TREEN: A Multilingual Treebank Project on Environmental Discourse

Adriana Pagano
Universidade Federal de Minas Gerais
apagano@ufmg.br

Patricia Chiril
University of Chicago
pchiril@uchicago.edu

Elisa Chierchiello
Università degli Studi di Torino
elisa.chierchiello@unito.it

Cristina Bosco
Università degli Studi di Torino
cristina.bosco@unito.it

Abstract

The past few decades have seen a simultaneous increase in both the complexity of environmental debates in the media and the language people use to engage with them. While linguistic and communication studies have been pursued on this discourse, the development of computational linguistic tools and resources dedicated to support its analysis and interpretation is still very incipient. For one, no morphological and syntactic resources specific to the environmental domain can be found on major platforms and repositories. This paper introduces TREEN, a multilingual treebank project in progress which compiles texts on environmental discourse produced in different communication contexts, focusing on a set of non-governmental reports included in the first release of TREEN. It reports on the parallel component of the project and discusses issues faced during sentence-level alignment between original and translated texts, annotation of texts following UD guidelines, and labeling entities drawing on an ontology of environmental-related topics. This novel resource is expected to support environmental discourse analysis by providing morphological and syntactic data to enable cross-language and cross-cultural comparisons based on the semantics of the entities annotated in the treebank.

1 Introduction

In 1894, the Swedish Nobel Prize-winning chemist and physicist Svante Arrhenius formulated a climate model that correlated rising atmospheric CO_2 levels with glacier melting, pointing for the first time to human activity as the cause of global warming (Kolbert, 2024). More than a century later, Arrhenius' model remains the subject of heated debate, not only in academia but also in communication and social media, yielding a massive amount of environmental discourse and posing a challenge to analysts and consumers alike.

Environmental discourse is often characterized by its grammatical intricacy (Halliday, 1992), accountable for the emergence of its domain-specialized lexicon. It encompasses different genres and registers, whereby meaning is construed by different stakeholders. Moreover, there is the impact of languages in contact, as translation from major into minor languages is both a frequent means of production and a main source of variation.

As far as the computational analysis of environmental discourse is concerned, resources have only recently begun to be developed, driven by the poor performance of Large Language Models (LLMs) in this domain (Webersinke et al., 2022) and the need for dedicated pre-trained models to efficiently process such texts (Thulke et al., 2024). However, corpora and datasets remain scarce. Stede and Patz (2021) reviewed available datasets on climate-change discourse and highlighted the predominance of news and social media texts in NLP analysis, with approaches primarily relying on corpus linguistics (raw frequency counts, collocations) and out-of-the box tools for topic modelling, sentiment analysis and network analysis. Notably, the authors did not report the existence of any annotated resources, such as treebanks. A more recent

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 80–96

initiative, however, introduced a multimodal corpus of academic articles and texts from websites (including the International Panel on Climate Change (IPCC), Greenpeace International, and Greenpeace Germany), enriched with metadata and annotations in order to allow for more in-depth discourse analysis (Bartsch et al., 2023).

In line with Bartsch et al. (2023), our work pursues the development of annotated corpora to support the study of environmental discourse. We adopt a multilingual perspective which can inform both theoretical and applied studies in the fields of language comparison and typology, translation studies, discourse analysis, to name but a few. It also provides a valuable resource for computational linguists seeking to fine-tune general-purpose models and create a benchmark for evaluating LLMs.

As part of this wider project, this paper introduces TREEN (Treebanks for Environment), the first multilingual treebank project, comprising treebanks of comparable and parallel texts on environmental discourse. More specifically, it describes the parallel component of TREEN currently including texts extracted from a non-governmental report in four different languages (English, Brazilian Portuguese, Italian and Romanian), annotated for morphology and syntax following the Universal Dependencies (UD) guidelines and enriched with entity annotations drawing on the GEneral Multilingual Environmental Thesaurus (GEMET).[1]

The remainder of the paper is organized as follows. Section 2 provides a brief survey of related work. Section 3 describes the data collection process for the creation of the parallel component of TREEN and outlines the methodological decisions made during text alignment. It also details the annotation guidelines applied to this resource, namely UD and Universal Named Entity Recognition (UNER). Section 4 presents the statistics of our parallel treebanks and characterizes each of the four languages compiled in TREEN, highlighting key insights from a cross-linguistic comparison. Finally, Section 5 presents concluding remarks and outlines directions for future work.

2 Related Work

Most of the resources hitherto developed for the analysis of environmental discourse consist of raw corpora or datasets labelled for classification tasks such as Sentiment Analysis. Nevertheless, as noted by Ibrohim et al. (2023), Sentiment Analysis resources for environmental texts are currently extremely limited in all languages. While some corpora on environmental topics are available for English, resources in other languages are scarce. Even in dedicated emerging disciplines, as in Ecological Discourse Analysis (EAD), studies and resources are still at an early stage and mostly concentrated in specific countries and languages, as highlighted in a very recent systematic review by Song et al. (2025). Interestingly, some of the countries that are key players in environmental debates, as is the case of Brazil, are not represented in EAD at all.

In computational linguistics, there is growing interest in **environmental data**, as evidenced by the increasing number of events on this topic in recent years, such as the Workshop on *Ecology, Environment and Natural Language Processing*[2] (Basile et al., 2025), *ClimateNLP* (2024 - 2025),[3] and the *Tackling Climate Change with Machine Learning Workshop*,[4] which has been held regularly since 2019 in top-tier conferences. In addition, smaller-scale challenges and data competitions (such as those hosted on platforms like Kaggle[5]) or individual efforts by researchers have made annotated social media corpora related to climate and environmental issues available. These corpora cover a range of tasks related to environmental discourse, with a particular focus on aspects of climate change. Some works focus on verifying whether a given text contains environmental claims (or pertains to climate change) (Varini et al., 2020; Diggelmann et al., 2020; Stammbach et al., 2023), or facilitate expert-informed retrieval from corporate climate disclosures (Schimanski et al., 2024), while others examine the topics discussed within such narratives (Dahal et al., 2019; Duong et al., 2022; Effrosynidis et al., 2022; Vaid et al., 2022). Other studies focus on detecting and analyzing attitudes towards global warming (Luo et al., 2020), as well as identifying neutralization techniques (i.e., rhetorical strategies used in climate change skepticism to justify inaction or promote alternative views) (Bhatia et al.,

[1] https://www.eionet.europa.eu/gemet/en/about/

[2] https://econlpws2025.di.unito.it/

[3] https://nlp4climate.github.io/

[4] https://www.climatechange.ai/events#past-events

[5] One such dataset aggregates climate-related tweets, each manually labeled by three annotators as: *news* (factual news about climate change), *pro* (supports the human-caused climate change belief), *neutral* (no stance), or *anti* (rejects the human-caused climate change) belief: https://www.kaggle.com/competitions/climate-change-edsa2020-21/overview.

2021). Another important line of research targets climate change contrarianism, denial, and disinformation (Coan et al., 2021; Piskorski et al., 2022; Bhatia et al., 2020).

As previously mentioned, one of the few **annotated resources** reported in the literature is the InsightsNet Climate Change Corpus (ICCC) introduced by Bartsch et al. (2023). The ICCC comprises academic articles and texts from websites, and is annotated for metadata, morphology, syntax, and Named Entities. The authors used spaCy and StanzaCoreNLP for automatic annotation and do not report any manual check of the output. While the corpus includes texts in English along with their German translations (for most of the texts), no alignment between the originals and translations was carried out.

The availability of parallel texts is relevant for several NLP tasks and linguistic fields. With regard to **parallel treebanks**, there are few available in the UD repository. Among them, the Parallel Universal Dependencies (PUD) treebanks, created for the CoNLL 2017 shared task on *Multilingual Parsing from Raw Text to Universal Dependencies*[6] and including 15 languages (Zeman et al., 2017). For English and Swedish, we can cite LinEs (Ahrenberg, 2015), while Italian, English and French are aligned in ParTUT (Sanguinetti and Bosco, 2014). The fact that these treebanks have been made compliant with the UD guidelines enhances their comparability and makes them a particularly valuable resource for linguists and computer scientists. None of these treebanks, however, is dedicated to environmental discourse.

Several **unannotated parallel corpora** are also available through the OPUS project platform.[7] Among them, some parallel corpora can incidentally include language related to environmental topics, as is the case of EUROPARL,[8] which includes debates on environmental issues held in the European Parliament. However, these corpora are mostly unannotated and no dedicated subcorpus is readily available.

Another important line of work related to multilingual resources for comparability analysis is **entity recognition**, which is currently being promoted by the Universal NER (UNER) project which especially focuses on named entities, usually introduced in text by proper nouns. This community-driven initiative provides gold-standard named entity recognition annotations in a wide variety of typologically and geographically different languages, adding a consistent NER layer to UD corpora (Mayhew et al., 2024). This provides a valuable benchmark for multilingual NER, especially for low-resource languages and for testing inter-annotator agreement and tag distribution across domains.

This general purpose initiative is especially inspiring for our project since a key challenge in **domain-specific NER**—and particularly in **environmental NER**— is the lack of annotated corpora and benchmarks dedicated to the environmental domain. As in most specific domains, entities other than persons, organizations, and locations are relevant in environmental discourse and they are realized by common nouns and noun phrases, some entities being nested inside larger ones, which are not covered by general-purpose NER tools.

To map domain entities, researchers usually turn to **domain ontologies**, which provide a common ground for naming classes and subclasses and the relations holding between them. In the environmental domain, GEMET (General Multilingual Environmental Thesaurus)[9] provides structured multilingual identifiers and relations between environmental concepts (European Environment Agency, 2024). GEMET has been used for tasks such as concept mapping and normalization of environmental terminology; however, its integration into NER pipelines is still experimental. Nevertheless, GEMET is a solid framework for multilingual annotation, which accounts for our decision to adopt it in our treebank annotation, as detailed in the following section.

3 Corpus Compilation and Annotation

To compile our parallel corpus, we downloaded the original version of *WWF's*[10] *2024 Living Planet Report*, available in multiple languages, including English,[11] Brazilian Portuguese,[12] Italian,[13] and Romanian,[14] the languages currently annotated in

[6] http://universaldependencies.org/conll17/

[7] https://opus.nlpl.eu/legacy/

[8] https://opus.nlpl.eu/legacy/Europarl.php

[9] https://www.eionet.europa.eu/gemet/en/about/

[10] The *World Wildlife Found for Nature* is the global leading organization for the nature conservation with offices in many countries.

[11] https://livingplanet.panda.org/en-US/

[12] https://livingplanet.panda.org/pt-BR/

[13] https://www.wwf.it/cosa-facciamo/pubblicazioni/living-planet-report/

[14] https://wwf.ro/campanii/raportul-planeta-vie-2024/

ALIGN	ENGLISH	BRAZILIAN PORTUGUESE
1-2	For example, in Colombia, a 2019 law requires that solar and wind projects transfer a percentage of their sales to communities within the project's area of influence while the Philippines' Renewable Energy Act requires that 80% of project royalties be directed toward subsidizing power costs in affected communities.	Por exemplo, na Colômbia, uma lei de 2019 exige que os projetos de energia solar e eólica transfiram um percentual de suas vendas para as comunidades dentro da "área de influência" do projeto. Já a Lei das Energias Renováveis das Filipinas exige que 80% dos royalties dos projetos sejam direcionados para subsidiar os custos de energia nas comunidades afetadas.
2-1	The most recently published plan includes areas of intact ecosystems for human livelihoods and well-being. Care was taken to ensure the expansion of protected areas would contribute to South Africa's development goals by providing important ecosystem services to people.	O plano recém-publicado inclui áreas de ecossistemas intactos para a subsistência e o bem-estar humanos e teve o cuidado de garantir que a expansão das áreas protegidas contribua para os objetivos de desenvolvimento da África do Sul, fornecendo importantes serviços ecossistêmicos às pessoas.

Table 1: Examples of one-to-many and many-to-one sentence alignment.

our parallel treebanks. Other languages for which the WWF report is available are expected to be included in our resource using and the same methodology in a future expansion of the TREEN project.

A comparison of the four versions of *WWF's 2024 Living Planet Report* showed that they all featured the same images and text content. For the purposes of treebank development, images and infographics were removed and plain text files were created. All bibliographic references and footnotes were also removed.

3.1 Alignment

In general, the development of aligned parallel resources is considered a very time-consuming task, mostly due to alignment issues. This poses a challenge when alignment is pursued at sentence or more fine-grained level, as meanings can be realized at different levels in different languages.

In TREEN, alignment was performed at sentence level. Sentences were extracted from each file and pasted in their sequential order onto a spreadsheet, a language pair (original and translated text) per tab. Sentence alignment was manually checked and when no one-to-one alignment was obtained, extra rows were added. This was the case when a single sentence was split into two or more sentences in the translated text or the opposite, i.e. two or more sentences in the original in English were translated as a single one in one of the other languages. Table 1 shows how these two cases were dealt with. The first line (1-2) shows a sentence in English that is split in two sentences in Brazilian Portuguese, while the second line (2-1) shows a couple of English sentences encompassed in a single sentence in Brazilian Portuguese. Often the diversity in segmentation is triggered by punctuation and how this is deployed in the translated text. For example, the two English sentences *"Nature narratives."* and *"Using indicators to understand change over different timescales"*, which are adjacent in the original text of the *WWF's 2024 Report*, are encompassed in the Italian *"Raccontare il declino della natura: utilizzo di indicatori per comprendere il cambiamento su scale temporali diverse."* since the translator interpreted the end of the first sentence as colons.

Sequential id numbers assigned to source sentences in English were used as an alignment reference. This means that in cases of no one-to-one sentence correspondence, the id of the English sentence is assigned to the corresponding sentences in the other languages. Each sequential id number is, moreover, preceded by three further digits. Following the same encoding strategy applied in PUD, we used the digits on the second and third position of each id to encode the original language of the sentence (e.g. *01* for English) and used the first digit to encode the genre of text. In our treebanks, the texts are extracted from a non-governmental report, as they were produced by an activist organization. Hence, the first digit *a* stands for activism. We are also planning a future extension of our resource to other types of texts, namely generated by institutions (governments and policy-makers) and social media users. *i* will be used for governmental institutions and *s* for social media texts.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Transforming	transform	VERB	_	_	0	root	_	_
2	the	the	DET	_	_	4	det	_	_
3	energy	energy	NOUN	_	_	4	compound	_	gemet_id=2712 group=ENERGY
4	system	system	NOUN	_	_	1	obj	_	_

Table 2: Example sentence in CoNLL-U format with ontology-based annotation.

3.2 Morphological and syntactic annotation

As already mentioned, there are only a few resources compiling texts on environmental issues and, fewer still, annotated for morphology and syntax. To provide a comprehensive resource for linguistic analysis and computational uses, we annotated our texts following the UD guidelines (Nivre et al., 2020), which is the *de facto standard* for treebanks. Table 2 shows an example from the `WWF24-Eng` in CoNLL-U format.

To parse our texts following UD guidelines, we used UDPipe parser (Straka, 2018) trained on the best models available for the four languages currently included in our dataset. For English, we used `english-gum-ud-2.15-241121`, for Italian we used `italian-isdt-ud-2.15-241121`, for Romanian we used `romanian-rrt-ud-2.15-241121` and for Brazilian Portuguese, `portuguese-petrogold-ud-2.15-241121`. A sample of 116 sentences of the output generated by UDPipe (corresponding to the preamble to the 2024 WWF Living Planet Report) was uploaded to the Arborator Grew (Guibon et al., 2020) tool and manually checked to have some hints about the correctness of the data and enable preliminary evaluation of their quality. While this validation is suitable for the purpose of the first release of TREEN, it must be extended to a larger set of data in the future development of the treebank.

The annotated sentences will be submitted to be part of UD repository.

3.3 Entity annotation

To annotate environmental content at the level of specific concepts, we designed a pipeline that combines NER with ontology-based matching. Rather than relying on general-purpose resources, our approach is centered entirely on the GEMET ontology, as it provides unique identifiers and thematic groupings for environmental concepts. Environmental discourse involves a wide range of entities that go beyond traditional Named Entities such as *places* or *organizations*. In this domain, the most relevant elements are often abstract or thematic concepts — such as *biodiversity*, *climate change*, or *ecosystem resilience* — which are central to how environmental issues are discussed and framed.

A key insight of our method is the strong connection between syntactic structure and an ontology-based annotation. The UD framework provides detailed syntactic information that allows us to extract meaningful expressions in a principled way. By focusing on noun phrases and their modifiers — as determined by dependency relations like `compound`, `amod`, and `nmod` — we are able to detect linguistically grounded candidate entities.

This syntactic scaffolding is essential: it ensures that entities are not just found, but are consistently defined across different languages and sentence structures. Once identified, these expressions are matched to GEMET concepts through exact or partial alignment, adding an ontology-based layer to the treebank. This enriched structure opens the door to a wide range of multilingual applications, including cross-lingual entity propagation and concept-level analysis (see Figure 1).

Entity Extraction. As a first step, we identified potential entities in each sentence by extracting *noun phrases* (NPs). This was done by combining part-of-speech tagging with syntactic dependency information from the CoNLL-U files. We took into account both expressions composed of a nominal head and some modifiers (not properly multi-word expressions in the formal sense, such as "climate change") and *single nouns or proper nouns* (like "biodiversity"). To extract meaningful phrases, we used a simple set of heuristics that detect noun heads along with their modifiers (e.g., `compound`, `amod`, `nmod` dependencies), while discarding generic or weak elements such as quanti-

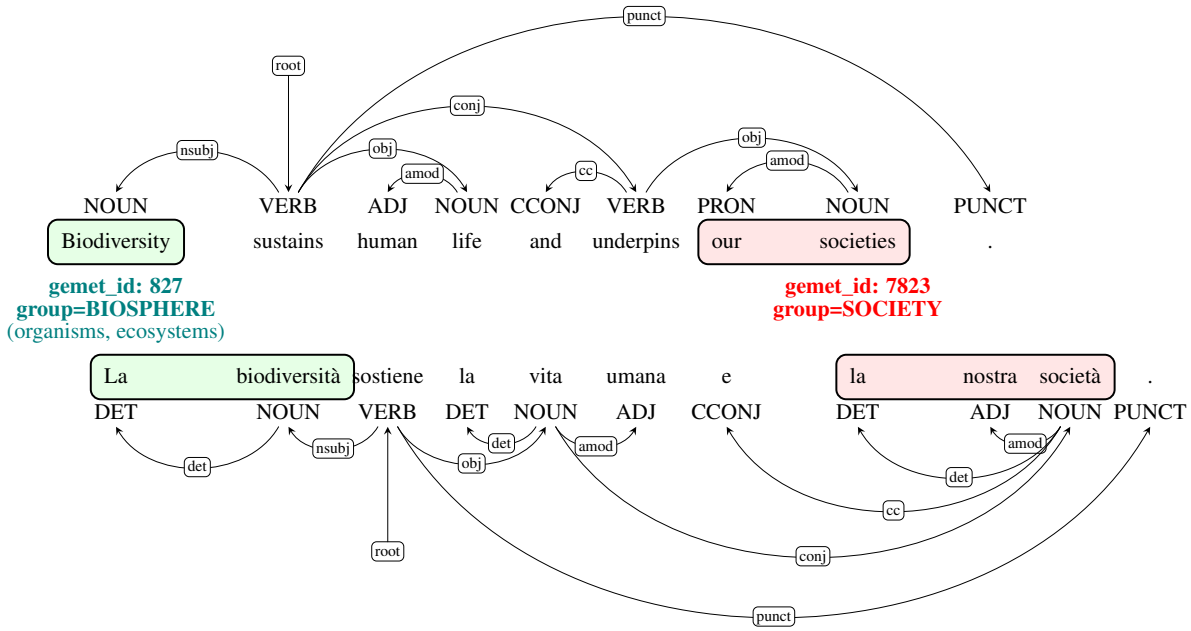

Figure 1: Example of dependency-based annotation of two parallel sentences, in English and Italian, enriched with GEMET ontology labels in the CoNLL-U format.

fiers (e.g., many, several) and evaluative adjectives (e.g., important, good).

Ontology Matching in English. Each candidate noun phrase was first normalized and then compared against the set of preferred labels provided by the GEMET ontology, using exact string matching. When a match was found, we associated the corresponding `concept ID` and its `group name` to the phrase. These annotations were added to the `MISC` column in the CoNLL-U file, like in Table 2 (other examples are in Table 4, 5 and 6 in the Appendix). In the case of multiword expressions (e.g., *"climate change"*), the annotation is consistently propagated across all tokens that make up the GEMET entity, ensuring that the concept is clearly represented throughout the span (see e.g. *"greenhouse gas"* in Table 6).

Multilingual Propagation. After annotating the English sentences, we transferred the entity annotations to the aligned sentences in the other languages in our dataset (Italian, Brazilian Portuguese, and Romanian). This projection process relies on sentence-level alignment between English and the target language. For each GEMET concept identified in the English sentence, we verified whether a corresponding noun or noun phrase appeared in the aligned sentence. We relied on two strategies: (i) direct translation of the English label into the target language using the `Google Translate API`[15], and (ii) lexical overlap, where we checked whether the translated term (or its components) matched any of the extracted noun phrases or their tokens in the target sentence. This allowed us to transfer GEMET annotations across languages even when the syntactic realization differs slightly. When a match was found, we assigned the same GEMET concept ID to the relevant tokens in the target language, ensuring consistent cross-lingual annotation.

These experiments were implemented in **Python**,[16] using the `spaCy`[17] and `pandas`[18] libraries for linguistic processing and data manipulation. The GEMET ontology was used as the sole reference for concept matching and a semantics oriented labelling.

An example of the annotation of two corresponding sentences, respectively from WWF-ENG and WWF-ITA, is shown in Figure 1 (for the CoNLL-U of these sentences see Table 4 and 5 in the Appendix). The two sentences were parsed following the UD framework and then annotated for the entities. The upper and lower parts of the figure show

[15] https://cloud.google.com/translate
[16] https://www.python.org/
[17] https://spacy.io/
[18] https://pandas.pydata.org/

the sentence in English and in Italian respectively, both associated with their parts of speech. The entities extracted according to GEMET can be seen as a semantic bridge between the syntactic structures. In the example, "*biodiversity*", "*life*", and "*societies*" were matched against the GEMET ontology. "*biodiversity*" was linked to concept ID `827` in the `BIOSPHERE` group, while "*society*" was matched to concept ID `7823` in the `SOCIETY` group. On the other hand, not all terms and meaningful expressions extracted as a candidate noun phrase have a direct match in the ontology: for example, "*human life*" did not match any GEMET concept.

Once there is matching, information is added to the `MISC` column of the CoNLL-U file and can be visualized in the syntactic tree. This enriched representation brings together syntactic structure and ontological information, creating a solid foundation for further tasks like recognizing environmental entities or transferring annotations across languages. Table 2 shows how the annotation of an entity is integrated in the CoNLL-U format.

4 Results

We computed the number of sentences, words, and lemmas in the WWF24 parallel corpora using a custom `Python` pipeline built on `pandas` and `spaCy` language-specific models (`{en,pt,it,ro}_core_web_sm`). Table 3 presents the resulting distributions.

As previously mentioned, some of the sentences in the original text in English were split in the translated texts, which accounts for the higher number of sentences in Italian and much higher in Brazilian Portuguese when compared with the source text in English. Regarding number of tokens, the translated texts exhibit a higher number, also accountable for by lexical rendition of some English terms and typological differences between the languages, such as the strong reliance of Romance languages on determiners and prepositions to construe noun phrases. This can be seen, for instance, in terms such as *"greenhouse gas emissions"* (3 tokens) translated into Italian as *"le emissioni di gas serra"* (5 tokens), Brazilian Portuguese as *"as emissões de gases de efeito estufa"* (7 tokens) and Romanian as *"emisiile de gaze cu efect de seră"* (7 tokens). The higher number of unique words and lemmas in the translated texts is also impacted by typological differences as well as by translators' choices in rendering the same word and lemma in English by different synonyms of the equivalent word in their languages. With regard to average sentence length, results point to relatively long sentences, which can be explored as a variable contributing to the overall complexity of this kind of discourse.

4.1 Annotated samples

We analyzed the human-curated annotated sample of 116 sentences from the original text in English and their translation into Romanian, Italian and Brazilian Portuguese (for the full distribution of POS and dependency relations see Table 7 and Table 8 in the Appendix). Both `POS` and `deprel` tags reflect analogous and different annotation guidelines in the four languages. Regarding POS tags, for instance, a source of variation is the annotation of geographical names and institutions as proper nouns (`PROPN`) in some of the languages, the relation *flat* holding between them, and as common nouns (`NOUN`) in others with *nmod* relations holding between them. This result shows different guidelines for English and the Romance languages. With regard to analogous guidelines, differences in the count of `DET`, for example, show typological differences, with English using relatively few determiners when compared to Italian and Brazilian Portuguese, and Romanian expressing determinateness through different morphological markers.

Regarding deprels, most noun phrases in English are annotated with the deprel *compound* while the Romance languages annotate them with the *nmod* tag. Finally, it is worth noting that the high number of *conj* relations in the four languages is an indicator of more complex syntax, also correlated to the average sentence length mentioned above.

4.2 Annotation challenges

Manual check of the parsed output revealed some of the main challenges in automatic annotation. These have to do with nested noun phrases and coordinated phrases and clauses. Our results are in line with those of the studies about the dissemination and communication about environmental matters. The recent literature has indeed problematized the effectiveness of such discourse due to the complexity in the content of this kind of texts (see e.g. (Bosco et al., 2023)). By providing a fine-grained analysis of the language used to communicate environmental issues, computational linguistics can provide crucial information on how individuals, groups of people or entire societies are coping with environmental issues.

Language	sentences	tokens	unique words	lemmas	average sentence length
English	1,042	23,462	3,163	2,346	22,5
Brazilian-Portuguese	1,063	26,471	3,790	2,598	24,9
Italian	1,048	26,976	3,988	2,783	25,7
Romanian	1,042	25,830	4,699	2,968	24,8

Table 3: Basic statistics per language.

As far as more precise syntactic phenomena, nested noun phrases are very common in scientific domains and consist of a noun phrase nested into another, sometimes having a coordinated relation as well. An example retrieved from our corpus is *"a combination of pine bark beetle infestation and more frequent and ferocious forest fires"*. In this example (as shown in Figure 2 in the Appendix), there is coordination between *"infestation"* and *"fires"*, which are both heads of noun phrases. The noun *"infestation"* has three nouns as pre-modifiers, which need to be annotated as a sequence of *compound* relations, where *"infestation"* stands in a *compound* relation with *"beetle"*, which, in turn, is in a *compound* relation with *"bark"*, in turn, in a *compound* relation with *"pine"*. In the case of Portuguese, Italian, and Romanian texts, as seen in Figure 3, 4, and 5, noun modification is established through the dependency relation *nmod*. Interestingly, all four language models used in UDpipe failed to assign the correct *conj* relation between *"infestation"* and *"fires"* and had to be manually edited.

Another challenging case for parsers is discriminating between *obl* and *nmod* relations. For example, in the English sentence *"In the biosphere, the mass die-off of coral reefs would destroy fisheries and storm protection for hundreds of millions of people living on the coasts"*, the UDpipe parser assigned an *obl* relation between *"destroy"* and *"for hundreds of millions of people [...]"*, instead of an *nmod* relation between *"protection"* and *"for hundreds of people"*, which is semantically more plausible in this context.

5 Conclusion and Future Work

This paper presented the first steps in the development of a novel set of aligned treebanks within the Multilingual Treebank Project TREEN, which compiles texts on environmental discourse produced by different stakeholders (institutions, citizens and activists) in different communication contexts. The data included in the treebanks were extracted from the *WWF's 2024 Living Planet Report* in English, Italian, Romanian and Brazilian Portuguese. They were annotated for morphology and syntax in UD format and enriched with the annotation of the entities involved in the discourse according to the GEMET ontology for environmental topics. This methodology is designed to apply to other texts in the ongoing expansion of TREEN also considering e.g. data alignment at fine-grained levels and an ontology matching that takes into account form variation (such as synonyms or paraphrases).

These treebanks and the future development of the project are intended to fill a gap regarding resources dedicated to the environmental discourse in order to offer the community the possibility of conducting a more in-depth analysis of environmental discourse. This is expected to open new perspectives to more reliably study the different narratives present in different languages, mirroring different cultures and perceptions in the discourse about environmental issues. The availability of morphological and syntactic annotation can be very helpful for detecting different narratives, e.g. allowing the identification of different tenses and forms of verbs as indicators of different ways of narrating events (realis vs. irealis). Complementarily, annotating named entities and linking the syntactic structures in the different languages to a common ontological framework such as GEMET helps us in the semantics oriented characterization of the discourse about the environment in different cultures.

Last but not least, we expect to use morphological and syntactic annotation and a deeper analysis of the annotated data according to several different perspectives to explore the level of intricacy and complexity of environmental discourse produced by the different stakeholders, which can yield insightful and actionable information to develop more inclusive forms of environmental discourse.

Limitations

The main aim of this study is to introduce the first release of a novel resource. This release comprises 116 sentences, with human-curated alignment and annotation, the remaining sentences being currently under human validation. This resource is part of a wider project for the study of the discourse about environmental matters. The texts collected in the component of TREEN here reported are drawn from a single source: one non-governmental association (WWF) and in four languages only. We are working on the expansion of the parallel treebanks to include more languages and to incorporate other stakeholders.

While the proposed methodology yields promising results, several aspects of the pipeline can be improved in future work.

In particular, as far as the matching between cited concepts and entities in GEMET is concerned, this is based on exact string matching, that is, the ontology matching relies on exact string comparison, which can miss variations, synonyms or paraphrases.

Also, the coverage of the ontology is partial: although GEMET is a rich resource, it does not cover all relevant environmental expressions, especially newer or more specific domain terms.

The translation strategy applied for the multilingual propagation of entity annotation may also produce inaccurate or overly generic results, affecting the mapping accuracy. The quality of the automatic translation may be improved using tools other than Google Translate.

Additionally, our entity extraction procedure is restricted to noun phrases. As a result, relevant information conveyed through verbs, adjectives, or multi-clause constructions may be overlooked. Incorporating broader syntactic patterns or using semantic role labeling could help address this limitation.

Acknowledgments

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

References

Lars Ahrenberg. 2015. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 10–19, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Sabine Bartsch, Changxu Duan, Sherry Tan, Elena Volkanovska, and Wolfgang Stille. 2023. The insightsnet climate change corpus (iccc). In *BTW 2023*, pages 887–900. Gesellschaft für Informatik e.V., Bonn.

Valerio Basile, Cristina Bosco, Francesca Grasso, Muhammad Okky Ibrohim, Maria Skeppstedt, and Manfred Stede, editors. 2025. *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*. University of Tartu Library, Tallinn, Estonia.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2020. You are right. i am alarmed–but by climate change counter movement. *arXiv preprint arXiv:2004.14907*.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021. Automatic classification of neutralization techniques in the narrative of climate change scepticism. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2167–2175, Online. Association for Computational Linguistics.

Cristina Bosco, Muhammad Okky Ibrohim, Valerio Basile, and Indra Budi. 2023. How green is sentiment analysis? environmental topics in corpora at the university of turin. In *Proceeding of CLiC-it 2023*, Venice, Italy.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9:1–20.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. 2022. Saving earth one tweet at a time through the lens of artificial intelligence. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Dimitrios Effrosynidis, Alexandros I Karasakalidis, Georgios Sylaios, and Avi Arampatzis. 2022. The

climate change twitter dataset. *Expert Systems with Applications*, 204:117541.

European Environment Agency. 2024. Gemet - general multilingual environmental thesaurus. https://www.eionet.europa.eu/gemet/en/about/. Accessed: 2025-04-11.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Michael AK Halliday. 1992. New ways of meaning: The challenge to applied linguistics. *Thirty years of linguistic evolution*, pages 59–95.

Muhammad Okky Ibrohim, Cristina Bosco, and Valerio Basile. 2023. Sentiment analysis for the natural environment: A systematic review. *ACM Computing Surveys*, 56(4).

Elizabeth Kolbert. 2024. *H is for Hope: Climate Change from A to Z.* Oneworld Publications.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, and other UD contributors. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4034–4043. European Language Resources Association (ELRA).

Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, Jens P Linge, and 1 others. 2022. Exploring data augmentation for classification of climate change denial: Preliminary study.

Manuela Sanguinetti and Cristina Bosco. 2014. PartTUT: The Turin University Parallel Treebank. In Basili, Bosco, Delmonte, Moschitti, and Simi, editors, *Harmonization and development of resources and tools for Italian Natural Language Processing within the PARLI project*. Springer Verlag.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

Xu Song, Kesumawati A Bakar, Azlan Abas, and Wan Fatimah Solihah Wan Abdul Halim. 2025. A systematic literature review on ecological discourse analysis (2014–2023). *Journal of World Languages*, (0).

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Milan Straka. 2018. 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*, pages 197–207.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, and 7 others. 2024. ClimateGPT: towards AI synthetizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. Climatext: A dataset for climate change topic detection. *arXiv preprint arXiv:2012.00483*.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A pretrained language model for climate-related text. *ArXiv*, abs/2110.12010.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava

Hlavacova, Václava Kettnerová, Zdenka Uresova, and 43 others. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Biodiversity	biodiversity	NOUN	_	_	2	nsubj	_	gemet_id: 827 group=BIOSPHERE (organisms, ecosystems)
2	sustains	sustain	VERB	_	_	0	root	_	_
3	human	human	ADJ	_	_	4	amod	_	_
4	life	life	NOUN	_	_	2	obj	_	
5	and	and	CCONJ	_	_	6	cc	_	_
6	underpins	underpin	VERB	_	_	2	conj	_	_
7	our	our	ADJ	_	_	8	nmod:poss	_	_
8	societies	society	NOUN	_	_	6	obj	_	gemet_id: 7823 group=SOCIETY
9	.	.	PUNCT	_	_	2	punct	_	_

Table 4: Annotation of GEMET entities in the English sentence in Figure 1.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	La	il	DET	_	_	2	det	_	_
2	biodiversità	biodiversità	NOUN	_	_	3	nsubj	_	gemet_id: 827 group=BIOSPHERE (organisms, ecosystems)
3	sostiene	sostenere	VERB	_	_	0	root	_	_
4	la	il	DET	_	_	5	det	_	_
5	vita	vita	NOUN	_	_	3	obj	_	_
6	umana	umano	ADJ	_	_	5	amod	_	
7	e	e	CCONJ	_	_	10	cc	_	_
8	la	il	DET	_	_	10	det	_	_
9	nostra	nostro	DET	_	_	10	det:poss	_	_
10	società	società	NOUN	_	_	5	conj	_	gemet_id: 7823 group=SOCIETY
11	.	.	PUNCT	_	_	3	punct	_	_

Table 5: Annotation of propagated GEMET entities in the Italian sentence in Figure 1.

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Nature	nature	NOUN	_	_	3	compound	_	_
2	-	-	PUNCT	_	_	1	punct	_	_
3	based	base	VERB	_	_	4	amod	_	_
4	solutions	solution	NOUN	_	_	8	nsubj	_	_
5	for	for	ADP	_	_	7	case	_	_
6	climate	climate	NOUN	_	_	7	compound	_	gemet_id=1462 group=ATMOSPHERE (air, climate)
7	mitigation	mitigation	NOUN	_	_	4	nmod	_	_
8	have	have	VERB	_	_	0	root	_	_
9	the	the	DET	_	_	10	det	_	_
10	potential	potential	NOUN	_	_	8	obj	_	_
11	to	to	PART	_	_	12	mark	_	_
12	reduce	reduce	VERB	_	_	10	acl	_	_
13	annual	annual	ADJ	_	_	16	amod	_	_
14	greenhouse	greenhouse	NOUN	_	_	15	compound	_	gemet_id=3763 group=WASTES, POLLUTANTS, POLLUTION
15	gas	gas	NOUN	_	_	16	compound	_	gemet_id=3763 group=WASTES, POLLUTANTS, POLLUTION
16	emissions	emission	NOUN	_	_	12	obj	_	gemet_id=2663 group=WASTES, POLLUTANTS, POLLUTION
17	by	by	ADP	_	_	21	case	_	_
18	10	10	NUM	_	_	21	nummod	_	_
19	-	-	SYM	_	_	20	case	_	_
20	19	19	NUM	_	_	18	nmod	_	_
21	%	%	SYM	_	_	12	obl	_	_
22	,	,	PUNCT	_	_	25	punct	_	_
23	while	while	SCONJ	_	_	25	mark	_	_
24	also	also	ADV	_	_	25	advmod	_	_
25	benefiting	benefit	VERB	_	_	12	advcl	_	_
26	ecosystems	ecosystem	NOUN	_	_	25	obj	_	gemet_id=2519 group=BIOSPHERE (organisms, ecosystems)
27	and	and	CCONJ	_	_	28	cc	_	_
28	improving	improve	VERB	_	_	25	conj	_	_
29	livelihoods	livelihood	NOUN	_	_	28	obj	_	_
30	.	.	PUNCT	_	_	8	punct	_	_

Table 6: Annotation of propagated GEMET entities made up by more than one word.

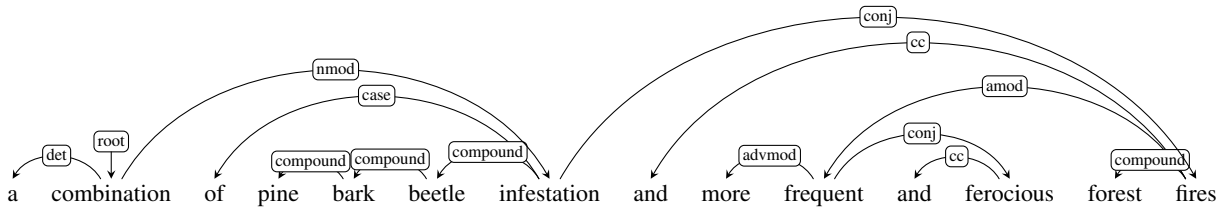

Figure 2: Dependency relations for the English noun phrase *"a combination of pine bark beetle infestation and more frequent and ferocious forest fires"*.

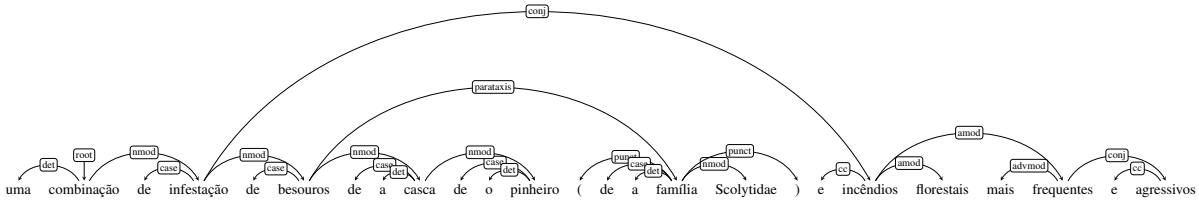

Figure 3: Dependency relations for the Brazilian Portuguese noun phrase *"uma combinação de infestação de besouros de a casca de o pinheiro (de a família Scolytidae) e incêndios florestais mais frequentes e agressivos"*.

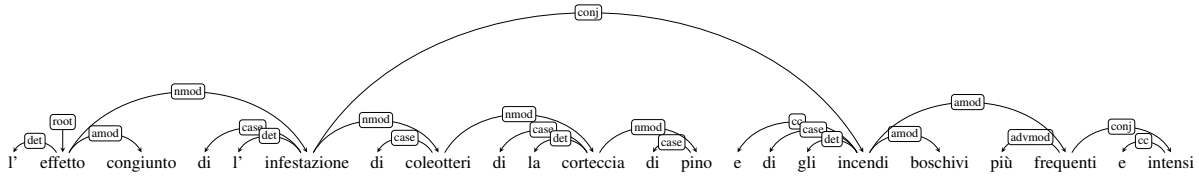

Figure 4: Dependency relations for the Italian noun phrase *"l'effetto congiunto di l'infestazione di coleotteri di la corteccia di pino e di gli incendi boschivi più frequenti e intensi"*.

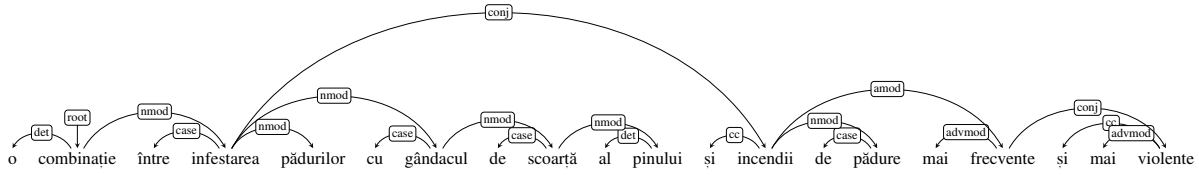

Figure 5: Dependency relations for the Romanian noun phrase *"o combinație între infestarea pădurilor cu gândacul de scoarță al pinului și incendii de pădure mai frecvente și mai violente"*.

POS tag	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
ADJ	270	366	399	325
ADP	280	489	553	564
ADV	109	117	117	134
AUX	127	97	104	80
CCONJ	160	154	152	163
DET	219	175	658	572
NOUN	746	896	795	763
NUM	101	93	83	98
PART	55	78	0	0
PRON	86	94	82	74
PROPN	80	40	62	75
PUNCT	353	323	347	327
SCONJ	43	23	32	30
SYM	48	0	29	45
VERB	349	294	294	317
X	0	6	11	1

Table 7: Distribution of POS tags across the four treebanks.

DepRel	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
acl	27	66	33	42
acl:relcl	23	0	29	27
advcl	57	44	47	59
advcl:relcl	4	0	0	0
advmod	115	140	109	122
amod	271	307	338	262
appos	10	16	14	10
aux	65	43	55	17
aux:pass	22	21	14	20
case	287	421	519	494
cc	159	159	152	168
cc:preconj	1	0	0	0
ccomp	6	30	17	14
ccomp:pmod	0	1	0	0
compound	180	3	21	0
compound:prt	3	0	0	0

Continued on next page

Table 8 – continued from previous page

DepRel	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
conj	206	217	217	221
cop	38	33	35	42
csubj	7	11	11	9
csubj:pass	0	1	0	0
dep	2	0	0	0
det	213	169	637	558
det:poss	0	0	17	0
det:predet	1	0	4	0
discourse	0	0	0	2
expl	1	4	16	0
expl:impers	0	0	5	0
expl:pass	0	2	1	3
expl:poss	0	1	0	0
expl:pv	0	15	0	7
fixed	2	96	9	45
flat	27	14	0	14
flat:foreign	0	0	5	0
flat:name	0	0	14	43
goeswith	0	0	0	1
iobj	0	15	4	2
mark	86	109	66	69
nmod	149	367	285	278
nmod:poss	22	0	0	0
nsubj	144	134	123	132
nsubj:pass	24	11	12	15
nummod	55	61	60	63
obj	167	123	175	155
obl	122	127	179	155
obl:agent	8	11	10	7
obl:arg	0	0	0	15
obl:pmod	0	12	0	0
obl:unmarked	4	0	0	0
parataxis	14	10	5	13
punct	352	323	347	325
root	116	116	121	117

Continued on next page

Table 8 – continued from previous page

DepRel	English.conllu	Romanian.conllu	Italian.conllu	Br_Portuguese.conllu
xcomp	34	12	12	38

Table 8: Distribution of dependency relations across treebanks.