Presenting an Annotation Pipeline for Fine-grained Linguistic Analyses of Multimodal Corpora

Elena Volkanovska **Sherry Tan** **Changxu Duan**
Debajyoti Paul Chowdhury **Sabine Bartsch**
Technische Universität Darmstadt, Germany
{firstname.lastname}@tu-darmstadt.de

Abstract

With the increasing availability of multimodal documents, it is becoming more difficult for researchers to not only find relevant information within documents in various modalities and media formats, but to also explore potential semantic relationships between data objects of two different modalities embedded in a single document. This paper proposes a method rooted in an annotation pipeline that takes as input text data objects that are either native text objects, or textual descriptions of a multimodal object, such as an image or video, and generates as output an attribute-rich document that unites four levels of annotation in a single framework. The annotated files generated by this pipeline lend themselves to exploration either in a non-programmatic way, by using the Corpus Query Language (CQL) in the web-based graphical user interface (GUI) of the IMS Open Corpus Workbench (CWB), or programmatically, using Python and a Jupyter Notebook. We present some preliminary results of analyses performed on the corpus.

1 Introduction

The means of communicating different areas of knowledge have expanded through the use of different modalities in documents in addition to text, such as images, videos, interactive maps, tables and equations, to name a few. Even documents that do not classify as natively digital content often contain some type of multimodal (MM) data objects. Depending on the genre the document belongs to, MM data objects can serve a range of purposes, from providing additional knowledge to triggering a certain emotional response in the reader (Bednarek and Caple, 2012). Some MM data objects, such as images, may be accompanied by textual descriptions or captions, whose goal is to contextualise the image within the document where it is embedded (Tan et al., 2020).

This paper explores the question of what it takes in terms of corpus annotation to allow for revealing potentially interesting connections between text objects (TOs) of two types: texts of documents (hereinafter: principal text objects, PTOs) and texts that serve as descriptions to multimodal data objects embedded in the document (hereinafter: descriptive text objects, DTOs). We propose an annotation pipeline that integrates existing libraries for natural language processing (NLP) and creates an annotation framework with linguistic and semantic attributes extracted from texts in English, which can be either PTOs or DTOs. The annotation process generates attributes that complement the inherent properties of each document, and allow for performing complex data queries on the document's body text and on texts describing multimodal objects. The attributes are generated at four levels: token, sentence, paragraph, and full-text (document) level. The goal is to create an annotated multimodal corpus with contents in English from a topic area where multimodal objects are natively used to communicate information; one such example is the topic of *climate change*. Thus, the output of the annotation pipeline should satisfy a twofold objective: (1) enriching the corpus with attributes that allow for thorough linguistic exploration of PTOs and DTOs in a non-programmatic manner, using queries performed with the Corpus Query Language (CQL) (Christ, 1994) within CQPweb (Hardie, 2012)[1], and (2) enriching the corpus with linguistic and semantic attributes which can be used to programmatically perform complex analyses on the interaction between text and images using Python and a Jupyter Notebook. Objective (1) should exemplify one way of making data available to researchers who do not necessarily have the skills to use natural language processing (NLP) libraries on a dataset, but who we believe could benefit from insights made available from annotated corpora.

[1] https://cwb.sourceforge.io/

Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing, pages 47–57

2 Related work

Discourse analysis in multimodal contexts is not a novel topic in corpus linguistics. In 1996, Kress and Van Leeuwen (1996) presented a descriptive framework entitled *Grammar of Visual Design*, whose goal was to equip researchers with a tool that would allow them to "read" visual modalities by applying a set of formal rules. The idea was to support efforts to examine the effect data objects in a format other than text, such as images, might have on composing and conveying meaning. Linguists have since approached images in multimodal corpora from several angles, including: (1) analysing and labelling the image itself; (2) conducting linguistic analysis on a caption accompanying the image; (3) simultaneously analysing an image both as a stand-alone artefact and a data object further explained by its caption, and (4) treating image captions as part of the PTO rather than a description of a multimodal object.

Various combinations of the aforementioned approaches can be found in the analysis of austerity discourse in the British press conducted by Tan et al. (2020) using a multimodal image-text corpus. Tan et al. (2020) first categorise images in four superordinate categories, before further classifying them across sixteen subcategories. Images are thus treated as independent data objects that are labeled and categorised as belonging to a certain type; the authors then look into the image-type distribution in the corpus and the associations between image-types and article-types. The analysis of Christiansen et al. (2020) distinguishes between image reference (IR) and image-text reference (ITR). Meanwhile, Bateman and Paris (2020) treat image descriptions, which are essentially DTOs, as part of the PTO when preprocessing the data for their study on changing ideological positions.

Conducting linguistic analysis on texts and captions of images embedded in texts raises the need to preprocess and ingest data in a tool that supports linguistic queries. For example, Griebel et al. (2020) preprocess textual data by annotating it with the Stanford CoreNLP pipeline (Manning et al., 2014), using its processors for tokenization[2], lemmatization, part-of-speech (POS) tagging, and named entity extraction. The linguistic annotation in this case is conducted with a single NLP library, and image captions are pointed to with the markers "captions" and "graphic". Once annotated, the data is ingested in CQPweb and made accessible to researchers of several disciplines.

The applicability of any of these methods for integrating images in discourse analysis driven by corpus linguistics is highly dependent on how images, or any other multimodal objects, are represented in a corpus. For example, an image unaccompanied by a caption cannot in itself be the subject of linguistic analysis, since there is no DTO on which such analysis would be conducted. While devising categories for images allows for both direct interaction with the data and substantial human input in its analysis, this method has limited practicality, since manual categorisation of images is both time- and resource-intensive.

This paper builds on work done by Griebel et al. (2020) and expands the coverage of DTOs to include not only image but also video descriptions. We use the markers "img" for DTOs referring to images and "vid_description" and "vid_summary" for DTOs referring to videos. In addition to presenting the potential for various corpus analyses, the paper elaborates on the steps taken to process the data, since the feasibility of various analyses and the types of questions that may be answered using a given dataset are strongly influenced by decisions made in the data processing stage. This is especially relevant if we take into account that not all researchers can access a corpus programmatically. We propose a linguistic annotation pipeline that uses multiple NLP libraries to extract attributes at token, sentence, paragraph and full-text (document) level. Section 5 showcases how attributes extracted with the linguistic processing pipeline can be used to unlock the potential for conducting corpus analyses both non-programmatically, via CQPweb, and programmatically, with Python and a Jupyter Notebook. Section 6 discusses the benefits and shortcomings of the proposed pipeline, and pinpoints areas for improvement in future work.

3 Corpus

The annotation framework has been developed and tested on the Greenpeace International subcorpus of the InsightsNet Climate Change Corpus (ICCC), a multimodal corpus on climate change described in Volkanovska et al. (2023)[3]. In the ICCC, a docu-

[2] In English texts processed with Stanford CoreNLP, a *token* is usually a word, a number, or a punctuation mark, where the boundary is the white space before and after it.

[3] Permission to use the corpus data for research purposes has been duly obtained.

ment that is *multimodal* would contain data objects in at least one modality that is not natural language text, such as video or image, either embedded in the document text or being referenced by it. The Greenpeace International subcorpus contains documents in English (n=698) from the website of Greenpeace International, of which 446 are documents with embedded images or videos; of these, 375 have images only, 3 have videos only, and 68 have both images and videos. There are 2057 images, of which 1906 are accompanied by a DTO (a caption or an alternative image description), while 151 are not. Of the 123 videos in the corpus, 117 are accompanied by a DTO. Each corpus document contains a set of properties, of which *keywords* and *keyphrases* are of special interest to the annotation pipeline. The corpus has 676879 tokens. The data objects of each document are saved as paragraphs that preserve the original HTML tag and each paragraph's order of appearance in the data source. The data object saved as a paragraph can consist of different modalities, with text, image, and video data objects making up the majority. As such, they stand in the focus of the annotation framework presented in this paper. Anchor links and iframes[4] are also types of paragraphs available in the corpus. Section 5.2 shows how this detailed structure can contribute to gaining various insights from the corpus.

Supplementing the corpus In order to provide a point of comparison and to exemplify better how the approach described in this paper can be used to analyse multimodal data, we supplement the corpus with a dataset that is of the same genre and on the same topic as the Greenpeace International subcorpus. Using the approach employed in the design of ICCC's Greenpeace International subcorpus, we collect multimodal documents on the topic of climate change from the website of the non-governmental organisation (NGO) Climate Analytics[5]. The newly-created dataset has 517 articles, of which 405 are multimodal, with 392 containing images only, one containing videos only, and 12 containing both images and videos. The total number of images is 894, of which 256 are accompanied by a caption. There are video descriptions for 31 of the 33 videos in the corpus. The corpus has 414308 tokens. Anchor links and iframes are accounted for and saved as consecutive paragraphs in the corpus structure, similarly to the Greenpeace International corpus.

[4]An iframe is an element in a webpage that embeds another webpage into the original one. The embedded webpage can also include content from social media, such as Twitter and Instagram posts.

[5]https://climateanalytics.org/

4 Annotation pipeline

As mentioned in Section 1, the annotation framework extracts linguistic and semantic information from a text object, which in this case is either a PTO or a DTO. The annotation pipeline builds on work done in Volkanovska et al. (2023), but entails a clearer delineation between the stages of annotation, generating attributes at four levels of text processing: token, sentence, paragraph, and full-text. Token-level attributes are used as CQL search criteria in CQPweb, while sentence, paragraph, and full-text attributes are utilized in programmatic data analyses.

Document keywords and keyphrases are treated as inherent attributes and used to augment annotation at paragraph, sentence, and token level. The annotation pipeline is implemented as a two-step process, comprised of main annotation and extended annotation. The former generates basic attributes (BA) and derived attributes (DA), while the latter results in extended attributes (EA). Figure 1 gives an overview of the attributes extracted at each level of annotation.

4.1 Main annotation

This section describes the libraries used to implement the main annotation and explains how basic and derived attributes for each annotated text object are obtained.

NLP libraries and processors For the annotation process, some of the NLP libraries applied in previous annotation work were used to extract linguistic attributes and named entities. The libraries include spacy-stanza[6] and Stanford CoreNLP (Manning et al., 2014)[7]. The pipeline includes the following processors: tokenization, part-of-speech (POS) tagging, lemmatization, dependency parsing, and named-entity recognition (NER). We opted for using stanza's models through spaCy's architecture because the latter allows for the application of various language models through a single NLP library.

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

[7]version 4.4.0

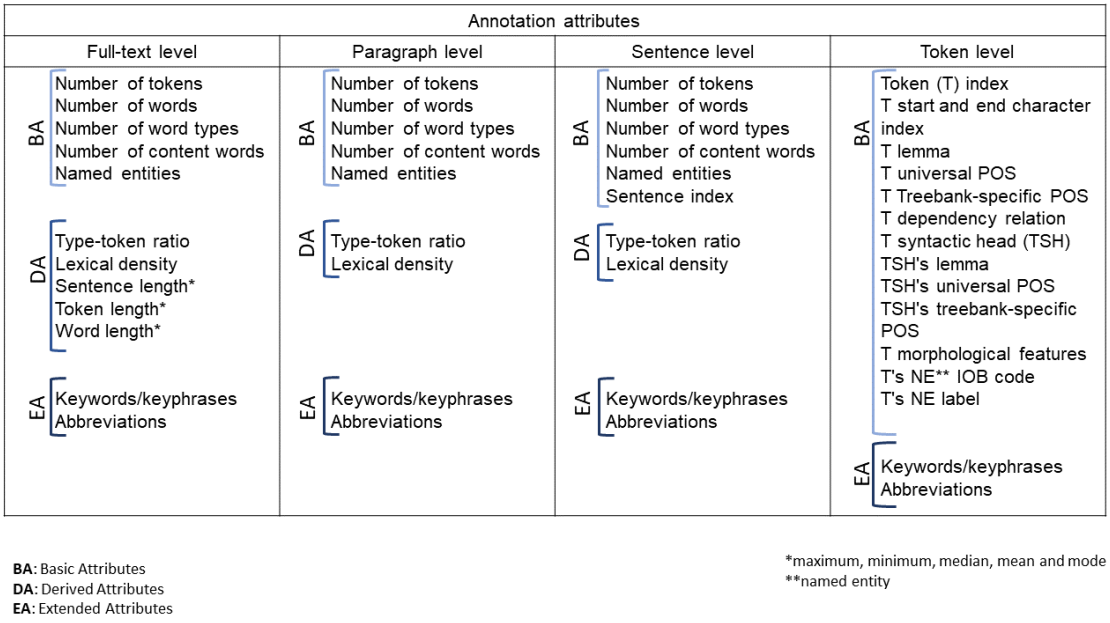


Figure 1: Attributes extracted at each level of annotation

Basic attributes Basic attributes (BAs) are retrieved either directly from the annotation output, or by applying minimum post-processing to it. Minimum post-processing refers to performing simple counts on basic attributes. Figure 1 provides an overview of BAs extracted at each annotation level. For each named entity (NE) at full-text, paragraph and sentence level we extract the properties: NE label, NE text, and frequency and position in the annotated text. At token level, we extract the token's NE inside-outside-beginning (IOB) code, and the token's NE label.

Derived attributes Derived attributes are attributes obtained by performing calculations using the previously extracted BAs at each level of annotation. At **full-text**, **paragraph**, and **sentence** level, we calculate type-token ratio and lexical density. At **full-text** level we also include statistical information about sentence, token, and word length, by calculating the maximum, minimum, median, mean, and mode length values for sentences, tokens and words of the document text.

4.2 Extended annotation

Extended annotation generates custom corpus-relevant attributes and encompasses integration of keywords and keyphrases, which are available for each document of the corpus, in paragraph-, sentence-, and token-level annotation, and extraction of abbreviations. The former is conducted with spaCy's PhraseMatcher tool, while for the latter we used the library SciSpacy (Neumann et al., 2019)[8].

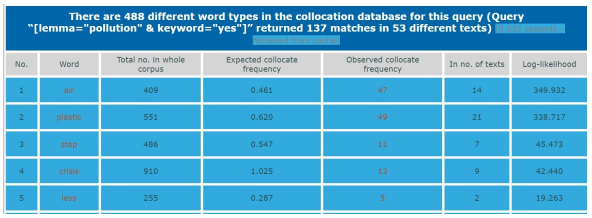
Integration of keyword/keyphrase information Each document of the corpus comes with a set of keywords and keyphrases, which we use to extend the annotations at paragraph, sentence, and token level. At paragraph level, we check if any of the given keywords/keyphrases are present and, if yes, mark their frequency. At sentence level, we annotate the keyword/keyphrase, the index or indices of the token(s) comprising it, and the start and end character index of the respective token(s). At the token level, we add the attribute "keyword" and set it to *yes* or *no* accordingly.

Abbreviation extraction At document and paragraph level, we extract abbreviations, their full form, and their frequency in the annotated text; at sentence level, we extract the token indices, and the start- and end-character index of the abbreviation in addition to its full form. At the token level, we add the attribute "abbreviation" and set it to either *yes* or *no*.

4.3 Saving the annotation pipeline output

The annotation output is saved at several stages of the annotation process. The raw output of the main annotation pipeline is serialized as a pickle file and a spaCy object. Once the basic, derived, and extended attributes are extracted, we save them within

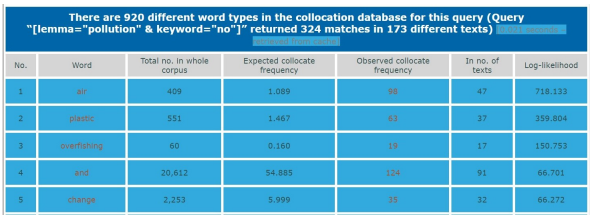
[8]https://github.com/allenai/scispacy



There are 488 different word types in the collocation database for this query (Query "[lemma="pollution" & keyword="yes"]" returned 137 matches in 53 different texts) 0.013 seconds – retrieved from cache

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	air	409	0.461	47	14	349.932
2	plastic	551	0.620	49	21	338.717
3	stop	486	0.547	11	7	45.473
4	crisis	910	1.025	13	9	42.440
5	less	255	0.287	5	2	19.263

(a)

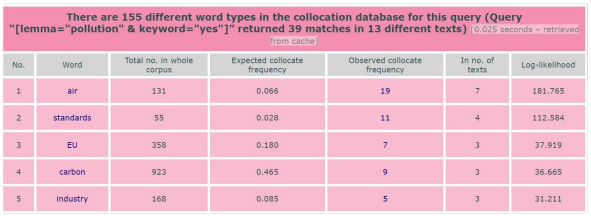


There are 920 different word types in the collocation database for this query (Query "[lemma="pollution" & keyword="no"]" returned 324 matches in 173 different texts) 0.013 seconds – retrieved from cache

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	air	409	1.089	98	47	718.133
2	plastic	551	1.467	63	37	359.804
3	overfishing	60	0.160	19	17	150.753
4	and	20,612	54.885	124	91	66.701
5	change	2,253	5.999	35	32	66.272

(b)

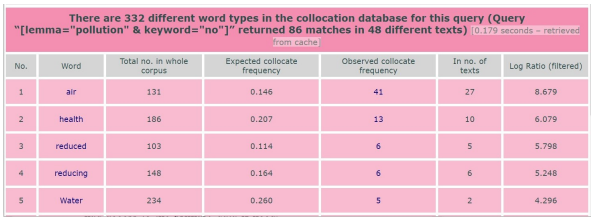
Figure 2: Collocations of the term *pollution* when the term is a keyword (2a) and when it is not a keyword (2b) in Greenpeace International.



There are 155 different word types in the collocation database for this query (Query "[lemma="pollution" & keyword="yes"]" returned 39 matches in 13 different texts) 0.025 seconds – retrieved from cache

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	air	131	0.066	19	7	181.765
2	standards	55	0.028	11	4	112.584
3	EU	358	0.180	7	3	37.919
4	carbon	923	0.465	9	3	36.665
5	industry	168	0.085	5	3	31.211

(a)



There are 332 different word types in the collocation database for this query (Query "[lemma="pollution" & keyword="no"]" returned 86 matches in 48 different texts) 0.179 seconds – retrieved from cache

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log Ratio (filtered)
1	air	131	0.146	41	27	8.679
2	health	186	0.207	13	10	6.079
3	reduced	103	0.114	6	5	5.798
4	reducing	148	0.164	6	6	5.248
5	Water	234	0.260	5	2	4.296

(b)

Figure 3: Collocations of the term *pollution* when the term is a keyword (3a) and when it is not a keyword (3b) in Climate Analytics.

the corpus document under the key *annotated content* and export the complete output as a JSON file. This file serves as a repository containing the attributes at all four levels and as such represents a source file from which files in a CQPweb-specific format can be easily created.

5 Use cases

This section exemplifies how the attributes extracted with the annotation pipeline of Section 4 can be used for performing corpus queries with the CQL and CQPweb, or to conduct deeper corpus exploration with Python and a Jupyter Notebook.

5.1 Corpus exploration with CQPweb

The annotation pipeline described in Section 4 generates an annotated corpus in a format suitable for ingestion and indexing with CQPweb [9]. According to Davies (2005), the option to query large collections of data with extensive annotations using CQL via CQPweb makes CQPweb a powerful query tool. Search queries with CQPweb can be simple, when a user enters a search term or phrase in a similar way as one would in any of the popular search engines, such as *pollute* or *forest fires*, or complex, when queries are defined with CQL using the token-level attributes listed in the column "Token level" of Figure 1. Results can be returned in different formats, such as Key Word in Context (KWiC) concordances, word frequency lists, or collocation tables. The wider textual context of the search query can also be retrieved for further examination. The objective of this use case is to test whether the detailed and extended token attributes can be indexed and searched with CQPweb, and whether we can distinguish between queries done on PTOs and DTOs.

With the basic token-level attributes listed in Section 4 and CQL, researchers can explore questions such as *Which organisations have been explicitly named as culprits of pollution in this corpus?* by extracting all sentences where the verb *pollute* is the syntactic head of a named entity with the label *ORG*[10], whose dependency relation to the verb *pollute* is that of a nominal subject[11], [12]. Another query along these lines would be to compare the number of passive sentences associated with the verb *pollute* in which the passive agent is explicitly stated to the number of agentless passive sentences. Such a query could shed a light on the circumstances in which the agent of a passive sentence is omitted[13]. Using the above-mentioned queries, we found that in the Greenpeace International corpus, only one organisation, Glencore, was openly mentioned as an organisation polluting the environment.

[9] CQPweb v3.3.17

[10] organisation

[11] CQL query: [entType="ORG" & dep="nsubj" & headLemma="pollute"]

[12] It should be borne in mind that linguistic features are extracted automatically, and careful examination of the output is necessary before making definitive claims or conclusions.

[13] CQL query for all passive sentences (1) and for passive sentences in which the agent is mentioned (2): (1) [dep="aux:pass" & headLemma="pollute"]; (2) [dep="aux:pass" & headLemma="pollute"][]*[dep="obl:agent"]

The query did not return any results from the Climate Analytics corpus. Greenpeace International had five passive sentences with the verb *pollute*, which were all agentless. Climate Analytics had three passive sentences with the same verb, which were also agentless.

Using a combination of basic and extended token-level attributes, we compare the collocates of the word *pollution* in documents in which it has been labelled as a keyword, against its collocates in documents where it is not a keyword. This can be done with CQL queries[14] and CQPweb's built-in collocation finder, which allows us to examine the queried term's collocates using one of the eight available association measures[15]. These queries can be conducted on PTOs or on DTOs; for the latter, we would need to add *within img*, *within vid_description* or *within vid_summary* in the CQL query[16]. When *pollution* is a keyword in Greenpeace International, its top-five collocates are *air, plastic, stop, crisis, less*; when it is not a keyword, it collocates with *air, plastic, overfishing, and, change*. In Climate Analytics, *pollution* as a keyword collocates with *air, standards, EU, carbon, industry* and as a non-keyword with *air, health, reduced, reducing, water*. Figures 2a and 2b, and 3a and 3b provide an overview of the query output from Greenpeace International and Climate Analytics respectively.

5.2 Corpus exploration with Python and a Jupyter Notebook

The structure yielded by the annotation pipeline described in Section 4 along with the metadata provided by the ICCC, combined into a JSON file, allows for corpus exploration by applying programmatic methods. Combining metadata and annotations can help researchers to quickly get an overview of the average statistical information contained in the DA of the annotation as well as a general overview of the metadata information; such as a plot containing years and the frequency of articles. The goal of having such a tool is to allow users to answer questions such as: *What are the keywords/keyphrases involved in Greenpeace International articles versus Climate Analytics articles in the years between 2019 and 2020? And which of those keywords/keyphrases appear in image or video DTOs and what is the link to the image/video?* Such a query is made possible by the annotation attributes and the embedded corpus structure. To answer the first question, one can count the number of keyword/keyphrase occurrences in documents belonging to the specified years of publication and compare the differences between the respective documents from each corpus, as seen in Figure 4.

The second question can be answered by choosing one of the keywords/keyphrases shown in Figure 4 and looking for the specific keyword/keyphrase that was annotated in image and video DTOs. The result with the example keyphrase *climate change* can be seen in Appendix A. The user is able to view the unique filename, the multimodal data type (image, video description or video summary), the paragraph text in which the keyphrase appears and the link to view the image or the video.

The same type of analysis can be done with the extracted entities. Figure 5 shows the comparison between organisations extracted in Greenpeace International and Climate Analytics. If the user is interested, a list of contexts where a specific entity occurs can also be obtained similar to that of Appendix A.

Accessing anchor links and iframe objects Multimodal data objects embedded in a document, such as images and videos, are usually accompanied by captions or video transcriptions. However, data that are obtained from the web, such as the corpora that are being explored in this paper, may also contain other types of data objects, such as anchor links and iframes, embedded in a document's text. These data objects are usually tricky to query as they are not accompanied by textual data of their own. One way to solve this problem would be to query for anchor links and iframes based on their context text; implying that when an anchor link or an iframe is found between two text paragraphs, it is likely that they are related to the context text rather than being standalone corpus elements. Such a query can be made possible due to the structure of the annotation and the preserved order of the data objects in which the document was obtained from the web. Another more general way to query would be to take all documents in the Greenpeace Interna-

[14] CQL queries: [lemma="pollution" & keyword="yes"], [lemma="pollution" & keyword="no"]

[15] Mutual information, MI3, Z-score, T-score, Log-likelihood, Dice-coefficient, Log-Ratio (filtered), and Conservative LR

[16] CQL query: [lemma="pollution" & keyword="yes"] within img ("img" can be replaced with "vid_description" or "vid_summary" depending on the DTO of interest).

Figure 4: Keyword/keyphrase comparison between Greenpeace International (left) and Climate Analytics (right) between the years 2019 and 2020 with top 10 keywords/keyphrases and their frequencies.

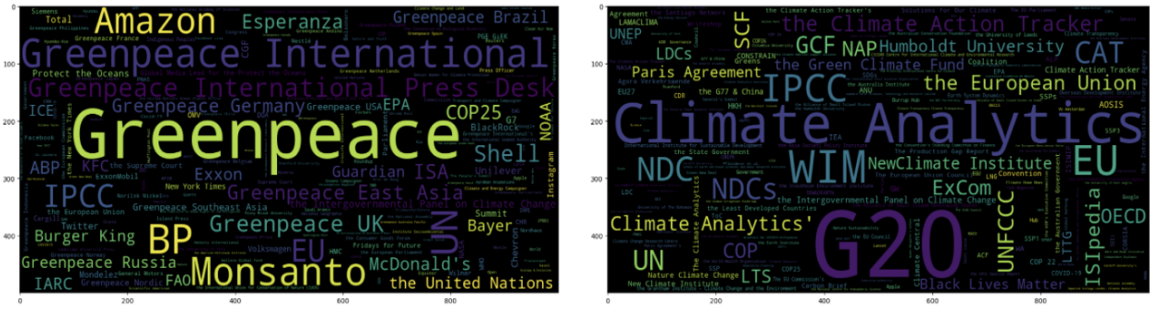


Figure 5: Entity: ORG comparison between Greenpeace International subcorpora (left) and Climate Analytics (right) between the years 2019 and 2020.

tional subcorpus with a specific keyword/keyphrase (e.g. *climate change*) within a specific year (e.g. 2019 and 2020). The tool will yield a list of anchor and iframe links and their corresponding contextual texts that satisfy the query requirements (see Appendix B for example output).

6 Conclusions

This paper demonstrates how a linguistic annotation pipeline can be applied to a multimodal corpus containing text, images, and videos, where images and videos are accompanied by textual descriptions, and how the attributes generated at various stages of annotation can support corpus analyses. Rather than introducing modality-specific attributes, the pipeline extends linguistic annotations to given descriptions of image and video data objects, thus making them accessible through the same query approach used for a document's text. We also show how a dataset annotated using our pipeline can be made available to researchers who are familiar with corpus querying techniques, but possess limited programming skills. In this section, we give a brief overview on some of the lessons learned during the annotation process, and how these can pave the way for future research in this field.

NLP researchers working with English texts have a myriad of NLP libraries at their disposal. Annotating a corpus by combining several NLP tools could generate a highly-detailed profile of a dataset, with many attributes to be used as query criteria. However, neither combining NLP tools nor making token-level attributes accessible is an easy task. For example, NLP tools could employ various tokenizers with differing interpretations of what a token is. In the context of our study, it proved challenging to reap the benefits of some Transformer-based language processing tools, whose success in tackling unseen words is to an extent due to the use of subword units[17]. In the future, we would like to explore ways of integrating annotations obtained with Transformer-based NLP libraries in the available token-level attributes. Having data of a certain size is also paramount to performing analyses. In Section 5.1 we attempted to compare the number of passive sentences with and without an agent involving a specific verb, but did not manage to retrieve a representative number of examples to analyse further due to the relatively small size of our corpus. This proved that the more fine-grained a query is, the more important the size of the corpus becomes. Finally, future work might consider storing metadata information about the annotation pipeline presented in this paper in formats that could promote the pipeline's integration in existing collections of tools for natural language processing[18].

7 Limitations

This paper presents a complex annotation framework that might not translate well into languages with fewer processing resources. It is highly likely that this type of linguistic analysis would not be fully reproducible for low-resource languages, which poses a hindrance to the transferability of this methodology at least in its full scope.

In Section 3 it was underscored that the annotation framework is only applicable to multimodal objects (images and videos) accompanied by textual descriptions. There is a marginal number of instances in which such descriptions were not readily available; consequently, it would not be possible to integrate these objects in the final analysis. This limitation could be overcome by applying image and video captioning tools, or by introducing modality-specific attributes, such as the output of object recognition techniques for images and videos. However, this is a layer of data processing that is beyond the scope of this paper.

The annotation pipeline was executed on a dedicated Nvidia GPU server. The annotation of the two corpora took approximately 360 minutes to run. The development and the running of the pipeline proved to be a computationally expensive process, which makes it potentially forbidding for researchers with limited access to such resources.

In Section 4.3 it is mentioned that the raw output of NLP libraries is serialized for the purpose of ensuring reusability of annotated texts. Loading serialized files in the respective NLP libraries and extracting additional attributes is dependent on the availability of the same version of the language model that was used in the NLP library that generated the serialized file. This could pose a limitation to reusability should the same language model no longer be available.

[17] For example, Devlin et al. (2019) use wordpieces, which are neither purely word-based nor character-based units

[18] One such example would be the XML Metadata Interchange (XMI), which is in use in DKPro, a community of projects for re-usable NLP pipelines.

Ethics Statement

An ethical consideration in this research was respecting and duly acknowledging the rights of owners of data and resources. This meant observing the conditions laid out in copyright regulations governing usage of the contents stipulated by the entity holding intellectual property rights over the data. It is necessary to point out that various data holders may apply differing constraints on data use, especially with regard to text on the one hand, and multimodal file formats on the other.

We acknowledge that using GPU computing leaves a carbon footprint. While this study does not include a training step, as is the case with the development of large language models (LLMs), we recognise the environmental consequences of GPU usage and commit to using these resources responsibly.

References

John A Bateman and Cécile L Paris. 2020. Searching for 'austerity': Using semantic shifts in word embeddings as indicators of changing ideological positions. In *Multimodal Approaches to Media Discourses*, pages 11–41. Routledge.

Monika Bednarek and Helen Caple. 2012. *News discourse*, volume 46. A&C Black.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, pages 23–32, Budapest, Hungary. tt cmp-lg: tt 9408005.

Alex Christiansen, William Dance, and Alexander Wild. 2020. Constructing corpora from images and text. *Corpus approaches to social media*, pages 149–174.

Mark Davies. 2005. The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics*, 10:307–334.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tim Griebel, Stefan Evert, and Philipp Heinrich. 2020. *Multimodal approaches to media discourses: Reconstructing the age of austerity in the United Kingdom*. Routledge.

Andrew Hardie. 2012. CQPWeb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Gunther R Kress and Theo Van Leeuwen. 1996. *Reading images: The grammar of visual design*. Psychology Press.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Sabine Tan, Kay O'Halloran, Peter Wignell, and Katharina Lobinger. 2020. Images of austerity in the british press and in online media. *Multimodal Approaches to Media Discourses: Reconstructing the Age of Austerity in the United Kingdom*, pages 134–62.

Elena Volkanovska, Sherry Tan, Changxu Duan, Sabine Bartsch, and Wolfgang Stille. 2023. The InsightsNet Climate Change Corpus (ICCC). *BTW 2023*.

Appendix

A Keyword/keyphrase in context with links to multimodal objects

cc_gp_int_204 img Hundreds of young protesters march through Central Tokyo to demand urgent action to prevent climate change (November 2019). The demonstration is part of the global movement known as Fridays for Future.
https://www.greenpeace.org/static/planet4-international-stateless/2020/04/bf786b62-gp0stugqt_medium_res_with_credit_line-1024x736.jpg
cc_gp_int_314 img SYDNEY, AUSTRALIA - MARCH 15: Protesters during a Climate Change Awareness March on March 15, 2019 outside Sydney Town Hall, Australia. The protests are part of a global climate strike, urging politicians to take urgent action on climate change. James Gourley/Getty Images
https://www.greenpeace.org/static/planet4-international-stateless/2019/03/2dd60f1c-gettyimages-1135884196.jpg
cc_gp_int_314 img PARIS, FRANCE - MARCH 16: A protester holds a sign reading "Game over" as he takes part in the "March of The Century" (La Marche du Siecle) to demand answers to climate change on March 16, 2019 in Paris, France. Several thousand people demonstrated in Paris to denounce the government's inaction on climate. Chesnot/Getty Images
https://www.greenpeace.org/static/planet4-international-stateless/2019/03/4cbe4794-gettyimages-1136214712.jpg
cc_gp_int_314 img TOKYO, JAPAN - MARCH 15: Participants hold signs and shout slogans during the Fridays for Future march on March 15, 2019 in Tokyo, Japan. Students around the world took to the streets on March 15 to protest a lack of climate awareness and demand that elected officials take action on climate change. Inspired by Greta Thunberg, the 16-year-old environmental activist who started skipping school since August 2018 to protest outside Sweden's parliament, school and university students worldwide have followed her lead and shared her alarm and anger. Takashi Aoyama/Getty Images
https://www.greenpeace.org/static/planet4-international-stateless/2019/03/5be5f84d-gettyimages-1135912723.jpg
cc_gp_int_402 img Thousands of Belgian students, for the seventh Thursday in a row, march through Brussels in order to draw attention to climate change.
https://www.greenpeace.org/static/planet4-international-stateless/2019/08/b08d4d69-gp0stt1dd_medium_res.jpg
cc_gp_int_402 img In a peaceful protest Greenpeace activists from Norway, Sweden, Denmark and Germany climb the oil rig West Hercules, located near Rypefjord village in the north of Norway, and display a banner reading "Ban New Oil". While a growing movement calling for real action on climate change is happening all over the world, Equinor's rig is preparing for a season of oil drilling in the Arctic waters of the Barents Sea.
https://www.greenpeace.org/static/planet4-international-stateless/2019/08/923c5b21-gp0stt9g6_medium_res.jpg

Figure 6: Keyphrase *climate change* in Greenpeace International subcorpus with corresponding links to multimodal objects for the years 2019 and 2020.

cc_ca_en_28 img Rural communities in the Horn of Africa Drylands like these farmers in Eritrea, depend on seasonal rainfall to sustain agriculture and are especially vulnerable to droughts which are becoming more severe due to climate change.
https://climateanalytics.org/images/w693/africa-2363380_1920.jpg
cc_ca_en_371 video -description In this webinar, the second in a series on land-climate interactions under the LAMACLIMA project, Dr Wim Thiery of the Vrije Universiteit Brussel (VUB) and Kashif Salik of the Sustainable Development Policy Institute (SDPI) provide insights into irrigation's effect on climate change and its benefits and trade-offs for local people, and discuss how LAMACLIMA, a European research project coordinated by Climate Analytics, seeks to inform the drafting of sustainable land-based adaptation and mitigation measures.
https://youtu.be/lQqvzOudNwE
cc_ca_en_382 video -summary I really want to understand what could be really helpful in tackling climate change versus what could actually just be greenwashing .
https://www.youtube-nocookie.com/embed/8i6FZqJD_mQ
cc_ca_en_382 video -summary I'm angry that some governments aren't taking their responsibilities for climate change seriously seriously .
https://www.youtube-nocookie.com/embed/8i6FZqJD_mQ
cc_ca_en_382 video -description Climate Analytics celebrates International Women's Day - Jessie Schleypen and Dr Anne Zimmer are economists, looking at climate change from different angles.
https://www.youtube-nocookie.com/embed/nvZHKtQEeIw
cc_ca_en_382 video -description Here, they tell us in their own language (Filipino and German) about their work providing evidence to persuade governments that tackling climate change is both necessary and in their own interest.
https://www.youtube-nocookie.com/embed/nvZHKtQEeIw
cc_ca_en_382 video -summary developing countries have done the least to cause climate change but have the least means to deal with its impacts .
https://www.youtube-nocookie.com/embed/-ei3fUsqxi0
cc_ca_en_382 video -summary Many of those countries have ambitious climate plans to help them to develop sustainably well adapt into the embeds of climate change so they must not face any challenge while trying to access resources from the different climate funds .
https://www.youtube-nocookie.com/embed/-ei3fUsqxi0

Figure 7: Keyphrase *climate change* in Climate Analytics subcorpus with corresponding links to multimodal objects for the years 2019 and 2020.

B Anchor links and iframes with contexual text

```
cc_gp_int_1 anchorLink https://twitter.com/intent/tweet?url=https://twitter.com/Greenpeace/status/14007844025
06870786&text=%23LetsGreenOurCities
Text paragraph before the link:  Tag your mayor
Text paragraph after the link:  Use the hashtag #LetsGreenOurCities on Twitter @tagging your mayor to demand
a greener city
--------------
cc_gp_int_1 anchorLink https://www.instagram.com/greenpeace/
Text paragraph before the link:  Spread the word
Text paragraph after the link:  Use the hashtag #LetsGreenOurCitieson Instagram stories/posts to tell us why
you do it and why is it important to have green spaces in our cities
--------------
cc_gp_int_1 anchorLink https://es.greenpeace.org/es/wp-content/uploads/sites/3/2021/05/Greening-the-City_Gree
npeace.pdf
Text paragraph before the link:  Read the report
Text paragraph after the link:  Cities should be designed and planned, taking into account the benefits of na
ture. Mayors, urban planners and public officials must share this same goal.
--------------
cc_gp_int_3 anchorLink https://www.greenpeace.org/international/act/curso-internacional-de-liderazgo-en-el-vo
luntariado/
Text paragraph before the link:  ¿Prefieres unirte en español?
Text paragraph after the link:  Who can take part?
--------------
cc_gp_int_3 anchorLink /international/act/volunteer-leadership-training/#form
Text paragraph before the link:  Registration for the April 2020 training has closed. To be informed about th
e next opportunity to join please provide us with your contact details using the form above.
Text paragraph after the link:  Questions?
--------------
cc_gp_int_4 iframe - youtube video https://www.youtube.com/embed/videoseries?list=PLClXnL5aHwxXDRjfBIK8lpohmb
09dXwJF
Text paragraph before the link:  Watch and share these eye-opening films that explain how big oil and agricul
ture firms are deceiving us through offsetting scams
Text paragraph after the link:  Offsets distort land and livelihoods
--------------
```

Figure 8: Anchor links and iframes and corresponding contextual texts for documents containing the keyphrase *climate change* in the Greenpeace International corpus between the years 2019 and 2020.