Communicating urgency to prevent environmental damage: insights from a linguistic analysis of the WWF24 multilingual corpus

Cristina Bosco
Dipartimento di Informatica
Università di Torino
Torino (Italy)
cristina.bosco@unito.it

Adriana Silvina Pagano
Università di Torino /
Universidade Federal de Minas Gerais
Belo Horizonte (Brazil)
apagano@ufmg.br

Elisa Chierchiello
Dipartimento di Informatica
Università di Torino
Torino (Italy)
elisa.chierchiello@unito.it

Abstract

Contemporary environmental discourse focuses on effectively communicating ecological vulnerability to raise public awareness and encourage positive actions. Hence there is a need for studies to support accurate and adequate discourse production, both by humans and computers. Two main challenges need to be tackled. On the one hand, the language used to communicate about environment issues can be very complex for human and automatic analysis, there being few resources to train and test NLP tools. On the other hand, in the current international scenario, most texts are written in multiple languages or translated from a major to minor language, resulting in different meanings in different languages and cultural contexts.

This paper presents a novel parallel corpus comprising the text of World Wide Fund (WWF) 2024 Annual Report in English and its translations into Italian and Brazilian Portuguese, and analyses their linguistic features.

1 Introduction

Environmental issues, such as biodiversity loss, global climate and sustainability, have an important social relevance today and are increasingly debated in all countries, in a variety of communication channels and media.

Nevertheless, notwithstanding the great amount of dissemination and communication about environmental matters, the recent literature has problematized the effectiveness of such discourse due to the complexity in the content of this kind of texts. For instance, Italian dissemination texts discussing issues related to the environment, published by the European Agency and journals, have been evaluated as posing readability challenges for people who do not have at least a high school degree (Bosco et al., 2023).

Moreover, in the current international scenario in which disseminating clear and homogeneous messages about the environmental crisis is crucial, texts are often written in multiple languages or translated from a major (in most of cases English) to minor languages, resulting in the construal of different meanings in different countries.

Governmental entities have also detected this kind of difficulty and are trying to address it. For example, in 2020 the European Commission published a study showing that more than half of the environmental claims examined in the European countries were vague, misleading or unfounded, while 40% were completely unfounded. To promote more accurate and timely information, in March 2023, the European Parliament published the *Green Claims Directive*[1].

In countries especially involved in the environmental crisis, such as Brazil, a large variety of educational and informative initiatives are promoted to foster the correct dissemination of information (see, among other, *Climate Change and Public Perception in Brazil*[2], a project for measuring knowledge and concern of Brazilians on climate change and the yearly forest fires in the country.

The studies and political interventions above reported clearly point to the importance of language in raising public awareness to save the planet. In addition to the life sciences, disciplines such as computational linguistics are also expected to address the challenges posed by environmental discourse from the perspective of their theoretical and methodological approaches and applying different strategies that can impact on discourse and, ultimately, on societal moves towards sustainability. By providing a fine-grained analysis of the

[1] https://environment.ec.europa.eu/topics/circular-economy/green-claims_en

[2] https://en.percepcaoclimatica.com.br/

Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025), pages 34–43

language used to communicate environmental issues, computational linguistics can indeed collaborate with crucial information on how individuals, groups of people or entire societies are coping with environmental issues, their attitudes, awareness and willingness to work towards more sustainable life patterns for the planet.

Nonetheless, machines also are facing challenges to analyse texts on environmental discourse, first of all because they use a specialized lexicon even when written with a popularization intent. This motivates the development of pretrained models (Thulke et al., 2024) dedicated to address texts about environmental issues which include specific expressions that general language LMs can not represent accurately (Webersinke et al., 2022) and different languages. But there are other challenges when dealing with environmental texts. For example, the frequent inclusion of infographics, images and tables that integrate the text content can make automatic analysis particularly difficult (Saha et al., 2024; Mishra et al., 2024).
A further and also more serious challenge is the scarcity of available datasets and corpora to train tools for automatic analysis and evaluate their performance on texts about environmental issues. For example, as reported in (Ibrohim et al., 2023), resources for Sentiment Analysis of environmental texts are currently very poor for all languages. Some corpora about environmental topics are available for English and very few for other languages. In general, it is striking that the development of tools and resources to tackle the problem, including by computational linguistics, has not yet involved languages spoken in countries which have long been among key stakeholders in environmental phenomena, as is Brazil, for example.
On top of that, only topics related to climate change seem to have attracted most of the research community interest (probably due to the higher visibility and quantifiability of its effects), while others are less studied in the context of computational linguistics. This is the case of biodiversity loss, one of the most debated environmental topics at present.

In this paper, we introduce a novel multilingual parallel corpus WWF24 comprising texts about biodiversity loss and other environmental issues published within a report in 2024 by the World Wildlife Fund (WWF)[3], and we provide the results of the language analysis we applied to it. In line with the literature, the results reveal differences in the way meanings are construed in each language. They may have significant impact on the ultimate effectiveness in the way meanings are construed to communicate ecological vulnerability and raise public awareness and encourage positive actions accordingly. We have included three languages in the preliminary release of our corpus, namely English, Brazilian Portuguese and Italian, also taking into account issues related to translation and different renditions for the key concepts.
The research questions we want to investigate are therefore as follows:

- Are meanings pertaining to environmental issues construed analogously in different languages?
- What do language patterns show about how the notion of biodiversity and nature is construed in Italian, Portuguese and English?
- What impact does translation have on meaning construal in the Italian and Brazilian texts?

Our main contribution includes the preliminary release of a novel multilingual parallel corpus WWF24 of environmental discourse[4] and the reporting of results of a set of analyses.
Our main goal is to provide data that support the development of LMs especially dedicated to the topics related to biodiversity. Drawing on the premise that our language construes our experience of the world (Halliday and Matthiessen, 2006) examining the language used in environmental discourse allows us, not only to gather insights into how different languages construe different meanings, but also to better characterize the complexity of environmental discourse.

The paper is organized as follows. The next section briefly surveys related work. Section 3 introduces the WWF24 corpus providing the details about the texts compiled. Section 4 and 5 are devoted to describing the analysis performed on the corpus and the results obtained. In Section 6 we

[3]The WWF is the leading wordlwide organization in wildlife conservation and endangered species (`https://www.worldwildlife.org/`).

[4]The corpus will be made publicly available and downloadable upon paper acceptance

provide a discussion of the results of our analysis, conclusions and envisioned steps for further work.

2 Related Work

NLP approaches to the analysis of environmental discourse have mainly focused on Sentiment Analysis.

A systematic survey of the application of Sentiment Analysis on environment related topics is presented in (Ibrohim et al., 2023). The paper shows that few projects on the subject have been carried out, even for major languages such as English, and these have used fairly rough techniques. In (Stede and Patz, 2021), a review is conducted to explore the application of Sentiment Analysis in the debate about climate change. The authors show how different communities (general public, policy-makers and scientists) use different genres, registers, and terminologies to communicate with each other and with other communities about this issue, pointing to the potential of NLP to assist them to assess in which direction the debate may evolve and to respond accordingly.

(Du et al., 2020) explore the use of Sentiment Analysis for examining opinions on several smart city issues like climate change, urban policy, energy, and traffic drawing on social media texts.

(Stede and Patz, 2021) review approaches to the analysis of climate change discourse both from the perspective of NLP studies and social science studies, arguing for potential enhancement of studies if both fields take each other's perspectives into account. Both these surveys, (Stede and Patz, 2021) and (Du et al., 2020) are enlightening in the results reported but do not provide an in-depth exploration of NLP techniques needed to apply Sentiment Analysis on natural environment topics and are restricted to a specific topic, not considering many other topics related to nature and environmental issues.

In the last year, the application of NLP to environmental issues has been more reported on in scientific events and, in particular, in workshops such as ClimateNLP 2024[5] or ICLR 2024 - Tackling Climate Change with Machine Learning[6]. Both events will be held again in 2025.

The development of language models dedicated to climate change topics, such as ClimateBERT (Webersinke et al., 2022) and the above mentioned events underline how climate change is currently raising a specific attention within computational linguistics.

As far as corpora to be used in NLP experiments, a corpus for Italian has been recently published and is described in (Grasso et al., 2024a), while efforts made for building a multilingual corpus including Italian, Indonesian and English are described in (Bosco et al., 2023). Among the most recent corpora for English we can cite (Grasso et al., 2024b), which is focused on issues related to climate change observed from a diachronic perspective.

One of the aims of our contribution is expanding the variety of environment-related topics and number of languages, in this particular paper including Brazilian Portuguese, which is currently under represented.

3 Data

This section describes the corpus we collected for the purpose of analysing linguistic features of environmental discourse.

To compile our corpus WWF24, we downloaded the original version of *WWF's 2024 Living Planet Report*, available in English at `https://livingplanet.panda.org/en-US/`. The Brazilian version was downloaded from `https://livingplanet.panda.org/pt-BR/` and the Italian version from `https://www.wwf.it/cosa-facciamo/pubblicazioni/living-planet-report/`. Henceforth the three versions will be referred as WWF24-Eng, WWF24-Ita and WWF24-Bra. They can be considered as a starting point for a resource in which more languages will be included and the same methodology applied.

A comparison of the three versions of *WWF's 2024 Living Planet Report* showed that they all featured the same images and text content. In both the Italian and Brazilian versions, the names of the translators are acknowledged. However, unlike the Brazilian version, in which all infographics are translated, most infographics in the Italian version are partially translated or not translated at all. For the purposes of language analysis, images and infographics were removed and plain text files were created as part of the novel corpus WWF24.

As *WWF's 2024 Living Planet Report* is a technical report, drawing on data and domain termi-

[5] `https://nlp4climate.github.io/climatenlp2024/`

[6] `https://www.climatechange.ai/events/iclr2024`

nology used by WWF in accordance to other major stakeholders and at the same time aiming at a wider readership, we expected the use of technical terms as well as lay explanations, which knowledgeably poses a challenge in translations. An example of one such challenge is the terminology in*WWF's 2024 Living Planet Report*, for instance "tipping point" and "nature-positive" (see section 4), which are used as technical terms in the environmental domain and which need to be clarified to a lay audience.

Our analysis thus began by exploring general characteristics of the texts and finally focused on specific terms. To query the corpus we relied on text analysis software: Sketch Engine[7] and Voyant Tools[8]. For a fine-grained analysis of two particular terms - 'tipping point' and 'nature-positive'- we performed manual alignment of the sentences in which the terms occurred.

The corpus compiled is deemed a valuable contribution for an area in which there are only very few datasets. Moreover, the first release of our corpus compiles data for Italian, Brazilian Portuguese and English, but and expansion is scheduled to include in WWF24 texts from other language families, as is the case of Turkish. Some considerations related to the debate about the environment and languages in general motivated our initial choice of three languages.

English is the language in which the WWF's reports are originally written, assumed as a medium of communication and the most spoken language around the world.

Portuguese (including European and Brazilian Portuguese) is among the 8 most spoken languages with 264 million speakers, and most of them being in Brazil[9]. Brazil is one of the countries most impacted by environmental phenomena and one of the main stakeholders in environmental discussions with the greatest biodiversity of flora and fauna on the planet. Nevertheless, based on our knowledge, we are not aware of corpora available for this language about the discussion of biodiversity loss.

Finally, **Italian** has been selected because of our expertise and because it underlines a culture that is enough different from those represented by the other two languages.

[7] https://www.sketchengine.eu/

[8] https://voyant-tools.org/

[9] The country for which the report is published by the WWF is Brazil and the language is Brazilian Portuguese.

Linguistic analysis is expected to pave the way for the development of annotation schemes that can be later applied on the data for building, not only LMs, but also benchmarks that are crucial for the evaluation of results provided by LMs.

4 Linguistic Analysis

We performed different techniques aimed at extracting the most important lexical and semantic features in a cross language perspective. We computed the number of sentences, words and lemmas in the WWF24-Eng corpus using the text analysis software Sketch Engine. Table 1 shows distributions in our corpus.

	Eng	**Ita**	**Bra**
tokens	29,996	33,437	32,914
words	25,359	29,087	28,467
different words	4,344	4,977	4,866
lemmas	3,036	3,479	3,354

Table 1: Distribution of tokens, words, different words and lemmas in the three subcorpora of the WWF24 corpus .

As can be seen in Table 1, figures are higher in the Italian and Brazilian Portuguese texts than in the original in English. This may be accounted for by typological features of the languages. Italian and Brazilian Portuguese, for instance, make use of prepositional phrases to realize many noun phrases in English, which adds to the number of tokens. Also, Italian and Brazilian Portuguese make use of more lemmas to realize a single lemma in English. For example, 'to see' is rendered by several lemmas, among them 'vedere' and 'osservare' in Italian and 'ver' e 'observar' in Brazilian Portuguese.

In order to explore lexical characteristics, using Sketch Engine we extracted the most frequent lemmas for nouns in each subcorpus. Table 2 shows similarities and differences between the three languages regarding the 10 **most frequent nouns**. For example, 'nature'/'natura'/'natureza' and 'biodiversity'/'biodiversità'/'biodiversidade' rank among the first most frequent nouns in all three languages. However, 'species' does not rank amid the first ten positions in English (it can be found at the 16th position with 82 occurrences), while unlike in English, 'planet' does not rank in Italian and Brazilian Portuguese (it is in the 70th position with 25 occurrences in Italian and

English	Italian	Brazilian
nature (196)	natura (224)	natureza (226)
climate (180)	cambiamento (165)	mudança (158)
change (150)	specie (124)	espécie (127)
food (138)	sistema (117)	ecossistema (116)
ecosystem (126)	popolazione (116)	área (115)
energy (124)	ecosistema (113)	sistema (113)
system (109)	obiettivo (107)	água (102)
population (105)	biodiversità (99)	população (100)
biodiversity (103)	acqua (94)	biodiversidade (99)
planet (103)	persona (91)	energia (98)

Table 2: 10 most frequent nouns in WWF24-Eng, WWF24-Ita and WWF24-Bra (frequency of each word in brackets).

in the 22th position with 62 occurrences in Brazilian Portuguese).
Comparing the three subcorpora, we can observe that only the following 6 nouns occur in all three of them (with comparable number of occurrences):
'nature'/'natura'/'natureza'
'change'/'cambiamento'/'mudança'
'ecosystem'/'ecosistema'/'ecossistema'
'population'/'popolazione'/'população'
'biodiversity'/'biodiversità'/'biodiversidade'
'system'/'sistema'/'sistema'.

These **6 nouns** were selected in order to examine their **co-occurrence with verbs**. It should be noted that the results of the following analyses (created with the Word Sketch function of Sketch Engine) are not identical for all three languages. Morpho-syntactic annotation tools are available for Italian and English, allowing Sketch Engine to build on its results to obtain accurate information about word behaviour, whereas they are not available for Brazilian Portuguese. This means, for example, that for a noun N, Sketch Engine can distinguish between verbs where N occurs as a subject and those where N occurs as an object complement, whereas for Brazilian Portuguese it can only recognise verbs with which N co-occurs.

In the upper part of the figure 1 we can see some important differences for the noun 'nature'/'natura'/'natureza', i.e. those that occur most frequently in all three languages: In Italian, the most frequently used verbs with 'natura' as object are 'ripristinare' (to restore) and 'proteggere' (to protect); they also occur in the other languages, but the link between the words 'nature' and 'natureza' and these verbs is less strong in the data for English and Brazilian.

In the lower part of the figure 1, the diagrams for the nouns 'biodiversity'/'biodiversità'/'biodiversidade' are shown. We can see that in all languages the concept of biodiversity is linked to the action of preserving ('conserve', 'conservare' and 'conservar'). However, the fact that in Italian biodiversity is in almost all cases the object of active verb forms underlines that the responsibility for its conservation is perceived as the task of a specific person or entity (subject of these verbs). For the other 4 most frequent word groups mentioned above, the analysis is given in the diagrams in the Appendix-A.

We can also observe that different attitudes towards environmental issues emerge from the data if we focus on the **types of verbs** used in their discourses. If we read the lists of verbs used in the three subcorpora of WWF24, we can see a significant difference in the use of modal verbs and thus an underlying expression of a different intention to describe actions as possible and their resolution as obligatory. Modal verbs are used more frequently in the Italian subcorpus than in the other two ones. In WWF24-Ita, among 558 different verbs, the modal verb 'potere' (can) is the second most frequently used (169 occurrences) and 'dovere' (must) the fourth most frequently used (115 occurrences). In WWF24-Bra, over 577 different verbs, 'poder' is the second most used verb (169 occurrences) and 'dever' the ninth (52 occurrences). In WWF24-Eng, modal verbs occur less than ten times in 557 different verbs.

The detection of **keywords**[10] from the three

[10] According to the approach applied by Sketch Engine, keywords are the words that are more frequent in the observed

Figure 1: Behavior of 'nature'/'natura'/'natureza' and 'biodiversity'/'biodiversità'/'biodiversidade' with verbs.

subcorpora also shows that there are important differences among them and that the focus of the discourse is not exactly the same in the three versions of the WWF's report.

We started from the list of the ten more characteristic keywords extracted from WWF24-Eng, i.e. 'wwf', 'lip', 'nature-positive', 'nature-based', 'overexploitation', 'ipbes', biodiversity', 'gbf', 'oecm', 'deforestation'. Then we observed whether the corresponding keywords occur also in the lists for Italian and Brazilian Portuguese respectively drawn from WWF24-Ita and WWF24-Bra. Only six of the ten keywords occur in the three subcorpora, but only four with comparable weight (rank in the list). It must be noted that among these four, three are acronyms which are not translatable expressions, i.e. 'lpi' (living planet index), 'ipbes' (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services) and 'gdf' (global diversity framework), while the remaining one is 'biodiversity'/'biodiversidade'/'biodiversità'.

The keywords related to 'overexploitation' ('superexploração' and 'sovrasfruttamento') are similarly ranked in the three subcorpora. It is particularly striking that 'deforestation', the keyword ranking as tenth in the list from WWF24-Eng and as sixth in the list from WWF24-Ita, does not appear in the list from WWF24-Bra.

With regard to **technical terms**, some of which are likely candidates to multi-word expressions[11], two in particular were explored in our analysis. The first one is 'tipping point', a term which is pivotal in the report and which is amply used and defined in several websites [12]. The original 'tipping point' has 74 occurrences in the WWF24-

corpus with respect to a very large reference corpus for the same language. The reference corpora used in our analysis of keywords are: enTenTen21 for English, itTenTen20 for Italian and ptTenTen23 for Portuguese.

[11] For a definition of MWE, see (Bhatia et al., 2024).

[12] https://www.reteclima.it/tipping-points-ambientali-e-riscaldamento-climatico/

Eng. The Italian version introduces the term in quotes and provides a translation: "'tipping point' o punto critico di non ritorno" (lit.: critical point of no return). Subsequently, it alternates between the term in English (41 occurrences) and its Italian translation (13 occurrences). Unlike the Italian text, the Brazilian version uses only 'ponto de não retorno' (73 occurrences).

The term 'tipping point' is actually a concept developed originally from a physics perspective, which later came to be adopted in other domains and introduced to the lay public through science journalism.[13] A quick query in online publications shows that academic publications in Italian and Portuguese use the English term.

The second term is 'nature-positive', technically defined in some websites[14] and which poses a challenge to translations into Italian and Brazilian Portuguese as in English is used mostly attributively in pre-modifying position, i.e. as an attribute to a noun, requiring renditions by qualifiers in post-modifying position. There are 20 occurrences of 'nature-positive' in the WWF24-Eng, collocating with 'production'(7), 'food systems'(4), 'practices' (3), 'businesses and enterprises' (2), 'future' (2), 'food production' (1), and 'energy transformation' (1).

The WWF24-Ita uses different renditions to translate 'nature-positive', namely, qualifiers such as 'rispettose della natura' (lit.: respectful of nature), 'positivo per la natura' ((lit.: positive for nature), 'nature-positive' as a borrowed term operating as a qualifying adjective 'pratiche nature-positive' (lit.: practice nature-positive); and a few adjectival clauses such as 'che rispettano la natura' (lit.: which respect nature).

Finally, the WWF24-Bra uses 'nature-positive' as a noun in prepositional phrases qualifying another noun, e.g., 'produção de natureza-positiva' (lit.: production of nature-positive); 'focado em natureza-positiva' (lit.: focused on nature-positive); and qualifiers such as 'positivo para a natureza' (lit.: positive for nature).

In environmental organizations webpages in Italian (cf. https://www.reteclima.it/nature-positive-un-mondo-equo-a-zero-emissioni-e-a-favore-della-natura/), the concept of "nature-positive" is presented as a term borrowed from English and is used in English. So is the case in publications by WWF Italy (cf. https://www.wwf.it/cosa-facciamo/pubblicazioni/biodiversita-fragile-maneggiare-con-cura/). In publications by WWF Brazil, the concept is used in Portuguese as "natureza positiva" (lit.: nature positive).

[13] cf. (Blaustein, 2015)

[14] https://blog.3bee.com/guida-al-concetto-di-nature-positive/

5 Sentiment Analysis and Topic Modeling of the WWF24 Corpus

The availability of tools and models for general language allowed us to apply two types of semantics oriented analysis to the English data. The application of the same analyis is scheduled for the other sections of the WWWF.

To **analyze the sentiment** of the WWF24-Eng, we employed the pre-trained BERT-based model *'distilbert-base-uncased-finetuned-sst-2-english'* (Sanh et al., 2019). This model, optimized for sentiment classification, categorizes text into Positive, Negative, or Neutral. Given the corpus's length (25,359 words) and the BERT token limit (512 tokens per input), the text was divided into overlapping chunks of 510 tokens, with a stride of 50 tokens to maintain contextual continuity. Each segment was individually analyzed and the results were aggregated. The analysis revealed the aggregated distribution in figure 2 and more precisely:

- Positive Chunks: 63
- Negative Chunks: 103
- Neutral Chunks: 0

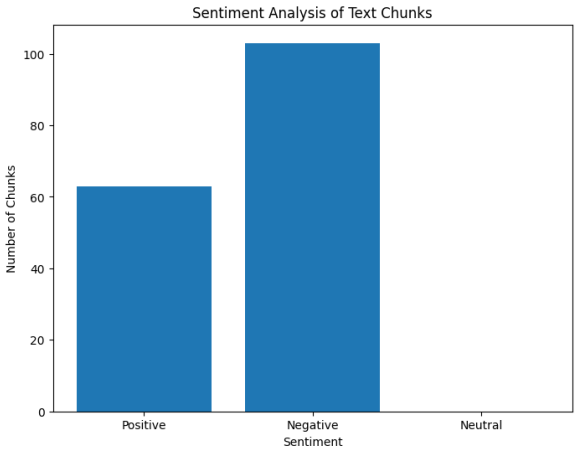

Figure 2: Sentiment distribution across text chunks in the WWF24 Corpus.

These findings indicate a predominance of negative sentiment throughout the corpus, suggesting a focus on challenges, critical issues or alarming environmental concerns. Positive sentiment, while present, appears significantly less frequent and no neutral segments were detected.

To uncover key themes within the WWF24-Eng, according to **topic modeling** approach, we employed the BERTopic algorithm (Grootendorst, 2022), a technique that uses transformer-based embeddings and clustering. The text was preprocessed to remove stop-words, non-alphanumeric characters and excessive white spaces. Then it was divided into manageable chunks of approximately 200 words to ensure compatibility with the model input size requirements. The resulting blocks were analyzed using BERTopic, generating topic clusters based on semantic similarities.
The analysis revealed three main clusters:

- Topic -1: outliers or segments not clearly assignable to a specific theme. Keywords: *tipping, points, change, energy*.
- Topic 0: global environmental challenges. Keywords: *nature, climate, energy, global, finance*.
- Topic 1: biodiversity and species conservation. Keywords: *species, populations, decline, index, change*.

These topics reflect a strong emphasis on global environmental challenges, energy sustainability and biodiversity conservation. Nevertheless, reporting the results of these analyses we are conscious that pre-training on general language works very well for common language, but its results are not as reliable for particular domain languages, such as texts about environment and we will work in the development of novel resources for it.

6 Discussion, Conclusions and Future Work

This paper presents the first release of the novel multilingual corpus WWF24. It compiles texts published in 2024 by the WWF for reporting on the evolution of environmental crisis all around the world and focuses on three languages, i.e. English, Brazilian and Italian.
Both the data and analyses provided in this paper are under several respects preliminary, but useful for drawing the methodology we will apply in the future development of the resource.
Our preliminary analysis points to some interesting avenues for studying diversity in approaches to environmental issues, in particular how meanings are construed in different languages regarding basic notions in the debate about the environment, such as biodiversity and nature. The expression of different intents and attitudes towards the crisis emerge instead from the different verbs used in the three subcorpora we observed. By contrast, the use of technical terms highlights that the language used in the WWF reports is featured by a certain degree of complexity and may be challenging for a significant part of the readers, as terms used to refer to technical concepts, such as 'tipping point' and 'nature-positive' have varied renditions in Italian and Portuguese, which runs counter the univocality of technical terms. However, the choice of varied renditions may be accounted for by the characteristics of the WWF report referred to in our Introduction, namely, the need to present technical information to a lay audience.

Our observations may be seen as the starting point for future work since they help us in formulating hypotheses to be validated (or refuted) following several possible directions. First of all, the availability of the WWF's reports for several other languages will allows us to expand our corpus by including more languages and comparing a larger set of different cultures.
As mentioned above, like most documents for the dissemination and public discussion of environmental issues, the reports published by the WWF also include pictures and infographics that integrate the textual content. By extending the analysis to multimodal information we will be able to collect more insights into the debate about the environmental crisis.

Finally, other forms of analysis, such as sentiment analysis and topic modelling (which we applied only on English data but will be applied on the other languages when more resources will be available for them also), but also frame extraction, can be helpful for collecting the different facets of the ongoing debate and how it varies across different cultures and countries.

Acknowledgments

The work of E. Chierchiello is funded by the International project *CN-HPC-Spoke1-Future HPC & Big Data, PNRR MUR-M4C2*. A. S. Pagano

has a grant from FAPEMIG (Minas Gerais State Agency for Research and Development), Brazil, to develop research at Dipartimento di Informatica, Università di Torino, Italy.

References

Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, and Alexandre Rademaker, editors. 2024. *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Richard Blaustein. 2015. Predicting tipping points. *World Policy Journal*, 32(1):32–41.

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 *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, Venice, Italy. CEUR.

Xu Du, Matthew Kowalski, Aparna S. Varde, Gerard de Melo, and Robert W. Taylor. 2020. Public opinion matters: mining social media text for environmental management. *SIGWEB Newsl.*, 2019(Autumn).

Francesca Grasso, Stefano Locci, Giovanni Siragusa, and Luigi Di Caro. 2024a. Ecoverse: An annotated twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024*, pages 5461–5472. ELRA and ICCL.

Francesca Grasso, Ronny Patz, and Manfred Stede. 2024b. NYTAC-CC: A climate change subcorpus based on new york times articles. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa (Italy). CEUR.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

M.A.K. Halliday and C. Matthiessen. 2006. *Construing Experience Through Meaning: A Language-Based Approach to Cognition*. Open linguistics series. Bloomsbury Academic.

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

Lokesh Mishra, Sohayl Dhibi, Yusik Kim, Cesar Berrospi Ramis, Shubham Gupta, Michele Dolfi, and Peter Staar. 2024. Statements: Universal information extraction from tables with large language models for ESG KPIs. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)@ACL2024*, pages 193–214, Bangkok, Thailand. ACL.

Diya Saha, Manjira Sinha, and Tirthankar Dasgupta. 2024. EnClaim: A style augmented transformer architecture for environmental claim detection. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)@ACL2024*, pages 123–132, Bangkok, Thailand. ACL.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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@ACL-IJCNLP 2021*, pages 8–18, Online. ACL.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha 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, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. ClimateGPT: towards AI synthetizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

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

Appendix-A: Behavior of the Most Frequent Nouns with Verbs