Datenbank-Spektrum (2023) 23:177–188
https://doi.org/10.1007/s13222-023-00454-1

FACHBEITRAG

The InsightsNet Climate Change Corpus (ICCC)

Compiling a Multimodal Corpus of Discourses in a Multi-Disciplinary Domain

Elena Volkanovska[1] · Sherry Tan[1] · Changxu Duan[1] · Sabine Bartsch[1] · Wolfgang Stille[1,2]

Received: 31 May 2023 / Accepted: 19 August 2023 / Published online: 11 September 2023

Abstract
The discourse on climate change has become a centerpiece of public debate, thereby creating a pressing need to analyze the multitude of messages created by the participants in this communication process. In addition to text, information on this topic is conveyed multimodally, through images, videos, tables and other data objects that are embedded within documents and accompany the text. This paper presents the process of building a multimodal pilot corpus to the InsightsNet Climate Change Corpus (ICCC) and using natural language processing (NLP) tools to enrich corpus (meta)data, thus creating a dataset that lends itself to the exploration of the interplay between the various modalities that constitute the discourse on climate change. We demonstrate how the pilot corpus can be queried for relevant information in two types of databases, and how the proposed data model promotes a more comprehensive sentiment analysis approach.

Keywords Corpus, Climate change, Computational linguistics, Annotation, Metadata

1 Introduction

In recent years, the topic of climate change has taken center stage in discourses across diverse segments of society through different channels, media, and publications. While climate scientists are in agreement that climate change is ongoing and real, debates on this topic as well as its influences on policy-makers remain highly controversial [30].

With the surge of published data on climate change, linguistics and other related disciplines have identified the study of data representing discourses on climate change as a research desiderate in order to gain a better understanding of this multidisciplinary field and the role played by a diverse set of participants with different scientific and political backgrounds who are assuming a number of roles and interests. In order to enable such studies, research is needed to collect and organize suitable corpora in a comprehensive and meaningful way to inform the different communities engaging and interested in relevant discourses as well as processes concomitant with their roles as scientists, laypersons, politicians, managers and many others involved in the relevant debates and policy making processes.

According to [17], the climate change related topic of global warming "has received little attention in natural language processing [NLP] despite its real world urgency". One plausible reason for this may be attributed to the lack of available corpora focusing primarily on climate change. Additionally, as a topic – like many topics with a multidisciplinary coverage – climate change is represented in many publications from multiple domains and not merely by means of natural language text, but also by means of

Sherry Tan, Changxu Duan, Sabine Bartsch and Wolfgang Stille contributed equally to this work.

✉ Elena Volkanovska
elena.volkanovska@tu-darmstadt.de

Sherry Tan
sherry.tan@tu-darmstadt.de

Changxu Duan
changxu.duan@tu-darmstadt.de

Sabine Bartsch
sabine.bartsch@tu-darmstadt.de

Wolfgang Stille
wolfgang.stille@hessian.ai

[1] Technische Universität Darmstadt,
Residenzschloss 1, 64283 Darmstadt,
Germany

[2] Hessian Center for Artificial Intelligence,
Mornewegstraße 30, 64293 Darmstadt, Germany

a multitude of modalities such as images, maps, data tables and visualizations that are hardly captured, let alone systematically analysed for their contribution at all. While there may be an abundant volume of digital text to be potentially included in corpora, the representation of textual and embedded multimodal data objects extracted and stored together in a corpus on the topic of climate change is still lacking. Efforts to conduct discourse analysis on this topic which would account for messages conveyed in more than one modality are therefore constrained.

The research reported in this paper aims to fill this gap by building a multimodal corpus representing discourses from the domain of climate change across different genres. The central objective is to illustrate how a multimodal corpus on a specific topic has been compiled and annotated in a manner conducive to discourse analysis on a specific topic that takes into account messages conveyed by more than one modality. The process starts by selecting genres of interest and corresponding data sources; this is followed by a demonstration of some exemplary methods from corpus and computational linguistics for collecting, parsing, and annotating the data with the purpose of enriching the corpus with metadata and linguistic annotations. A corpus compiled and processed in this way allows for more in-depth analyses that explore a wider range of connecting points within and between documents. This is exemplified through two use cases: one that shows how relevant information from the corpus can be retrieved by performing linguistic and document queries in a relational and a graph database respectively[1] using corpus metadata and annotations, and a second one that shows how the data model proposed in this paper supports multimodal sentiment analysis.

We believe that the compilation and curation of multimodal corpora, alongside the study of the interlinking between the multimodal objects with their textual counterparts, can create new insights into the topic of climate change and drive new discussions across various communities. Such corpora are also instrumental in shedding light on the respective discourses ensuing in different communicative contexts.

2 Overview of corpora for discourse analysis on climate change

Prior to embarking on corpus-building, we explored existing corpora and datasets that have been used in previous studies on the climate change discourse. A good overview of datasets used to investigate the debate on climate change by practitioners in the community of NLP and social sciences is provided in [30]; unfortunately, none of these studies takes multimodality into account. A further potentially relevant climate-change-related NLP resource is the Science Daily Climate Change (SciDCC) dataset, presented in [21], which includes approximately 11 000 news articles on the topics "Earth and Climate" and "Plant and Animals" scraped from the Science Daily website. Yet, this is a text-only resource as well. There is a limited number of studies on the topic of climate change conducted on multimodal corpora, but these are largely combinations of texts and photographic illustrations only (see [1] and [33]).

The exploration of existing corpora on the topic of interest revealed that while they are well-suited for text-only discourse analysis, none of them can fully address the needs of a study aiming to analyse the climate change discourse as an interaction between various modalities. The corpora that we inspected do not store data objects of different formats in a single corpus in a manner that lends itself to the study of the interplay between a document's text and any multimedia content embedded in it. In addition, existing multimodal corpora take into consideration a set number of media types, which does not allow for the exploration of the range of embedded media types. Rather than moulding our research to fit the data that was readily available at the time this study began, we decided to build a multimodal corpus from authentic data that would allow us (1) to examine the type of modalities embedded in a document, and (2) to explore methods of querying different modalities and their contribution to the discourse on climate change.

3 Developing a pilot corpus

The pilot corpus described in this section is a precursor to ICCC. The objective is to explore the possibilities of developing a multimodal corpus on climate change and to systematically learn more about the challenges before expanding it. At the onset of the corpus-building process, two main criteria were devised: the corpus had to contain content in both English and German, and any collected multimedia content had to be embedded in the document. We refrained from incorporating stand-alone collections of single-modality data such as collections of tables, images, videos etc. We did not set a limit on the types of multimodal data to be collected with the expectation that data objects other than images and videos would be encountered. Beyond this, we adhered to a fairly standard corpus-design procedure, which included the following steps: (1) identifying genres of interest and data sources that contain suitable content, (2) contacting copyright holders to obtain their approval to collect and use the data, (3) defining metadata properties to store relevant information, (4) collecting the data from

[1] Databases of this type allow us to switch between a bottom-up i.e. word-to-document and a top-down i.e. document-to-word analysis.

each data source, (5) parsing the data in a project-specific corpus structure.

3.1 Identifying genres, data sources, and obtaining copyright permissions

The objective in this step was to ensure that each genre included in the corpus represents various entities or members of society that actively take part in the public discourse on climate change. Hence, the pilot corpus entails content from three sources: academic papers on climate change, reports of the Intergovernmental Panel on Climate Change (IPCC), and content published on the websites of Greenpeace International and Greenpeace Germany (Non-Governmental Organisations (NGOs)). We thus established a modular corpus structure representing three distinct genres: academic literature, content produced by international or intergovernmental organisations, and content produced by NGOs. Setting up a corpus in this way allows for both easy extension of its contents, for example by adding more academic papers or including data from additional international organisations and NGOs, and for a simple narrow-down step by selecting a single genre as per research requirements.

Academic papers can be found either under a free open access (OA) policy, which does not require specific copyright permissions, or hidden behind a paywall, in which case the rules for content use are governed by the specific publisher. IPCC reports can be downloaded from the official website of IPCC[2] and used for personal, non-commercial purposes as long as the source is duly acknowledged. Translation of IPCC reports into German is managed by the German IPCC Coordination Office[3] and the translated content can be retrieved from their website. Content published on the two Greenpeace websites posed the most complex copyright case, mostly because of the different copyright rules applicable to text on the one hand, and multimedia content on the other. Greenpeace has granted us approval to use images and videos that have been created by and are sole property of Greenpeace, as long as the content is used for research purposes exclusively[4,5]

3.2 Establishing and implementing a metadata schema

Metadata support corpus management and exploitation and constitute an integral part of linguistic research. They can be retrieved from the content description provided by the publisher (corpus/document properties), or obtained through data post-processing, including linguistic processing and information extraction (corpus/document annotation). An example of the former would be the year in which an article was published, and of the latter the number of tokens in an article. We refer to corpus and article information obtained in this way with the umbrella term *metadata*.

The pilot corpus metadata framework uses properties from the Dublin Core Metadata Initiative (DCMI Metadata Terms) as its backbone. We opted for the DCMI framework because it provides descriptive terms for data objects of different formats and constitutes a widely acknowledged standard that has been used in the description of both web and physical collections. This allows us to use the same schema for digitised collections which were not primarily designed to serve as web content.

At the time of property selection, DCMI Metadata Terms entailed 55 properties [4], accompanied by a set of datatypes and vocabulary encoding schemes for the description of digital resources of various formats (including image, video, and audio). We selected 14 DCMI metadata terms: title, type, subject, publisher, contributor, identifier, rights, format, bibliographicCitation, rightsHolder, license, extent, created, accrualMethod. For a more detailed description of each term see [4]. This information should be retrievable for each document in the corpus.

While the DCMI Metadata Terms provide a good selection of descriptive elements, they do not include fields for encoding all information of relevance to the project. Two containers of metadata properties were added to address this shortcoming: *linguisticInformation* and *mediaInformation*. The former is a container for project-relevant linguistic information gathered from both the given metadata, that is, metadata provided by the publisher, and for metadata derived by performing linguistic processing on the corpus, described in Sect. 6. The latter stores information about the number and type of multimedia data objects embedded in a document. The two metadata containers are flexible and more properties can be added as necessary. At the moment, *linguisticInformation* holds information about genre, language, text type, status of content (archived or not, for clarification see Sect. 4.3), number of tokens, number of words, word types, content words, type-token ratio, lexical density, information about sentence, word, and token length, named entities and abbreviations.

Once the metadata schema was established, each document was given a *filename* according to an agreed workable convention. In this way, various media types can be linked to the document in which they are embedded. The collected metadata was added to each document and helped us build a profile of the whole corpus. The intention is to apply this

[2] https://www.ipcc.ch/.

[3] https://www.de-ipcc.de/index.php.

[4] Greenpeace International e-mail to Elena Volkanovska, 13 April 2022.

[5] Greenpeace International e-mail to Elena Volkanovska, 15 March 2022.

schema to every document collected from the three data sources described in Sect. 3.1.

4 Data collection

This section elaborates on the data collection process from the three sources identified in Sect. 3.1, which include: academic papers, reports of the Intergovernmental Panel on Climate Change (IPCC), and relevant content published on the websites of Greenpeace International and Greenpeace Germany.

4.1 Academic papers

As a starting point for data collection for the pilot corpus, we used an article published by CarbonBrief titled "The most influential climate change papers of all time" [25]. The article highlights eight academic papers [2, 3, 7, 8, 11, 14, 18, 24] as the most "cited" papers, which is a measure and an indication of how much impact the paper has in the scientific world. The publication dates of the seed papers lie between the years 1896 to 2012, thus providing a wide range of different climate change perspectives as the topic has evolved over time. We coined the eight papers our "seed papers"; they provided a way for us to extract information from them that would link our search path to other related academic works along the same topic lines across different years, providing a method of building a more comprehensive and transparent corpus.

4.1.1 Building a corpus with the seed papers

We explored two methods for building a corpus using the eight seed papers defined earlier: (1) checking the overlap of references between the seed papers, (2) extracting keywords and keyphrases from the academic papers and using them as *seed terms* to search for more academic papers on similar topics in Google Scholar.

The first approach did not reveal an overlap between the references of the seed papers. Therefore, we did a search on Dimensions[6] for a list of the top citation references for each seed paper and from there we looked for overlapping citations. If a paper referenced to at least two seed papers, then that paper was taken to be included in the corpus. Based on this method, a total of 84 papers were initially collected.

The second method was based on information extraction. The text content of the seed papers was extracted and analyzed with KeyBERT [6]. KeyBERT provides integration of different pre-trained language models and since we only have academic papers in the English language, we opted for the model *all-MiniLM-L6-v2* developed initially by [26].

The top 10 keywords/keyphrases from each paper were extracted and grouped according to semantic similarity. The first iteration of extracting and grouping keywords/keyphrases from seed papers resulted in 9 clusters of keywords/keyphrases. Items of each cluster were used as seed terms to search Google Scholar with the *AND* operator between the terms and the top 20 results were taken and added to our collection. This iterative process was completed when we evaluated the corpus and found that we had obtained 1812 academic papers using this method. The total number of academic papers downloaded was 1887, ranging from the years 1895 to 2022.

4.2 Reports published by the Intergovernmental Panel on Climate Change (IPCC)

The pilot corpus contains IPCC synthesis reports from each reporting period. These include:"Climate Change: The IPCC 1990 and 1992 Assessments", "SAR Climate Change 1995: Synthesis Report", "TAR Climate Change 2001: Synthesis Report", "AR4 Climate Change 2007: Synthesis Report", "AR5 Synthesis Report: Climate Change 2014", "Synthesis Report (SYR) of the IPCC Sixth Assessment Report (AR6)"[7]. IPCC reports are originally published in English and were collected as such; translations into German were collected when available[8]. In the pilot corpus, we included 6 synthesis reports in English and 3 full or partial translations of synthesis reports in German.

4.3 Greenpeace International and Greenpeace Germany

The webpages from Greenpeace International and Greenpeace Germany relevant to our project were retrieved by entering the prompt "climate change" and "Klimawandel" respectively in the search bar on each organisation's website[9,10]. The search, performed in March 2022, returned 4057 links to webpages from Greenpeace International, of which 698 were hosted on the domain of Greenpeace International, while 3359 were archived and hosted on the

[6] https://www.dimensions.ai/.

[7] At the time of writing this paper, the Synthesis Report (SYR) of the IPCC Sixth Assessment Report (AR6), published on 20 March 2023, remains subject to final copy editing and layout changes. The pilot corpus will be updated with the final version of the respective report, and potentially its German translation, once the documents become available.

[8] At the moment, there are full translations of the synthesis reports for the years 2007 and 2014, and a translation of the Summary for Policymakers from 2001.

[9] https://www.greenpeace.org/international/.

[10] https://www.greenpeace.de.

Table 1 Summary of the pilot corpus with number of extracted contents

Data Source	Docs without MC*	Docs with MC*	Multimedia Content			# of Tokens
			Imgs/Figs	Videos	Other	
Academic Papers	100	1787	15 461	–	3302	43 152 714
IPCC reports EN	0	6	354	–	111	556 646
IPCC reports DE	0	3	135	–	31	170 617
Greenpeace International	228	470	2066	188	458	676 879
Greenpeace Germany	229	1052	2463	14	463	645 962

*MC: Multimedia Content

domain of the Wayback Machine – Internet Archive[11]. In the pilot corpus we included only the 698 webpages hosted on the Greenpeace International domain in order to have a balanced number of tokens in each language (see Table 1 for corpus size). From Greenpeace Germany, the search returned 1281 links to webpages.

5 Data parsing

The process described in Sect. 4 resulted in files in two formats: PDF and HTML. This section discusses the tools used to parse the documents and extract relevant information. It should be pointed out that texts accompanying and describing multimodal objects, such as captions, were extracted and saved together with the multimodal object. This allowed us to conduct linguistic annotation and information extraction on these texts, too. Sect. 5.2.1 zeroes in on the post-parsing processing applied to videos, as none of the extracted videos was accompanied by text descriptions.

5.1 Academic papers and IPCC reports

All academic papers and IPCC reports were saved and parsed as PDF files. For the scanned versions of PDF files, we used Tesseract [13] for OCR on the text and appended the results as transparent text layers to the original PDF pages.

We combined VILA [29] and Resnet101 [10] models that were trained on DocBank [16] to parse PDF Files. VILA is a model for token sequence prediction, which does not predict the images in the document. Resnet101 takes as input the rendered image of each page of the document and identifies only the location of the figures on the page. The output label set is *Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table* and *Title*.

The models for parsing PDF files were run in an Online Learning [12] framework and were loaded in Label Studio [31] as a machine learning backend service. We imported the documents as a rectangular label object detection task into the front end, used the original models to make predictions for a small subset of documents, corrected the prediction manually, and then fine-tuned the models. Finally, we parsed all documents using the updated models. The goal is not only to extract the text and other data objects from the PDF files, but also to retain the layout information of the documents so that we can examine the interactions between the data objects.

5.2 Greenpeace International and Greenpeace Germany

Since many of the webpages that we needed to download and parse were dynamic, we used Selenium[12] to retrieve the HTML from the collected links and BeautifulSoup [28] to parse the content. When extracting the relevant data objects, we made sure to preserve the order of appearance of HTML elements containing important information, to extract their position on the webpage, and to extract all HTML elements that might contain links to data objects in any modality other than text. This resulted in documents that mirror the output of the PDF parsing process described in Sect. 5.1.

5.2.1 Adding text descriptions to videos

As none of the videos of the Greenpeace corpus was accompanied by a text description, we extracted the title and description of the video, if available, using the Python library *pytube*[13]. We also obtained the transcription of the videos (where the original language is English or German) using the Python library *youtube-transcript-api*[14]. This resulted in descriptions for all videos in the English collection and for 13 of the 14 videos of the German collection. This text was run through the pipeline described in the next section.

[11] https://archive.org/web/.

[12] https://github.com/SeleniumHQ/selenium, v.3.14.0.

[13] https://github.com/pytube/pytube.

[14] https://pypi.org/project/youtube-transcript-api/.

6 Data annotation and information extraction

This section explains some of the methods used to annotate the parsed data and extract relevant information from it. We elaborate on the steps for: (1) linguistic annotation, (2) named-entity recognition and (3) keywords/keyphrases extraction. The three processing steps were applied to the main text (body) of each document and to texts describing multimodal objects. Sect. 6.1 describes (1) and (2) and Sect. 6.2 describes (3). Several libraries for automatic text processing were used to perform linguistic annotation and information extraction. Before diving into the details of these processes, it is important to underline that there are many NLP libraries that can be used for this type of text processing, and that each of them has its strengths and weaknesses. The objective of this paper is to demonstrate the application of known computational tools for linguistic annotation of and information extraction from authentic data, rather than assess or improve the performance of the used NLP tools. We also point out that while ongoing advancements have increased the reliability of the output of tools for automatic text processing, one should not consider them a single source of truth and should refrain from expecting 100% accuracy.

6.1 Linguistic annotation and named-entity recognition (NER)

The linguistic annotation was carried out by splitting a document's text into sentences and running the sentences through an annotation pipeline that extracted features at token and sentence level. For both languages, the linguistic annotation pipeline entailed tokenization, lemmatization, part-of-speech (POS) tagging, dependency parsing and named entity recognition (NER). From English texts we also extracted abbreviations and obtained their long forms with the help of an adequate NLP tool. The annotation pipeline was implemented fully on the two Greenpeace subcorpora and on samples from the subcorpora of academic papers and reports from the IPCC[15].

Three libraries constitute the annotation pipeline for English texts: spacy-stanza[16], Stanford CoreNLP [19][17], and SciSpacy[23][18] running on language model *en_core_sci_lg*. For several processors, including part-of-speech tagging and dependency parsing, stanza offers a combined language model for English, which has been trained on multiple datasets and as such should provide better coverage of the language[19]. This is important given the three genres constituting the corpus. Stanza language models were used through a spaCy pipeline because this allowed us to complement the 18 named entity (NE) categories offered by stanza's NER model with additional NE categories from CoreNLP in a single pipeline, thus avoiding differences in character offsets when extracting token-level features.

In addition to using linguistic annotations to calculate features such as lexical density or type-token ratio, linguistic information is saved at token level, which paves the way for linguistic queries to be executed in a relational database, a process illustrated in Sect. 8.1. The extracted named entities from English texts belong to the following 24 categories: PERSON, NORP (nationalities, religious or political groups), FAC (facilities such as airports, buildings, highways, bridges), ORG (organisations including companies, agencies, institutions), GPE (geo-political entities), LOC (non-GPE locations, mountain ranges, bodies of water), PRODUCT (objects, vehicles, foods, etc. but not services), EVENT (named hurricanes, battles, wars, sports events, etc.), WORK-OF-ART (titles of books, songs, etc.), LAW (named documents made into laws), LANGUAGE (any named language), DATE (absolute or relative dates or periods), TIME (times smaller than a day), PERCENT (percentage), MONEY (monetary values, including unit), QUANTITY (measurements, such as weight or distance), ORDINAL (first, second, etc.), CARDINAL (numerals that do not fall under another type), TITLE (administrative, legislative, institutional, etc.), CITY (cities only), IDEOLOGY, RELIGION, CRIMINAL-CHARGE, and CAUSE-OF-DEATH (includes diseases and natural disasters). With SciSpacy we extracted abbreviations and their long forms from each sentence. The German documents were processed with stanza only, since both stanza and Stanford CoreNLP distinguish only four categories of named entities for German language texts (ORG, PERSON, LOC, MISC).

In addition to saving the extracted annotations in a JSON file, each document with linguistic annotations was serialized as a pickle file (German content) and both pickle file and spaCy object (English content). This step should ensure consistency should we decide to extract additional linguistic features.

The linguistic annotation served as the backbone of the linguistic information extracted and calculated for each document. The result of this process is fed back into the meta-

[15] This allows us to test the transferability of the pipeline across the three genres at a lower computing cost.

[16] https://spacy.io/universe/project/spacy-stanza, running on stanza language model 1.4.1.

[17] Version 4.4.0.

[18] https://github.com/allenai/scispacy.

[19] Information about the performance of the models is available at: https://stanfordnlp.github.io/stanza/performance.html (stanza) and https://allenai.github.io/scispacy/ (SciSpacy).

data, where the information is saved in the metadata container *linguisticInformation*, described in Sect. 3.2.

6.2 Keywords/Keyphrases extraction

As mentioned in Sect. 4.1, KeyBERT was implemented to capture keywords and keyphrases that are semantically similar to the document content. The same approach was used to annotate the documents in our corpus. Since our corpus contains documents both in English and German, using pretrained language models in English is not enough. Therefore, the multilingual model *paraphrase-multilingual-MiniLM-L12-v2* developed by [27] was used for German texts.

Through our implementation and experimentation of using KeyBERT for identifying seed terms as shown in Sect. 4.1, we found that KeyBERT has the tendency to create noisy results, which did not have much effect on our results when using them as seed terms to search on Google Scholar, but could have a more profound effect on keywords/keyphrases annotation of the data. Therefore, we combined the KeyBERT approach with the textrank approach proposed by [20]. We implemented textrank using PyTextRank [22] in the spaCy pipeline. Both English[20] and German[21] models were used through spaCy.

The keywords and keyphrases that had a semantic similarity score of 0.7 or higher were extracted using KeyBERT and the list was compared to the set that was extracted using PyTextRank; overlapping results were discarded and the final list of keywords and keyphrases was added to the metadata term *subject*.

7 Results: Data collection and parsing

We present the results obtained from the data collection process of the pilot corpus, and the types and number of multimodal data objects retrieved by parsing the documents collected from each of the three data sources.

7.1 Academic papers

A total of 1887 academic papers were collected and 15 461 images and figures were extracted from the academic PDF files. 1095 equations and 2207 tables were extracted and these are listed under "Other" in Table 1. Over 43 million tokens were extracted from these documents.

7.2 IPCC reports

The 6 English IPCC reports contained 354 images and figures and 111 tables/equations. A total of 556 646 tokens were extracted. From the 3 reports in German, 135 images and figures were extracted, alongside 31 tables/equations and a total of 170 617 tokens.

7.3 Greenpeace International and Greenpeace Germany

Of the 698 documents of Greenpeace International, 470 are multimodal; these have 2066 embedded images, 123 embedded videos, 67 videos added to the content as hyperlinks, and 458 other types of multimedia objects. In Table 1, "Other" entails iframes, which are webpages embedded within another webpage. In the context of the Greenpeace International corpus, iframes store videos, text, images, animations, dynamic charts, tweets, Facebook posts, Instagram posts, and PDF files. Of the 1281 documents of Greenpeace Germany, 1052 are multimodal, with 2463 images and 14 videos. We retrieved 188 YouTube videos from Greenpeace International, whose total duration is 25.55 hours. The 14 videos of Greenpeace Germany amount to 3.35 hours. We counted 157 115 tokens in the transcripts of Greenpeace International, and 27 586 tokens in the transcripts of Greenpeace Germany.

It is evident that of the total number of collected documents (3874), the majority have embedded multimedia content (3317), compared to text-only documents (557). This finding underpins the need for awareness of the various media types existing alongside textual content, and for incorporating processing techniques that would enable researchers to analyse media content in the context of a document as a whole.

8 Use cases

This section aims to show how the annotated corpus can be used for discourse analysis through two different use cases. For the first one we ingest the corpus in a relational and a graph database and demonstrate how the linguistic annotations and extracted information could promote the interlinking of text and multimodal objects within documents, and between documents within a corpus. The second use case utilizes the corpus structure to obtain a better understanding of sentiments expressed in images and various sections of text belonging to a document. The two use cases exemplify how a corpus collected, parsed, and annotated using the methodology described in this paper can support a study on the climate change discourse that incorporates a multimodal perspective.

[20] *en_core_web_sm* and *en_core_web_trf*.

[21] *de_core_news_sm* and *de_dep_news_trf*.

```
[word="climate"][pos="NOUN" & word="change"]   (1)

[pos="NOUN" & word="Klimawandel"]   (2)
```

Fig. 1 A CQL query to retrieve sentences containing *climate change* (1) or *Klimawandel* (2)

8.1 Querying the corpus with a relational and a graph database

To demonstrate the types of queries that are made available through the linguistic annotations and the extracted information, as described in Sects. 5 and 6, we ingest the Greenpeace subcorpora in two types of databases: A relational database and a graph database.

Relational databases have been used as the backbone of tools for linguistic queries, such as CQPweb [9] or SketchEngine [15]. Such databases help establish connections between documents by querying the fine-grained linguistic features. The annotated data can be adapted for linguistic analysis in such a tool by constructing tables where each row represents a linguistic feature of interest (such as a POS tag) [9][22]. Since the annotated corpus contains token-level linguistic features, it is possible for us to perform linguistic queries using the Corpus Query Language (CQL) as implemented, for example, in the IMS Open Corpus Workbench and its webfrontend CQPweb. We complement this linguistic data modelling with document-level modelling by means of the graph database ArangoDB[23]. Graph databases can promote the understanding of how documents relate to each other beyond their linguistic features by providing data inter-connectivity and uncovering inherent relations between documents. In our use case, a graph database helps to unveil new connections between documents of interest: whether through the document properties, such as date of publication, the multimedia objects they contain, such as the same image being used in two different documents, hence in possibly differing contexts, or through the extracted keywords, entities or abbreviations.

A simple CQL query, such as the one presented in Fig. 1, would retrieve all sentences containing *climate change* or *Klimawandel*, as well as the identifiers of the documents where the sentences appear. From here on, discourse analysis can be conducted by either zooming in on the sentences containing the queried word or phrase, or by zooming out of the sentence level and looking at the documents and their embedded data objects as constituents of a corpus represented as a graph. Using the Corpus Annotation Graph Builder (CAG) [5], we ingest the Greenpeace subcorpora in ArangoDB and build a graph with the documents of interest using the identifiers obtained through the CQL query. Fig. 2 is a simplified version of such a graph, showing how the documents of interest are related between each other beyond the queried phrase, i.e. word, and through overlapping named entities, images, videos, or other multimedia objects. The search can be further refined to show only documents that are multimodal or that are also connected by a specific named entity. The two query systems are complementary and help build a better understanding of the research question at hand.

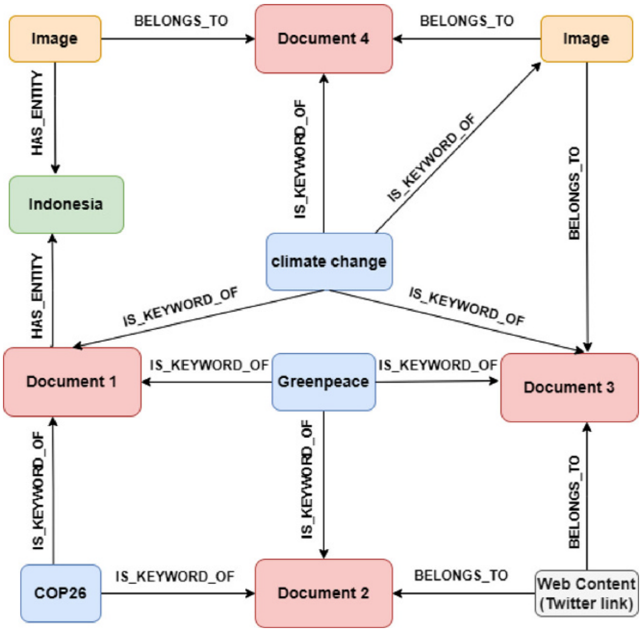


Fig. 2 A simplified view of a graph database output for documents and data objects connected through a keyword/keyphrase (*climate change*, *Greenpeace*, *COP26*) or a named entity (*Indonesia*). A graph could also tell us if two documents make use of the same image

8.2 Comparing sentiments in texts and embedded images

The present data model, which embeds images in a document's text and stores them in the order in which they appear in the original document, allows us to extract information about the sentiment expressed in embedded images on the one hand, and in context texts related to those images on the other. This helps us to examine whether the sentiment detected in images aligns with the sentiment detected in the text, or whether there is a mismatch in the sentiments detected in the different modalities.

We use text from various contexts to ensure that we consider not only image-approximate context text, such as paragraphs preceding and following the image, but also the

[22] This type of tables, known as “vertical files” among the linguistic community, can be created easily from our data model with a Python script.

[23] https://www.arangodb.com.

Table 2 Sentiment analysis of images and texts in Greenpeace International

Img Sent	Img Pos	Img Neg	Img Neutral
Text Sent			
Title Pos	273	333	431
Title Neg	104	134	202
Article Text Pos	61	77	111
Article Text Neg	316	390	522
Context Text Pos	182	189	268
Context Text Neg	56	77	117

Table 3 Sentiment analysis of images and texts in Greenpeace Germany

Img Sent	Img Pos	Img Neg	Img Neutral
Text Sent			
Title Pos	53	60	78
Title Neg	168	215	298
Article Text Pos	0	0	2
Article Text Neg	29	39	41
Context Text Pos	0	0	1
Context Text Neg	201	267	294

text of the document as a whole and the text of the title only. This resulted in a total of 1477 images from Greenpeace International and 1488 images from Greenpeace Germany that were extracted with their corresponding context texts, article titles, and article texts. We then used pre-trained models[24] from HuggingFace[25] to extract sentiment from the text and the visual sentiment analysis pre-trained model developed by [32] to extract sentiment from images.

Tables 2 and 3 summarize the results of the sentiment analysis. Sentiments of context texts were extracted separately, meaning that the results shown in the tables consist only of context texts that had the same sentiment (i.e. the sentiment of the paragraph before the image matches the sentiment of the paragraph after the image). It should be noted that the results from the English model did not have any texts that were classified with neutral sentiment while the German model has 269 titles, 555 article texts and 128 context texts that were classified as neutral. Neutral results in texts of Greenpeace Germany were discarded to make the results comparable with Greenpeace International texts. Overall, Greenpeace International has more neutral sentiment images with positive sentiment titles compared to Greenpeace Germany, which has more neutral sentiment images with negative sentiment titles. Greenpeace International has more positive and negative sentiment article texts compared to Greenpeace Germany, whose texts are mostly classified as neutral.

Of particular interest to us were cases of mismatch of sentiments in images and texts belonging to the same document. One example is between the image and the text shown in Fig. 3, hosted on the website of Greenpeace International[26]. The image itself is dark with white letters in all capitals and its sentiment has been classified as negative. This is in contrast to the context text which says, “We’ve got a few bold suggestions for how to make this happen in a socially just, green and collective way”. The text itself gives a positive sentiment including phrases and words such as “socially just”, “green” and “collective”, which is in stark contrast to the sentiment detected in the image. An example of sentiment mismatch where the text is classified as negative, and the image positive, is provided in Fig. 4. The paragraph under the image, classified as negative, reads: ““Jeder hat das Recht, in Freiheit und Sicherheit zu leben.” Das ist ein Menschenrecht. “Jeder darf seinen Aufenthaltsort selber wählen” auch. Und auch: “Niemand darf willkürlich seines Eigentums beraubt werden.” Doch die Erderhitzung wird Millionen Menschen dieser Rechte berauben, wenn es der Weltgemeinschaft nicht gelingt, den Klimaschutz so voranzutreiben, wie es angesichts des Klimanotstandes dringend nötig wäre. Heute ist der Internationale Tag der Menschenrechte. Doch auch auf der diesjährigen Klimakonferenz in Madrid fehlt es an echtem Engagement, die drohende Katastrophe noch aufzuhalten”[27,28]. The image in Fig. 4 could be interpreted as implicitly negative, as it hints that climate change exacerbates problems of people who are already struggling with access to clean water. Nevertheless, its label stands in contrast to the sentiment expressed in the text, which entails phrases and words such as “Erderhitzung” (global warming), “Klimanotstandes” (climate emergency), “fehlt es an echtem Engagement” (a lack of real commitment), “drohende Katastrophe” (impeding

[24] EN: distilbert-base-uncased-finetuned-sst-2-english, DE: oliverguhr/german-sentiment-bert.

[25] https://huggingface.co.

[26] https://www.greenpeace.org/international/story/45030/time-to-transform-transport/.

[27] The original content is hosted on the website of Greenpeace Germany: https://www.greenpeace.de/ueber-uns/leitbild/recht-schutz.

[28] Translation into English, unofficial: “Everyone has the right to live in freedom and security”. This is a human right. “Everyone may choose their place of residence”, too. And also: “No person shall be arbitrarily deprived of their property”. But global warming will deprive millions of people of these rights, should the global community fail to advance climate protection with the urgency that is needed in the face of climate emergency. Today is the International Human Rights Day. But even this year’s climate conference in Madrid sees a lack of real commitment to stop the impending catastrophe..

catastrophe). Detecting mismatched sentiments across different modalities could set the scene for a qualitative and multimodal approach to discourse analysis, which could examine, for example, if there is a pattern of document-level topics where mismatches in sentiments occur[29].

9 Discussion

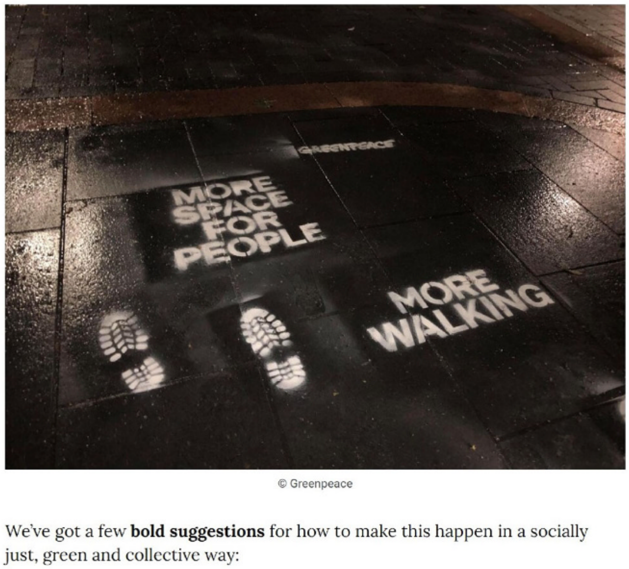
This paper describes the process of building a multimodal pilot corpus representing discourses on climate change from three genres, comprising both a substantial number and a wide range of data types in documents. It also exemplifies how various NLP techniques can be employed to augment the corpus (meta)data. The use cases of Sect. 8 demonstrate two potential approaches to gaining insights from various data objects in a multimodal corpus and their respective annotations and could serve as inspiration for future work. The pilot corpus is the starting point for developing methodologies that would allow us to better design and curate the ICCC.

Prior to discussing the challenges encountered and lessons learned throughout this work, it is important to reiterate that the goal of this paper is to present a methodology for creating and curating a multimodal corpus using known tools for computational analysis. Improving the performance of the NLP tools used for the corpus annotation is beyond the scope of this paper. However, we briefly report our observations on the advantages and disadvantages of using out-of-the-box NLP tools for various genres so as to reflect on the work done so far and to contribute to the development of a methodology that supports the extraction of meaningful links between data objects in a multimodal corpus.

9.1 Lessons learned and future steps

We present some of the lessons learned in terms of data modelling of multimodal corpora, application of data annotation and information retrieval techniques, and challenges of working with a bilingual corpus.

It is evident that discarding multimedia data objects, as is common practice in corpus development, results in the possible loss of relevant information and eliminates the opportunity to investigate the interaction between data objects of different modalities. As seen in this paper, creating a data model that lends itself to modelling and analyzing interactions between different types of data objects presents another layer of complexity in the process of collecting, parsing, and analysing multimodal data, especially when the objective is to build a dataset that represents more than one genre. In the case of the ICCC, we ensured that the subcorpora comprising the pilot corpus have the same metadata schema and document structure. However, the type of multimodal data objects found in each subcorpus is idiosyncratic to the respective data source: academic papers are bound to have a *references* section, while webpages embed *videos* and *iframes*. As our intention was to explore the types of data objects available in different genres, represented by subcorpora from three data sources, we refrained from developing a corpus-wide wording convention for the extracted multimodal data objects. Data objects in the pilot corpus are thus distinguished by relying on existing labeling systems, such as the ones mentioned in Sect. 5. While the complexity of maintaining subcorpus-specific labelling conventions for data objects could be partly removed by developing corpus-level labels and genre-specific sub-labels, a step that could cater to a more unified corpus structure, such a step could also lead to the loss of the comprehensive structural granularity of the current corpus.



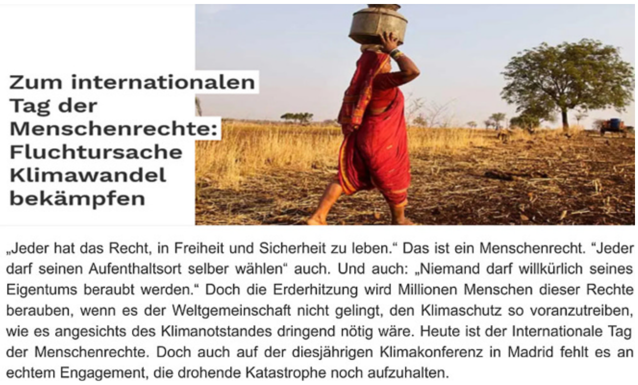
© Greenpeace

We've got a few **bold suggestions** for how to make this happen in a socially just, green and collective way:

Fig. 3 An example of an image/context text mismatch. The image in this figure is the sole property of Greenpeace International and is used with their permission. This is a screenshot of the webpage where the image is hosted

Another important task was to enrich the corpus metadata by incorporating NLP tools for corpus annotation and information retrieval. While some of these techniques evidently enriched the metadata of the corpus, others performed well on one type of data, but not on another, highlighting the necessity of employing tools or models built for specific tasks in the specific target domain. To be more precise, we found that the extraction of linguistic features

[29] It should be noted that automatically generated labels should be subject to closer scrutiny, which is why in this instance we would recommend to further pursue this study with a qualitative rather than a quantitative discourse analysis approach.



Zum internationalen Tag der Menschenrechte: Fluchtursache Klimawandel bekämpfen

„Jeder hat das Recht, in Freiheit und Sicherheit zu leben.“ Das ist ein Menschenrecht. “Jeder darf seinen Aufenthaltsort selber wählen“ auch. Und auch: „Niemand darf willkürlich seines Eigentums beraubt werden.“ Doch die Erderhitzung wird Millionen Menschen dieser Rechte berauben, wenn es der Weltgemeinschaft nicht gelingt, den Klimaschutz so voranzutreiben, wie es angesichts des Klimanotstandes dringend nötig wäre. Heute ist der Internationale Tag der Menschenrechte. Doch auch auf der diesjährigen Klimakonferenz in Madrid fehlt es an echtem Engagement, die drohende Katastrophe noch aufzuhalten.

Fig. 4 An example of an image/context text mismatch. The image in this figure is the sole property of Greenpeace Germany and is used with their permission. This is a screenshot of the webpage where the image is hosted

is less genre-sensitive compared to the extraction of named entities. While SciSpacy's abbreviation extraction tool [23] used with the model *en_core_sci_lg* did extract relevant abbreviations from scientific texts, the library's NER tool does not provide an option to extract named entities belonging to categories relevant to our envisaged discourse analysis task. More work will need to be done in this respect to seek out the appropriate tools and models and to fine-tune them to our domain-specific documents.

We also experienced difficulties in trying to achieve entirely matching annotations for English and for German corpora, mostly because existing tools and models for processing German texts do not offer the same level of granularity and linguistic detail. Improving this situation is identified as a research desiderate in our future work.

In Sect. 8 we presented two potential use cases of the pilot corpus, namely: (1) investigating connections between documents through data objects, metadata properties and linguistic features, and (2) investigating the mismatch in sentiments between two types of modalities (image and text). In the context of discourse analysis, the strengths of use case (1) lie within its power to reveal novel connections among documents, which can expand the context of discourse analysis beyond a sentence- and text-level approach. Use case (2) could open the way for a comprehensive qualitative analysis, which could shed some light on how sentiments in documents are used and whether a computational interpretation of sentiments could benefit discourse analysis. While it is beyond the scope of this paper to discuss the sentiment analysis model performance, we believe that this type of data structure could reveal the strengths and weaknesses of such models, which is paramount given the amount of automatic processing of texts and multimedia contents being done every day. Future analysis may incorporate explainable sentiment analysis models, and the multimodal corpus itself may be used to improve existing sentiment analysis models.

As stated previously, the goal of this paper on the creation of ICCC pilot corpus is to explore the design and curation process of a multimodal corpus, and to propose a corpus development methodology that is conducive to a multifaceted discourse analysis on the topic of climate change. The next step of the research is to further improve the proposed methodology, to expand the ICCC by looking further into collecting data from other sources in the climate change domain, and to test domain-specific NLP tools for data processing.

Acknowledgements The research reported is this paper was conducted within the research project InsightsNet (https://insightsnet.org/) which is funded by the Federal Ministry of Education and Research (BMBF) under grant no. 01UG2130A.
We would like to thank Rasmus Beckmann and Jasper Korte of the Institute for Software Technology, German Aerospace Centre (DLR) for their assistance in identifying sources of data relevant to our research.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. DiFrancesco AD, Young N (2011) Seeing climate change: the visual construction of global warming in canadian national print media. Cult Geogr 18(4):517–536
2. Arrhenius S, Holden ES (1897) On the influence of carbonic acid in the air upon the temperature of the earth. Publ Astron Soc Pac 9(54):14–24
3. Callendar GS (1938) The artificial production of carbon dioxide and its influence on temperature. QJR Meteorol Soc 64(275):223–240
4. DublinCore (2020) Dcmi metadata terms
5. El Baff R, Hecking T, Hamm A et al (2023) Corpus annotation graph builder (CAG): an architectural framework to create and annotate a multi-source graph. In: The 17th conference of the European chapter of the association for computational linguistics (EACL 2023) (System Demonstrations)
6. Grootendorst M (2020) Keybert: Minimal keyword extraction with bert https://doi.org/10.5281/zenodo.4461265
7. Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Model 135(2-3):147–186
8. Hansen J, Sato M, Ruedy R (2012) Perception of climate change. Proc Natl Acad Sci USA 109(37):E2415–E2423

9. Hardie A (2012) CQPWeb—combining power, flexibility and usability in a corpus analysis tool. Int J Corpus Linguist 17(3):380–409
10. He K, Zhang X, Ren S et al (2015) Deep residual learning for image recognition. arXiv:1512.03385
11. Held IM, Soden BJ (2006) Robust responses of the hydrological cycle to global warming. J Climate 19(21):5686–5699
12. Hoi SCH, Sahoo D, Lu J et al (2018) Online learning: a comprehensive survey. arXiv:1802.02871
13. Kay A (2007) Tesseract: An open-source optical character recognition engine. Linux J 2007(159):2
14. Keeling CD, Bacastow RB, Bainbridge AE et al (1976) Atmospheric carbon dioxide variations at mauna loa observatory, Hawaii. Tellus 28(6):538–551
15. Kilgarriff A, Baisa V, Bušta J et al (2014) The Sketch Engine: ten years on. Lexicography 1(1):7–36
16. Li M, Xu Y, Cui L et al (2020) Docbank: A benchmark dataset for document layout analysis vol 2006.01038
17. Luo Y, Card D, Jurafsky D (2020) Detecting stance in media on global warming. arXiv preprint arXiv:201015149
18. Manabe S, Wetherald RT (1967) Thermal equilibrium of the atmosphere with a given distribution of relative humidity
19. Manning CD, Surdeanu M, Bauer J et al (2014) The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
20. Mihalcea R, Tarau P (2004) Textrank: bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing In, pp 404–411
21. Mishra P, Mittal R (2021) Neuralnere: Neural named entity relationship extraction for end-to-end climate change knowledge graph construction. In: Tackling climate change with machine learning workshop at ICML
22. Nathan P (2016) Pytextrank, a python implementation of textrank for phrase extraction and summarization of text documents https://doi.org/10.5281/zenodo.4637885
23. Neumann M, King D, Beltagy I et al (2019) ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th bioNLP workshop and shared task. Association for Computational Linguistics, Florence https://doi.org/10.18653/v1/W19-5034
24. Nordhaus WD (1991) To slow or not to slow: the economics of the greenhouse effect. Econ J 101(407):920–937
25. Pidcock R (2015) The most influential climate change papers of all time. Available at: https://www.carbonbrief.org/the-most-influential-climate-change-papers-of-all-time/. Accessed: 13.10.2021.
26. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. In: Association for computational linguistics (ed) Proceedings of the 2019 conference on empirical methods in natural language processing. (http://arxiv.org/abs/1908.10084)
27. Reimers N, Gurevych I (2020) Making monolingual sentence embeddings multilingual using knowledge distillation. In: Association for computational linguistics (ed) Proceedings of the 2020 conference on empirical methods in natural language processing (https://arxiv.org/abs/2004.09813)
28. Richardson L (2007) Beautiful soup documentation. Available at: https://beautiful-soup-4.readthedocs.io/en/latest/. Accessed: 18.02.2022.
29. Shen Z, Lo K, Wang LL et al (2022) Vila: Improving structured content extraction from scientific pdfs using visual layout groups. Trans Assoc Comput Linguist 10:376–392
30. Stede M, Patz R (2021) The climate change debate and natural language processing. In: Proceedings of the 1st workshop on NLP for positive impact, pp 8–18
31. Tkachenko M, Malyuk M, Holmanyuk A et al (2022)) Label Studio: Data labeling software. Available at: https://github.com/heartexlabs/label-studio. Accessed: 29.09.2021.
32. Vadicamo L, Carrara F, Cimino A et al (2017) Cross-media learning for image sentiment analysis in the wild. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp 308–317 https://doi.org/10.1109/ICCVW.2017.45
33. Wessler H, Wozniak A, Hofer L et al (2016) Global multimodal news frames on climate change: a comparison of five democracies around the world. Int J Press Polit 21(4):423–445