierchiello is funded by the International project *CN-HPC-Spoke1-Future HPC & Big Data*, *PNRR MUR-M4C2*. Adriana Pagano has a grant from Brazil’s National Council for Scientific and Technological Development (CNPq 404722/2024-5; 313103/2021-6) and Minas Gerais State Agency for Research and Development (FAPEMIG) to conduct research in collaboration with the University of Turin.

## References

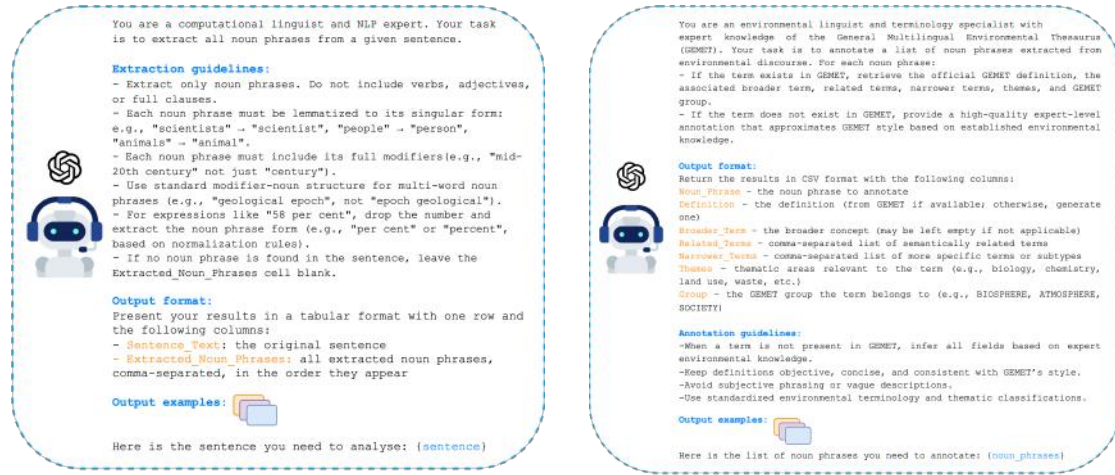
- [1] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, *Lingvisticae Investigationes* 30 (2007) 3–26.
- [2] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2018, pp. 2145–2158.
- [3] X. Zhang, Y. Jiang, X. Wang, X. Hu, Y. Sun, P. Xie, M. Zhang, Domain-specific ner via retrieving correlated samples, in: *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, 2022, pp. 2398–2404.
- [4] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, J. M. Gómez-Berbís, Named entity recognition: Fallacies, challenges and opportunities, *Computer Standards & Interfaces* 35 (2013) 482–489.
- [5] A. García-Silva, Corcho, B. Villazón-Terrazas, Ontology-based information extraction: A case study on environmental data, *Knowledge and Information Systems* 62 (2020) 449–471.
- [6] P. Zhou, N. El-Gohary, Ontology-based information extraction from environmental regulations for supporting environmental compliance checking, in: *Proceedings of the International Workshop on Computing in Civil Engineering 2015, ASCE*, 2015, pp. 190–198. doi:10.1061/9780784479247.024.
- [7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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 (NAACL-HLT)*, 2019.
- [8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al., Language models are few-shot learners, in: *Advances in Neural Information Processing Systems*, volume 33, 2020, pp. 1877–1901.
- [9] C.-H. Wei, R. Leaman, Z. Lu, Ontology attention layer for medical named entity recognition, *Journal of Biomedical Informatics* 141 (2023) 104385.
- [10] J.-D. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, Biomed-

<sup>12</sup><https://www.eionet.europa.eu/gemet/en/about/>



- ical entity recognition: A systematic review of approaches and challenges, *Briefings in Bioinformatics* 22 (2021) bbaa180.
- [11] S. P. Villacorta Chambi, M. Lindsay, J. Klump, K. Gessner, E. Gray, H. McFarlane, Assessing named entity recognition by using geoscience domain schemas: the case of mineral systems, *Frontiers in Earth Science* 13 (2025) 1530004.
- [12] X. Dai, S. Karimi, C. Paris, Building domain-specific knowledge graphs for named entity linking: A case study of cancer research literature, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7460–7471.
- [13] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial Intelligence* 193 (2012) 217–250. doi:10.1016/j.artint.2012.07.001.
- [14] European Environment Agency, Gemet — general multilingual environmental thesaurus, <https://www.eionet.europa.eu/gemet/en/themes/>, 2025. Accessed: 2025-06-15.
- [15] J. Ryu, J. Lee, J. Kang, Cross-lingual entity linking with multilingual bert and knowledge graph embedding, volume 546, 2021, pp. 663–674.
- [16] Y. Zhao, W. Chen, X. Xie, Z. Liu, J. Li, Entity-aware neural machine translation, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2858–2866.
- [17] J. S. Dryzek, *The Politics of the Earth: Environmental Discourses*, Oxford University Press, 2013.
- [18] J. Doyle, *Media and the Environment*, Polity Press, 2020.
- [19] B. Nerlich, N. Koteyko, Competing representations of the 'climate change' frame in uk news media, *Nature Climate Change* 3 (2009) 423–427.
- [20] P. S. Jørgensen, colleagues, Machine learning and natural language processing in environmental research, *Environmental Research Letters* 17 (2022) 023003.
- [21] Z. Chen, T. Zhang, H. Su, Analyzing climate change discourse with nlp: A review, *Current Opinion in Environmental Sustainability* 61 (2023) 101237.
- [22] A. Pagano, P. Chiril, E. Chierchiello, C. Bosco, Treen: A multilingual treebank project on environmental discourse, in: *Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, 2025, p. 80.
- [23] M. Straka, UDPipe 2.0 prototype at CoNLL 2018 UD shared task, in: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 197–207. URL: <https://www.aclweb.org/anthology/K18-2020>. doi:10.18653/v1/K18-2020.
- [24] S. Li, Z. Zhou, L. Huang, F. Wu, A survey on ontology-based named entity recognition, *IEEE Access* 10 (2022) 113192–113210.
- [25] S. Tedeschi, R. Navigli, MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation), in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, Seattle, United States, 2022, pp. 801–812. URL: <https://aclanthology.org/2022.findings-naacl.60/>. doi:10.18653/v1/2022.findings-naacl.60.
- [26] F. Martelli, A. S. Bejgu, C. Campagnano, J. Čibej, R. Costa, A. Gantar, J. Kallas, S. P. Koeva, K. Koppel, S. Krek, M. Langemets, V. Lipp, S. Nimb, S. Olsen, B. Sanford Pedersen, V. Quochi, A. Salgado, L. Simon, C. Tiberius, R.-J. Ureña-Ruiz, R. Navigli, XLWA: a gold evaluation benchmark for word alignment in 14 language pairs, in: F. Boschetti, G. E. Lebani, B. Magnini, N. Novielli (Eds.), *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 272–280. URL: <https://aclanthology.org/2023.clicit-1.34/>.
- [27] G. Martinelli, F. Molfese, S. Tedeschi, A. Fernández-Castro, R. Navigli, CNER: Concept and named entity recognition, in: K. Duh, H. Gomez, S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 8336–8351.

## A. Appendix



**Figure 4:** GPT-o3 prompt used to *extract noun phrases from English sentences* (left), and *annotate English noun phrases according to the GEMET ontology* (right).

WWF report	Language	noun phrases	in GEMET		partial in GEMET		in BabelNet		partial in BabelNet	
2014	English	778	232	29.82%	157	20.18%	483	62.08%	318	40.87%
	Italian	815	64	7.85%	80	9.82%	450	55.21%	343	42.09%
2016	English	1,198	309	25.79%	289	24.12%	699	58.35%	537	44.82%
	Italian	1,146	99	8.64%	144	12.57%	615	53.66%	514	44.85%
2018	English	875	220	25.14%	196	22.40%	513	58.63%	398	45.49%
	Italian	883	46	5.21%	90	10.19%	456	51.64%	384	43.49%
2020	English	1,283	349	27.20%	312	24.32%	770	60.02%	556	43.34%
	Italian	1,207	110	9.11%	159	13.17%	696	57.66%	475	39.35%
2022	English	2,926	787	26.90%	743	25.39%	1,719	58.75%	1,349	46.10%
	Italian	2,632	64	2.43%	195	7.41%	1,343	51.03%	1,195	45.40%
2024	English	2,926	625	21.36%	934	31.92%	1,565	53.49%	1,444	49.35%
	Italian	3,240	136	4.20%	802	24.75%	1,088	33.58%	2,073	63.98%

**Table 2**

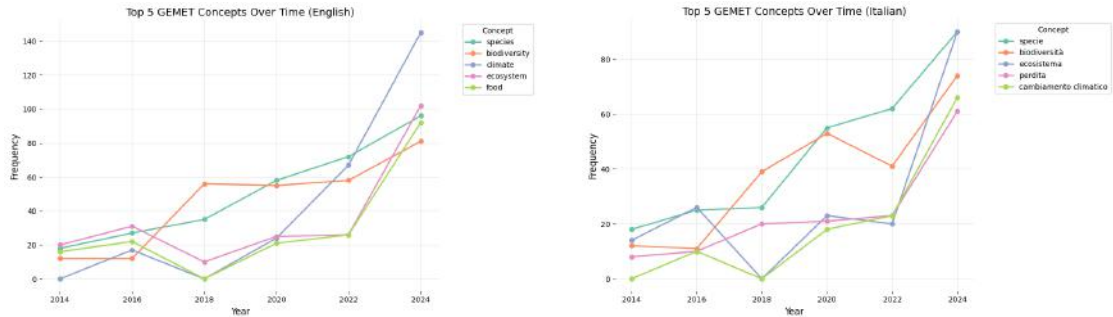
Coverage of (unique) noun phrases extracted through the rule-based method from WWF Reports in GEMET and BabelNet (2014–2024).

WWF report	Language	noun phrases	in GEMET		partial in GEMET		in BabelNet		partial in BabelNet	
2014	English	691	207	29.96%	160	23.15%	442	63.97%	280	40.52%
	Italian	791	156	19.72%	143	18.08%	228	28.82%	403	50.95%
2016	English	1,093	280	25.62%	293	26.81%	656	60.02%	484	44.28%
	Italian	1,264	185	14.64%	309	24.45%	255	20.17%	707	55.93%
2018	English	799	179	22.40%	250	31.29%	398	49.81%	394	49.31%
	Italian	885	127	14.35%	184	20.79%	237	26.78%	472	53.33%
2020	English	1,214	294	24.22%	402	33.11%	638	52.55%	620	51.07%
	Italian	1,224	191	15.60%	334	27.29%	275	22.47%	632	51.63%
2022	English	2,950	725	24.58%	1,009	34.20%	821	27.83%	1,921	65.12%
	Italian	3,364	424	12.60%	478	14.21%	623	18.52%	2,085	61.98%
2024	English	3,050	574	18.82%	1,192	39.08%	636	20.85%	2,020	66.23%
	Italian	2,827	343	12.13%	870	30.77%	459	16.24%	1,623	57.41%

**Table 3**  
Coverage of (unique) noun phrases extracted through the LLM-based method from WWF Reports in GEMET and BabelNet (2014–2024).

WWF report	Language	Top five GEMET Concepts (with frequency)
2014	English	ecosystem (20), world (20), species (18), energy (18), resource (16)
	Italian	specie (18), ecosistema (14), energia (13), mondo (13), biodiversità (12)
2016	English	ecosystem (31), species (27), resource (22), food (22), energy (18)
	Italian	ecosistema (26), specie (25), risorsa (23), habitat (15), consumo (15)
2018	English	biodiversity (56), species (35), loss (26), indicator (15), land (14)
	Italian	biodiversità (39), specie (26), perdita (20), indicatore (10), conservazione (9)
2020	English	species (58), biodiversity (55), ecosystem (25), climate (24), world (23)
	Italian	specie (55), biodiversità (53), ecosistema (23), perdita (21), mondo (20)
2022	English	species (72), climate (67), biodiversity (58), loss (38), climate change (30)
	Italian	specie (62), biodiversità (41), perdita (23), cambiamento climatico (23), foresta (21)
2024	English	climate (145), ecosystem (102), species (96), food (92), energy (91)
	Italian	specie (90), ecosistema (90), biodiversità (74), cambiamento climatico (66), clima (64)

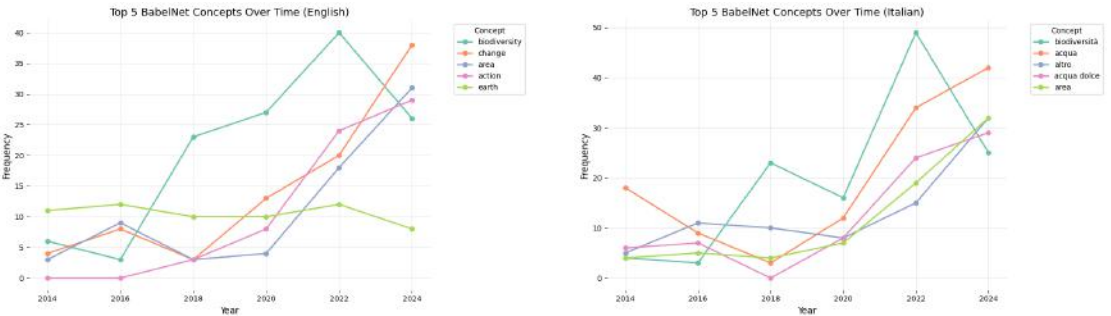
**Table 4**  
Most frequent GEMET concepts extracted from WWF Reports (2014–2024) in English and Italian, with corresponding frequency counts.



**Figure 5:** Top five GEMET concepts across WWF Reports (2014–2024) in English (left) and Italian (right).

WWF report	Language	Top five BabelNet Concepts (with frequency)
2014	English Italian	earth (11), country (10), lpi (9), development (8), biodiversity (6) acqua (18), ambientale (15), alto (10), declino (8), anno (8)
2016	English Italian	lpi (15), earth (12), anthropocene (10), area (9), consumption (9) ambientale (11), altro (11), anno (10), acqua (9), antropocene (8)
2018	English Italian	biodiversity (23), index (11), earth (10), lpi (9), abundance (8) biodiversità (23), altro (10), anno (8), accordo (7), agricoltura (7)
2020	English Italian	biodiversity (27), change (13), index (11), earth (10), action (8) biodiversità (16), acqua (12), agricolo (10), alimentare (9), abbondanza (9)
2022	English Italian	biodiversity (40), action (24), amazon (21), change (20), area (18) biodiversità (49), acqua (34), abbondanza (28), acqua dolce (24), approccio (20)
2024	English Italian	change (38), area (31), action (29), biodiversity (26), lpi (22) acqua (42), alimentare (40), altro (32), area (32), acqua dolce (29)

**Table 5**  
Most frequent BabelNet concepts extracted from WWF Reports (2014–2024) in English and Italian, with corresponding frequency counts.



**Figure 6:** Top 5 BabelNet concepts across WWF Reports (2014–2024) in English (left) and Italian (right).