Conception of a Discourse Glossary

Enrichment of the Glossary and the Corpus Data

Anna-Janina Goecke
Universität Potsdam
Matriculation number: 777707
goecke@uni-potsdam.de

1 Motivation

This project report aims to give a brief overview of the current status of the climate change glossary project that has started in 2021. Based on the existence of other online glossaries, for instance with respect to Covid-19 terms, we wanted to create a glossary, that maps the discourse of climate change. In our project from 2021[1], we investigated the discourse actors of the current climate change discourse and built two corpora containing text data from websites which are associated with those discourse actors. The project relies on the idea that we can identify two opposing sides of discourse actors: so-called climate change *skeptics* - i.e. persons who question climate change as a result of human activities - and climate change activists. On the basis of these corpora we then extracted words of the form "KlimaX" (en. "climateX") which assembled our set of potential climate change compound candidates. In 2022, Simmel constructed a web app to display the final climate change compound words. Within the course of her project, the list of "KlimaX" words has been cleaned and reduced. The web app provides a web version of our glossary and can be found on `http://www.klimadiskurs.info`.

On the basis of what we initiated in 2021 and the work of Simmel (2022), we now created R scripts to enrich the corpora and to integrate new compound creations into our glossary data base. With these scripts we can easily extract the JSON file from the web app and add information of new compound proposals.

2 Corpus-Based Methods

For the current project[2], we re-evaluated some of the results found by Goecke (2021) and retrieved further information for the climate change compounds. One thing that changed with respect to the evaluation from 2021 is the fact, that we could now subset the results of our TF-IDF and Keyword-In-Context analysis to the climate change compounds that are contained in the final glossary. Previously, we only reported information about all words of the form "KlimaX" since the word list was not cleaned and not limited to noun-noun compound words yet. Accordingly, we can now extract sample sentences from relevant compounds and manually check the context of those.

2.1 Preprocessing

The corpus-based methods that we applied to the corpus data back in 2021 made use of `spacyr`'s lemmatization (Benoit and Matsuo, 2022) function within the preprocessing step. Since the lemma forms provided by `spacyr` did not cover our climate change compounds, we could not provide any lemmatization of the climate change compounds at the time. To tackle this problem, we now semi-automatically created a table of the lemma form and its associated word forms[3]. Table 1 shows a small subset of the full table. With the help of this table, we could now convert the compound tokens into their associated lemma forms. Consequently, the following analyses provide us with more accurate information since all occurrences of a compound word are now taken into consideration.

2.2 Frequency Analysis

Given that we now have a final list of climate change compounds, we can have a closer look at

[1] Project report and corpus files can be found here: `https://github.com/ajgoecke/climate_change_glossary`

[2] See: `https://github.com/ajgoecke/im_project`

[3] See `files/compounds.csv`. This was done in Python within the scope of another project by Goecke (2021). To retrieve the word forms, the compounds were split into the two noun constituents. The second noun's word forms were looked up via the German `WordNet` library. Next, the first constituent "Klima" was reapplied in front of each word form to generate the full compound word forms.

Lemma	Word Forms
Klimaapokalyptiker	Klimaapokalyptiker, Klimaapokalyptikern, Klimapokalyptikers
Klimaasyl	Klimaasyl, Klimaasyls, Klimaasyle, Klimaasylen
Klimafreund	Klimafreundes, Klimafreunds, Klimafreunde, Klimafreunden, Klimafreund
Klimahysterie	Klimahysterie, Klimahysterien

Table 1: The table shows four example compounds with its lemma form and the associated word forms.

the most frequent compounds that are contained in our glossary. We identify words such as *Klimawandel* (en. "climate change"), *Klimaschutz* (en. "climate protection") and *Klimapolitik* (en. "climate policy") to be present in both corpora because of their informative nature, i.e. these compounds do not carry any specific connotation. Since we want to ensure that certain words are very common in texts about climate change, we look up the word frequencies from the *Webkorpus* of the DWDS[4]. We retrieve the frequencies for the ten most frequent climate change compounds of each corpus and the findings prove our assumption: Common climate change compounds (such as the above mentioned words) are very frequent in texts about climate change, while the words that remained in our glossary carry a rather specific connotation and are only used by one of the two communities or only in certain situations.

Additionally, we re-computed the TF-IDF scores for the climate change compounds that also appear in our glossary and added a normalization to the data. The TF-IDF scores now range from 0 to 1. Since both corpora consist of a different number of total tokens, this is an important step to be able to compare the relevance of the compounds for each corpus. Figure 1 gives an overview.

[4] `https://www.dwds.de/d/korpora/web`

2.3 Context Analysis

To provide an evaluation of an example compound, we consider the word *Klimaleugner* (en. "climate change denier") here. This compound was chosen because it appeared within the Top 20 TF-IDF scores with a higher relevance for the *skeptics* corpus. We found this fact to be surprising since one could assume this compound to be primarily used in the climate change activists discourse to refer to the opposing group. Consequently, we retrieved sentences, containing this compound via the Keyword-In-Context (KWIC) function of the `quanteda` library (Benoit et al., 2018). The KWIC analysis of the keyword sentences revealed that the word itself is commonly used with quotation marks by the *skeptics* community. We interpret this fact as a form of sarcasm (see example 1).

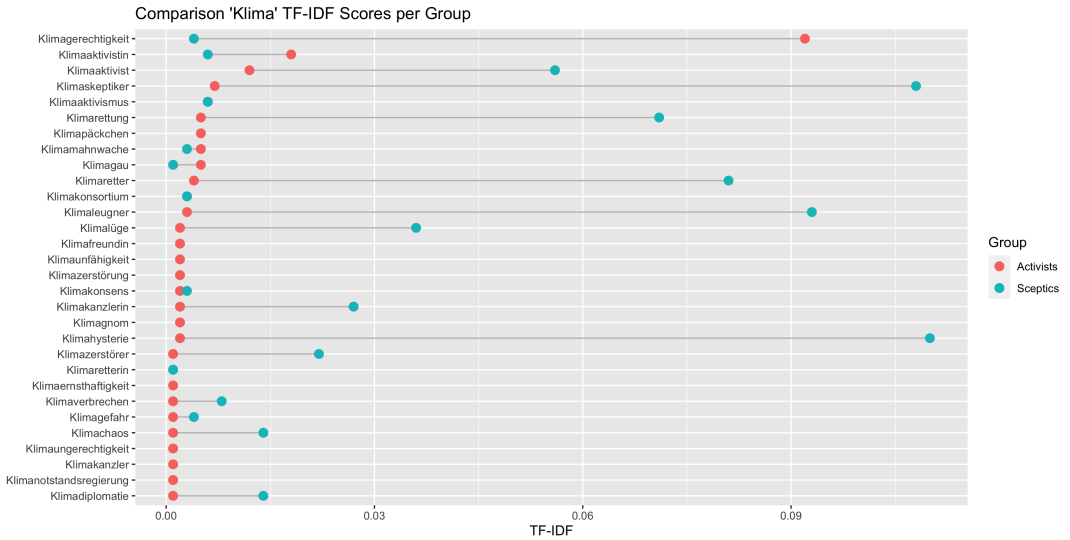

Figure 1: Normalized TF-IDF scores of the Top 20 compound words that are also part of our final glossary.

(1) Das führt dazu, dass die Vertreter abweichender Meinungen gern als "Klimaleugner" diffamiert werden.

Additionally, a more detailed bigram analysis of another example compound *Klimakrise* (en. "climate change crisis") revealed that collocations for the activists corpus are expected words such as *betroffen* (en. "affected"), *aufmerksam* (en. "mindful") and *menschengemacht* (en. "anthropogenic"). In contrast, within collocations found for the *skeptics* corpus we identify the words *angeblich* (en. "so-called"), *dramatisch* (en. "dramatic") and *imaginär* (en. "pseudo")[5]. This finding is in line with what we already conceptualized in section 1: Climate change *skeptics* tend to rather understate the urgency of climate actions.

3 Adding Data

3.1 New Corpus Data

One aim of this project is also to be able to add more data to our corpus and to enrich our glossary. The corpora that we created in 2021 consisted each of 2000 texts. While the corpus of the climate change activists (P2000) was already constructed with texts from multiple websites to increase the sample size, the *skeptics* corpus only included texts from the EIKE website[6]. Given that we want our corpus data to be diverse and to display the discourse of multiple views on the topic, more data was added to our corpora. For the *skeptics* corpus we found an online blog called *Klimaschwindel*[7] which added 14 texts to the corpus. Besides, we manually extracted articles from the *Compact-Spezial 15 Magazin* about climate change to enrich the C2000 corpus with 31 texts. The P2000 received 297 more texts from a website called *Farn*[8]. The 2022 version of the corpora now consist of 2297 texts with a total number of 1.235.021 tokens (P2022) and 2.045 texts with a total of 3.190.338 tokens (C2022). Table 2 gives an overview of the current corpus sizes and sources of the texts. To easily add new corpus data to the current corpus, we created an R script[9]. With the help of the script, textual data can be extracted from websites via the `trafilatura` tool (Barbaresi, 2021). The new texts can then be converted into a `quanteda` corpus object and be added to the corpus.

Furthermore we manually evaluated the types of text that are predominant for each corpus. To do this, a sample of 50 texts per corpus were extracted and manually assessed. While the P2022 corpus primarily consists of calls, short statements and reports, the C2022 is built up of rather scientific articles and reader's letters. This is also reflected by the fact, that the texts *skeptics* corpus have an average length of 75.9 sentences per text, while the texts in the P2022 corpus are about 24.5 sentences long.

3.2 Glossary Enrichment

The very first word list that was presented as the final result of the project from 2021 consisted of 2.967 words of the pattern "KlimaX". The terms of this list were then considered as potential candidates for our final glossary. For the glossary we seek to display compounds made up of noun components only. Since the list also contained compound words involving inter alia adjectival components, we semi-automatically cleaned the list in a few steps: Hyphens, special characters, numbers and URL formats were removed from the items. Additionally, Simmel (2022) discarded any compound words that were already existent in the online version of the German Duden[10], except for *Klimaaktivist*, *Klimaaktivistin* and *Klimawandel*[11]. Furthermore, we only considered words appearing at least twice in our corpora or are mentioned on Twitter. We manually removed compound words carrying a neutral connotation and additional word forms of the compounds. The final word list that we use as basis for our discourse glossary includes 248 climate change compounds. The web app created by Simmel (2022) provides the option for a user to propose new climate change compound words that are not in the glossary yet. To be able to work with these proposals and to easily enrich the glossary with new data, we created another R script[12]. The script takes the new compound word as an input and then retrieves multiple information from the corpus data to create a new glossary entry. It automatically detects different spellings of the word and retrieves example sentences from the corpora.

[5] For the collocation analysis we considered combinations of the form "KlimaX" + word as well as word + "KlimaX".

[6] `https://eike-klima-energie.eu`

[7] `https://klimaschwindel.net`

[8] `https://www.nf-farn.de`

[9] see `notebooks/im_add_corpus`

[10] `https://www.duden.de`

[11] Those words are particularly important for our glossary, since they provide the counterpart of the *skeptics* community.

[12] see `notebooks/im_add_compounds`

Corpus	Group	Tokens	Sentences	Source	Texts from Source
P2022	Activists	1.235.021	24,5	IKEM	1.312
				Gerechte 1 Komma 5	18
				Fridays for Future (DE)	506
				Klimafakten	36
				Klimareporter	82
				German Zero	46
				Farn	297
C2022	Skeptics	3.190.338	75,9	EIKE	2000
				Compact-Spezial 15	31
				Klimaschwindel	14

Table 2: Overview of the new corpora including the average amount of sentences per text (see column Sentences).

Moreover, it captures the community that the word can be associated with by checking occurrences of the compound in both corpora. So far, the remaining information such as ”Definition” and ”Sources” and ”Related words” stay empty until we find a way to generate these (semi-)automatically. The new glossary entry is then saved to the JSON file that we retrieve from the web app and the glossary can be updated.

4 Outlook

The scope of this project was to describe the conception of our final glossary and its current status after having performed some minor configurations. So far, we can only provide definition texts for nine of the compound words, since those texts are created manually and currently still require a lot of manual research and evaluation of corpus-based methods. In the future, it would be desirable to generate scripts that automatically retrieve more information about each compound. This could potentially be done by creating a text pattern with placeholders which then could be filled with facts from our corpus data. For instance, sentences of the keyword could reveal a sentiment that can be associated with the compound word. Moreover, co-occurrences of the compound, for instance of adjectives, could be counted and displayed for each climate change compound. Also the evaluation of entities could give us information about persons being involved in certain actions related with the compound or persons being associated with one of our two discourse actors. Such a semi-automatic approach to generate definition texts requires a lot of evaluation of the current corpus data but could automate the definition extraction for our glossary. This would be particularly useful for the proposal of new compound words and the extension of the glossary.

References

A. Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

K. Benoit and A. Matsuo. 2022. *spacyr: Wrapper to the 'spaCy' 'NLP' Library*. R package version 1.2.1.

K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo. 2018. Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

A.-J. Goecke. 2021. Discourse-oriented German Climate Change Glossary. Project report, University of Potsdam - Department of Linguistics.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

N. Simmel. 2022. Klimaretter oder Klimaspinner? Entwicklung einer Web-App zum Klimawandeldiskurs. Bachelor thesis, University of Potsdam - Department of Linguistics.