Master Thesis

Semi-Automated Definition Phrasing

Creating a Discourse Glossary for Climate Change Compounds

Anna-Janina Goecke

Department of Linguistics at University of Potsdam

February 13, 2023

Master Thesis

Semi-Automated Definition Phrasing

Creating a Discourse Glossary for Climate Change Compounds

by

Anna-Janina Goecke

Supervisors

Prof. Dr. Manfred Stede

Professorship of Applied Computational Linguistics

Prof. Dr. Birgit Schneider

Professorship of Knowledge Cultures and Media Environments

Department of Linguistics at University of Potsdam

February 13, 2023

Zusammenfassung

Der aktuelle Diskurs um den Klimawandel kann in zwei Hauptlager kategorisiert werden: Klimaforschungsvertreter und Klimaforschungsskeptiker. Während die erste Gruppe an Erkenntnisse der Klimaforschung glaubt und den Einfluss von menschlichem Handeln an der globalen Erwärmung als unmittelbar zu lösendes Problem sieht, steht die zweite Gruppe der Klimawandelforschung skeptisch gegenüber und bedient sich diverser Kommunikationsstrategien, um Zweifel am menschengemachten Klimawandel zu streuen. In einem früheren Projekt wurde mit Hilfe von zwei Textkorpora - einen für jeden der genannten Subdiskurse - ein Diskursglossar für deutsche Klimawandelkomposita erstellt. Dieses Glossar bestand bis dato aus einer Liste von 248 Komposita und Definitionstexten für ein kleines Sample der Glossarbegriffe. Die vorliegende Arbeit erstellt auf der Grundlage von Textkorpora unter Anwendung von computerbasierten, linguistischen Techniken in einem halbautomatischen Ansatz Definitionstexte für alle Begriffe des Glossars. Die Applikation von Textanalyseverfahren wie beispielsweise die Extraktion von Entitäten und die syntaktische und quantitative Untersuchung des Kontexts der Komposita resultiert in Definitionstexte, die das Glossar vervollständigen. Die explorativ angewandten Techniken erweisen sich als geeignet für die Formulierung und Erstellung der Definitionstexte und ermöglichen durch ihren halbautomatischen Charakter eine effiziente Anpassung der Glossartexte für künftige Wortneuschöpfungen im Klimawandeldiskurs.

Contents

1 Introduction

The discourse of climate change is formed by various opinions ranging from the strong belief of need for action, to neglect and the active denial of the existence of climate change resulting from human activities. The topic of climate change is very contemporary and affects all of us. The alternating theories concerning the origin of climate change that arose during the global warming controversy in the 1980's is still very relevant and the ongoing debate is present in all kinds of domains such as social media, politics and research itself. While climate research clearly provides a set of facts about the causes and actors of global warming, there is still an undeniable and rising community of people contradicting scientific evidence about climate change and refusing to acknowledge humans to be one the main contributors to global warming. Consequently, two (main) opposing positions formed, with one discourse group approving scientific evidence and facts while the other discourse community is characterised by a truth of reason.

In a previous project, text corpora were created that attempt to reflect the current discourse of climate change and can be identified as representatives of the two sub-discourses. The scope of this projects was to build a discourse-oriented glossary of German climate change compounds that is accessible online. Also, this project builds the starting point for the present work which seeks to finalise the discourse glossary by providing definition texts for each of the glossary terms. The creation of the definition texts will be addressed in a semi-automatic approach to allow for future extensions of the glossary by the addition of new terms without the need of manual definition phrasing. The nature of this project is rather exploratory in that it seeks to apply various text mining techniques and corpus-based approaches to identify methods that are appropriate for building reliable definitions for the glossary entries. Hereby this work not only concentrates on quantitative methods but also on qualitative observations that can be drawn from the context of the compounds by manual annotation and evaluation. The exploratory investigation of computer-based techniques to retrieve useful information from the text corpora is a very promising road as it (in our approach) combines quantitative and qualitative metrics.

Coupling human knowledge with automatically retrieved data from a collection of texts certainly increases the validity of findings and eventually results into solid definition texts that can be easily adapted in future projects to fit upcoming tasks. The present work will show that the computer-based techniques that are applied to the corpus data result in a knowledge base which contains various information about the glossary terms. On the basis of this knowledge base we are then able to create valuable definition texts for the compounds that can now be used to enrich the online glossary and to provide full glossary entries to the users.

This paper will give a brief overview of the origins of the climate change debate and the attempt to determine two sub-discourses within this controversy. This distinction is essential for the discourse glossary - as it builds the basis of the text corpora - and the implementation of the methodology of this project. To illustrate the development of the corpora and the glossary, chapter 2 gives a brief summary of what was done in the previous project that provides the starting point of the present work.

The upcoming sections seek to guide through the complete process of creating a knowledge base of the glossary terms by applying a broad range of corpus-based methods and text mining techniques to filter the corpora for relevant information. We report, inter alia, the application of collocation measures and the identification of concordances of the climate change compounds. The extraction of named entities is used to identify persons and organisations that can be associated with the glossary terms. Dependencies of the compounds are parsed to determine adjectives that further specify the use of the compounds in discourse. Given that the focus of this project lies on the evaluation of the context with respect to discourse, a manual annotation will be carried out to draw conclusions about whether the compounds are used in terms of a self-attribution or to refer to the opposing discourse community.

We will finally elaborate on how the information of the resulting knowledge base is arranged to construct full definition texts containing unique pieces of knowledge for each glossary term. Further, the validity of the definitions and the applied methods to retrieve information from the corpora will be discussed and evaluated. A summary of the complete project, followed by an outlook on further improvements and potential additions to the work concludes the thesis paper.

2 Glossary

2.1 Conception of a Discourse Glossary

There are various German online glossaries to cover specific domains. For instance, the *DWDS-Themenglossar for Covid-19*[1] which consists of terms that gained particular interest during the pandemic. Current issues of the world can trigger a semantic change in words or even induce new creations, so-called neologisms. Also, words that were previously known by particular communities or exclusively used in technical language can become subject to a broader audience. Accordingly, domain-specific glossaries are constantly built whenever there is the need to fill a gap. Those glossaries typically not only aim to give information about the meaning and spelling of a term, but also seek to record the changes in meaning and use over time. Regarding Covid-19, there was a prompt need of a rephrasing of certain definitions of words and the integration of terms of which the use changed due to social and political transitions during the pandemic.

The conception of our discourse glossary is strongly oriented on already existing thematic glossaries of the DWDS. While a lot of glossaries related to the topic of climate change can be found online, for instance from the NDR,[2] WELT[3] or EWE,[4] we still see the demand to propose another climate change glossary. In contrast to what other resources offer, the purpose of our glossary is to illustrate the use of intensely connoted terms in the political discourse of climate change. The upcoming section will give an overview of the terms that form the discourse glossary.

[1] https://www.dwds.de/themenglossar/Corona (Last accessed 17 Oct 2022).

[2] https://www.ndr.de/ratgeber/klimawandel/Klimawandel-Das-Glossar-von-A-bis-Z,glossar124.html (Last accessed 17 Oct 2022).

[3] https://www.welt.de/wissenschaft/article181807952/Glossar-zum-Klimawandel-Klimawandel-verstehen-das-muessen-Sie-wissen.html (Last accessed 17 Oct 2022).

[4] https://www.ewe.com/de/zukunft-gestalten/klimaschutz/klimaglossar (Last accessed 17 Oct 2022).

2.2 Climate Change Compounds

Compared to other Germanic languages, German plays a special role concerning the composition of words (Schlücker, 2012, 1). In German, composition, i.e. the building of compounds, is the most productive way to extend the vocabulary. Hence, it is obvious that we choose to incorporate this phenomenon into our thematic discourse glossary. Determinative compounds consist of two or more constituents where the final one is the morphological *head* and carries the semantic core (Glück and Rödel, 2016; Schlücker, 2012). Accordingly, the head not only contains information about the ground meaning of the compound, but is also the constituent that is being inflected and carries the genus information of the compound.[5] The left constituent, the *modificator*, specifies the meaning of the head. One characterisation of determinative compounds of the form N-N (noun-noun compounds) is the possibility to recursively apply the process of composition to form more complex elements (Schlücker, 2012, 7). According to Ortner and Müller-Bollhagen (1991) just around 10% of N-N compounds consist of three parts and only 1.5% of four parts. The first constituent of N-N compounds mostly remains in its nominative form. To derive the exact meaning of a compound it is usually not sufficient to simply evaluate the head. The exact relation of meaning can be of various types, e.g compounds such as "Haustür" (en: "door") denote a part-of relation and "Morgenkaffee" (en: "morning coffee") express a temporal relation (Schlücker, 2012, 12). Respectively, it is not always straightforward to identify the meaning of the head to derive the semantic features of the complete compound word. Gagné and Spalding (2006) investigated the relation of meaning between the constituents. They came to the conclusion that the meaning of newly-built, non-lexicalised compounds is strongly connected to the relations that the single constituents already have in existing, lexicalised compounds. Here, the lexicalisation process basically fixes one of the potential interpretations of the compound word.

With focus on the role of the terms in discourse, we particularly concentrated on compound words starting with the modificator "Klima" (en: "climate") followed by any head to build the glossary. The extraction of compounds with a noun as their first constituent is very straightforward and eases the processes for obtaining them, since we do not have any inflected forms that have to be regarded and can easily extract words starting with the desired modificator string (Langer, 1998, 2). In a previous project, corpora[6] were

[5] We will make use of this knowledge in section 5.1.1.

[6] We will elaborate on the corpora in more detail in section 3.4.

filtered for words of the pattern "KlimaX" (en: "climateX") to identify climate change compound words. The identification of those words resulted in a list of 2.967 words with the prefix *Klima* that constituted the set of candidates for the discourse glossary. Since the filter that were applied to collect those terms were simply based on the regular expression *klima**, the set of compound candidates also consisted of compound words that were not of the form N-N. For instance, terms such as *klimapolitisch* (en: "climate political") and *klimaneutral* (en: "climate neutral") were contained in the list. Thus, several cleaning steps were performed to obtain the final list of terms that compose our glossary. The cleaning procedure which was performed by Simmel (2022) inter alia tackled the automatic removal of special characters (except for hyphens). Within this step, URLs such as *klimaretter.info* and occurrences of gender-sensitive formats *klimaschützer_innen* and quotations marks within a word were removed. Words that contained digits were discarded as well. While the lemmatisation process that was implemented back then did not provide the desired output[7] this task has been carried out manually to retrieve the nominative form of each compound. Within this step all inflected forms were discarded. The goal was to primarily display words that are used in the heated debate of the opposing positions towards climate research. Accordingly, the discourse glossary aims to concentrate on climate change compounds that carry a loaded connotation. For the identification of the term's use in discourse, we seek to focus on rather specific information, e.g. the context in which a compound occurs and its use in terms of a self-attribution or to refer to the opposing discourse community. For this reason we removed words that, to us, seemed to rather carry a neutral sentiment. Furthermore, our glossary aims to fill the gap of newly composed compounds that are not already lexicalised in German. Therefore, the list was compared to the online version of the Duden[8]: Matches, i.e. words that are already included in the Duden, are considered lexicalised and were discarded from our list[9]

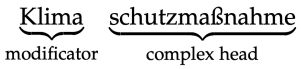

Figure 2.1: Example of the compound word *Klimaschutzmaßnahme* (en: "climate protective measures") consisting of a complex head.

[7] This will be further addressed in section 5.1.2.

[8] `https://www.duden.de` (Last accessed 17 Oct 2022).

[9] Except for the compounds *Klimaaktivismus* (en: "climate activism"), *Klimaaktivistin* (en: "female climate activist") and *Klimaaktivist* (en: "male climate activist"). Those words were kept in the list as they are an essential counterpart to the climate skeptics group.

Moreover, some of the compound candidates were built of a complex head, i.e. the head constituent consisted of more than one noun as illustrated by figure 2.1. To reduce the number of final candidates, we decided to only consider compounds that consist of not more than two parts. More complex compounds, such as the one shown in figure 2.1 were compared to the Duden again and discarded if the part following the modificator was not lexicalised. The final list that forms the basis of the climate glossary consists of 248 compound words.[10] Figure 2.2 shows a subset of the final glossary terms.

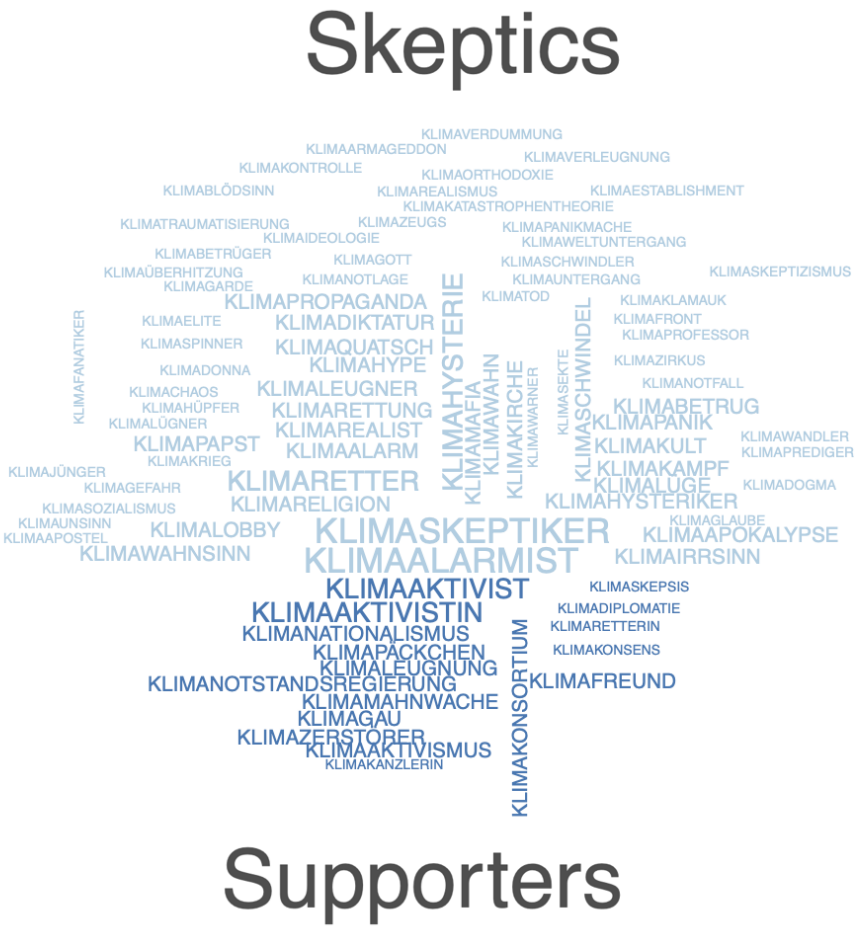

Figure 2.2: Wordcloud of a sample of the glossary terms compared by the sub-discourse in which they occur.

[10] A full list of all glossary terms can be found in the appendix.

3 The Discourse of Climate Change

3.1 Global Warming Debate

Climate change is one of the most severe environmental issues that has concerned our world not just recently, but also over the last decades. The topic firstly entered the political (and also public) sphere during the energy-policy debates of the 1970's (Dessler and Parson, 2006, 42) and was treated as a more serious problem in the 1980's when researchers determined the impact of fossil fuels on greenhouse gases. As a result, the Intergovernmental Panel on Climate Change (IPCC) was founded to collect scientific observations on global warming by gathering hundreds of scientists (Dessler and Parson, 2006, 44). This was the government's first action to address climate change, followed by the introduction of more concrete climate actions. Global warming is not easy to manage, since the actions from which the problem arises are mainly essential for economic progress. As an example, the burning of fossil fuels generates energy which increases the production of greenhouse gases. Moreover, the fact that knowledge about climate change is primarily uncertain (Dessler and Parson, 2006, 21ff.) results inter alia into disagreements on the origin of global warming and fragility of the public opinions towards the actual causes of climate change. Even though the wide-ranging discussion about predictions of what to expect is characterised by ambiguous statements, the urge of climate actions remains indisputable and emerges across all generations and social communities. The Eurobarometer poll of the European Commission in 2021,[11] for instance, reports that 92% of the surveyed persons consider climate change to be a "fairly serious problem" or even "very serious problem". Only 8% of the respondents do not acknowledge climate change as a serious problem. The Potsdam Institute of Climate Change Research (PIK) states that a survey of IPSOS Mori and the Global Commons Alliance in 2021, investigating the attitude of the public towards necessary adjustments, reveals that the majority of

[11] https://ec.europa.eu/clima/system/files/2021-06/de_climate_2021_en.pdf (Last accessed 17 Oct 2022).

respondents (73%) is aware of the urge of prompt climate actions and about 83% of the persons are willing to do more.[12] Another poll from Statista in 2022[13] concerning the commitment towards climate protection in Germany presents the following statistics: 13% of the interviewed persons believe that current climate actions are excessive, while 21% consider the actions as appropriate and 63% would like to increase the number of climate actions. The Umweltbundesamt in Germany in 2020 surveyed the environmental awareness of respondents and reported 4% of the interviewees to renounce global warming resulting from human activities and 1% of the interviewed persons to completely deny the existence of climate change.[14]

The political debate of climate change, the so-called *Global Warming Controversy*, emerged from the growing body of scientific explanations and recommendations on how to take action against global warming. The IPCC, for instance, consists of multiple researchers who originally come from various sciences somehow related to climate. Accordingly, the diverse group of scientists who investigate the consequences and origins of climate change does not always agree in opinion. As a result, citizens consistently question certain aspects of global warming. The controversy has been proven to be rather politically motivated than scientifically and has heated up over the last decade. A clear majority of people recognises the fact that climate change is a phenomenon that largely evolves from human activities and acknowledge the facts proposed by current climate research. The relatively small group of persons denying the anthropogenic nature of global warming will be described in the upcoming section.

3.2 Climate (Research) Skeptics and Deniers

The number of persons rejecting the existence of climate change resulting from human activities received very little attention for years. Recently, the community of climate change *skeptics* and *deniers* is involved more and more in the challenging discourse of global warming (Goeminne, 2012, 3). Self-appointed climate skeptics and deniers form the group of individuals who do not consider global warming to be caused by humans. This

[12] https://www.pik-potsdam.de/de/aktuelles/nachrichten/73-der-menschen-glauben-laut-einer-neuen-umfrage-dass-sich-die-erde-einem-kipppunkt-naehert (Last accessed 17 Oct 2022).

[13] https://de.statista.com/statistik/daten/studie/784401/umfrage/umfrage-zum-engagement-fuer-den-klimaschutz-in-deutschland/ (Last accessed 17 Oct 2022).

[14] https://www.umweltbundesamt.de/themen/nachhaltigkeit-strategien-internationales/umweltbewusstsein-in-deutschland (Last accessed 17 Oct 2022).

community seeks to actively spread doubt on the findings of climate research. They make use of strategies such as *framing*, which basically consists in omitting certain knowledge, and the discreditation of climate scientists and their work (Schneider, 2018, 241). Climate skeptics and deniers argue that the political engagement with this issue aims to cause fear or, simply put, is just a big lie from the government. A discussion that firstly focused on the denial of global warming turned more and more into criticism on certain climate actions (Brunnengräber, 2013, 4). According to Brunnengräber (2013), this ideologically characterised debate has its origin in the USA, where more than 60% of the citizens are convinced that global warming is not threatening humanity. While in 2012, persons such as Regine Günther (WWF Germany) still think that the group of climate skeptics can be characterised as a chaotic splinter group (Brunnengräber, 2013, 8), the trend shows that climate skepticism tries to settle down in Germany: "Die Stimmen der Skeptiker, die den menschlichen Einfluss auf das Klima abstreiten oder als unproblematisch erachten, waren und sind stets zu hören"[15] (Volken, 2010, 1). Hornschuh (2008) reports that debates of the skepticism community are continuously increasing and becoming more intense due to the arising energy revolution. According to an article of Welt Online in 2007, "Die Bewegung der Klima-Skeptiker formiert sich",[16] the community of climate skeptics primarily consists of individuals of bloggers and lobbyists of the oil and energy industry, but also of journalists and scientists. The growth of supporters of climate skepticism was also fueled by journalism. In that balanced journalism typically seeks to include opposing perspectives in their reporting to give a full picture of the discourse, it in fact made room for climate skeptics in public discussions without considering the fact that climate skepticism is not based on science itself. As stated by Germanwatch, this type of skepticism is not the constructive skepticism that is known and very welcome in scientific discourse, but rather a thorough denial of anthropogenic global warming. In the context of the climate change debate, skepticism appears uncoupled from scientific processes and rather seeks to serve the purpose of a misguiding communication tool (Schneider, 2018, 245). In some cases, the discourse of climate skeptics is characterised by hostility, intimidation and threats (Brunnengräber, 2013, 11). Stefan Rahmstorf (PIK) also supposes

[15] En: "The voices of skeptics who renounce or play down the impact of humans on climate change are constantly present."

[16] En: "The movement of climate skeptics arises", see `http://www.welt.de/wissenschaft/article1158499/Die-Bewegung-der-Klima-Skeptiker-formiert-sich.html` (Last accessed 17 Oct 2022).

the term *skeptics* to be inappropriate to refer to this community given that it is almost impossible to argue with members of the group on a factual basis (Rahmstorf, 2007).

While Dunlap (2013, 5) proposes "to think of skepticism-denial as a continuum", Brunnengräber (2013) seeks to further differentiate the notion of climate skepticism. The very first distinction between climate skeptics and climate deniers tackles the existence of global warming at all. Climate deniers reject the presence of climate change, whereas climate skeptics rather question the impact of human activities on global warming. Climate skeptics indeed acknowledge climate change but actively doubt anthropogenic CO_2 emissions to cause a majority of global climate change. He further distinguishes between climate skeptics who question the anthropogenic nature of climate change, and climate skeptics who agree on the opinion of climate change resulting from human activities but completely reject the resulting climate hysteria and climate alarmism. The latter states that reactions and claims from climate activists are exaggerated. Respectively, the term climate skeptics somewhat denotes the hypernym for the two subgroups of skeptics with positions on a continuous scale. From the nonexistent climate change, to the assumption that the sun is a key actor of global warming and the impact of the ocean on CO_2 emissions (Brunnengräber, 2013, 20). Brunnengräber (2013) even makes a more detailed subdivision of this community into *climate research skeptics*, *climate instruments skeptics* and *climate politics skeptics*. The first one is of special interest for the separation into sub-discourses that we attempt to clarify in section 3.3. The community of *climate research skeptics* is characterised by the strong rejection of scientific findings regarding effects and causes of global warming. They consistently deny statements of international climate politics and the IPCC. Rather than questioning climate change itself, *climate research skeptics* or *deniers* refuse to acknowledge the involvement of human actions in global warming. Detached from the division of climate skeptics into further sub groups is the common strategy of climate skeptics to trivialise effects of global warming and to gain public attention by spreading simple but effective information to involve people being irritated and overwhelmed by the political debate about climate change and the scientific facts provided by researchers.

3.3 Separation into Sub-Discourses

The attempt to draw a clear line between climate deniers and climate skeptics was done for the sake of completeness (with respect to the discourse of climate change). For the purpose of the glossary we decided to split the discourse into two groups of key

actors which we identify as **climate research supporters** or **climate science supporters** and **climate research skeptics** or **climate science skeptics**.[17] The latter, different from what was elaborated by Brunnengräber in 2013, also involves the community of climate research (or science) deniers as being opposing to the group of climate research supporters. Nevertheless, we decided to aggregate the two sub groups of climate research skeptics and climate research deniers to the discourse actor *climate research skeptics*, since the large proportion of information we gathered for the corpus data can rather be identified as representative for climate skepticism than denial. For simplicity, we determined the two groups of climate research supporters and climate research skeptics given that the sample texts we retrieved originally come from websites that can be linked to one of those two discourse actors. Also, the scene of skeptics is more represented in current debates about climate change.

One very well known representative of the climate skeptics scene in Germany is the "Europäisches Institut für Klima und Energie" (EIKE) which officially is not a real scientific institute but an association that is funded by private individuals and that formulates its political proclamation as follows:

> "EIKE (European Institute for Climate and Energy) is an association of a growing number of nature, humanities and economists, engineers, journalists and politicians who regard the assertion of climate change as solely „man-made“ as not scientifically rigorous and neglects known solar and other natural influences. EIKE, therefore, opposes „climate policy“ which is based solely on GHG reduction because of its negative impacts on the economy and the wider population and the consequent tax burden it creates on people, especially those in energy poverty. Within the framework of its tasks, EIKE offers members and partners a platform for the discussion and publication of scientific findings. EIKE produces appraisals at national and international levels, organizes symposia and congresses. In addition, EIKE participates in the education and education of the population and supports the establishment of political initiatives by providing scientific expertise."[18]

[17] Throughout the paper and implementation notebooks we will interchangeably use the short notations *climate skeptics* or *skeptics* and *climate supporters* or *supporters* to refer to the opposing discourse actors. However, the short notations always refer to the groups of climate **research** skeptics and supporters which build up to the two discourse communities.

[18] Retrieved from `https://eike-klima-energie.eu/about-us/` (Last accessed 17 Oct 2022).

The self-appointed institute yearly organises an international conference for energy and climate, where it gathers scientists and lobbyists from all around the world to educate public about the actual causes of climate change and to discuss political instruments related to the energy revolution in Germany.[19] The *Publications* section on the EIKE website[20] reveals multiple connections to persons who can clearly be allocated to the group of climate skeptics, among them Prof. Horst-Joachim Lüdecke and Michael Limburg. Furthermore, former members of the German party *Alternative für Deutschland* (AfD), such as Beatrix von Storch (and her husband Sven von Storch) draw clear connections between the AfD and EIKE (Brunnengräber, 2018, 285).

Opposite to the group of climate research skeptics we determine the community of climate research supporters. This community is formed by climate activists that recognise climate change as one of the most serious issues of this decade. They see the urge to react to climate change and to take actions to reduce the human impact on CO_2 emissions. Climate activists actively seek to inform public about potential effects of climate change and risks of global warming. According to Marris (2019, 471), the movement of climate activists benefits from the large proportion of young adults and children being involved. Dana Fisher states that "young people are getting so much attention that it draws more young people into the movement" and engage with climate change as a concern of global justice (Marris, 2019, 472). A movement of climate activists that is known worldwide is Fridays for Future.[21] This movement which is composed of predominantly young adults changed the public discourse on climate change significantly (Maier, 2). The movement started with the protest actions of Greta Thunberg back in 2018. Thereupon many young people follow the example she has set and participate in demonstrations to demand the fulfilment of climate protection goals. Meanwhile, in 2021 one of the largest climate strike demonstrations took place in Berlin with more than 100.000 participants. The discourse of climate research supporters, as we identify it for the purpose of the glossary, not only consist of climate activists but is rather formed by the community that approves climate research.

[19] `https://eike-klima-energie.eu/eikeik14/` (Last accessed 17 Oct 2022).
[20] See `https://eike-klima-energie.eu/publikationen/` (Last accessed 17 Oct 2022).
[21] `https://fridaysforfuture.de` (Last accessed 17 Oct 2022).

3.4 The Discourse Corpora

One essential component of this project are the discourse corpora that were initially created in 2021 and extended in 2022. In the course of a project in 2021, we created corpora for each of the key actors of the discourse on climate change which we identified in section 3.3. The corpora were built by recursively extracting texts from the websites for Fridays for Future[22] (German version) and EIKE.[23] More than 14.000 texts were gathered from EIKE and around 500 texts could be extracted from Fridays for Future. Since the size of both sources differed significantly, we decided to add more texts from websites that can be associated with the community of climate supporters, among them websites of organisations such as *GermanZero*,[24] *Gerechte 1 Komma 5*,[25] *Farn*,[26] and the institute of climate protection and mobility (*IKEM*).[27] Furthermore, websites from the journalistic projects *Klimareporter*[28] and *Klimafakten*,[29] were considered for the extension of the activists corpus. Also, to not only provide contents from a single origin for the group of climate skeptics, we determined additional sources of climate skepticism. Accordingly, the corpus of the skeptics community was enriched by material from the magazine *Compact-Spezial no. 15* which contains texts on the discourse of climate change, and an online blog named *Klimaschwindel*.[30]

The final version of the discourse corpora looks as follows: The supporter's corpus (P2022) consists of 2.297 texts with a total of 1.235.021 tokens and an average text length of 24.5 sentences. The skeptic's corpus (C2022) contains 2.045 texts with a total number of 3.190.338 tokens and an average of 75.9 sentences per text.[31] We identified the use of particular text types in both corpora to cause the unequal count of tokens and average

[22] `https://fridaysforfuture.de` (Last accessed 17 Oct 2022).
[23] `https://eike-klima-energie.eu` (Last accessed 17 Oct 2022).
[24] `https://www.germanzero.de` (Last accessed 17 Oct 2022).
[25] `https://wiki.gerechte1komma5.de/tiki-index.php` (Last accessed 17 Oct 2022).
[26] `https://www.nf-farn.de` (Last accessed 17 Oct 2022).
[27] `https://www.ikem.de` (Last accessed 17 Oct 2022).
[28] `https://www.klimareporter.de` (Last accessed 17 Oct 2022).
[29] `https://www.klimafakten.de` (Last accessed 17 Oct 2022).
[30] `https://klimaschwindel.net` (Last accessed 17 Oct 2022).
[31] We will constantly refer to both corpora using the short labels P2022, C2022 or the terms *contra* and *pro* as this is the standard notation that is used throughout the implementation part of this project. However, the use of the terms *pro* and *contra* (predominantly used in the implementation notebooks) does not reflect a rating of the author towards the two sub-discourses. This notation rather arises from the simplified attitude of the groups towards the findings of climate research.

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

sentences per text?[32] While the texts in P2022 can be allocated to text types such as calls, short statements and reports, C2022 predominantly contains scientific articles and reader's letters. Table 3.1 gives detailed information about the components of the corpora.

The scope of the corpora is to display the current discourse of climate change from both key perspectives. Correspondingly, the corpora form the basis for all terms included in the discourse glossary and, for each entry, the example sentences that are retrieved from the corpora. They will also play a big role for the work of definition phrasing (see section 7) and information retrieval (see section 5) in the upcoming implementation.

[32]This was evaluated manually on a sample of 50 texts per corpus in the previous project.

4 Theoretical Background

4.1 Named Entity Recognition

One of the key tasks of text mining is the automatic extraction of knowledge, so-called information extraction, from unstructured or semi-structured text (Aggarwal and Zhai, 2012; Jiang, 2012; Jurafsky and Martin, 2008; Mansouri et al., 2008). The notion of information extraction was firstly brought up in early Message Understanding Conferences (MUC) in the 1990's where it was denoted as a method to fill predefined templates with placeholder slots (Jiang, 2012, 12-13). While the task of template filling can be very complicated and template-dependent, in MUC-6 various subtasks of information extraction were determined to dissolve template-specificity. Accordingly, the task of information extraction nowadays rather consists of various sub-methodologies, such as relation extraction and named entity recognition, to extract knowledge from text. These components can be adapted to a particular domain and then all together build a domain-specific system of information extraction. Furthermore, the decomposition into subtasks makes it possible to treat the single tasks as simple classification problems which can be fed into supervised learning algorithms (Jiang, 2012, 14). A very well known and important subtask of information extraction is the extraction of entities. The main goal of named entity recognition (NER) is to identify proper nouns and classify them into predefined entity labels, e.g. persons, locations and institutions[33] (Liu et al., 2022; Jiang, 2012; Jurafsky and Martin, 2008; Mansouri et al., 2008). Named entities are basically sequences of words that refer to a real-word entity, e.g. "Greta Thunberg" (as a person), "Fridays for Future" (as an organisation) and "Berlin" (as a location). Since the set of entities is not closed and can be highly context-dependent, this task requires more than just simple string matching.

[33] Nowadays the extraction of named entities includes further entities such as date, currency, events and much more. Nevertheless, the very first implementations of NER base on the extraction of persons, organisations. These types are mostly language and domain-independent and also build the subset of entities we want to focus on in this work.

Generally said, we can distinguish three approaches of NER: i) Rule-based (and template-based) systems, ii) machine learning approaches and iii) deep learning systems (Liu et al., 2022, 65). Rule-based methods to extract entities rely on manually constructed rules and templates to identify and label entities. While the accuracy of this method usually exceeds human annotation with accuracy scores higher than 90%, it is a very time-consuming approach that highly relies on domain experts to generate proper rules. In a rule-based approach, each token of a text is defined by a set of features. According to Jiang (2012), these patterns, e.g. a regular expression, is matched against a sequence of tokens and triggers an action, e.g. the labeling of the sequence, once a match is found. For the second approach which relies on the machine learning domain experts are redundant since this technique does not require manually constructed rules but rather depends on annotated corpora to train a model (Liu et al., 2022, 66). With this statistical learning approach the machine learning algorithm has to deal with a sequence labeling problem where each token is treated as an observation. Here, the BIO notation (Ramshaw and Marcus, 1995) comes into play: each entity type *T* obtains the labels *B-T* (to indicate the beginning of an entity) and *I-T* (to refer to the inside of an entity); *O* marks the outside of an entity and is therefore used for tokens that do not belong to any entity (Jiang, 2012, 18). The BIO method, illustrated by table 5.4 in section 5.3.2, identifies the span of the entity and its type and is the type of method that we use in this project.

The third method which exploits deep learning algorithms arose over the last years to account for the error propagation problem that could be identified for machine learning methods. Here, Convolutional Neural Networks and Recurring Neural Networks are the predominant network choices to perform NER (Liu et al., 2022, 66).

As specified by Cunningham (2005), named entity recognition arrives at up to 95% accuracy and is therefore comparable to human performance levels given that even human annotators do not reach 100% of the accuracy level (as indicated by innotator comparisons of MUC). This fact renders NER a promising method of knowledge extraction for the present project. Due to the easy application of pretrained NER algorithms, we can simply extract entities from key word sentences. The identification of entities is useful to provide information about persons or organisations being potentially associated with a given compound word. However, given that the extraction of entities in German is not as straightforward as in English, the accuracy of current named entity models on German text is less reliable. In German, not only proper nouns are capitalised. That fact highly

increases the number of potential entity candidates. Furthermore, the relatively free word order in German can impede a precise identification of entities (Rössler, 2004).

4.2 Dependency Parsing

Dependency-based techniques to retrieve information from text has become very popular over the last decade which is also displayed by the fact that there are numerous dependency parsing tools available (Bird et al., 2009; Honnibal and Montani, 2017; Chen and Manning, 2014). The notion of syntactic dependency originates in the theoretical work of dependency grammars and is commonly used in tasks of information extraction (Nivre, 2010, 138). Dependency parsing basically denotes the automated task of extracting the syntactic relation of an input sentence to retrieve binary grammatical relationships between words (Jurafsky and Martin, 2008; Kübler et al., 2009). Those relations are commonly illustrated in a tree-like structure. The so-called dependency parse consists of *heads* and *dependents* which are connected by directed, labeled edges, and a *root* node which serves as the head of the complete dependency parse (Nivre, 2010, 140). The head-dependent relationship is essential for dependency-based approaches with the head being the central constituent of a syntactic structure and the dependents being formed by all remaining constituents that are directly or indirectly linked to the head (Jurafsky and Martin, 2008, 281). Head-dependent relationships are usually classified into various types of grammatical function, e.g. *direct object* and *nominal modifier*. A set of *Universal Dependencies* has been elaborated by Nivre et al. (2016) to provide a set of cross-linguistic annotation labels. Dependency parses can be very beneficial for languages such as German, where word order is rather flexible, since they can handle free word order and capture the relationships between words. Another feature is given by the fact that dependency relations can often be linked to semantic relationships and are therefore useful for information extraction tasks (Jurafsky and Martin, 2008, 281). While Nivre (2010) discriminates between four types of dependency parsers, i.e. *context-free dependency parsing*, *constraint dependency parsing*, *graph-based dependency parsing* and *transition-based dependency parsing*, we will only apply the latter to our corpus data. Transition-based parsers depend on machine learning techniques to derive a parse tree and achieve state-of-the-art accuracy (Nivre, 2010, 147).

In the application of dependency parsing for information extraction, Gamallo et al. (2012) used a rule-based dependency system to build semantic representations from text corpora. They rely on a three-step process to transform unstructured texts into structured

information. Firstly, the head-dependent relations are retrieved by running a dependency parser on the texts. In a next step, they extract verb clauses of the parsed phrases to classify the syntactic functions of the dependents. Then, a set of rules is applied to derive verb-based triplets. Those triplets could, for instance, be used to retrieve knowledge for specific domains regarding question answering applications or definition extraction systems.

More work on dependency parsing for definition extraction has been made by Espinosa-Anke and Saggion (2014). They seek to automate the process of manually defined pattern-matching rules to extract definitional sentences. The approach is based on only syntactic features from head-dependent relationships which already have been proven in previous information extraction tasks and rely solely on machine learning algorithms. They consider two types of dependence subtrees which allow to extract sequences that are of particular interest for the retrieval of hypernym indicators (e.g. "X *is a type of* Y") and multiword terminology (e.g. "X is a *segmental writing system*"). With their work they could prove the assumption that syntactic features are essential to classify phrases as definitional or non-definitional.

Regarding question answering tasks, systems tend to retrieve incorrect passages due to the lack of taking dependency relations into consideration. To tackle this issue, Cui et al. (2005) propose statistically-based "fuzzy" relation detection. Different to other approaches where exact dependency relation matches were used to extract answers, they use a statistical method to extract relations in candidate phrases that prefers sentences with similar semantic content between question terms. Their approach seems to exploit the nature of language to be variable which could probably cause misdetections in question answering systems. With the approach to also involve broader matches they obtain more sentence candidates which are then being ranked with respect to their likelihood. This fuzzy matching system indeed outperforms exact-matching approaches for answer extraction.

In the upcoming implementation we will apply dependency parsing to determine specific relations between a compound words and its dependents. Correspondingly, sentences containing the key word will be parsed to retrieve all dependents of the key word. Similar to the above mentioned work on dependency parsers, we seek to use the knowledge that we retrieve via dependency parsing to enrich our glossary and provide further information of the syntactic context of the compound. Even though we do not attempt to pursue a classic question answering task, we indeed want to fill information

gaps in our glossary for which we will use dependency parses of the sentences. As we will see in section 4.3, our data does not consist of typical definitional phrases. We therefore do not have to tackle a classification problem here but aim to simply extract information from the key word phrases. Likewise Gamallo et al. (2012), we are going to retrieve dependents of heads that are interesting to us and look for recurring patterns.

4.3 Glossary Building and Definition Extraction

A glossary is "an alphabetical list of difficult, technical, or foreign words in a text along with explanations of their meaning"?[34] According to Velardi et al. (2008), glossaries are a very useful tool to connect information and to solve semantic ambiguities of a specific term. To infer a meaning of a term, we often refer to the context of a word. For this reason, it just makes sense to follow a corpus-based approach for the detection of meaning and extraction of information from a word's context. The manual construction of glossaries can be very time-consuming. Therefore, a corpus-based approach to extract significant terms is a contemporary solution to save hours of manual work by automating the process (Velardi et al., 2008, 18). Furthermore, it simplifies and homogenises the recurring process of extracting terms from textual data as new terms may appear over time. Respectively, the automated process of extracting terms from a corpus can significantly speed up the procedure by minimising the temporal aspect (Velardi et al., 2008, 18). In the work of Velardi et al. (2008), the glossary generation process consists of two phases. Firstly, to extract the terms, relevant documents are gathered and fed into a so-called *TermExtractor* system which returns a list of term candidates. In a second step the *GlossExtractor* retrieves information about each term from the web to provide potential definitions. In contrast to other term extraction systems, Velardi et al. (2008) add a stylistic filter to ensure the selection of well-formed definitions, containing information about the concepts a term can be associated with and a specification of the concept.

In another approach of glossary building, Rambousek et al. (2014) present the methodology to generate a terminologically-oriented thesaurus from a domain corpus. They use a domain-specific corpus to guarantee that extracted terms are strongly connected to the scientific area that they want to explore and extend with new terms. They describe the tools to build a corpus text data retrieved from websites related to the domain. Similar to the approach that was used to construct the discourse corpora that we will use in

[34] `https://dictionary.cambridge.org/dictionary/english/glossary` (Last accessed 17 Oct 2022).

the present project, Rambousek et al. (2014, 2) identify a set of websites of the domain which they recursively scrape for additional text data. In a next step, the corpus is used to retrieve key term candidates associated with the subject. From these candidates, they build a domain-specific thesaurus by integrating several resources and giving information about the meaning and translation of each term.

One thing that all glossary creation techniques based on corpora have in common is the essential role of the corpus data. The sources which are used to build a corpus are of great relevance. For that reason, the generation of the climate change corpora described in section 3.4 relied on websites being representative for the sub-discourses. This ensured that the desired climate change compounds are over-represented in the corpora. In contrast to what was done by Rambousek et al. (2014) and Velardi et al. (2008), the extraction of terms for our glossary did not start with *unknown* terms. Exact patterns, namely words of the form *KlimaX*, were used to identify term candidates. Different from the above mentioned approaches, we could make use of key word extraction techniques to filter the corpora for words of the desired pattern. Nevertheless, since this process was performed automatically, a manual cleaning step was inevitable to ensure the extracted compounds consist of exclusively noun components (see section 2 for more information).

Also, the extraction of definitions from a corpus can be a helpful tool to automate the glossary generation process by extending entries with their meaning or use in context. By definition extraction, we refer to the subtopic of information extraction that seeks to in determine sentences of the data which contain a potential definition for a given term (Veyseh et al., 2020). To tackle the problem of automatic definition extraction, Klavans and Muresan (2000) establish the rule-based system *DEFINDER*. Their system is originally thought to extract medical terminology and their associated definitions from websites but they also propose to use it for several tasks such as summarisation and text categorisation. An important aspect of their work is the separation into two stages: The output is being generated by extracting text patterns which are then input of a natural language parser to check for more complex linguistic structures. Muresan and Klavans (2002) identify definitions to be the most crucial part of any dictionary-like application. While most glossaries are manually build by human experts it is desirable to develop processes to automate the creation of glossaries and the according definition extraction.

More work on this has been done by Westerhout (2010) who adopted the pattern-based technique by Muresan and Klavans (2002) and combined it with a machine learning system to create a glossary for eLearning platforms. She identifies the distinction between

definitional and non-definitional phrases as the biggest issue of definition extraction tasks. Accordingly, she introduces a sequence of pattern-based and machine learning based techniques to tackle the determination of proper definition strings. For the first step, text phrases are matched against definition patterns to retrieve a set of definitional sentences. Given that the patterns she uses for the classification task, e.g. "X is a", may also occur in non-definitional text passages, the sentence candidates from the first step are fed into a machine learning algorithm to filter out non-definitions. For the pattern-based approach, Westerhout (2010) makes use of linguistic information such as part-of-speech tags, lemma forms and morpho-syntactic knowledge.

To automatically extract definitions in portuguese, Del Gaudio and Branco (2007) develop a rule-based system. Like Westerhout (2010) their system exploits morpho-syntactic information such as part-of-speech tags and inflectional details to construct a glossary of for eLearning purposes. They manually annotate definitions in the data and assigned specific definition types. Also, they applied machine learning algorithms to increase the number of correctly identified definition phrases.

Spala et al. (2019) build the DEFT corpus which is specifically attuned to the purpose of definition extraction. It consists of sentences from contract filings, with a coverage of 22% of the phrases being definitions, and sentences from open source textbooks. 28% of the sentences of the latter contain definitional phrases. With DEFT, they provide one of the largest and most complex corpora used for definition extraction. One issue they encountered within their definition extraction task was the fact that about 50% of the term-definition pairs are spread across sentences boundaries or are part of complex linguistic structures. Furthermore definition may consist of complete sentences or ony smaller NP phrases. One particular information that we will make use in our upcoming application of information extraction is the detail about the composition of the DEFT corpus. It not only consists of single sentences incorporating the definition phrase but is composed of so-called "context-windows" containing the definition sentence, the preceding sentence and the following sentence. Similar to this, we are also going to explore the surrounding context of a given key word for the extraction of information. While the context of the DEFT corpus is made up of *3-sentence-pairs*, we go one step further and examine context-windows within five preceding sentences, the sentence containing the key word itself, and the five sentences following the key word phrase. The larger context-window is chosen to increase the chance of retrieving more information about the compound words. This is particularly useful for tasks such as sentiment analysis. Given that very

often, the polarity a concept is associated with can hardly be determined within a single sentence boundary. Spala et al. (2020) mention that the likelihood of definitions or defining elements being further away than just the *3-sentence window* is actually very high. They decided on the *3-sentence window* because of the manual annotation that is carried out on DEFT. The annotation of larger sentence windows could easily get very time-consuming and expensive. Since we predominantly apply automated text mining and corpus-based techniques on the corpora to retrieve information, a window of five sentences before and after the key word phrase is beneficial to observe long-distance relationships of the key words and to obtain additional information from the text data.

While there is a wide range of approaches that have been made on glossary creation (Velardi et al., 2008; Rambousek et al., 2014) and in particular on definition extraction (Westerhout, 2010; Klavans and Muresan, 2000; Muresan and Klavans, 2002; Spala et al., 2019), our approach differs in some points. Although we are also interested in automating the process of definition extraction and key word retrieval for our glossary, the data from which the corpora were built is actually diverse from the data that was used in the procedures we elaborated above. The corpora are built to fulfill the main purpose of illustrating the discourse of specific key words. Since they were obtained via web scraping they contain a lot of noisy text data and a larger proportion of texts does not consist of definitional phrases. In truth, the corpus texts represent the climate compound words in their *natural* discourse.

4.4 Word Similarity

One approach to obtain the semantic relatedness of two words is the computation of the similarity of those words. A similarity measure basically evaluates how identical two concepts are with respect to their relation in an hierarchy such as the one provided by the lexical database WordNet (Fellbaum, 2010). This database is commonly used in text mining applications given that it maps nouns and verbs into tree-like, hierarchical structures consisting of *is-a* connections. With over 80.000 entries for noun concepts divided into nine hierarchies, WordNet is one of the largest resources for the creation of knowledge bases and the retrieval of word senses and relations. The *is-a* relations do not overcome part of speech boundaries. Therefore, the similarity measures offered by WordNet are restricted to words belonging to the same part of speech category (Pedersen et al., 2004, 1024). In our case, since we focus on compound words, we only need to

compare and compute similarities of nouns. The word similarities that can be obtained via WordNet are, inter alia, computed on the base of path lengths of the hierarchical structure.[35] One of the measures, `path`, computes the inverse of the shortest path length between the two concepts, as illustrated in equation 4.1, where p denotes the length of the shortest path between the two concepts. The score ranges between 0.0 and 1.0 with the latter denoting a high similarity and a score of 0.0 indicating that there is no path between the two concepts.

$$\frac{1}{p+1} \tag{4.1}$$

The Wu-Palmer similarity (Wu and Palmer, 1994), also called `WUP`, searches a path to the root node from the least common subsumer (LCS) of both words which equals to the lowest common hypernym between the two concepts (Pedersen et al., 2004). The exact computation of the Wu-Palmer similarity is shown in equation 4.2, where i is the shortest path distance from the first concept to LCS, j is the shortest path from the second concept to the LCS and k is the number of nodes (distance + 1) from the LCS to the root node.

$$\frac{2k}{i+j+2k} \tag{4.2}$$

Accordingly, the Wu-Palmer similarity evaluates the similarity of two concepts based on the relative path in the hierarchical structure. Similar to the `path` function, the score of the `WUP` computation is greater than 0.0 and smaller or equal to 1.0. The score cannot be equal to 0.0 since the depth of the LCS is never zero. These two measures are known as taxonomy-based metrics since they seek to find a path, or better to say distances, with respect to the hypernym-hyponym relations.[36]

The computation of such similarities is particularly useful for the upcoming implementation section. Since we only have a list of compound words so far for which we want to determine additional information, discovering potential relations between the compounds is fundamental for building a proper glossary. We are going to investigate the relationships between the concepts of the compound words obtained by the `path` and `WUP` measures. With the help of these measures we are able to identify words that are strongly

[35] The following equations and information on the similarity computations are taken from `https://wn.readthedocs.io/en/latest/api/wn.similarity.html#path-similarity` (Last accessed 17 Oct 2022).

[36] There is also another measure available by Leacock and Chodorow (1998) which we do not consider here as it finds the shortest path and scales it by the maximum path length. Since we want our computation to be in a comparable range (0-1) we only evaluate the `path` and Wu-Palmer similarity.

semantically connected. Obtaining this kind of knowledge is beneficial for generating appropriate definition texts for each glossary entry. Approaches by Li et al. (2006) even adapt the computation of similarities between words to the finding of similarity between complete sentences.

4.5 Sentiment Analysis

The growing body of literature on sentiment classification reflects its current status in research: Sentiment analysis is a very popular text mining technique to evaluate data with respect to its polarity which can be positive, negative or neutral. For many companies it is a useful tool to automatically derive peoples feelings towards a new product in almost real time. Since sentiment analysis relies on emotionally charged expressions in the input text, its main application is on subjective content. The identification of subjectivity in text can be challenging. Commonly, subjectivity is a rather implied feature of discourse that cannot always be determined unambiguously. Simplistic approaches of sentiment analysis aim to classify the input data into two predominant polarities: positive and negative. The output of the analysis can either be a score on a continuous scale, indicating a range of polarity, or appears as the discrete labels *positive* and *negative* (Mejova, 2009; Ahmad et al., 2017). The input units can either be complete documents or single paragraphs or phrases. Given that oftentimes not every piece of text in a document carries a sentiment or is subjective at all, it is indispensable to make a decision on the sentiment units to be analysed with respect to the type of text. Furthermore, the polarity expressed in a document is not always homogeneous: according to Mejova (2009), the distinction between explicitly subjective phrases (ex. 1) and text parts containing implicit subjective statements (ex. 2) can be crucial for the labeling procedure.

(1) "I really loved the weather today."

(2) "The camera is too big."

While in example 1 the word "loved" clearly expresses a positive sentiment towards the object "weather", the sentence in example 2 does not carry any word that obviously triggers a sentiment. The characteristic of subjectivity to be highly context-sensitive (Mejova, 2009, 5) is illustrated by this example. Indeed, example 2 seeks to express negative sentiment towards the product. The use of "too" suggests that something is not in the expected

state and through world knowledge we understand that in the context of a camera, the adjective "big" is probably not a desired characteristic. The *trigger words* that are typically contained in subjective phrases such as the one in example 1, this type of sentence is usually easier to evaluate with respect to its polarity. Another linguistic marker that can increase the level of difficulty when obtaining polarity labels is the use of negations. They have to be handled somehow to make sure sentences, such as example 3 are not evaluated as being positive due to the occurrence of the adjective "pretty".

(3) "The flower is not pretty."

Most of the previous studies on sentiment use data from product or movie reviews (Hu and Liu, 2012). This arises from the fact that this type of data either contains a lot of polarity cues and the fact that algorithms can easily be trained on the star rating which is oftentimes included in product review systems. In contrast to this, evaluating the sentiment of politically oriented documents is usually not as easy. Different to the text type of reviews, political texts commonly contain pragmatic categories such as sarcasm, irony or metaphors. Furthermore, quotations of references to other organisations make the correct evaluation of sentiment very difficult (Mejova, 2009, 7). Gamon et al. (2008) for instance, also refer to the argumentative structure of political texts and whether the quotation or reference was added in support of the author's argument, or in opposition to it.

Most techniques that tackle the identification of sentiment are based on statistical approaches including the application of machine learning algorithms (Ahmad et al., 2017, 2). These algorithms treat the task of sentiment analysis as a classification problem. In most cases the algorithm is trained on labeled data and generates predictions on newly encountered data based on the labeled training set. Another very popular approach has been made with lexicon-based techniques (Nielsen, 2011; Hutto and Gilbert, 2014; Young and Soroka, 2012). However, this approach is very sensitive to the domain of the text and requires a lot of manual work to build proper sentiment lexicons.

For the upcoming implementation, we make use of a very simplistic approach that differentiates between the discrete polarity labels *positive*, *negative* and *neutral*. The latter indicates that a text either does not carry a specific sentiment and is therefore considered to be objective or the text consists of mixed polarity without any label being predominant and is therefore evaluated as neutral. The application of sentiment analysis to the corpora is very interesting with regard to the context of the compound words. Obtaining the

polarity of the context of a word could potentially provide further details about the connotation of a key word which can then be used as an additional information in the discourse glossary. The level of analysis in our case is formed by the context window of the key words which we defined by the five preceding and the five following sentences. As mentioned before, identifying the sentiment of phrases consisting of implicit subjectivity is not always straightforward. Accordingly, the sentiment labels that we obtain in section 5.3.4 should be treated with caution. The corpora contain a large amount of scientific papers and articles in general, that is why a major part of the texts is rather objective and potentially does not carry any polarity cues, i.e. sentiment-laden terms. The diffused use of quotations, references to other articles or to the opposing community complicates the attribution of authorship and the correct classification of polarity labels. As stated by Mejova (2009), the sentiment for the type of discourse we seek to evaluate is particularly challenging. We discovered that for the discourse corpora the evaluation of adjectives and adverbs may be of particular interest for obtaining cues about the feeling of an author towards a concept. Accordingly, we will also focus on these types of part of speech for the upcoming implementation part.

4.6 Corpus-Based Methods

While the above mentioned techniques can be counted towards the research field of text mining, corpus-based methods exploit the idea of working with a corpus directly to derive and analyse information (Stefanowitsch, 2020, 22). With information, we refer to systematic patterns and the use of linguistic features within a text corpus (Biber, 2012, 1). Stefanowitsch (2020, 49) defines *Corpus Linguistics* inter alia as the "investigation of linguistic phenomena on the basis of linguistic corpora" and Cheng (2013) refers to *Corpus Linguistics* as a research field exploiting corpora as the center of linguistic analysis. Krishnamurthy and Teubert (2007, 6) describe *Corpus Linguistics* as a bottom-up approach which "tries to accommodate the full evidence of the corpus [and] analyses the evidence with the aim of finding probabilities, trends, patterns, co-occurrences of elements, features or groupings of features". Accordingly, the research field of *Corpus Linguistics* not only depends on qualitative methodologies but also on the application of quantitative processes. It seeks to explore the distribution of grammatical features and has gained more attention in the research of discourse which makes use of the tools originating from *Corpus Linguistics* (Biber, 2012, 3). Also, its research methodologies (statistically) investigate the

creation and analysis of frequency-based word lists, key words, collocations and concordances. However, corpus-based methods primarily concentrate on quantitative techniques to evaluate the context of words (Cheng, 2013, 1), whereas methods of discourse analysis focus "on the contents expressed by language" (Flowerdew, 2012, 175). For the purpose of this project, we want to use corpus-based methods to further analyse the context of terms being present in the glossary.

4.6.1 Collocations

Very popular for the investigation of linguistic phenomena under the umbrella of corpus-based methods is the examination of collocations. The concept of collocations was first introduced by Firth (1957) to refer to commonly reoccurring word combinations. Since then, the notion of collocations has spread widely and has become established within computational lexicographers for the creation of corpus-based dictionaries (Evert, 2009, 2). Just like Evert (2009) we will use the term *collocations* in its narrow and initial sense, where it refers to a sequence of two (or more) words that very likely occur close to each other in natural discourse. As specified by Grundmann and Krishnamurthy (2010), collocations can be specified as a linguistic feature for which the selection of one expression favours the co-appearance of other words within a specific context window. The so-called co-occurrence of words can be investigated in several ways (Evert, 2009, 4f.):

1. *Surface co-occurrence*, for instance, denotes an approach where words are said to be co-occurring if they are commonly found to be next to each other in a text.
2. *Textual co-occurrence* rather concentrates on the co-occurrence of words within the same clause, paragraph or document.
3. *Syntactic co-occurrence* deals with syntactic relations between words that appear adjacently, e.g. the co-occurrence of nouns and adjectives.

The determination of co-occurring words, i.e. collocates, of a term typically involves frequency lists of the words occurring adjacent to the term of interest. These lists of collocations can then be used to draw conclusions about the circumstances in which a key word appears, or even to conceptualise the context of an expression (Stefanowitsch, 2020, 50f.). Accordingly, collocation lists are a quantitative approach to evaluate the distribution of a key word and a straightforward way to immediately capture the context in which the key word typically occurs. The sequence of words that we examine in

collocations is restricted by grammatical, semantic and topical factors: A determiner typically precedes a noun (grammatical factor), a verb such as *drink* triggers a direct object that refers to a liquid entity (semantic factor), and co-occurring words typically refer to related contents and therefore stem from the same discourse domain (Stefanowitsch, 2020, 215f.). Correspondingly, co-occurring words can be of particular relevance for the discourse or they can be somehow specific to the discourse and therefore provide further information about the discourse. Collocation analysis denotes the discovery of expressions that frequently occur adjacent to each other, i.e. the distribution of words. While there is no common agreement on how to identify collocations in current research, we are going to treat co-occurrence as a purely structural phenomenon (Stefanowitsch, 2020, 220) to extract words appearing next to our key words.

Current applications of corpus-based collocation analysis have been made by Orenha-Ottaiano et al. (2022). They identify co-occurrences to generate a multilingual collocations dictionary (PLATCOL). The approach combines the automated retrieval of potential collocation candidates with a manual annotation of lexicographers. With a focus on the morpho-syntactic classes *noun*, *verb* and *adjective* they firstly extract the according lemma forms from a corpus which are then being manually validated by humans to reduce the candidate collocations to the most frequent lemmas.

Furthermore, Grundmann and Krishnamurthy (2010) investigate the discourse of climate change by applying corpus-based collocation measures. The work compares the discourse of four countries: the UK, the US, Germany and France. Collocation lists are generated for a context window of five words occurring to the left or right of a key word (e.g. "change") in a corpus consisting of articles related to the topic of climate change. The results show a difference in the usage of the words between the US and the UK, with the US reporting about climate change in a rather neutral tone with special regard to the scientific and political aspects of climate change. In contrast, the articles from the UK reveal a particular emphasis on climate actions. Further findings consist in the observable difference in the style of media reporting across the considered countries. The articles obtained from French and German sources express the urge of political actions towards global warming, whereas the US discourse is characterised by its scientific setting.

With respect to the upcoming work of definition phrasing, we seek to identify words that commonly co-occur with the compounds of the glossary. Those collocations can provide information about the context in which a word is used. The analysis of co-occurrences can be beneficial for the process of information extraction, since the quantitative nature

of this corpus-based method is an automated way of obtaining co-occurring words and their frequencies that potentially renders the need of manual annotation steps superfluous. Besides, we can easily hand out very frequent collocations to the user by integrating it into the glossary web app. Considering the purpose of the glossary to reflect the discourse and contextual use of each compound word, this information can be crucial to further determine the context of the word. Different to what was proposed by Stefanowitsch (2020), namely the observation of differences for the frequencies of co-occurrences within a text, to us the comparison of compounds and their collocates between both discourse corpora is essential to draw conclusions from the use of the words for each of the discourse actors. Nevertheless, the evaluation process of this method still requires a lot of manual annotation work since the actual importance of certain collocations cannot be retrieved otherwise.

4.6.2 Concordances

Another method that arose from the field of *Corpus Linguistics* is the analysis of concordance, i.e. the "collection of the occurrences of a word-form, each in its textual environment" (Sinclair and Carter, 1991, 32). Concordance setups are one of the main methods of *Corpus Linguistics* (Conrad, 2002; McEnery and Hardie, 2011). By concordances, we refer to a *Key Word in Context* (KWIC) analysis which basically arranges a key word in its adjacent context. This context can either consist of a window of adjacent sentences which occur to the left or right of the key word, or a predefined window of individual words appearing on either side of the key word (Stefanowitsch, 2020; McEnery and Hardie, 2011). The advantage of the identification of *concordancers*, i.e. words or sentences appearing in the context window, lies in the immediate overview of the typical use of the key word one can gain. Since these concordancers can be automatically sorted according to their frequency, a list of frequently co-occurring items can be obtained. Stefanowitsch (2020) describes the KWIC concordances as a visual presentation of the distribution of a key word within different contexts. This visual presentation is given by a concordance line that aligns the key word and its context in a format which makes it possible to easily capture specific properties of the context.

Displaying the context of target words in this way can be crucial for the verification and identification of contextual characteristics. Barlow (2004) defines this type of visualisation procedure as a transformation of text that enables different perspectives on the data

and eases the process of finding patterns of co-occurrences from which we can draw conclusions about associations of the key word. A classic example has been made by Gries (2001) who compares adjectives of the form *-ic* and *-ical*.

The type of concordances which we seek to analyse is rather diverse from the ones that are commonly examined by researchers of the field. We not only investigate the direct context of a target word but decided for a window of five sentences to the left and right of the key word phrase. Accordingly, we are not primarily interested in the beneficial alignment of concordances but attempt to explore the wider context of each key word to check for peculiarities and to use the output of the alignment as an input for the application of the above mentioned text mining techniques.

5 Implementation

For the implementation of the computer-based methods that were elaborated and reviewed in section 4, we will follow the outline that is shown in figure 5.1. A general overview of the coding procedure will be given in the following. The upcoming sections provide a detailed insight in the exact processes that are involved in the individual methodologies.

Firstly, the list of final glossary terms and the most current version of the discourse corpora is loaded to Python and R to be able to perform several techniques on the data. The data is going to be preprocessed to bring the compounds and the corpora into a normalised format that can be used as an input for the upcoming steps. The output of the preprocessing can then easily be fed into text mining and corpus-based methods. This step is highly domain-dependent and is non-identical for the list of glossary terms and the corpora. Both approaches require different standards for the implementation. The normalisation of the glossary terms takes place in Python to account for the following text mining methods. The corpora are preprocessed in R for the upcoming corpus-based techniques. After the preprocessing we will perform several corpus-based methods (see figure 5.1, shown by the path in red): The collection and frequency-based evaluation of common collocations and the retrieval of concordances (KWIC), as well as the retrieval of term frequencies and TF-IDF scores for each compound. For the extraction of KWIC tables we rely on a window of five sentences to the left and right of the phrase containing the key word. The output of the KWIC extraction is then used as input to some of the text mining techniques. With respect to the text mining methods that are carried out in Python (see figure 5.1, illustrated by the blue-colored path) we firstly retrieve definitions and hypernyms and compute word similarities for the compound words. Also, we are going to perform stemming to retrieve words that share the same stem, e.g. *Klimalügner* (en: "climate liar") and *Klimalüge* (en: "climate lie"). Afterwards, based on the output of the KWIC windows that were previously created in R, we will extract named entities from the 5-sentence context-window of each glossary term. In addition, we use a dependency parser to extract significant dependencies and modifiers of the compounds. In a last step,

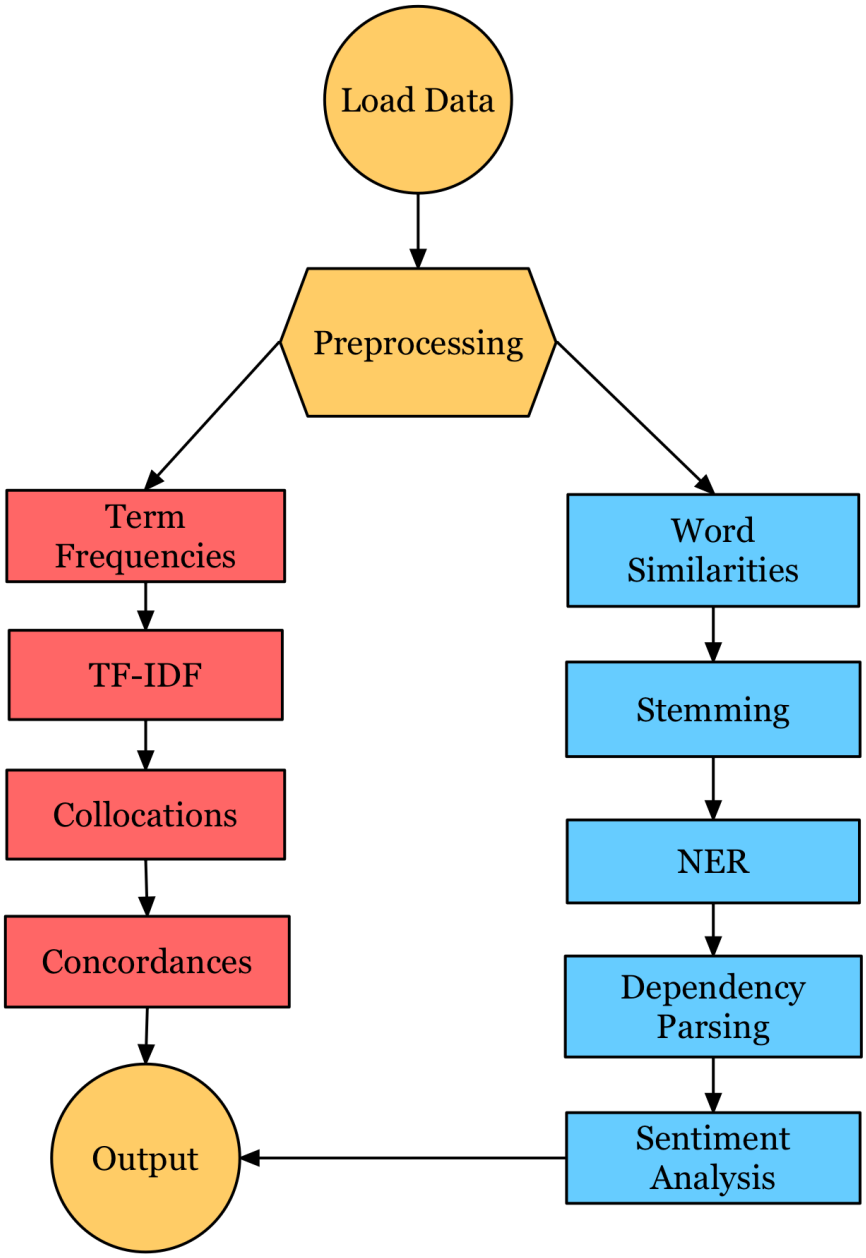

Figure 5.1: Outline of the coding process. Steps marked in red are related to corpus-based approaches (carried out in R), blue-colored boxes denote text mining methods (carried out in Python) and yellow colored fields display processes that are done in both Python and R.

we determine the sentiment of the context-windows of the words to look for specific polarity patterns. The output of the text mining and the corpus-based techniques is then saved to a final table. The complete implementation of this project is distributed over walkthrough code notebooks[37] which contain the code and its documentation.

5.1 Preprocessing

Data preparation is a very important starting point for any type of computational text analysis (Welbers et al., 2017, 248). To account for the specific domains, it is essential to make choices that positively affect the accuracy or are beneficial for the validity of the implementation processes. Similar to Welbers et al. (2017), we will distinguish between the following steps for the preparation of our data:

1. Importing text
2. Normalisation
3. Preprocessing

Considering that the necessary preparation steps are different for the two types of data that we use for our implementation, we will differentiate between the following preprocessing pipelines:

- **Compound list**: Splitting of compound words, declension of nouns, lemma and genus retrieval, and normalisation.
- **Discourse corpora**: Lemmatisation, normalisation, and preprocessing

Each of the data types will be explained in more detail in the according sections. Generally said, the preparation of data seeks to produce consistency across the input data to simplify and potentially reduce text representations for upcoming analytical tasks (Hu and Liu, 2012, 388).

[37] The notebooks can be found in the folder `implementation` in the GitHub repository under `https://github.com/ajgoecke/thesis`. Please see the `README` for more detailed information on the notebooks.

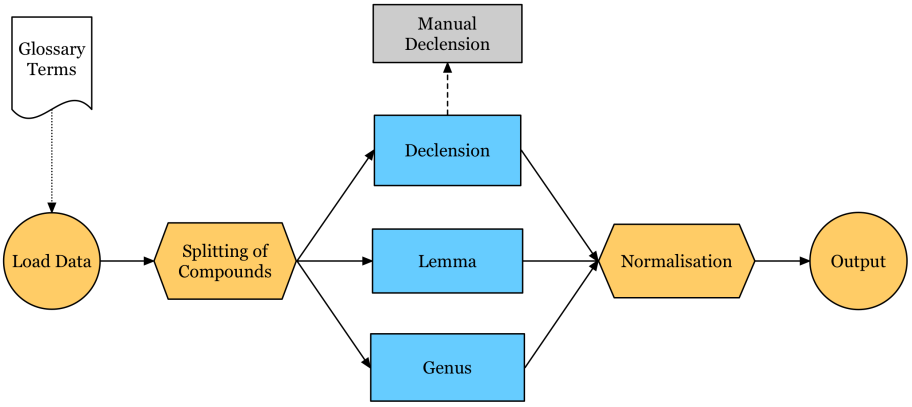

Figure 5.2: Preprocessing pipeline of the glossary terms.

5.1.1 Preprocessing of Compound Words

The preprocessing of the glossary terms is carried out in Python. To retrieve the declension forms of the nouns the Python library `german-nouns`[38] is used. This library offers a list of about 100.000 German nouns including their grammatical properties which are obtained via the German version of Wiktionary. In a first step, we load the text file containing all glossary terms[39] into a data frame. Due to the fact that most of the compounds are not lexicalised and therefore cannot be found as a whole in the list of the `german-nouns` library, we had to come up with a different solution to retrieve the noun forms of the glossary terms. As we elaborated in section 2, the second constituent of a compound word carries the morpho-syntactic information of the complete word. Respectively, it is sufficient to simply retrieve the declension forms of the second constituents to generate the declension forms of the complete compound words. For this reason we segment the compounds into their two components: E.g. the compound *Klimalügner* is split into the first constituent *Klima* (en: "climate") and the second part *lügner* (en: "liar"). The second constituents of each compound are then saved to a separate column named `second_part`.

The library `german-nouns` offers a function which basically looks up all declension forms of a noun. Accordingly, we use this function to obtain the declension forms of the second constituents. Also, we retrieve further information offered by this library,

[38] Further information and the official documentation of the library can be found here: `https://github.com/gambolputty/german-nouns` (Last accessed 17 Oct 2022).

[39] The most current file can be obtained here: `http://www.klimadiskurs.info/download/wordlist` (Last accessed 17 Oct 2022).

such as lemma forms and the genus of all second constituents and save this information to the data frame. For some of the nouns no information is available. This appears to be the case for a total of 15 compounds of the 248 glossary terms. Consequently, we extract a list of those terms.[40] The declension forms, lemma and genus information is then manually retrieved from the Duden website.[41] This file is loaded back into Python and the information is merged to the data frame. Next, we normalise the data by lowering all strings in the data frame. Also, given that we only retrieve declension forms of the second constituents of the compound, we re-append the *Klima* prefix back to the beginning of all declension forms to obtain the complete compound forms. A subset of the final data frame of the compound forms and the supplementary information is illustrated by table 5.1.

original	second_part	noun_forms	lemma	genus	compound_forms
klimaabzockerei	abzockerei	[abzockerei, abzockereien]	abzockerei	f	[klimaabzockerei, klimaabzockereien]
klimablödsinn	blödsinn	[blödsinn, blödsinns]	blödsinn	m	[klimablödsinn, klimablödsinns]
klimacrash	crash	[crash, crashes, crashs]	crash	m	[klimacrash, klimacrashes, klimacrashs]
klimafanatiker	fanatiker	[fanatiker, fanatikers, fanatikern]	fanatiker	m	[klimafanatiker, klimafanatikers, klimafanatikern]
klimahype	hype	[hype, hypes]	hype	m	[klimahype, klimahypes]
klimahysteriker	hysteriker	[hysteriker, hysterikers, hysterikern]	hysteriker	m	[klimahysteriker, klimahysterikers, klimahysterikern]
klimakompetenz	kompetenz	[kompetenz, kompetenzen]	kompetenz	f	[klimakompetenz, klimakompetenzen]
klimatyrranei	tyrannei	[tyrannei, tyranneien]	tyrannei	f	[klimatyrannei, klimatyranneien]
klimaverbrechen	verbrechen	[verbrechen, verbrechens]	verbrechen	n	[klimaverbrechen, klimaverbrechens]
klimazerstörer	zerstörer	[zerstörer, zerstörern, zerstörers]	zerstörer	m	[klimazerstörer, klimazerstörern, klimazerstörers]

Table 5.1: Sample of 10 of the preprocessed compound words.

5.1.2 Preprocessing of the Corpora

Given that most of the original work to create the corpora was carried out in R (RStudio Team, 2020) and the fact that it provides well-documented libraries such as `quanteda` (Benoit et al., 2018) to work with big corpus data, we decided to adhere to this programming language and apply the corpus-based methods in R. The following data is loaded into R to prepare the corpora:

- **Discourse Corpora**: P2022 and C2022
- **Glossary terms**: in form of a text file

[40] The file is available under `implementation/Pyton/evaluation/declension_forms.csv`.
[41] `https://www.duden.de` (Last accessed 17 Oct 2022). Please see table A.1 the appendix for the table of words.

- **Compound forms**: here we use a simplified version of the output data frame (see table 5.1). We use information of the columns `original` and `compound_forms`
- **Stop words**: we retrieve a stop word list from `Snowball`[42] and a list of custom stop words that was already created within previous projects on the glossary.[43]

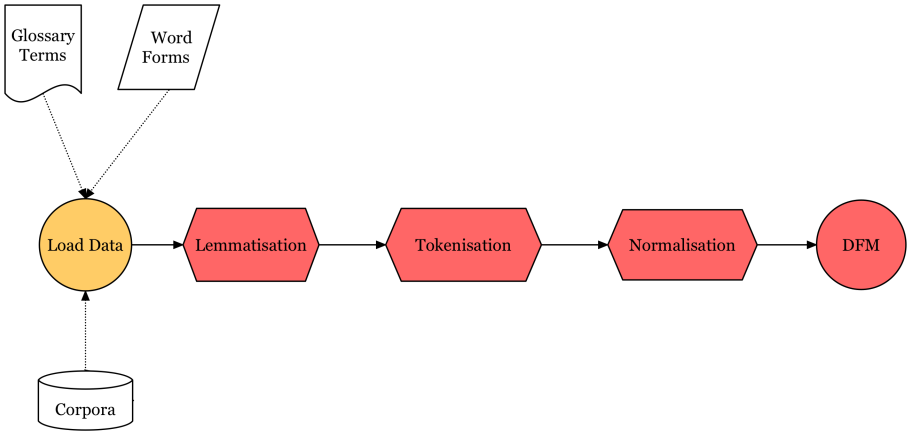

Figure 5.3: Preprocessing pipeline of the corpora.

The table containing the compound word forms has proven crucial for the upcoming lemmatisation procedure. To lemmatise the corpora, the `spacyr` library (Benoit and Matsuo, 2022) is used to retrieve the lemma forms of the words. The previous project on the glossary already discussed the inappropriateness of `spacyr`'s lemmatisation function on German compound words. This most likely arises from the fact that `spacyr` may not able to handle German compound words properly. To address this issue, we use table 5.1 to replace all forms that appear within the column `compound_forms` by their lemma form which is given in the column `original`. To give an example, the genitive form "Klimafanatikern" (en: "climate fanatics") is replaced by its lemma form *Klimafanatiker*. Also, as part of the lemmatisation procedure, the texts are split into tokens. Very frequently, German compounds appear with the notation of a hyphen that connects the first and the second constituent. To normalise the data, we identify climate change compounds that contain a hyphen, e.g. "Klima-Skeptiker" (en: "climate-skeptics") and remove the hyphen

[42] Available via the `quanteda` library, for documentation please see `http://stopwords.quanteda.io` (Last accessed 17 Oct 2022).

[43] For more information please proceed to section 1.1 in the R notebook `implementation/R/notebooks/preprocessing.Rmd`.

from those words. For the upcoming frequency counts and identification of concordances and collocations, it is important to standardise this particular spelling such that we can identify the compound words by simply searching for their lemma form. Also, the corpus texts are all changed to a lowercase format since the corpora not only consist of technical texts but also of comments and reader's letters where nouns may also appear in a lower case format.

Next, we take care of special characters, numbers and URLs. Those are removed by applying `quanteda`'s `tokens` function. Furthermore, very frequent but uninformative words to which we refer as *stop words* here, are removed within this step. We use the stop word list offered by `Snowball` and a list of custom stop words. In a last step, we generate a Document-Feature-Matrix (DFM) from the tokens object that was obtained previously. The DFM basically maps the tokens into a matrix-like format where the rows consist of the document names and the columns show all the features, i.e. tokens, of the according document. For each feature, there is a count of how often this feature appears in any of the documents. Respectively, DFMs are usually very sparse since most of the features will receive a count of zero in the matrix. Figures 5.4 and 5.5 illustrate examples of how the DFMs of our corpora look like.

```
Document-feature matrix of: 6 documents, 56,748 features (99.77% sparse) and 0 docvars.
                 features
docs              mehr uhr mensch ikem jahr sollen deutschland geben neu energie
  ikem_00467.txt     2   0      1    1    0      0           0     0   0       0
  ikem_00746.txt     2   0      1    1    0      0           0     0   0       0
  fff_de_00139.txt   6   0      6    0    6      0           2     1   1       0
  fff_de_00311.txt   4   0      1    0    2      4           2     2   0       1
  ikem_00449.txt     2   0      1    1    0      0           0     0   0       0
  ikem_01109.txt     0   0      0    0    1      0           1     0   0       1
[ reached max_nfeat ... 56,738 more features ]
```

Figure 5.4: Sample of the Document-Feature-Matrix of P2022.

```
Document-feature matrix of: 6 documents, 127,629 features (99.33% sparse) and 0 docvars.
                features
docs             jahr geben mehr sollen schon immer kommen gut % gehen
  eike_00744.txt    4     1    3      2     3     5      3   2 0     0
  eike_03192.txt    0     0    0      0     0     0      0   0 0     0
  eike_06606.txt    6     0    3      1     5     0      1   1 0     0
  eike_00911.txt    4     2    1      2     2     1      1   5 1     2
  eike_01046.txt   15    40   28     42    34    32     32  16 8    30
  eike_12137.txt    0     0    0      0     0     0      0   0 0     0
[ reached max_nfeat ... 127,619 more features ]
```

Figure 5.5: Sample of the Document-Feature-Matrix of C2022.

5.2 Corpus-Based Methods

5.2.1 Term Frequencies

One information that we want to pass to our climate change glossary is the number of occurrences for each of the terms. We retrieve the term frequencies, i.e. how often each compound occurs in each of the two corpora of the sub-discourses, in R by simply creating a Document-Feature-Matrix of the lemmatised tokens of each discourse corpus. The matrix is then filtered for the compound words and the final output is saved to a table.[44] Since both corpora differ in the number of tokens we additionally compute the relative frequencies of each compound term given by the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a common metric in *Corpus Linguistics* to obtain the relevance of a term. Its weight increases for terms being very specific to a corpus and decreases for very common terms that can be found in a large proportion of the texts. We normalise the scores to a range of 0 to 1 to be able to compare the TF-IDF scores of compounds occurring in both corpora. A score of 1 indicates a high relevance of a term for the corpus and a score of 0 is obtained for terms that occur very frequently in a large subset of texts from a given corpus. Figure 5.6 shows the TF-IDF scores for a sample of 30 glossary terms in the direct comparison of both sub-discourses.

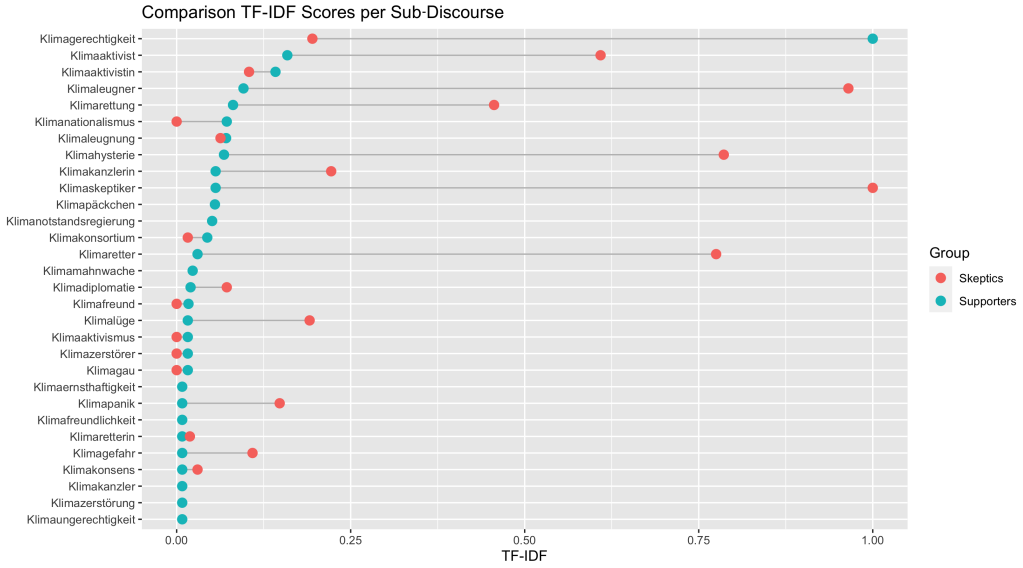

Figure 5.6: Comparison plot of the TF-IDF scores for a sample of 30 glossary terms.

[44] The full table can be found in `implementation/R/output/tf_complete.csv`.

5.2.2 Collocations

To retrieve the immediate collocations, i.e. one token to the left or right of each compound word, we use the `kwic` function offered by `quanteda`. With this function we can easily choose a window of items that we want to retrieve for a specific key word. In our case, we decide for a window of one and our glossary terms as the key word. As an example, listing 5.1 shows the code that we use to extract the collocations for the glossary term *Klimaleugner* (en: "climate denier").

Listing 5.1: Code example to retrieve the top 5 collocations for the key word *Klimaleugner*.

```
1  # define key word
2  word = "klimaleugner"
3
4  # retrieve collocations for key word
5  kwic_con <- kwic(sp_c2022_tokens, pattern=word, window=1, valuetype="fixed") %>%
6    as_tibble()
7
8  kwic_con %>%
9   count(pre) %>% # count collocates to left of key word
10   arrange(desc(n)) %>% # sort descending
11   head(n=5) # give top 5
12
13 kwic_con %>%
14  count(post) %>% # count collocates to right of key word
15   arrange(desc(n)) %>% # sort descending
16   head(n=5) # give top 5
```

In listing 5.1, we start by applying the `kwic` function to the preprocessed C2022 corpus (`sp_c2022_tokens`, see line 5).[45] Then we retrieve the counts for each collocate and arrange the output in descending order. Since we only seek to retrieve the five most frequent collocates to the left and the right of the key word, we specify the number of collocates. Applying this code to both corpora provides us with the output shown in table 5.2.

In a next step, we apply the code of listing 5.1 to all compounds to retrieve the top collocations of all glossary terms. To do this, we initiate an empty data frame to which we iteratively append the top-5 preceding and following collocates (if available).[46] We realise that a large proportion of the collocates occur only once in each corpus. Due to the fact that we only retrieve the first five words that are given in the descending Keyword-In-Context

[45] I.e. the corpus is already lemmatised, split into tokens and normalised as described in section 5.1.2.

[46] We will continue referring to the collocates as `pre` if they occur to the left of the key word and `post` if they appear to the right of the key word.

(KWIC) table, there may be more than just those five words which appear as collocates of the compound. We actually only draw a random subset of five words that appear just once (if there are no words that have a count higher than 1). Since the collocates that have a count of 1 are not very informative with respect to each glossary term, those collocates are excluded from our analysis. All collocates receive a tag, according to their appearance to the left (pre) or right (post). Table 5.2 shows a small subset of collocations for P2022 and C2022.[47]

pre	n	post	n	corpus
1.	1	afd	2	P2022
angabe	1	behaupten	1	P2022
argument	1	erwähnt	1	P2022
bestrebung	1	fall	1	P2022
esoteriker*inn	1	ganz	1	P2022
todesstrafe	5	bezeichnen	3	C2022
aussage	2	befinden	2	C2022
behaupten	2	darstellen	2	C2022
beispiel	2	diffamieren	2	C2022
deutsch	2	eike	2	C2022

Table 5.2: Top-5 `pre` and `post` collocations (and their counts `n`) of the compound *Klimaleugner* for both corpora.

Due to the removal of collocates with a count of 1, we end up with tables that may not contain any collocates for some of the glossary terms at all. Furthermore, the cleaning of the final data frames is required to remove peculiarities such as empty strings, words containing special characters, and repetitions of the key word.[48] Since this method is thought to be one of the automated techniques that we apply to our glossary data, we only perform very simplified cleaning steps on the data. Words that contain special characters or words that are not in their lemma form, are removed from the collocations list. Moreover, we retrieve the output in form of a table to merge this information with the upcoming outputs. The results in table 5.2 already indicate that we need to apply more

[47] The full (cleaned) list of most frequent collocations can be found in the files `top_colls_con_cleaned.csv` and `top_colls_pro_cleaned.csv` in `implementation/R/output/`.

[48] Please see appendix table A.2 for the rows that are removed from the list of collocations.

profound techniques to our data to obtain relevant collocations of the compound words. We will address this issue with the application of a dependency parser in section 5.3.3.

5.2.3 Concordances

We adapt the approach of common concordance extraction (Stefanowitsch, 2020) to fit our objective of context retrieval for the glossary terms. A manual analysis of all concordances of the compound words would exceed the capacity of our semi-automatic definition extraction procedure. Given that we are interested in approaching a definition for the terms, we are not only attentive on the immediate context that we actually already elaborated in section 5.2.2 but also in the broader context of each compound to potentially find patterns that can be used for the investigation of the semi-automatic definition extraction of our glossary terms. Respectively, we consider sentences surrounding a phrase containing the key word as potential candidates that carry additional information of the compounds. It is not always straightforward to exclusively analyse the key phrase to detect sentiment. We hope for supplementary clues about a term's meaning and function in discourse by exploring a window of five sentences.

To be able to identify the concordances it is necessary to adapt and reapply some of the previously mentioned preprocessing steps. For this technique, we do not want a lemmatised version of the tokens, given that we want to have a look at the complete context in text form in the end. Respectively, to extract the concordances to either side of the key phrase, i.e. the sentence containing the key word, on sentence-level it is necessary to tokenise the corpus for sentences. This can be easily done with the help of the `tokens` function of `quanteda` by changing the parameter `what` from *word* to *sentence*. We do not want a lemmatised version of the tokens for this method. For this reason we re-apply some selected preprocessing steps such as the removal of hyphens. To do this, we split the corpora into word tokens to which we again apply a sequence of commands to remove the hyphenation. Since we then want to split the corpora into sentence tokens, we have to re-merge the word tokens back into a corpus object. Then we can proceed with the tokenisation of those corpora into sentence-level tokens. To obtain the concordances for each compound word, we can iteratively create a data frame (for each of the corpora) that contains the results of the `KWIC` function for a window of five sentences to either side of the key phrase and the key phrase itself. The choice to include the removal of the hyphens leads to an increased number of concordances that we can obtain in total. For the

P2022 we obtain 803 instead of 743 concordances and for the C2022 corpus we increase the number of 1508 concordances to a total of 1911.

Given that we are now working with sentence tokens, it is necessary to modify the `KWIC` function by setting its parameter `valuetype` to *regex* to identify complete phrases containing the key word. However, this also leads to the identification of concordances for words that are not in the list of glossary terms. More complex forms are also matched. To give an example, while searching for concordances of the key word *Klimagerechtigkeit* (en: "climate justice") also the concordances of the more complex compound *Klimagerechtigkeit-spolitik* (en: "climate justice politics") are triggered. The same problem occurs for adjectival forms such as "klimaalarmistisch" (en: "climate alarmist") which is triggered by the regular expression of the compound *Klimaalarm* (en: "climate alarm"). To discard the concordances not containing exact matches of our compound declension forms, the output of the `KWIC` function is loaded into Python to filter for all rows containing one of the exact declension forms. Within this step, the table is further reduced from 1911 to 1390 concordances (C2022) and from 803 to 510 concordances (P2022). This final table will now be used for the following steps.[49]

(4) "Dasselbe gälte für Teile Nordfrankreichs und Teile Südfrankreichs, insbesondere der Provence. Überschwemmungen gäbe es auch in Großbritannien. Teile Londons wären ebenfalls betroffen. Wir haben hierzu einen Sonderbericht verfasst. Den findet ihr hier. **Das zeigt: *Klimagerechtigkeit* ist wichtig.** Das Handeln hierzu muss jetzt erfolgen, nicht erst irgendwann. Gipfel der Hoffnung - Leader Summit on Climate Der US-amerikanische Präsident Joe Biden hat das erkannt. Am ersten Tag seiner Amtszeit führte er die USA zurück in das Pariser Abkommen. Nun lädt er 40 Staats- und Regierungschefs der Welt zum Leaders Summit on Climate ein, der am 22.04.-23.04.2021 stattfinden wird. Die USA wollen dort ein neues, ehrgeiziges Klimaziel für 2030 verkünden."[50]

An example of the concordances is illustrated in example 4. The text clearly shows that not only the immediate context given by the key phrase (printed in bold) is relevant to draw conclusions about the compound word (printed in italic). The broader context

[49] The full table can be found in `implementation/R/output` under `pro_context.csv` and `con_context.csv`. The reduced tables with additional information can be found in `implementation/Python/output/info.zip`.

[50] The example of the 5-sentence-window from KWIC for the compound *Klimagerechtigkeit* is taken from the P2022 corpus with the source ID fff_de_00002.

reveals additional, potentially useful knowledge about the term. In example 4 we have information about *why Klimagerechtigkeit* is important, e.g. "Überschwemmungen" (en: "flooding"), and the sentence to the right of the key phrase expresses the *urge* to take action and gives examples of climate actions: "Joe Biden [...] führte die USA zurück in das Pariser Abkommen"?[51]

5.3 Text Mining

5.3.1 Exploring Word Relations

To explore the relations between the terms of our glossary we make use of the lexical data base `WordNet` (Fellbaum, 2010). `WordNet` is a computational lexicon with over 117.000 concepts. This general-purpose ontology is arranged in a hierarchical tree structure as we will see in the upcoming examples. Siegel and Bond (2021) developed a German version, namely `OdeNet`, of the English `WordNet` data base which combines existing resources such as the OpenThesaurus German synonym lexicon[52] and the Open Multilingual WordNet?[53] The resulting `OdeNet` resource contains about 120.000 lexical entries?[54] In the upcoming sections we will make use of very simplistic information that we can obtain from the `WordNet` and `OdeNet` knowledge base to gather additional linguistic knowledge about the glossary terms.

5.3.1.1 Hypernyms

Firstly, we will retrieve the *synsets*, i.e. the sets of near-synonyms of a `WordNet` dictionary-styled entry, of each second constituent of the compounds. To retrieve the according synsets for each of the constituents, we will use the functions `get_synset` and `get_lemmas`?[55] Listing 5.2 shows an exemplary output of both functions for the word

[51] En: "Joe Biden [...] guided the USA back to the Paris Convention."

[52] `https://www.openthesaurus.de` (Last accessed 17 Oct 2022).

[53] In contrast to `GermaNet` (Hamp and Feldweg, 1997), another very well-known resource for German WordNet hierarchies, `OdeNet` is open-source and included in the Open Multilingual WordNet initiative (Bond and Paik, 2012; Bond et al., 2015). For this reason we decided to use the `OdeNet` library for the purpose of this project.

[54] The full documentation is available under `https://github.com/hdaSprachtechnologie/odenet` (Last accessed 17 Oct 2022).

[55] The full code can be found under `implementation/Python/notebooks/textmining.ipynb`.

Alarm. For each compound, we retrieve and save the output of `get_lemmas` to the column `related_words`.

Listing 5.2: Functions and output to retrieve synset information from WordNet.

```
get_synset("Alarm")
>>>[Synset('odenet-7941-n')]

get_lemmas("Alarm")
>>>['Alarm','Notruf','Alarmruf','Warnton','Warnsignal',
'Gefahrenmeldung','Alarmsignal']
```

With the help of the synsets we can now easily retrieve the hypernyms of each word. This information can be drawn from the hierarchical structure of `WordNet` which is basically organised in hypernym relations. To give an example, the hypernym relations of the `OdeNet` library for the word *Betrug* (en: "fraud") are illustrated in figure 5.7. All hypernyms that can be extracted for the compounds is saved to the column `hypernyms`.

```
Path 1
 Synset('odenet-10880-n') krimineller Akt
  Synset('odenet-15937-n') Frevel
   Synset('odenet-6999-n') Topf
    Synset('odenet-9850-n') Vermögen
     Synset('odenet-4667-n') Vermögen
      Synset('odenet-10390-n') Liegenschaft

Path 2
 Synset('odenet-10880-n') krimineller Akt
  Synset('odenet-5502-n') Handlung

Path 3
 Synset('odenet-8872-n') Rauheit
  Synset('odenet-25840-n') Unglück
```

Figure 5.7: Hypernym relations of the word *Betrug* drawn from `OdeNet`.

To enrich our own knowledge base of the glossary terms, we retrieve the root hypernyms, i.e. the hypernym which is at the highest position in the `OdeNet` synset tree, for the compound words by obtaining the hypernym paths as shown in figure 5.7 and collecting the very last synset of each path. For the example in figure 5.7 the root hypernyms are *Liegenschaft* (en: "property"), *Handlung* (en: "action") and *Unglück* (en: "misfortune"). Furthermore, we are interested in specifying the concept of each compound word, i.e. categorising whether the concept describes an action (e.g. *Betrug*) or a person (e.g. *Betrüger*, en: "fraudster"). Initially, this information was thought to be obtained by the root hypernyms provided by `OdeNet`. However, the root hypernyms of `OdeNet` have proven to be not as informative as expected and therefore not useful for the separation into the

two categories *person* and *action*. To overcome this problem and to be able to perform the separation automatically, we decided to work with the English data base `WordNet` itself. `WordNet` offers the possibility to translate synsets based on their ID into other languages. Knowing that the English lexicon `WordNet` offers a more elaborated data base, we make use of this property and translate the German synsets into their English counterparts and retrieve the hypernym paths once more for the English synset entries. One feature of `WordNet` is that the root concept for all entries is denoted as *entity*. In contrast to `OdeNet`, it is therefore not sufficient to only retrieve the root concept here since this would give us the exact same information for each compound word. Respectively, we now retrieve the complete English hypernym taxonomy paths for each compound word and return a list of all hypernyms contained in those paths. The taxonomy paths in figure 5.8 contain very straightforward information that we can use for our attempt to categorise the concepts of the compounds.

Based on the English hypernym paths that are extracted, we decide to go with the following simplified main categories (printed in bold) for the glossary terms in the upcoming part of definition building:

1. **Abstraction**: (rational) motive, psychological feature, state, attribute, phenomenon, process, cause, physical object, abstract entity, artifact
2. **Person**: person, soul, image, ideal, spiritual being
3. **Action**: activity, human action, wrongdoing
4. **Group**: grouping, people
5. **Location**: location, area

For *Betrüger* we can identify key words such as *person*, *being*, *soul*, and *physical entity*. For *Betrug* we find the key words *activity* and *human action*. The key words for all categories are saved to lists which are then used in the function `specify_concept` to check for matches between the key word list of each category and the list of English hypernyms that we obtained before within the English path retrieval. With the help of this knowledge we could obtain main categories for a large proportion (i.e. 80%) of the compound words.[56] For the remaining compounds (n = 48) we manually specify the categories of the concepts

[56]For a subset of 13 compounds the category had to be manually adjusted to fit our purpose. This involved the compounds *Klimabeeinflussung* (en: "climate manipulation"), *Klimageschrei* (en: "climate screaming"), *Klimatrash* (en: "climate trash"), *Klimatod* (en: "climate death"), *Klimakompetenz* (en: "climate competence"), *Klimajünger* (en: "climate followers"), *Klimamilliardär* (en: "climate billionaire"), *Klimaschwachsinn* (en: "climate

```
Betrüger - Hypernym Paths:
 Synset('oewn-09974494-n') chiseler
  Synset('oewn-10017621-n') slicker
   Synset('oewn-09657157-n') offender
    Synset('oewn-09851208-n') bad person
     Synset('oewn-00007846-n') soul
      Synset('oewn-00004475-n') being
       Synset('oewn-00004258-n') animate thing
        Synset('oewn-00003553-n') unit
         Synset('oewn-00002684-n') physical object
          Synset('oewn-00001930-n') physical entity
           Synset('oewn-00001740-n') entity
 Synset('oewn-09974494-n') chiseler
  Synset('oewn-10017621-n') slicker
   Synset('oewn-09657157-n') offender
    Synset('oewn-09851208-n') bad person
     Synset('oewn-00007846-n') soul
      Synset('oewn-00004475-n') being
       Synset('oewn-00007347-n') cause
        Synset('oewn-00001930-n') physical entity
         Synset('oewn-00001740-n') entity


--------------------------------------------------


Betrug - Hypernym Paths:
 Synset('oewn-00770581-n') fraud
  Synset('oewn-00767761-n') criminal offence
   Synset('oewn-00767587-n') offence
    Synset('oewn-00746303-n') transgression
     Synset('oewn-00734044-n') wrongdoing
      Synset('oewn-00408356-n') activity
       Synset('oewn-00030657-n') human action
        Synset('oewn-00029677-n') event
         Synset('oewn-00002137-n') abstraction
          Synset('oewn-00001740-n') entity
```

Figure 5.8: Hypernym relations of the words *Betrug* and *Betrüger* drawn from `WordNet`.

and added those to our data?[57] This piece of information is saved to the column `concept`. Additionally, we retrieve the definitions of each synset by iterating over all synsets of a compound and saving the list of definitions to a new column `definition`.

5.3.1.2 Word Similarities

To explore the relationships between the compound words, we make use of a similarity measure that are offered by the `WordNet` library. As stated in section 4.4, `WordNet` offers multiple options to compute the similarity of two input concepts. For our case we will focus on the two functions `PATH` and `WUP`. For both measures, a score of 1 indicates that the two input concepts are semantically close to each other, i.e. they carry a very similar meaning, and 0 denotes the case where the two concepts are not semantically related at all. The relation that we compute here is basically given by the hierarchical tree structure that we have already seen before. Respectively, the similarities are computed by following the tree structure and looking for common hypernyms. The two measures differ in the way they obtain the similarity score. While `PATH` returns a score that indicates how similar two word senses are, based on the shortest path that connects the two senses, `WUP` computes the relatedness of both concepts by taking the depths of the synsets in the tree structure into consideration. The depth is put into relation with the depth of the *Least Common Subsumer* (LCS). Then, the similarity is calculated based on *how* similar the word senses are and *where* the synsets occur relative to each other in the hypernym tree. The difference of both computation methods is also illustrated by the example in listing 5.3.

Listing 5.3: Application and output of both similarity measures on the word *Betrug*.

```
# PATH similarity
wn.similarity.path(get_synset("Betrug")[0], get_synset("Verbrechen")[0])
>>> 0.5

# WUP similarity
wn.similarity.wup(get_synset("Betrug")[0], get_synset("Verbrechen")[0], True)
>>> 0.923
```

idiocy"), *Klimakirche* (en: "climate church"), *Klimanotstandsregierung* (en: "climate emergency government"), *Klimabrandstifter* (en: "climate arsonist"), *Klimakonsortium* (en: "climate consortium") and *Klimaplanwirtschaft* (en: "climate planned economy").

[57] The manual annotated set of compounds can be found in `implementation/Python/evaluation/concept_manual.csv`.

Since we seek to work with the similarity scores on our compound words to see which compounds can be put into relation, we create a matrix-like data frame where all compounds appear as columns and as rows. For each cell in this data frame we compute the similarity scores of both methods. A subset of the matrix and the according scores is shown in table 5.3.[58]

	Alarm	**Betrug**	**Blödsinn**	**Chaos**	**Erzählung**	**Gefasel**	**Krieg**	**Lüge**	**Unfug**	**Zirkus**
Alarm	**1.0**	0.286	0.286	0.4	0.333	0.333	0.25	0.4	0.286	0.222
Betrug	0.286	**1.0**	0.25	0.333	0.286	0.286	0.222	0.333	0.25	0.2
Blödsinn	0.286	0.25	**1.0**	0.333	0.286	**0.8**	0.222	0.333	**1.0**	0.2
Chaos	0.4	0.333	0.333	**1.0**	0.4	0.4	0.286	0.5	0.333	0.25
Erzählung	0.333	0.286	0.286	0.4	**1.0**	0.333	0.25	**0.667**	0.286	0.222
Gefasel	0.333	0.286	**0.8**	0.4	0.333	**1.0**	0.25	0.4	0.8	0.222
Krieg	0.25	0.222	0.222	0.286	0.25	0.25	**1.0**	0.286	0.222	0.182
Lüge	0.4	0.333	0.333	0.5	**0.667**	0.4	0.286	**1.0**	0.333	0.25
Unfug	0.286	0.25	**1.0**	0.333	0.286	**0.8**	0.222	0.333	**1.0**	0.2
Zirkus	0.222	0.2	0.2	0.25	0.222	0.222	0.182	0.25	0.2	**1.0**

Table 5.3: Sample of the matrix containing the `WUP` similarities.

In a next step, we collect all cells with a value higher than 0.5. This threshold is chosen after performing the manual evaluation of the output given by the two similarity measures. Values higher than 0.5 commonly seem to indicate high relatedness between the two concepts. E.g. the words *Aktivist* (en: "activist") and *Alarm* (en: "alarm") got a score of 0.143 (`PATH`) and 0.25 (`WUP`). In our opinion, those words should not be connected to each other in the glossary. Further, we find the `PATH` similarity not appropriate for our purpose. Most of the words received a score of 0.25, 0.33 or lower, indicating that words are somehow related. Nevertheless, with the output of `PATH` it is difficult to draw a clear distinction between an intense semantic relationship and no semantic relationship at all. Accordingly, after the manual evaluation of the data, we decide to proceed with the similarity scores we obtained by the `WUP` method to make sure we only collect word relations with a high semantic similarity. In the `WUP` similarity matrix, the scores of *Bibel* (en: "bible") and *Abzockerei* (en: "rip-off") or *Erkrankung* (en: "disease") and *Alarm* (en: "alarm") is 0.4. *Besorgnis* (en: "concern") and *Ungerechtigkeit* (en: "injustice") or *Königin*

[58] The full matrices for both computation methods can be found in `implementations/Python/output/` under `nouns_sim.csv` (for the `PATH` similarity) and `nouns_wup.csv` (for the `WUP` similarity).

(en: "queen") and *Tod* (en: "death") received a score of 0.5. All those words should not be put into relation with each other. In contrast, strongly connected words can indeed be identified by very high scores such as *Erzählung* (en: "story") and *Lüge* (en: "lie") (0.667) or *Unfug* (en: "nonsense") and *Gefasel* (en: "drivel") (0.8) as shown in table 5.3. Consequently, we decide to retrieve very similar words with a similarity score higher than 0.5 and save those to the new column `wup` in our data frame.[59]

5.3.1.3 Stemming

During the procedure of computing the path similarity, we realise that many words sharing the same stem and carrying a very similar meaning are not included in the `OdeNet` data base and could not be put into relation within the computation of path similarities. This is, for instance, the case for the words *Aktivistin* (en: "female activist") and *Aktivismus* (en: "activism") which are apparently not contained in `OdeNet` and therefore did not receive any similarity score or information about hypernyms. To connect those words, we try a different approach: *Stemming* which reduces words to their common stem form. We attempt to use this procedure to obtain relations between words sharing a morphological stem. The `NLTK` library offers different stemmers, namely `Cistem`, `Porter`, `Snowball` and `Lancaster`. Since our goal is to detect as many word relations as possible, all of the stemmers are applied to the data frame with their outputs being saved to a new column for each stemmer (`stem_X`). After having applied each stemmer to our compound words, we check for compounds that received the exact same stem form. Since we consider the compounds sharing a stem to be semantically related to each other, we create a list of words that share a stem and save this list to a new column for each stemmer (`share_X`). Then, we check each of the stem columns for duplicates and save those to a list of compounds. The fact that we use different stemmers gives us as many combinations as possible since each stemmer reduces the words at a different position. This is, for instance, the case for the combination *Alarm - Alarmist*. While the `WUP` similarity does not find a relatedness between *Betrug* (en: "fraud") und *Betrüger* (en: "fraudster") those words are now put into relation thanks to the output of `Cistem` and `Snowball`.The same applies for the examples *Lüge* (en: "lie") and *Lügner* (en: "liar"), *Propaganda* (en: "propaganda") and *Propagandist* (en: "propagandist"). Within the application and evaluation of the stemmers, we found the `Porter` stemmer to perform worst on our data. While the other stemmers

[59] For comparison purposes this is done for both computation methods. Column `sim` contains the similarities obtained via `PATH`.

could identify up to 6 combinations, `Porter` only found 1 combination, namely *Kritiker* (en: "critic") and *Kritik* (en: "criticism"). In total we obtain additional 9 relations (i.e. binary relations between 18 compounds) by the application of the stemmers which could not be identified by the `WUP` similarity measure.

5.3.1.4 String Distance

To go one step further, we also examine relationships between compounds by computing the similarity of strings. `NLTK` offers a metric `Jaro Distance` (Jaro, 1989) which automatically computes the similarity between two strings.[60] The values range from 0 to 1, analogue to the path similarity, with 1 indicating that strings are identical and 0 denoting that no similarity can be found between both strings. Since the stemmers are not able to capture all words that imply a certain similarity, we apply the `Jaro Distance` metric to the stemmer columns to potentially obtain additional relations between the stem forms. Stems that are very similar, i.e. with a `Jaro` similarity score of higher than 0.87 are considered to be related to each other and are therefore added to a new column `dist_stemmer`. We compute the string similarity for the output of all stemmers and manually check the scores. We discard the results of the `Lancaster` stemmer for the computation of the string distance since it also gives combinations such as *Presse* and *Professor* which we do not consider to be semantically related. In a next step, we combine the information that we obtain via the stemming procedure and the computation of string distance with the output of the `WUP` similarity. This gives us a combined list of words which are semantically related - either based on the path similarity or the stem.[61]

5.3.2 Named Entity Recognition

To extract named entities that our compound words are associated with, we use the data frames containing the concordances, i.e. the context of the compounds, that we created in section 5.2.3 (`pro_context` and `con_context`) for the two corpora. Given that the context in those data frames is split into three columns - namely `pre`, `keyword` and `post` - it was first necessary to merge those three columns to a full text that consists of the key word phrase and the context to the left and to the right. The full text is saved in a new column

[60] Please see `https://www.nltk.org/api/nltk.metrics.distance.html` for the documentation (Last accessed 17 Oct 2022).

[61] See column `similar_words` in the final knowledge base.

`full` and contains the full context text passages for each compound occurrence. In a next step, we retrieve the entities by using `spacy`'s Named Entity Recognition (NER) pipeline.[62] The German model of `spacy` provides us with the main entities *person*, *organisation* and *location*. For the purpose of our project, we seek to look deeper into the first two entity groups to determine the persons and organisations that may be associated with one of the compound words. `Spacy`'s NER is a transition-based algorithm that uses the BIO annotation that was elaborated in section 4.1. Figure 5.9 shows the entity extraction of a sample sentence taken from our data.[63]

Figure 5.9: Visualisation of the entity extraction with `spacy`'s `displacy` function.

Token	BIO Tag	Entity Label
Dank	O	
Greta	B	PER
und	O	
FFF	B	ORG
ist	O	
endlich	O	
Bewegung	O	
in	O	
den	O	
Stillstand	O	
bei	O	
der	O	
Klimarettung	O	
gekommen	O	

Table 5.4: BIO tags and entity labels for the sample sentence.

Table 5.4 illustrates the BIO annotation for the same sample sentence. In this phrase the entities *Greta* and *FFF* are correctly identified as *person* (indicated by the label PER) and *organisation* (indicated by the label ORG). With respect to the BIO tag, both instances

[62] The full documentation can be found here: `https://spacy.io/api/entityrecognizer` (Last accessed 17 Oct 2022).

[63] En: "Thanks to Greta and FFF there is now progress in the stagnation of climate saving."

receive a **B** which denotes the *beginning* of an entity, while the remaining tokens are tagged with **O** given that they occur *outside* of an entity.[64]

Along with the application of the algorithm, three new columns are created in our data frame:

1. `entities`: contains all entities that could be extracted for this text
2. `persons`: shows all entities of the `entities` column with the PER tag for this text
3. `organisations`: shows all entities of the `entities` column with the ORG tag for this text

A quick look at the output in the `entities` column suggests that `spacy`'s entity recognition does not work properly for a large proportion of our text data. The *person* entities consist of more than just proper names. Here we can also find verbs such as "hilfst" (en: "help") and the pronoun "Du" (en: "you"). For *organisations* we obtain words such as "Covid19" and incorrectly tokenised words such as "zuKlimaschädlich". To clean our data and get rid of the undesired entities we create a list of unique persons and organisations that are extracted from both data frames. These lists are saved as a text file to manually clean the entities. The *person* entities are reduced by manually working through the list of unique persons and checking for person names consisting of a last name or a combination of a first name and a last name. Whenever we find a full name (or the last name) of a real person[65] we keep that person in the list. For cases where only the last name is available, the according first name is manually added via online research. Also, some persons are mentioned by using a nickname (e.g. "Greta Thunfisch"), epithet (e.g. "Nobelpreisträger Ivar Giaever"), title (e.g. "Prof Rahmstorf"), or different spelling (e.g. "Luise Neubauer", "Annalena Bockbär"[66]). As far as we can find those in the list of unique persons, the nicknames and spellings are added in a new column `spellings`. In some situations last names are ambiguous and refer to different persons, e.g. *Schmidt*. Since we cannot automatically solve the ambiguities here, cases where we have recurring last names are discarded from the final list of unique persons. This manual step is a very time-consuming but necessary procedure to render the list of persons valuable. The same procedure is applied to the list of unique *organisations*. We eventually reduce the list of

[64] An **I** would indicate the inside of an entity, e.g. the complete name "Greta Thunberg" would receive a **B** for "Greta" and **I** for "Thunberg".

[65] The name Pipi Langstrumpf, for instance, is excluded from the list.

[66] This misspelling is found in the C2022 corpus and is probably used to sarcastically refer to Annalena Baerbock.

unique persons from 2726 to 570 and the number of unique organisations from 2643 to 824.[67]

Once the entities are cleaned, the tables are loaded into Python and compared to the original data frame. The basic idea is to compare the strings of the entities that are identified by `spacy`'s NER to the table of cleaned persons and check for matches. For the *persons* entity, whenever a string of the last name or the full names match a string in the cleaned list, the entity is considered to be a match and the full name is retrieved and saved to a separate column. To be able to correctly retrieve as many persons as possible, we again perform normalisation steps within `spacy`'s entities: some of the entities include punctuation symbols or digits (e.g. "L .Neubauer", ".Dunlap", "Richard Branson30"). Those are removed from the string. Furthermore, white spaces on the left and right end of a string are eliminated. Also, some of the names appear in their genitive case form (e.g. "Herr Timmermanns" for "Timmermann"). To handle these occurrences, we also include a case distinction where the "s" on the very end of the string is removed and the resulting string is compared once more to the list of cleaned persons to check for matches.

Listing 5.4: Pseudocode of the case distinction for the person matching for NER.

```
For each entity in persons, do:

    if entity matches full name in cleaned list:
        full_name <- entity # entity is equal to full name

    elif entity matches last name in cleaned list:
        full_name <- full_name of cleaned list # use full name from list

    elif entity matches spellings in cleaned list:
        full_name <- full_name of cleaned list # use full name from list

    elif entity is genitive:
        entity <- entity.rstrip("s") # remove genitive "s" from string

        if entity matches full name in cleaned list:
            full_name <- entity # entity is equal to full name

        elif entity matches last name in cleaned list:
            full_name <- full_name of cleaned list # use full name from list
    else
        do nothing
```

[67] The cleaned list of unique persons and organisations can be found in `implementation/Python/evaluation/` under `persons_cleaned.csv` and `organisations_cleaned.csv`.

Listing 5.4 illustrates the case distinctions of the person matching process. Finally, for each match, the full name is retrieved and saved to a new column `PERS` and give us a final list of *person* entities that can be associated with the compound words. The same procedure is applied to the *organisations* entity label. For all organisations, we compare the strings in the `organisations` column with the cleaned list of organisation. This list also includes spellings for some of the organisations that we could obtain within the cleaning procedure. For instance, the entity "Fridays for Future" is spelled (and abbreviated) in various ways. We compare the strings with the original organisation name and the spellings and finally output the original name of each *organisation* entity. The cleaned version can be found in the column `ORG`.

5.3.3 Dependency Parsing

To extract dependency information from our corpus data we use `spacy`'s dependency parser. The dependencies are parsed for each occurrence of the glossary terms which are given in the two context data frames `pro_context` and `con_context`. Here we are particularly interested in the finding and evaluation of words that further specify the compounds. So-called *modifiers* are basically words that modify their sentences' meaning or, in our case, the meaning of the noun phrase. If we speak of modifiers on word level[68](which is the level on which we will operate for our compound words), we commonly refer to words that belong to the part of speech categories *adjectives* and *adverbs*. To obtain words that modify the glossary terms it is not sufficient to just extract adjectives or adverbs from the key word phrase since not all occurrences of adjectives and adverbs seek to modify the key word. Thus, to only extract the relevant modifiers it is beneficial to first look at the syntactic structure of the key word phrase. In syntactic lingo, words that are being modified by an adjective or adverb are commonly denoted as their *heads*. Hence, we will exploit the notion of *heads* to evaluate the modifiers of our compound words and have a look at the cases where one of the compounds appear to be the head of some token. Figure 5.10 shows the dependency structure for the sample phrase *"Über den weltweit bekanntesten und wohl aggressivsten Klimaaktivisten Bill McKibben"*[69] which we will further analyse in the following paragraph.

[68]Modifiers can also appear as adjective phrases or adverbial phrases. We will concentrate on the appearance at word level for our project.

[69]En: "About the worldwide most known and apparently most aggressive climate activist Bill McKibben". This phrase is taken from the C2022 corpus with the text ID `eike_13429`.

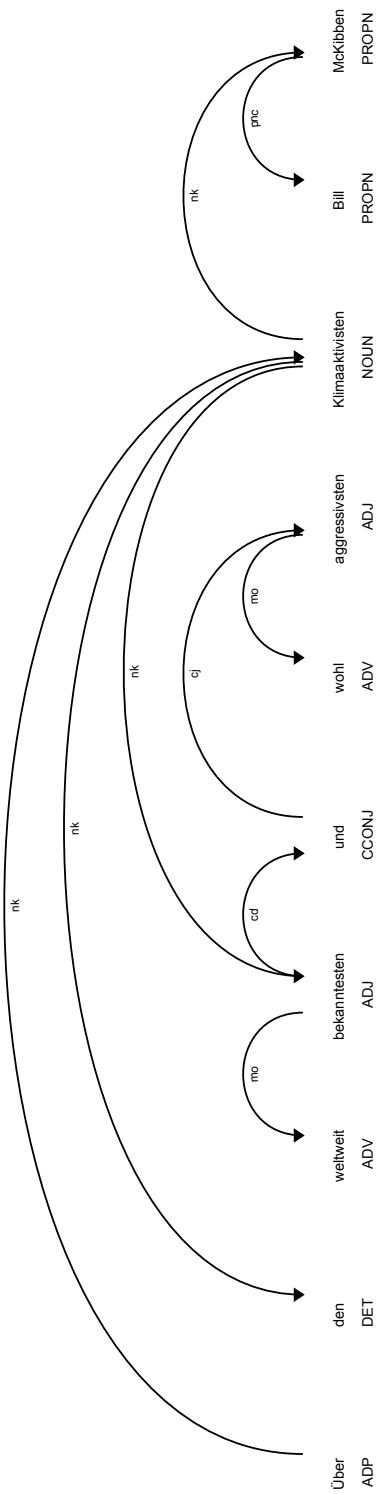

Figure 5.10: Dependency plot generated via `spacy`'s `displacy` function.

Token	Dependence	Head Text	Explanation
Über	ROOT =>	Über	root
den	nk =>	Klimaaktivisten	noun kernel element
weltweit	mo =>	bekanntesten	modifier
bekanntesten	nk =>	Klimaaktivisten	noun kernel element
und	cd =>	bekanntesten	coordinating conjunction
wohl	mo =>	aggressivsten	modifier
aggressivsten	cj =>	und	conjunct
Klimaaktivisten	nk =>	Über	noun kernel element
Bill	pnc =>	McKibben	proper noun component
McKibben	nk =>	Klimaaktivisten	noun kernel element

Table 5.5: Table of dependency labels, heads and explanation for the sample phrase.

In figure 5.10 the adverbs *weltweit* (en: "worldwide") and *wohl* (en: "apparently") (marked by the POS tag `ADV`) and the adjectives *bekanntesten* (en: "most known") and *aggressivsten* (en: "most aggressive") (marked by the POS tag `ADJ`), connected by the conjunction *und* (en: "and"), seek to further modify the noun phrase *Klimaaktivisten Bill McKibben* (en: "climate activist Bill McKibben"). For the reason that we are interested in the extraction of the above mentioned adjectives and adverbs we now attempt to automatically retrieve exactly those tokens for the key word *Klimaaktivist*. Those words act as modifiers for the key word and respectively seek to further specify the term. This is useful knowledge about the use of the term and its potential connotation that we wish to include in our glossary. Accordingly, for the extraction of the dependents of our climate compounds, i.e. words that modify the key word, we will retrieve the dependency labels for the key word phrase and analyse the dependencies to obtain specific dependency relations. With `spacy` we can not only retrieve the part of speech tags and dependency labels for each token but also extract the dependency relation *head* that is relevant for our approach of finding the dependent tokens. The example in figure 5.10 can also be displayed as a table to show the syntactic relations. Table 5.5 illustrates the head-token relation of the sample phrase. The `ROOT` label indicates the occurrence of the head of the complete phrase (in this case a prepositional phrase with the head *über*. The explanation of the dependency labels is given by the column `Explanation`. Looking at the table 5.5, we can easily see for which tokens our key word *Klimaaktivist* appears as a head. Listing 5.5 illustrates how the code iterates through the table to receive a final list of dependents for the key word. To start

from the top of the table *den* is the first token that has *Klimaaktivisten* as a head. Further below, *McKibben* also has *Klimaaktivisten* as its head?[70] Since both, *den* and *McKibben* are not relevant for us since they do not belong to the category of adjectives or adverbs, we discard those two findings and move on to the next tokens. Since the token *bekanntesten* has *Klimaaktivisten* as its head and is an adjective (see figure 5.10) it is added to our list of dependents (see line 14 in listing 5.5).

Listing 5.5: Pseudocode of the recursive parsing of dependencies for the compounds.

```
1  mods = [mo, mnr, nk] # list of modifiers
2  pos = [ADJA, ADJD, ADJ, ADV] # list of POS tags
3
4  for each compound do:
5
6      context <- keyword_phrase # set key word phrase as context
7      doc <- tokens(context) # create tokens object
8
9      for each token do:
10         if head of that token is key word do:
11             if token in mods and pos do:
12                 deps <- list( ) # initiate empty dependency list
13
14                 deps.append(token_information) # append token information
15
16                 new_head <- token
17
18                 for each token do:
19
20                     # if original head and new head are equal
21                     if head of token == new_head do:
22                         if token in mods and pos do:
23                             # append to token information
24                             deps.append(token_information)
25
26         if head of token is conjunct and token in mods and pos do:
27             deps.append(token_information)
28
29             next_head <- token
30
31             for each token do:
32                 if head of token == next_head do:
33                     if token in mods and pos do:
34                         deps.append(token_information)
```

Now we use this token as our new head (line 16) and recursively look for tokens being dependent on this new head (*bekanntesten)*. This leads us to *weltweit* which is an adjective and dependent on the new head and consequently also added to our list of dependents

[70] And recursively from here we could also derive *Bill* which has *McKibben* as its head.

(line 24). Now there are no tokens left that have our key word as their head (or are indirectly connected as it was the case for *weltweit*). In a next step, the tokens are now checked for modifiers that are potentially connected by a conjunct. We analyse tokens that appear to have the conjunct *und* as a head and are either adjectives or adverbs (line 26). This appears to be the case for the token *aggressivsten* which is now added to the list of dependents. Likewise we did before, we now use this last token as a new head (see line 29) to potentially obtain more modifiers being dependent on this token. We identify *wohl* as another modifier and add it to our list of dependents for the compound *Klimaaktivisten* (line 34). Finally, we end up with a list of the following modifiers for this specific occurrence of our key word: weltweit, bekanntesten, wohl, aggressivsten. This procedure is applied to all occurrences of each glossary term and combined to a list of dependency information and modifiers for each key word?[71]

5.3.4 Sentiment Analysis

The evaluation of sentiment of key word phrases can give various information about the connotation of a word. Respectively, we perform sentiment analysis on our corpus data to potentially obtain specific knowledge about the use of a term. In the following we apply two different sentiment models to the corpora and attempt to identify which fits better to our type of texts.

The first model that we apply to both corpora is the sentiment model of the `German Bert` library (Guhr et al., 2020). We choose `German Bert` as it is a very recent, general-purpose model for sentiment classification. To train their model, Guhr et al. (2020) combined several existing resources with new data into a data set of 5.4 million labeled samples. BERT classification models incorporate context to create word representations. Its application is very straightforward: we feed the input text into the model and obtain a polarity label for each text. By polarity label we refer to the three main polarities which are *neutral*, *negative* and *positive*. We retrieve polarities for the full context of a compound word, i.e. the `full` column of our context data frames which contains the preceding and following 5-sentence-windows for each key word phrase and save the polarity information to a new column `bert`.

[71] A full list of the modifier labels and the POS tags that are used here can be found in table A.3 in the appendix.

Then, in a next step we repeat the procedure for the second sentiment model, the German version of `TextBlob` (Loria, 2018),[72] that we want to apply to our data. `TextBlob` originally belongs to `NLTK` and the `pattern` library. It uses a rule-based approach for sentiment classification. `TextBlob` has the advantageous feature of including heuristics, such as the handling of modifiers and negations, to increase the accuracy of sentiment scores. Instead of discrete polarity labels `TextBlob` computes sentiment scores that have to be converted to polarity labels to be able to compare the output of both models. To convert those scores into discrete labels a threshold value that indicates the turning points of sentiment, i.e. at which point in the range from -1.0 to +1.0 do we consider sentiment to be negative or neutral or positive, must be chosen. This threshold value is implemented as a parameter of the function `convert_sentiment` which seeks to convert the continuous scores into discrete labels to make sure we can easily adjust it to our needs. To decide for a threshold we iterate over different threshold values ranging from 0.0 - 0.4 with a step size of 0.1. For each threshold value, we compare the polarity labels that we obtain from the `TextBlob` model to the ones that we computed via `GermanBert`. Then we count how many of the input texts received the exact same label for each threshold for both corpora. Tables 5.6 shows the difference counts for each threshold value for both models.

Threshold	**C2022**	**P2022**
0	679	295
0.1	543	221
0.2	472	187
0.3	430	172
0.4	415	173

Table 5.6: Difference counts for each corpus and each threshold value that is chosen for the conversion of continuous sentiment scores into discrete polarity labels.

So far, we computed the sentiment score for each occurrence of each key word. Given that we aim to provide a single prevailing sentiment label for each compound, we now retrieve the predominant label for each key word to obtain a final sentiment label.

[72] For the German distribution of `TextBlob` please see `https://github.com/markuskiller/textblob-de` (Last accessed 17 Oct 2022).

To evaluate which sentiment model fits better to our data, we run the `TextBlob` model with the threshold parameter that gave us the most common sentiment labels for the two models (x = 0.4). Also, we retrieve the most common polarity label for each compound. This approach is used to reduce the number of texts that have to be evaluated manually. Then we retrieve the subset of compounds that received differing sentiment scores from the two models to manually evaluate the sentiment of those compounds?[73] In this manual evaluation of the models, we subjectively label the texts as *positive*, *negative* or *neutral*. Then we compared the manual label with the ones that are given by the models. We were hoping to be able to clearly identify a model whose output is in line with the labels that are assigned manually. However, our manual annotation only partially agrees with the models. For P2022, we have an agreement (of manual and automatically derived sentiment labels) of 61.5% (`TextBlob`) and 38.5% (`GermanBert`) with respect to the subset of compounds that we extracted for the manual evaluation. For the C2022 corpus we have an agreement of 44.6% (`TextBlob`) and 55.4% (`GermanBert`). These results make it hard to decide for one of the models and lead us to the conclusion that the sentiment analysis procedure we applied to our data is inefficient and at its current status not appropriate for the compounds.

Nevertheless, to draw further conclusions from the data, we evaluate whether the sentiment of a compound can potentially be derived from its second constituent. Occasionally, the connotation of a word can be derived from the sentiment that the word is associated with. For instance, the word *Verbrechen* (en: "crime") suggests a negative action and respectively, the compound *Klimaverbrechen* (en: "climate crime") is very likely to carry a negative connotation as well. This is also in line with the characteristic of our type of compound words, as stated in section 2.2 in that the second constituent commonly carries the morpho-syntactic information and the core meaning of the compound. We re-apply the `TextBlob` model and the `GermanBert` model to the second constituent of each compound to obtain a polarity label. This approach exploits the idea of context-independence of sentiment. After the application of both models we find that a large proportion of compounds did not receive the expected polarity label (or sentiment score). *Klimaapokalypse* (en: "climate apocalypse"), for instance, is tagged as *positive* by `GermanBert` and as *neutral* by `TextBlob`. Thus, we repeat the procedure on the column containing the related words as we expect the sentiment output to be more accurate when we give more input to the models. For

[73] The tables containing the compounds, texts and manual labels can be found in `implementation/Python/evaluation/` under `con_sentiment_diff_manual.csv` and `pro_sentiment_diff_manual.csv`.

some compounds there are no related words available. In this case we simply obtain the sentiment label of the second constituent. To give an example, the related words of the compound *Klimaalarm* are *Alarm, Notruf, Warnton, Warnsignal* (en: "alarm", "emergency call", "alert", "warning signal"). Those words are now fed into both sentiment models and the most common label or average sentiment score is retrieved. This output is then again manually evaluated to see how the models perform on the single word without context. Actually both models give inconsistent polarities and none of the models could correctly identify a major part of the compounds?[74] At this point we decide to manually assign polarity labels to each of the compounds?[75] With the manual labels we can retrospectively evaluate the accuracy of both models. For the context-independent sentiment analysis the `GermanBert` model has an accuracy of about 0.55 for both types of input (the second constituent and the related words) and `TextBlob` has an accuracy of 0.47 for the second constituent and 0.58 for the related words?[76] The main takeaway of the evaluation of the sentiment models is the fact that both models do not work sufficiently on our data. A manual assignment of polarity is the preferred approach for the compound words to obtain knowledge about the connotation.

Another fact that we seek to evaluate is whether the attribution of sentiment is different for the two discourse groups. To do this, we create a table that contains the polarity labels of both models for the context of the supporter's discourse as well as the labels for both models for the context of the skeptic's discourse?[77] Due to the fact that none of the applied sentiment tools is considered to be sufficient for our data and given that most of the climate change compounds eventually receive a *neutral* polarity label by both or only one of the two models, we decide to discard the information given in the table for the final knowledge base. The results that can be found in the according file do not provide valuable findings from which we could draw any conclusions about differences with respect to the connotation of the compounds for both discourse groups?[78]

[74] With respect to the labels that are given by the manual evaluation of the previous steps.
[75] This is done in the file `implementation/Python/evaluation/compounds_sentiment.csv`.
[76] For `TextBlob` we use a threshold value of 0.001 here since the manual evaluation suggests that all values higher or lower than exactly 0.0 should be considered *positive* or *negative* to obtain optimal accuracy rates.
[77] Please see `implementation/Python/evaluation/sentiment_comparison.csv`.
[78] See `textmining.ipynb` notebook section 5.3.3 for additional information.

6 Attribution

The purpose of the glossary is to display the compound words in their natural discourse. For this, we attempt to determine whether a term is used in a corpus in terms of a *self-attribution* - meaning that it is used to refer to the own discourse group - or to refer to the opposing discourse group. We will refer to the latter as *external attribution*. This knowledge is very essential to be able to formulate definitions about the use of each word in discourse. In a first approach to tackle this issue we retrieve a random subset of the tables containing the concordances of each compound for each corpus. The sample consists of the key word phrases of a sample of 80% of each compound word. This is done because some of the compounds only occur twice and accordingly only have two appearances for the key word phrases. Thus we ensure that each of the compound words occurs at least once in the sample. The attribution question is rather interesting for the main concept categories *person* and *group*, respectively we will concentrate on those two categories. Next, we carry out a manual annotation in which we focus on whether the term is used to refer to the own sub-discourse or to the opposing group. Each of the occurrences of the compound in the sample receives a tag *self*, *external* or *none*. The latter is used for cases where no attribution can be determined on the basis of the key word phrase. Example 5 and 6 show key word phrases (with the compound word printed in bold) that are labeled as *external attribution*. Here we can clearly indicate an attribution to the opposing discourse either by the semantic content of the context or by the mentioning of certain organisations (AfD, Springerverlag).

(5) An alle **Klimahysteriker**: Alles läuft nach Plan, wir dürfen alle weiterleben, ausser Ihr entzieht den Menschen ihre wirklichen Lebensgrundlagen, indem Ihr ihnen Ihre dringend benötigten Energiequellen wegnehmt. (C2022: `eike_11614`)[79]

(6) Letztlich ein weiteres Greta-, oder FFFBashing, auf unterschiedlichsten Ebenen, incl. des irren Kampfes gerade der **Klimaleugner** (AfD) und Klimaskeptiker

[79]The translation of all examples in this chapter can be found in the appendix in the section Attribution.

(Springerverlag, die Welt) GEGEN eine Satire die ebenfalls letztlich ein Greta / FFFBashing zum Ausdruck brachte. (P2022: `fff_de_00199`)

In contrast, in the following examples 7 and 8 we have indicators such as personal pronouns "wir" (en: "we") and the verb form "bin" (en: "am") that suggest a self-attribution of the terms.

(7) Zukünftig werden wir **Klimarealisten** auch noch von den „Eisheiligen" aus Potsdam Unterstützung bekommen. (C2022: `eike_07029`)

(8) Hallo liebe Klimafreunde, bin ebenso ein junger **Klimafreund**, möchte nur ein paar Anmerkungen zu eurer Zusammenfassung zur GradStudie machen . . . Ihr schreibt: Im Sektor der Energiewirtschaft fand das Institut unter anderem heraus, dass erneuerbare Energien und explizit die Wind- und Solarenergie in den kommenden Jahren stark ausgebaut werden müssen. (P2022: `fff_de_00195`)

This distinction is carried out for the complete sample of compounds. During the manual annotation[80] we further notice that some of the compounds mostly appear in quotation marks. The contexts of those appearances suggest that this is very often the case to express sarcasm or irony. This is related to the notion of self- versus external attribution as illustrated in the examples 9 - 12 below.

(9) Also Klimaleugner wenden sich gegen eine Satire , die letztlich die Klimabewegung als „**Klimahysterie**" anprangert, dies deshalb, weil es diese verbohrten Ideologen von rechtsaußen intellektuell nicht verstanden haben, und einen Sündenbock, bzw. einen Aufhänger im Kampf gegen die „Lügenpresse" gesucht und gefunden haben. (P2022: `fff_de_00199`)

(10) Denn es gäbe dringlichere globale Probleme als die „**Klimahysterie**": Hunger etwa oder das Sicherstellen der Rente. (P2022: `farn168`)

(11) Dieser setzt sich klug mit der medialen Abwertung der so genannten „**Klimaleugner**" auseinander. (C2022: `eike_04522`)

(12) Mein persönlicher Eindruck: So viel besser hätten wir „**Klimaleugner**" das gar nicht hingekriegt!!! (C2022: `eike_02490.txt`)

[80] The full annotation of the *person* and *group* compounds can be found in `implementation/Python/evaluation/attr_manual.csv`.

Examples 11 and 12 illustrate that the term *Klimaleugner* that we expect to encounter within the sub-discourse of the climate activists appears to be frequently used in the discourse of climate skeptics to refer to themselves. The term is very often written in quotation marks, suggesting a rather sarcastic use. The same applies for the term *Klimahysterie* (en: "climate hysteria") with respect to the sub-discourse of the climate activists. As a result of the manual annotation of the attribution question we now additionally make the very simplistic attempt to further categorise the use of the compound words by looking for quotation marks on both ends of each occurrence of a compound. To come back to the example of *Klimaleugner* for the C2022 corpus we find this term to be used within quotation marks in almost 40% of the use cases, i.e. 49 times out of the total of 125 occurrences. For the term *Klimahysterie* which is usually used by the sub-discourse of climate skeptics to refer to climate activists, we identify its use in quotation marks for 7 out of 10 occurrences.

The information about the proportion of sarcastic use cases and the attribution tag is added to the data frame containing all knowledge that we gathered so far in chapter 5.

7 Definition Phrasing

7.1 Approach

The purpose of this work is to semi-automatically generate definition texts for all the terms of the discourse glossary. For this reason we gather the information we described in chapter 5 and 6 in form of a table that we will call our *knowledge base* from now on.[81] The information that we want to transfer to the glossary is diverse. We seek to not only present quantitative facts such as the term frequency (or TF-IDF) to the user but also qualitative insights that focus on the use of the term in discourse. For this reason we also attempted to identify the prevalent sentiment and addressed the attribution question in the previous chapters. To give an overview of the general information that should be included in a definition text, please see figure 7.1.

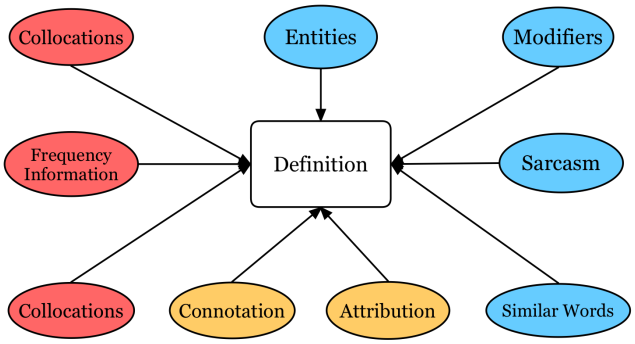

Figure 7.1: Overview of the information we want to include in the definition texts. Red boxes indicate the output of a corpus-based approach and blue-colored boxes denote the outputs of text mining methods. Yellow boxes display information obtained via manual annotation.

[81] The complete table (which is basically the output of the `textmining.ipynb` notebook) can be found in `implementation/Python/output/knowledge_base.csv`. However, for documentation purposes, the final knowledge base containing only the information that is used in the definition texts and also the final definition texts is saved in the file `knowledge_base_updated.csv`.

The basic idea of semi-automatically generating a full definition text from our knowledge base consists of the approach to create definition texts for each of the compound categories (*person*, *abstraction*, etc.). To do this, we identify a general structure and base phrase for each category. Also, we formulate phrases for each piece of information as shown in example 13. Those phrases contain place holders which are iteratively filled by the according information drawn from the knowledge base.[82] Example 13 shows an example of the full definition text for the category *persons*, including the place holder variables.[83]

(13) **{COMPOUND}**, **{ARTICLE}**

{COMPOUND} bezeichnet eine Person, die in einer gewissen Beziehung zum Klimawandel steht. **{COMPOUND}** wird in unserem Korpus **{CON_FREQ}** Mal von den Klimaforschungsskeptikern und **{PRO_FREQ}** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **{CON_TFIDF}** für die Skeptiker und **{PRO_TFIDF}** für die Vertreter. Verwendet wird **{COMPOUND}** hierbei im Sinne einer **{CON_ATTRIBUTION}** von Seiten der Skeptiker und als **{PRO_ATTRIBUTION}** im Vertreter Korpus. Die Verwendung wird in **{CON_SARCASM}** der Fälle für den Vertreter Diskurs und in **{PRO_SARCASM}** der Fälle für den Skeptiker Diskurs als sarkastisch eingestuft. In unserem Korpus Sample ist der Begriff tendenziell **{SENTIMENT}** konnotiert. Im Subdiskurs der Klimaforschungsvertreter wird der Begriff von Wörtern wie **{PRO_MODS}** modifiziert. Wörter wie **{CON_MODS}** treten auf, um den Begriff im Subdiskurs der Klimaforschungsskeptiker näher zu beschreiben. Im Zusammenhang mit dem Begriff erwähnt der Skeptiker Korpus die Person(en) **{CON_PERS}** und der Vertreter Korpus die Person(en) **{PRO_PERS}**. Im Kontext von **{COMPOUND}** erfolgt die Nennung folgender Organisation(en): **{CON_ORG}** (Skeptiker Korpus) und **{PRO_ORG}** (Vertreter Korpus).

Kollokationen: **{CON_COLLS}{PRO_COLLS}**

Siehe auch: **{SIMILAR_WORDS}**

[82] The complete coding procedure of this chapter can be found in the notebook `implementation/Python/notebooks/definitions.ipynb`.

[83] The final definition strings (with place holders) for all categories can be found in the appendix section A.3.1.1.

7.2 Preprocessing of the Knowledge Base

It is necessary to apply some minor preprocessing steps to the final knowledge base that we obtained from chapter 5 and 6. English polarity terms are replaced with their German counterparts and English attribution tags are translated into German as well. A counter is applied to the list of modifiers for each corpus. Here, for each compound we use (if available) the two most common modifiers of each corpus. The entities are also counted and reduced to (if available) the two most common *persons* and *organisations* for each corpus. For the collocations we choose a random sample of up to two collocations for each corpus to be displayed in the definition texts. Next, the genus information of each compound (see section 5.1.1) is used to create a new column in the knowledge base that contains the definite article of each compound according to the information that we obtained from the genus tag.[84] The sarcasm score which consists of a float number indicating the percentage of sarcastic use cases of the compound is translated into an integer value for the final definition.

The final pieces of information that are used to compose a full definition text consist of the following[85]:

- **Category**: person, group, location, abstraction, action
- ***Attribution**: only for *group* and *person* category
- ***Sarcasm**: only if attribution available
- **Definite Article**
- ***Term Frequencies**
- ***TF-IDF Scores**
- **Connotation**: positive, negative or neutral
- ***Modifiers**
- ***Entities**: persons and organisations
- ***Collocations**
- **Related Words**

[84] I.e. "der" for masculine nouns, "die" for feminine nouns and "das" for neutral nouns.
[85] The * indicates that this information is given for each sub-discourse separately.

7.3 Definition Patterns

To generate the definition patterns we clearly focused on the information that is given in our knowledge base and formulate phrases with place holders that seek to provide a verbalisation of the information. In a first iteration of the definition phrasing process we start to preprocess the information of the knowledge base (as described above) so that we can use it to fill the place holders. Then we check for missing or incomplete information. Some of the terms, for instance, do not have a list of similar words or entities. To address this issue, the phrases for each information and each corpus are saved to separate strings that we can then combine based on the information we have for each compound. The *base* string contains place holders for the information that we have for all of the compounds. This includes the compound, genus, term frequencies, TF-IDF scores, and sentiment. The base string for the category *action* is shown in example 14.

(14) Der Begriff '**{COMPOUND}**' wird in unserem Korpus **{CON_FREQ}** Mal von den Klimaforschungsskeptikern und **{PRO_FREQ}** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **{CON_TFIDF}** für die Skeptiker und **{PRO_TFIDF}** für die Vertreter.

Additionally, as stated before, we initiate separate strings for each of the information that is contained in the knowledge base. The place holders are always denoted in curly brackets and capital letters.

To create a final definition text for each glossary term we iterate over all the compound words and depending on which category it belongs to, we select the according base string. Then, with respect to the information that we can obtain from the knowledge base for the current compound word, we append the additional strings that contain place holders for those pieces of information to the definition text string. This results in a unique definition string with place holders for the unique information that we can retrieve for each glossary term. During this procedure we access the respective information from the knowledge base and assign it to the according place holders. Since some information pieces in our knowledge base are superfluous and not valuable for the definition text of the discourse glossary we only use a subset of those pieces. For the entities, for instance, we only retrieve a maximum of the two most mentioned persons and organisations for each discourse to be displayed in the final definition. Also, with respect to the modifiers and collocations

we only use up to two most occurring modifiers or collocations that appear at least twice for each discourse to make sure the information is not arbitrary.

After the first iteration of the definition phrasing process, we manually evaluate a subset of the definitions and check for improvements regarding the formulation as well as the retrieved information. As a result, we decided to clean the modifiers of each compound by manually discarding noisy words that are included. Here we remove words such as "Flashcrash" and "dieser" (en: "this") that lead to a poor definition string.[86]

In the final iteration, we then adapt the base strings of each main category and the single information strings. Furthermore, we work on the formulation of the strings for the entities for each discourse to render those more independent from one another and to ensure they also work with the cases where some information pieces are missing or only partially available (e.g. entities for only one discourse). A final definition for the term *Klimaleugner* which belongs to the category *person* and the term *Klimaelite* (en: "climate elite", category *group*) are given below in example 15 and example 16.[87]

(15) **Klimaleugner, der**

Der Begriff **'Klimaleugner'** bezeichnet eine Person, die in einer gewissen Beziehung zum Klimawandel steht. Der Begriff wird in unserem Korpus **123** Mal von den Klimaforschungsskeptikern und **37** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **0.965** für die Skeptiker und **0.096** für die Vertreter. Verwendet wird **'Klimaleugner'** hierbei im Sinne einer **Selbstzuschreibung** von Seiten der Skeptiker und als **Fremdzuschreibung** im Vertreter Korpus. Die Verwendung wird in **2%** der Fälle für den Vertreter Diskurs und in **44%** der Fälle für den Skeptiker Diskurs als sarkastisch eingestuft. In unserem Korpus Sample ist der Begriff tendenziell **negativ** konnotiert. Wörter wie **'kein' und 'genannt'** treten auf, um den Begriff im Subdiskurs der Klimaforschungsskeptiker näher zu beschreiben. Im Zusammenhang mit dem Begriff erwähnt der Skeptiker Korpus die Person(en) **Stefan Rahmstorf und Michael Limburg** und der Vertreter Korpus die Person(en) **Greta Thunberg und Tom Buhrow**. Im Kontext von **'Klimaleugner'** erfolgt die Nennung folgender Organisation(en): **EIKE, IPCC** (Skeptiker Korpus) und **AFD, FFF** (Vertreter Korpus).

[86] The cleaning step is retrospectively integrated into the `textmining.ipynb` (see section 4.3). For a full list of words that are removed within the cleaning step, please see table A.4 in the appendix.

[87] A translation of both definition texts can be found in the appendix in section A.3.2.

Kollokationen: **'Erfindung', 'EIKE'** (Skeptiker Korpus) und **'AfD'** (Vertreter Korpus)

Siehe auch: **Klimalüge**

(16) **Klimaelite, die**

Der Begriff **'Klimaelite'** bezeichnet einen Zusammenschluss von Personen im Bezug auf den Klimawandel. Der Begriff wird in unserem Korpus **4** Mal von den Klimaforschungsskeptikern und **0** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **0.044** für die Skeptiker und **0.0** für die Vertreter. Verwendet wird **'Klimaelite'** hierbei im Sinne einer **Fremdzuschreibung** von Seiten der Skeptiker. Die Verwendung wird in **25%** der Fälle im Skeptiker Diskurs als sarkastisch eingestuft. In unserem Korpus Sample ist der Begriff tendenziell **neutral** konnotiert. Im Zusammenhang mit dem Begriff erwähnt der Skeptiker Korpus die Person(en) **Maybritt Illner und Christiana Figueres**. Im Kontext von **'Klimaelite'** erfolgt die Nennung folgender Organisation(en): **EIKE, Die Grünen** (Skeptiker Korpus).

Kollokationen: **'alternativ', 'Behöre'** (Skeptiker Korpus)

One can see that the approach of using separate definition strings is suitable for a wide range of examples and information that are given by the knowledge base. The definition text for *Klimaleugner* contains the base string and additional information about the attribution and the sarcastic use cases of the term. In contrast, the definition text for *Klimaelite* does not contain the exact same information and results in a shorter text. The methodology that we used to generate proper definition texts actually turns out to be appropriate for the glossary terms and the knowledge base that was create to serve the purpose of definition phrasing. In section 8.1 we will elaborate in more detail on the applied methods and the suitability of the final definition texts.

To add the definition texts of the glossary terms a notebook is created.[88] The notebook provides a walkthrough explanation of how to integrate the definition texts into the web app. To do this, the `JSON` file has to be retrieved from the online glossary. The final definitions for each glossary term are then added to the `JSON` file.[89]

[88] Please see `implementation/Python/notebooks/json.ipynb`.

[89] The updated file containing the definition texts can be found under `implementation/Python/files/glossary_updated.json`.

8 Conclusion

8.1 Discussion

Overall we were able to enrich the knowledge base of our compound words significantly by retrieving information from the corpora. We could not only automatically collect quantitative information about a term but also qualitative insights about a term's use in discourse by combining automated approaches with manual annotation. Most of the information that was collected within this project was finally composed to build a knowledge base for the glossary terms and could be used to create definition texts for the discourse glossary. The main purpose of this project was to generate definition texts for all current glossary terms in a semi-automated approach. The corpora for both sub-discourses that were already built in previous projects have proven to be representative for each of the discourse parties. Moreover, they uncovered a lot of useful knowledge that was extracted in the chapter 5 and 6. With respect to the corpus-based methods that were applied to the corpora, the computation of the TF-IDF scores came in very handy for the illustration of major differences in the occurrence of terms for each of the discourses, as shown in figure 5.6. The retrieval of collocations has proven to be an interesting quantitative metric that however still requires manual post-processing and could not reveal exclusively relevant findings. That is why we decided to proceed with a more profound technique, namely dependency parsing, to obtain words that indeed act as modifiers for the compounds. The extracted concordances built the base for a large proportion of the text mining tools and techniques that were applied to gather additional knowledge about the compounds. During the manual annotation and evaluation of some of the methods, the 5-sentence-context-window appeared to be appropriate to display the broader context of a term. Regarding the text mining techniques we find the knowledge that was obtained from the `OdeNet` (and `WordNet`) library to be particularly valuable. The retrieval was very straightforward and can be easily applied to new unknown terms if

needed. However, the short definitions we retrieved for the synsets via `OdeNet` were only partially useful. For instance words such as *Klimabzockerei* (en: "climate rip-off") obtained the definition phrase "der Akt des Stehlens"[90] and *Klimaaktivist* the phrase "ein militanter Reformer"[91] which actually captures one of the meanings of the terms. In contrast, for the word *Klimafeind* (en: "climate enemy") the phrase "ein Kandidat gegen den man antritt"[92] and for *Klimafreundin* (en: "climate friend") "eine Frau, die mit einem wichtigen Mann zusammenlebt"[93] do not indicate a definition that is valuable for the definition phrasing part. During the implementation of the final definition phrases we identified a large proportion of cases where the short definitions could not be used to indicate the meaning of the compound accurately. Accordingly, we did not use the short definitions for the final definition texts and discarded this piece of information in the process of definition phrasing. Since this could be addressed and elaborated in future project, we keep the short definitions in our knowledge base.

Another method that was very crucial to mark relationships across the glossary terms, was the application of similarity measures to identify related terms. Also, the retrieval of the word forms in the preprocessing step using the `German-Nouns` library has proven essential for the lemmatisation process within the corpus-based methods to identify concordances and collocations. The combination of text mining and corpus-based approaches is very beneficial for the corpus data of both sub-discourses. Even though, some of the methods did not provide the expected results. For instance, the extraction of entities where still a lot of manual post-processing was required to obtain valuable entities (for both, persons and organisations). Also, the application of text mining tools to the context of the compounds did not prove to be suitable to access the connotation of our terms. Further, we have hoped to identify patterns in the context of a compound to address the question of whether a word is used in terms of a self-attribution or an external attribution in an automated approach. However, the manual analysis did not reveal any specific patterns that could be exploited for this task. Accordingly the identification of attribution resulted in a time-consuming manual annotation since this is one of the information pieces that we consider essential for building knowledge about a term's use in discourse. Still, the manual annotation unleashed the idea of identifying quotation marks an indicator of potentially sarcastic use of a term. The implementation of this interpretation provides

[90] En: "the act of stealing".
[91] En: "a militant reformist".
[92] En: "a candidate that one compete against".
[93] En: "a woman who lives together with an important man".

additional information about a term's use in discourse and was therefore added to the knowledge base. Even though some of the methods required a manual annotation or post-processing step to make sure they are appropriate for the glossary terms, the manual steps resulted in a starting point for the addition of new terms in the future.

The road that was taken to create definition phrasing patterns for the acquisition of full definition texts is very straightforward. The introduction of place holder has proven a useful tool to iteratively generate complete definition texts and fill the types of information that are unique to each compound word. Since the definition strings are formulated as separate phrases which either work independently or can be combined to full paragraphs, we cover a lot of cases for the definition phrasing of existing and new compounds. Otherwise, it would have been inevitable to manually retrieve all necessary information for the 248 glossary terms to build definitions. A very time-consuming task that would be required for each new term in the future. With this work, the addition of definition texts for new terms is automated to a certain degree. In consequence, the glossary can now easily be enriched with new terms.

8.2 Summary

As described earlier, the enrichment of the glossary entries is based on the exploratory application of different methods in the field of computational linguistics to give valuable information about the terms. This is a very beneficial approach to explore text mining techniques and corpus-based methods while also producing output that can be directly handed over to the user of the discourse glossary. The focus of this project lies on the exploratory application of the computational techniques to generate proper definition texts. However, the attempt to describe the two opposing constellations of the climate change discourse was necessary as it forms the basis of a previous project in which we construct the two corpora. Accordingly, we provide a simplistic insight on the current climate change discussion and the separation into the two discourse groups. Chapter 3 provides the reader with essential knowledge about the topic and distribution of positions towards climate research. It is out of the scope of this project to further specify and elaborate on the two opposing positions in the climate change controversy. In the course of this work we not only give a brief overview of how the corpora were constructed in a previous project and on the discourse of climate change, but also on literature related to the application of elemental text mining and corpus-based methods that are used to obtain knowledge from

the corpora. We apply a broad range of text mining methods to explore which of those are appropriate to build a knowledge base that eventually results in definition texts for the glossary. The preprocessing steps build the starting point to normalise the input data that is used. Quantitative methods are applied to retrieve term frequencies and TF-IDF scores from the corpora and to obtain a list of collocations for a subset of the compound terms. Acquiring concordances of the compound words gives us the context of each word that is necessary to derive various knowledge within the application of several automated and semi-automated computer-based techniques. Within the project we not only retrieve essential linguistic information such as declension forms and genus information of each compound, but also more elaborated insights on the semantic relationship across glossary terms by computing the word similarities. Furthermore, the application of sentiment tools sparked the idea of identifying the connotation of a word by looking at the second constituent of the compound. This significantly decreased the amount of manual work that work that would have been necessary otherwise. The extraction of entities such as organisations and persons is still in its initial stage and should be addressed in upcoming project as we will elaborate in section 8.3. The parsing of dependencies to identify words that modify the compounds was also a very straightforward automated technique which gave a lot of interesting insights on the context of the compounds. The last coding steps to identify discourse information consisted of the initialising of the notion of attribution by manually annotating the terms with respect to their use in discourse. This finally leads to the attempt of identifying sarcasm for our specific type of data which actually appears to work sufficiently. Overall, the way we apply text mining techniques to the compounds could retrospectively rather be identified as a rule-based application of various tools. Also, the final step of generating definition texts depends on a set of rules and constraints which lead to a complete definition text. Nevertheless, during the application of several tool we had to come up with adapted solutions to make the techniques fit to our purpose and to give valuable output. For instance, the detection of related words. The first idea was originally based on simply computing word similarities. This had to be adapted and improved by adding stemming and the computation of string distances. While the implementation part of this work focuses on the exploratory application of tools the final definition phrasing task does not consist of attempts to present the information of the knowledge base but provide elaborated definition texts that perfectly fit the discourse glossary. The primary goal of this project, namely the semi-automated phrasing of definition texts for the discourse glossary is fulfilled. We were able to enrich the glossary, which so far only

provided a list of climate change compounds, and finalise it by contributing definition texts. Given that language is changing and the vocabulary is constantly growing, the partial automation of the definition phrasing task is crucial for the addition of new terms to the glossary. Our work therefore constitutes a basis for upcoming projects regarding the glossary and fills the gap of how to proceed with new glossary terms. The complete project is created in such a way that it can easily be extended and adapted to apply to new terms and situations, for instance the retrieval of further information from the corpora.

8.3 Outlook

Future projects could prove wheter our approach of sarcasm detection is applicable to new potential glossary terms as well. Since we only rely on quotation marks as an indicator of sarcasm our approach is indeed very simplistic but seems to be sufficient for the kind of text data that the corpora contain. One idea to potentially improve the simplistic sarcasm detection that we attempted here, is the inclusion of specific modifiers such as "sogenannte/r" (en: "so-called") that could be additionally implemented to identify sarcastic utterances of the compounds. Also, the definitions could be further improved and extended by various information that are still present in the corpora that we were not able to capture within the scope of this project. For instance, we retrieved definitions from `OdeNet` for a large proportion of compounds which were not used for the definition phrasing part in the end. However, these short definitions could be adapted by manual cleaning to be appropriate and useful for the final definition texts. So far, the definitions from `OdeNet` are retrieved for the second constituent of the compound nouns and therefore do not capture the full meaning of the words in the context of climate change or global warming. It would be necessary to re-define the short phrases with respect to the first component of the compounds. This could be addressed in the future and easily implemented into the final definitions by adding an introductory sentence that gives information about the semantic content of each term. Further, the use of an API of an online dictionary (such as Duden[94] or DWDS[95]) could be used to retrieve definitions for the remaining words or to improve existing definitions. However, since for the glossary the focus is to display the use in discourse rather than the meaning of the compound itself, we did not address this in our project.

[94] `https://www.duden.de/api` (Last accessed 17 Oct 2022).
[95] `https://www.dwds.de/d/api` (Last accessed 17 Oct 2022).

Another approach to potentially improve the contents of the final definition texts and to capture the meaning of the compounds in the specific context of the corpora could be made by applying pattern-based techniques to the compound's context. This could be done to automatically retrieve further specific information from the context which could then be added to the definition texts. We would also wish for a categorisation of the entities that are extracted in this project. The identification of membership of the persons or organisations to either side of both discourse groups would be a promising path to be able to give elaborated information about mentioned entities. So far, we only name entities and the corpora they appeared in. For the illustration of the use in discourse it would indeed be beneficial to rather give information about the actual membership of the entities to one of the discourses. However, this is currently out of the scope of this project and could be addressed in upcoming work on the data.

Future projects should also address the enrichment of the corpora by adding further sources from a diverse set of domains. This would be particularly important for the corpus of the skeptics to make sure the glossary does not only display the discourse of the EIKE website (since this is the major source for the main proportion of the texts of the skeptics corpus).

One advantageous aspect of the definition texts and the knowledge base itself is the possibility to easily enrich and further improve the available information in future work. This renders the addition of new glossary terms and the integration of new information in the definition texts very straightforward. Moreover, it could be beneficial to display some of the information that are now contained in the definition texts by implementing graphical visualisations. For instance the frequency proportion of a compound in both corpora could be visualised in a pie chart or statistics plot to support the glossary on a graphical level. Also, the frequency of each term in the Twitter discourse could be displayed as well: How many hits do we have for each month of the year? Does the use of a term increase or decrease over time? The temporal aspect is not given yet in the glossary and would be a promising path to follow to render the discourse glossary even more valuable.

To finalise the implementation of the new definitions into the discourse glossary it is still necessary to load the updated `JSON` file into the web app.

Bibliography

C. C. Aggarwal and C. Zhai. 2012. *An Introduction to Text Mining*, pages 1–10. Springer US, Boston, MA.

M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad. 2017. Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng*, 8(3):27.

M. Barlow. 2004. Software for corpus access and analysis. *How to use corpora in language teaching*, pages 205–221.

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.

D. Biber. 2012. *Corpus-Based and Corpus-driven Analyses of Language Variation and Use*. Oxford University Press.

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

F. Bond, L. Morgado da Costa, and T. A. Lê. 2015. IMI — a multilingual semantic annotation environment. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 7–12, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

F. Bond and K. Paik. 2012. A survey of wordnets and their licenses.

A. Brunnengräber. 2018. *Klimaskeptiker im Aufwind*, pages 271–292. Springer Fachmedien Wiesbaden, Wiesbaden.

A. Brunnengräber. 2013. *Klimaskeptiker in Deutschland und ihr Kampf gegen die Energiewende*.

D. Chen and C. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

W. Cheng. 2013. *Corpus-Based Linguistic Approaches to Critical Discourse Analysis*, pages 1–8.

S. Conrad. 2002. Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22:75–95.

H. Cui, R. Sun, K. Li, M.-Y. Kan, and T.-S. Chua. 2005. Question answering passage retrieval using dependency relations. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 400–407, New York, NY, USA. Association for Computing Machinery.

H. Cunningham. 2005. Information extraction, automatic. *Encyclopedia of language and linguistics*,, 3(8):10.

R. Del Gaudio and A. Branco. 2007. Automatic extraction of definitions in portuguese: A rule-based approach. In *Progress in Artificial Intelligence*, pages 659–670, Berlin, Heidelberg. Springer Berlin Heidelberg.

A. E. Dessler and E. A. Parson. 2006. The science and politics of global climate change : A guide to the debate.

R. Dunlap. 2013. Climate change skepticism and denial: An introduction. *American Behavioral Scientist*, 57:691–698.

L. Espinosa-Anke and H. Saggion. 2014. Applying dependency relations to definition extraction. In *Natural Language Processing and Information Systems*, pages 63–74, Cham. Springer International Publishing.

S. Evert. 2009. Corpora and collocations. *Corpus Linguistics: An International Handbook*.

C. Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

J.R. Firth. 1957. Modes of meaning, papers in linguistics.

L. Flowerdew. 2012. *Corpus-based discourse analysis*, pages 174–187.

C.L. Gagné and T.L. Spalding. 2006. Conceptual combination: Implications for the mental lexicon in: Libben g, jarema g, editors. the representation and processing of compound words.

P. Gamallo, M. Garcia, and S. Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, Avignon, France. Association for Computational Linguistics.

M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. König. 2008. Blews: Using blogs to provide context for news articles.

H. Glück and M. Rödel, editors. 2016. *Metzler Lexikon Sprache*, 5 edition. SpringerLink : Bücher. J.B. Metzler, Stuttgart.

G. Goeminne. 2012. Lost in translation : climate denial and the return of the political. *Global Environmental Politics*, 12(2):1–8.

S. T. Gries. 2001. A corpus-linguistic analysis of english -ic vs. -ical adjectives. *Icame Journal*, 25(i):65–108.

R. Grundmann and R. Krishnamurthy. 2010. The discourse of climate change: A corpus-based approach.

O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme. 2020. Training a broad-coverage German sentiment classification model for dialog systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1627–1632, Marseille, France. European Language Resources Association.

B. Hamp and H. Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

M. Honnibal and I. Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

T. Hornschuh. 2008. Skeptische Kommunikation in der Klimadebatte. *Von der Hypothese zur Katastrophe: Der anthropogene Klimawandel im Diskurs zwischen Wissenschaft, Politik und Massenmedien*, pages 141–154.

X. Hu and H. Liu. 2012. Text analytics in social media. In *Mining text data*, pages 385–414. Springer.

C. Hutto and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

M. A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

J. Jiang. 2012. *Information Extraction from Text*, pages 11–41. Springer US, Boston, MA.

D. Jurafsky and J. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2.

J. L. Klavans and S. Muresan. 2000. Definder: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In *Proceedings of the AMIA Symposium*, page 1049. American Medical Informatics Association.

R. Krishnamurthy and W. Teubert. 2007. Introduction to corpus linguistics: Critical concepts in linguistics (6 volumes).

S. Kübler, R. McDonald, and J. Nivre. 2009. Dependency parsing. *Synthesis Lectures on Human Language Technologies*, 2(1):1–127.

S. Langer. 1998. Zur Morphologie und Semantik von Nominalkomposita.

C. Leacock and M. Chodorow. 1998. *Combining Local Context and WordNet Similarity for Word Sense Identification*, volume 49, pages 265–.

Y. Li, D. McLean, Z. Bandar, J. O'Shea, and K. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18:1138–1150.

X. Liu, H. Chen, and W. Xia. 2022. Overview of named entity recognition. *Journal of Contemporary Educational Research*, 6(5):65–68.

S. Loria. 2018. textblob documentation. *Release 0.15*, 2.

B. M. Maier. " no planet b": An analysis of the collective action framing of the social movement fridays for future. Master project, Jönköping University.

A. Mansouri, L. S. Affendey, and A. Mamat. 2008. Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344.

E. Marris. 2019. Why young climate activists have captured the world's attention. *Nature*, 573(7775):471–472.

T. McEnery and A. Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Y. Mejova. 2009. Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

S. Muresan and J. Klavans. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

F. Å. Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

J. Nivre. 2010. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152.

J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

A. Orenha-Ottaiano, M. Garcia, M. Olímpio, M.-C. L'Homme, M. Ramos, C. Valêncio, and W. Tenório. 2022. Corpus-based methodology for an online multilingual collocations dictionary: First steps.

L. Ortner and E. Müller-Bollhagen. 1991. *Deutsche Wortbildung: Typen und Tendenzen in der Gegenwartssprache: eine Bestandsaufnahme des Instituts für deutsche Sprache, Forschungsstelle Innsbruck. Substantivkomposita:(Komposita und kompositionsähnliche Strukturen 1). Hauptteil 4*, volume 4. Walter de Gruyter.

T. Pedersen, S. Patwardhan, J. Michelizzi, et al. 2004. Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI*, volume 4, pages 25–29.

S. Rahmstorf. 2007. Alles nur Klimahysterie? Wie „Klimaskeptiker "die Öffentlichkeit verschaukeln und wirksame Klimaschutzmaßnahmen verhindern. *Universitas*, 9:895–913.

A. Rambousek, A. Horák, V. Suchomel, and L. Kocincová. 2014. Semiautomatic building and extension of terminological thesaurus for land surveying domain. In *The 8th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2014, Karlova Studanka, Czech Republic, December 5-7, 2014*, pages 129–137. Tribun EU.

L. Ramshaw and M. Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

M. Rössler. 2004. Corpus-based learning of lexical resources for german named entity recognition. In *LREC*. Citeseer.

RStudio Team. 2020. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

B. Schlücker. 2012. *Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung*, pages 1–26. De Gruyter, Berlin, Boston.

B. Schneider. 2018. *Klimabilder*. Matthes & Seitz, Berlin.

M. Siegel and F. Bond. 2021. OdeNet: Compiling a GermanWordNet from other resources. In *Proceedings of the 11th Global Wordnet Conference*, pages 192–198, University of South Africa (UNISA). Global Wordnet Association.

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

J. Sinclair and R. Carter. 1991. *Corpus, Concordance, Collocation*. Describing English language. Oxford University Press.

S. Spala, N. A Miller, F. Dernoncourt, and C. Dockhorn. 2020. Semeval-2020 task 6: Definition extraction from free text with the deft corpus.

S. Spala, N. A. Miller, Y. Yang, F. Dernoncourt, and C. Dockhorn. 2019. DEFT: A corpus for definition extraction in free- and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 124–131, Florence, Italy. Association for Computational Linguistics.

A. Stefanowitsch. 2020. *Corpus linguistics*. Number 7 in Textbooks in Language Sciences. Language Science Press, Berlin.

P. Velardi, R. Navigli, and P. D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems*, 23(5):18–25.

A. Veyseh, F. Dernoncourt, D. Dou, and T. Nguyen. 2020. A joint model for definition extraction with syntactic connection and semantic consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9098–9105.

E. Volken. 2010. Die Argumente der Klimaskeptiker. In *ScNat ProClim-Forum for Climate and Global Change. Forum of the Swiss Academy of Science*, pages 1–8.

K. Welbers, W. Van Atteveldt, and K. Benoit. 2017. Text analysis in r. *Communication Methods and Measures*, 11(4):245–265.

E. Westerhout. 2010. Definition extraction for glossary creation: A study on extracting definitions for semi-automatic glossary creation in dutch.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. pages 133–138.

L. Young and S. Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

A Appendix

A.1 Implementation

Please note: The full implementation of the project can be found on GitHub under `https://github.com/ajgoecke/thesis`.

Lemma	Noun Forms	Genus
Glaubenslehre	Glaubenslehre, Glaubenslehren	f
Verdummung	Verdummung	f
Notstandsregierung	Notstandsregierung, Notstandsregierungen	f
Kasteiung	Kasteiung, Kasteiungen	f
Bremser	Bremser, Bremsers, Bremsern	m
Besoffenheit	Besoffenheit	f
Verblödung	Verblödung	f
Alarmist	Alarmist, Alarmisten	m
Hysteriker	Hysteriker, Hysterikers, Hysterikern	m
Donna	Donna, Donnas, Donnen	f
Aktivismus	Aktivismus	m
Gläubigkeit	Gläubigkeit	f
Abzockerei	Abzockerei, Abzockereien	f
Bewegtheit	Bewegtheit	f
Orthodoxie	Orthodoxie	f
Gnom	Gnom, Gnome, Gnoms, Gnomen	m
Virus	Virus, Viren	n

Table A.1: List of compound words that received a manually retrieved noun form (see section 5.1.1).

word	n	keyword	tag	corpus
innen	10	klimaaktivist	post	P2022
tan	2	klimaaktivistin	pre	P2022
|göppinge	2	klimamahnwache	post	P2022
alt68erökoopa	2	klimaretterin	post	P2022
klimaalarm	4	klimaalarm	pre	C2022
klimaalarm	4	klimaalarm	post	C2022
09.	2	klimaretter	post	C2022
info	51	klimaretter	post	C2022

Table A.2: List of collocations that are removed in the manual cleaning step (see section 5.2.2).

POS Tag	Explanation	Modifier Label	Explanation
ADJA	adjective, attributive	mo	modifier
ADJD	adjective, adverbial or predicative	mnr	postnominal modifier
ADV	adverb	nk	noun kernel element
ADJ	adjective		

Table A.3: Table of abbreviations for POS tags and modifiers used for dependency parsing in section 5.3.3.

A.2 Attribution

In the following, we provide the translation of the examples from chapter 6.

Ex. 5 To all **climate hysterics**: Everything is going to plan, we are all allowed to continue to live except you start to withdraw people's basis of life by taking away their much needed energy sources.

Ex. 6 Ultimately another Greta- or FFF-bashing on different levels, including the crazy fight of the **climate deniers** (AfD) and climate skeptics (Springerverlag, die Welt) AGAINST a satire which also ultimately expressed a Greta / FFF-bashing.

Ex. 7 In future we **climate realists** are going to receive support from the "ice-holy" from Potsdam.

Ex. 8 Hello dear climate friends, I am also a young **climate friend** and want to make a few remarks to your summary of the "GradStudie" ... you're writing: The energy industry the institute discovered, inter alia, that renewable energies, explicitly wind and solar energy, have to be expanded massively in the upcoming years.

Ex. 9 Thus, climate deniers stand against a satire which ultimately denounces the climate movement as "**climate hysteria**" because those stubborn right winged ideologists do not understand intellectually and because they searched for and found a scapegoat or rather anchor in their fight against the "lying press".

Ex. 10 Because there would be more urgent global problems than "**climate hysteria**": for instance famine or the assurance of retirement.

Ex. 11 This one wisely deals with the medial devaluation of the so-called "**climate deniers**".

Ex. 12 My personal impression: We "**climate deniers**" couldn't have done it so much better!!!

A.3 Definition Phrasing

auch	eben	ja	Kohleprojekt	trotzallem	Nimmer
durch	zu	nun	Dossier	sein	denn
Dekolonialer	eur	weiter	so	dieser	WORT
ihr	mein	wie	unser	TwitterAccount	hier
aller	Flashcrash	irgendwelcher	anderer	sprich	eu
ganz	daher	jeder	ebenso	nach	zurück

Table A.4: List of modifiers that are removed in the manual cleaning step (see section 7)

A.3.1 Definition Strings

A.3.1.1 German Strings

Person: Der Begriff '**{COMPOUND}**' bezeichnet eine Person, die in einer gewissen Beziehung zum Klimawandel steht. Der Begriff wird in unserem Korpus **{CON_FREQ}** Mal von den Klimaforschungsskeptikern und **{PRO_FREQ}** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **{CON_TFIDF}** für die Skeptiker und **{PRO_TFIDF}** für die Vertreter.

Location, Action, Abstraction: Der Begriff '**{COMPOUND}**' wird in unserem Korpus **{CON_FREQ}** Mal von den Klimaforschungsskeptikern und **{PRO_FREQ}** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **{CON_TFIDF}** für die Skeptiker und **{PRO_TFIDF}** für die Vertreter.

Group: Der Begriff '**{COMPOUND}**' bezeichnet einen Zusammenschluss von Personen im Bezug auf den Klimawandel. Der Begriff wird in unserem Korpus **{CON_FREQ}** Mal von den Klimaforschungsskeptikern und **{PRO_FREQ}** Mal von den Klimaforschungsvertretern verwendet. Auf den gesamten Korpus gesehen entspricht das einer relativen Häufigkeit (TF-IDF) von **{CON_TFIDF}** für die Skeptiker und **{PRO_TFIDF}** für die Vertreter.

Sentiment: In unserem Korpus Sample ist der Begriff tendenziell **{SENTIMENT}** konnotiert.

Attribution: Verwendet wird '**{COMPOUND}**' hierbei im Sinne einer **{ATTRIBUTION}**.

Sarcasm: Die Verwendung wird **{SARCASM}** als sarkastisch eingestuft.

P2022 Modifiers: Im Subdiskurs der Klimaforschungsvertreter wird der Begriff von Wörtern wie **{PRO_MODS}** modifiziert.

C2022 Modifiers: Wörter wie **{CON_MODS}** treten auf, um den Begriff im Subdiskurs der Klimaforschungsskeptiker näher zu beschreiben.

Person Base: Im Zusammenhang mit dem Begriff erwähnt

C2022 Persons: der Skeptiker Korpus die Person(en) **{CON_PERS}**

P2022 Persons: der Vertreter Korpus die Person(en) **{PRO_PERS}**

Organisations Base: Im Kontext von '**{COMPOUND}**' erfolgt die Nennung folgender Organisation(en):

C2022 Organisations: **{CON_ORG}** (Skeptiker Korpus)

P2022 Organisations: **{PRO_ORG}** (Vertreter Korpus)

Collocations: Kollokationen: **{CON_COLLS}** **{PRO_COLLS}**

Similar Words: Siehe auch: **{SIMILAR_WORDS}**

A.3.1.2 English Translation of the Definition Strings

Base Info: **{COMPOUND}**, **{ARTICLE}**

Person: The term **{COMPOUND}** refers to a person that adopts a certain attitude towards climate change. In our corpus the term is used **{CON_FREQ}** times by the climate research skeptics and **{PRO_FREQ}** times by the climate research supporters. With respect to the complete corpus this corresponds to a relative frequency (TF-IDF) of **{CON_TFIDF}** for the skeptics and **{PRO_TFIDF}** for the supporters.

Location, Action, Abstraction: In our corpus the term **{COMPOUND}** is used **{CON_FREQ}** times by the climate research skeptics and **{PRO_FREQ}** times by the climate research supporters. With respect to the complete corpus this corresponds to a relative frequency (TF-IDF) of **{CON_TFIDF}** for the skeptics and **{PRO_TFIDF}** for the supporters.

Group: The term **{COMPOUND}** refers to a group of persons in relation to climate change. In our corpus the term is used **{CON_FREQ}** times by the climate research skeptics and **{PRO_FREQ}** times by the climate research supporters. With respect to the complete corpus this corresponds to a relative frequency (TF-IDF) of **{CON_TFIDF}** for the skeptics and **{PRO_TFIDF}** for the supporters.

Sentiment: In our corpus sample the term is generally speaking **{SENTIMENT}** connoted.

Attribution: **{COMPOUND}** is used in terms of a **{ATTRIBUTION}**.

Sarcasm: Its use is classified as sarcastic for **{SARCASM}** of the cases.

P2022 Modifiers: In the sub-discourse of climate research supporters the term is being modified by words such as **{PRO_MODS}**.

C2022 Modifiers: Modifiers such as **{CON_MODS}** occur to specify the term in the sub-discourse of the climate research skeptics.

Person Base: In the context of the term

C2022 Persons: the skeptic's corpus mentions the person(s) **{CON_PERS}**

P2022 Persons: the supporter's corpus mentions the person(s) **{PRO_PERS}**

Organisations Base: The following organisations are referred to in the context of **{COMPOUND}**:

C2022 Organisations: **{CON_ORG}** (skeptic's corpus)

P2022 Organisations: **{PRO_ORG}** (supporter's corpus)

Collocations: Collocations: **{CON_COLLS}** **{PRO_COLLS}**

Similar Words: See also: **{SIMILAR_WORDS}**

A.3.2 Translation of the Definition Texts

Ex. 15 **Climate Denier**

The term **'climate denier'** refers to a person that adopts a certain attitude towards climate change. In our corpus the term is used **127** times by the climate research skeptics and **41** times by the climate research supporters. With respect to the complete corpus this corresponds to a relative frequency (TF-IDF) of **0.965** for the skeptics and **0.096** for the supporters. **'climate denier'** is used in terms of a **self-attribution** by the skeptics and an **external attribution** in the supporter's corpus. Its use is classified as sarcastic for **2%** of the cases in the supporter's discourse and for **44%** of the cases in the skeptic's discourse. In our corpus sample the term is generally speaking **negatively** connoted. Modifiers such as **'no' and 'called'** occur to specify the term in the sub-discourse of the climate research skeptics. In the context of the term the skeptic's corpus mentions the person(s) **Stefan Rahmstorf and Michael Limburg** and the supporter's corpus mentions the person(s) **Greta Thunberg and Tom Buhrow**. The following organisations are referred to in the context of **'climate denier'**: **EIKE, IPCC** (skeptic's corpus) und **AFD, FFF** (supporter's corpus).

Collocations: **'invention', 'EIKE'** (skeptic's corpus) and **'AfD'** (supporter's corpus)

See also: **Climate lie**

Ex. 16 **Climate Elite**

The term **'climate elite'** refers to a group of persons in relation to climate change. In our corpus the term is used **4** times by the climate research skeptics and **0** times by the climate research supporters. With respect to the complete corpus this corresponds to a relative frequency (TF-IDF) of **0.044** for the skeptics and **0.0** for the supporters. **'climate elite'** is used in terms of an **external attribution** by the skeptics. Its use is classified as sarcastic for **25%** of the cases in the skeptic's discourse. In our corpus sample the term is generally speaking **neutrally** connoted. In the context of the term the skeptic's corpus mentions the person(s) **Maybritt Illner and Christina Figueres**. The following organisations are referred to in the context of **'climate denier'**: **EIKE, Die Grünen** (skeptic's corpus).

Collocations: **'alternative', 'authority'** (skeptic's corpus)

A.4 List of Glossary Terms

Klimaabzockerei, Klimaaktivismus, Klimaaktivist, Klimaaktivistin, Klimaalarm, Klimaalarmist, Klimaanbeter, Klimaapokalypse, Klimaapokalyptiker, Klimaapostel, Klimaargument, Klimaarmageddon, Klimaasyl, Klimaaufschrei, Klimaaufstand, Klimabankrott, Klimabeeinflussung, Klimabefürworter, Klimabesoffenheit, Klimabesorgnis, Klimabetrug, Klimabetrüger, Klimabewegtheit, Klimabibel, Klimabigotterie, Klimabluff, Klimablödsinn, Klimabrandstifter, Klimabremser, Klimachaos, Klimacrash, Klimadampfer, Klimademagoge, Klimadepression, Klimadialektik, Klimadiktatur, Klimadiplomatie, Klimadogma, Klimadonna, Klimadramatik, Klimadürre, Klimaelite, Klimaerkrankung, Klimaernsthaftigkeit, Klimaerzählung, Klimaestablishment, Klimaethik, Klimaextremismus, Klimafachkraft, Klimafanatiker, Klimafeind, Klimafestung, Klimafieber, Klimafluch, Klimaflüsterer, Klimafreund, Klimafreundin, Klimafreundlichkeit, Klimafront, Klimaführerin, Klimagangster, Klimagarde, Klimagau, Klimagefahr, Klimagefasel, Klimagerechtigkeit, Klimageschrei, Klimageschäft, Klimagesinnung, Klimaglaube, Klimaglaubensbekenntnis, Klimaglaubenslehre, Klimagläubigkeit, Klimagnom, Klimagott, Klimagunst, Klimagöttin, Klimahardliner, Klimaherausforderung, Klimahoax, Klimahokuspokus, Klimahose, Klimahybris, Klimahype, Klimahypothese, Klimahysterie, Klimahysteriker, Klimahölle, Klimahüpfer, Klimaideologie, Klimaikone, Klimainsider, Klimaintelligenz, Klimairrsinn, Klimajünger, Klimakabarett, Klimakaiserin, Klimakamarilla, Klimakampf, Klimakanzler, Klimakanzlerin, Klimakarawane, Klimakasteiung, Klimakataklysmus, Klimakatastrophentheorie, Klimakatechismus, Klimaketzer, Klimakirche, Klimaklamauk, Klimaklempner, Klimakleriker, Klimakollekte, Klimakolonialismus, Klimakommissar, Klimakompetenz, Klimakonfusion, Klimakonklave, Klimakonsens, Klimakonsortium, Klimakontrolle, Klimakontroverse, Klimakreationismus, Klimakreuzzug, Klimakrieg, Klimakrieger, Klimakritik, Klimakritiker, Klimakult, Klimakäse, Klimakönigin, Klimalaie, Klimaleidenschaftlichkeit, Klimaleugner, Klimaleugnung, Klimalobby, Klimalösegeld, Klimalüge, Klimalügner, Klimamacher, Klimamafia, Klimamahnwache, Klimamanipulation, Klimamantra, Klimamilliardär, Klimamisere, Klimamonster, Klimamutti, Klimamärchenonkel, Klimanationalismus, Klimanonsens, Klimanotfall, Klimanotlage, Klimanotstandsregierung, Klimaorthodoxie, Klimapanik, Klimapanikmache, Klimapapst, Klimaparadies, Klimapflicht, Klimaplanwirtschaft, Klimapolemik, Klimapopulismus, Klimaprediger, Klimapresse, Klimaprofessor, Klimaprofiteur, Klimapropaganda, Klimapropagandafilm, Klimapropagandist, Klimaprotestler, Klimapseudowissenschaft, Klimapäckchen, Klimapäpstin, Kli-

maquatsch, Klimarealismus, Klimarealist, Klimareligion, Klimaretter, Klimaretterin, Klimarettung, Klimascharade, Klimaschiff, Klimaschmerz, Klimaschrecken, Klimaschuld, Klimaschwachsinn, Klimaschwindel, Klimaschwindler, Klimaschänder, Klimaschändung, Klimasekte, Klimasektierer, Klimashow, Klimaskepsis, Klimaskeptiker, Klimaskeptikerin, Klimaskeptizismus, Klimasozialismus, Klimaspinner, Klimastaat, Klimastopp, Klimastreit, Klimastreiter, Klimataliban, Klimateppich, Klimateufel, Klimatheater, Klimatheologie, Klimatod, Klimatrash, Klimatraumatisierung, Klimatrip, Klimatyrannei, Klimaunfug, Klimaunfähigkeit, Klimaungerechtigkeit, Klimaunsinn, Klimauntergang, Klimavatikan, Klimaverblödung, Klimaverbrechen, Klimaverbrecher, Klimaverdummung, Klimaverleugnung, Klimavernunft, Klimaverrücktheit, Klimaverstand, Klimaverteidigung, Klimaverweigerung, Klimaverwirrung, Klimavirus, Klimawahn, Klimawahnsinn, Klimawandler, Klimawarner, Klimaweltuntergang, Klimawendehals, Klimawirklichkeit, Klimazar, Klimazerrüttung, Klimazerstörer, Klimazerstörung, Klimazeugs, Klimazipfel, Klimazirkus, Klimazunft, Klimazwang, Klimaüberhitzung

Eidesstattliche Erklärung

Hiermit versichere ich, dass meine Master Thesis “Semi-Automated Definition Phrasing” (“Creating a Discourse Glossary for Climate Change Compounds”) selbständig verfasst wurde und dass keine anderen Quellen und Hilfsmittel als die angegebenen benutzt wurden. Diese Aussage trifft auch für alle Implementierungen und Dokumentationen im Rahmen dieses Projektes zu.

Potsdam, den 13. Februar 2023,

_____

(Anna-Janina Goecke)