An analysis of the development of the German touch verbs 'anfassen', 'angreifen', 'anlangen' with text data from `Common Crawl`. A comparison between Austria and Germany.

Project Report
194.147 Interdisciplinary Project in Data Science
Marina Sommer | 11778902 | TU Wien
November 17, 2024

Supervisor: Assistant Prof. Mag. Mag. Dr. Andreas Baumann, University of Vienna
Co-supervisor: Assistant Prof. Mag.a Dr.in Julia Neidhardt, TU Wien

Special thanks to the CD Lab for Recommender Systems (`https://recsys-lab.at`) for providing the resources and infrastructure for this project!

Abstract

Since natural language develops at any time, the aim of this project is to find out if the usage of the German touch verbs 'anfassen', 'angreifen' and 'anlangen' has changed over the last decade. This analysis includes a comparison of two varieties of German, one spoken in Austria and the other in Germany. The text data is collected from `Common Crawl` and is used to train one word embedding model per time period and variety. The word sets of the semantically related words to the target words are compared using the 'Jaccard index' in two different ways. Ultimately, broader insights can be obtained in relation to the findings of Ahlers and Fink (2017) due to the time aspect involved in the evaluation. The results show that the word sets for these verbs have changed over time, with more variation in Austria, and that the word sets of 'angreifen' are linked to actions of attack as anticipated by Ahlers and Fink (2017). Despite facing technical limitations in data collection and preparation, this exploratory study lays the groundwork for future research utilizing `Common Crawl` to examine linguistic trends over extended time periods.

Problem Statement

The intention of this project is to analyze if the German verbs 'anfassen', 'angreifen' and 'anlangen' are used as synonyms, i.e. in the same context, in everyday language. Initially, the following two research questions should be answered:

- How has the usage of the German touch verbs 'anfassen', 'angreifen', 'anlangen' changed over the last years?
- What differences can be observed by comparing the usage of these verbs in the German and Austrian everyday language?

I analyze the same set of words as done by Ahlers and Fink (2017) for their experimental study on motor concepts. However, I use a corpus-based data science approach and concentrate on a comparison between Austria and Germany. While working on this project, I realized that it is not easily possible to find out *how* the usage of these words has changed. For the evaluation, I make use of the Jaccard index (JI), which measures the similarity between finite sample sets, for instance, word sets of nearest neighbors of the target words. It returns a value between 0 and 1, which only gives an indication if the word sets are completely different (JI = 0) or equal (JI = 1). For that reason, I changed the research question to "To what extent regarding the Jaccard index has the usage of the German touch verbs 'anfassen', 'angreifen', 'anlangen' changed over the last years?".

Methods

Data Collection

As a first step, I have to collect text data from `Common Crawl`[1], an open-source platform which provides a huge amount of web data since 2008. The idea is to create a text corpus for multiple years per variety by searching for websites with the top-level domain (TLD) 'de' for Germany and 'at' for Austria.

The exploration of the `Common Crawl` data comprises the main part of this project, since it is essential to understand the unique data repository and data format before working with it. The data is hosted by `Amazon Web Services (AWS)` and can be processed directly in the cloud, but it can also be downloaded over a HTTP(S) connection. A new release of data, denoted as 'crawl', is available almost every month since 2014 and contains around three billion web pages. For my analysis, I use data from the early beginnings, from the latest crawl, and from one point of time in between.

The `Common Crawl` data repository is built upon a specific index, i.e. data format, which should make it easy and efficient to work with the data. Every crawl has one index file named `'cluster.idx'`. It lists every 3,000th record in alphabetical order of the top-level domain and gives information, where the data is stored. The following line represents an Austrian web page:

```
at,orf)/stories/3218646 20230925052948 cdx-00003.gz 522482295 181686 14613
```

The URL consists of the top-level domain `'at'`, the root domain `'orf'` and the path `'/stories/3218646'`. Another important information in this line is `'cdx-00003.gz'`, which is the name of the gzip-compressed `'.cdx'` file where more information of the web archive data can be found. Each crawl has exactly 300 of these files. Since they are quite large regarding memory consumption, it makes sense to download only the ones containing the selected TLDs and if there are multiple files per TLD available, to restrict it to one or two downloads per TLD. Every row in such a file provides the following information in JSON format:

```
at,orf)/stories/3218646 20230925052948
{"url": "https://orf.at/stories/3218646/",
"mime": "text/html",
"mime-detected": "text/html",
"status": "200",
"digest": "GDIUH3H5DG7KRBIL43C4YZXMX4FZY3OZ",
"length": "11815",
"offset": "490478697",
"filename": "crawl-data/CC-MAIN-2023-40/segments/1695233506686.80/warc/
CC-MAIN-20230925051501-20230925081501-00554.warc.gz",
"charset": "UTF-8",
"languages": "deu"}
```

For this project, relevant features are `'length'` and `'filename'`. The goal is to create a large text corpus for each selected TLD, that is why only web archives with length exceeding 10,000 characters are taken into account. The variable `'filename'` specifies the file where the exact content of the web archive is stored. The web data is stored in various formats[2]:

1. Web Archive format (WARC) files: store the raw crawl data

[1] `https://commoncrawl.org/`
[2] `https://www.commoncrawl.org/get-started`

2. Web Archive Transformation (WAT) files:
   store computed metadata for the data stored in the WARC

3. Web Extracted Text (WET) files:
   store extracted plaintext from the data stored in the WARC

For reducing memory space, it is preferable to use the smallest files available, which are the WET files. Each of these files contains the metadata and plain text data of various web archives. Unfortunately, these web pages are not sorted by TLD or language, which is inconvenient regarding memory consumption. Following is the content of the web archive from the example above:

```
WARC/1.0
WARC-Type: conversion
WARC-Target-URI: https://orf.at/stories/3218646/
WARC-Date: 2023-09-25T05:29:48Z
WARC-Record-ID: <urn:uuid:f7763db5-c3b9-4bd5-851c-81684e9f55a5>
WARC-Refers-To: <urn:uuid:33637f4e-6cb1-488c-ba21-415dc7231beb>
WARC-Block-Digest: sha1:ZLHCBPDFHLIRBF7YT5OVEHMONUUMJGQY
WARC-Identified-Content-Language: deu
Content-Type: text/plain
Content-Length: 7340

Ibiza-U-Ausschuss: Vom Vorsitzenden zur Auskunftsperson - news.ORF.at
Zum Inhalt [AK+1] / Zur ORF.at-Navigation [AK+3]
Fernsehen
TVthek
Sound
Topos
Debatte
Österreich
Wetter
Sport
News
ORF.at im Überblick
news
Navigation
News
Sport
Wetter
Österreich
Debatte
Sound
Topos
TVthek
Fernsehen
Alle ORF Angebote
ORF.at/Carina Kainz
Ibiza-U-Ausschuss
Vom Vorsitzenden zur Auskunftsperson
Im Ibiza-U-Ausschuss ist am Donnerstag zum zweiten Mal der Ausschussvorsitzende,
Parlamentspräsident Wolfgang Sobotka (ÖVP), befragt worden. Im Zentrum standen
unter anderem Fragen zum Alois-Mock-Institut und zu Gesprächen mit dem suspen-
```

```
dierten Sektionschef Christian Pilnacek. Dass Sobotka den Sitzplatz binnen weni-
ger Stunden wechselte, war nicht ganz friktionsfrei.
24.06.2021 20.06
24. Juni 2021, 20.06 Uhr
Dieser Artikel ist älter als ein Jahr.
[...]
```

After downloading a specific amount of WET files, the text data of each file has to be extracted and added to the final text corpus. This data usually includes all parts of a website, e.g header, navigation, sidebar and footer. But most importantly is the web content, since this is the part where everyday language can be collected. The following conditions should ensure that the text corpus consists of high-quality text data:

- TLD of `'WARC-Target-URI'` matches the selected TLD
- `'Content-Length'` $> 10{,}000$
- `'Content-Type'` = 'text/plain'
- number of stopwords > 0
- number of stopwords $\geq$ number of punctuation marks
- number of stopwords $\geq 2 \cdot$ number of line breaks

Stopwords are commonly used words like function words. They primarily serve grammatical or structural purposes rather than conveying significant meaning in the text. Nevertheless, their existence in a document can verify that it uses the right language and that it consists of connected text, i.e. sentences with a subject, a predicate and at least one object. As visible in the example above, the web archive can contain many nouns without context, followed by line breaks or punctuation marks, which represent headers on the original web page. This text data does not provide context information about verbs and should not be used for the final text corpus. That is why I use the number of stopwords in each record, as well as its relation to the number of punctuation marks and line breaks, as an important feature for eliminating low-quality text.

Data Preparation

After creating the raw text corpus, a few typical data preparation steps have to be done. For example, removing very short lines, as well as URLs and HTML tags. After sentence tokenization, word tokenization with lemmatization follows. Every punctuation mark and every token exceeding 15 characters is removed, as well as duplicated sequential lines and very short sentences with less than five words. For German lemmatization, the Python package `spaCy` provides various models[3], which differ in type and size. For this project, the model `'de_core_news_md'`, which is of medium size, is used. For future needs, the `spaCy` model can easily be changed. This entire procedure of creating a pre-processed text corpus has to be done for every crawl and every variety, i.e. top-level domain.

Training of `Word2Vec` Models

Subsequently, each text corpus is used to train a word embedding model. In particular, the skip-gram model 'Word2Vec'[4] from the `gensim` Python package is used for this purpose. This

[3] `https://spacy.io/models/de`
[4] `https://radimrehurek.com/gensim/models/word2vec.html`

NLP approach was initiated by Mikolov, Chen, et al. (2013) in January 2013 and extended by Mikolov, Sutskever, et al. (2013) in October 2013. The model needs a broad text corpus as input and outputs a vector representation of each word in the vocabulary of the training data. Hence, it is possible to predict the context from one specific word by looking at its nearest neighbors in the vector space.

Evaluation and Results

The idea is to analyze the sets of semantically related words, named as nearest neighbors, of the target words 'anfassen', 'angreifen' and 'anlangen'. I compare the differences between the both varieties, Austria and Germany, and study the changes over time if there are any. The 'Jaccard index'[5], which measures the similarity between finite sample sets, serves as key metric. It can be calculated between

- the word set of the first time period available and any given time period (per target word and variety) and
- the word sets of the two varieties (per target word and time period).

I expect similar findings as in the study conducted by Ahlers and Fink (2017) with respect to the differences between Austria and Germany. The analysis of the development of these specific words over time is a completely new approach and cannot be compared with other results. I use three data points ('2014-00', '2019-35' and '2024-38') to answer the first research question. Since there were not a lot of Austrian web pages collected when `Common Crawl` started, I have to acquire data from four different crawls in 2013 and 2014, and refer to their combination as '2014-00'. For the top-level domain 'de' only the crawl '2013-20' is used, but for better readability, it is also be denoted with '2014-00'. The analysis has been done with nearest neighbors of different size, e.g. 20, 50, 100 and 200 words. Since the results are similar, the evaluation in this report is presented for word sets of size 100.

The following graphs display the Jaccard index between the top-level domains 'at' and 'de' and between the first available crawl (denoted with '2014-00') and the crawls from '2019-35' and '2024-38'. Overall, the values on the y-axis, which represent the Jaccard similarity, are very low. This means that the word sets are almost completely different from each other. Regarding the comparison between the word sets from 2014 and 2019 in Figure 1, we can see that the target word 'angreifen' has the largest share of common words between the two periods. The similarity values for 'de' are higher than for 'at', except for 'anfassen', where there is no intersection of the word sets for both varieties. In terms of language evolution, this means that 'angreifen' shows less change than the other two target words, and that there is a greater shift in Austria than in Germany.

Figure 2 displays the similarity values between the words sets from 2014 and 2024 and shows a similar pattern. The blue line illustrates that the word sets of all target words completely changed for TLD 'at', since the Jaccard index is always zero. This reinforces our previous observation in Figure 1 that the usage of the target words in Austria has changed more than in Germany. Additionally, the JI is lower when comparing the word sets from a longer time difference (2014 vs. 2024) with a shorter one (2014 vs. 2019). In Figure 2, the JI for the target word 'anfassen' for 'de' is greater than in Figure 1, which is the only exception when comparing the concrete values of both figures.

[5] `https://www.sciencedirect.com/topics/computer-science/jaccard-similarity`

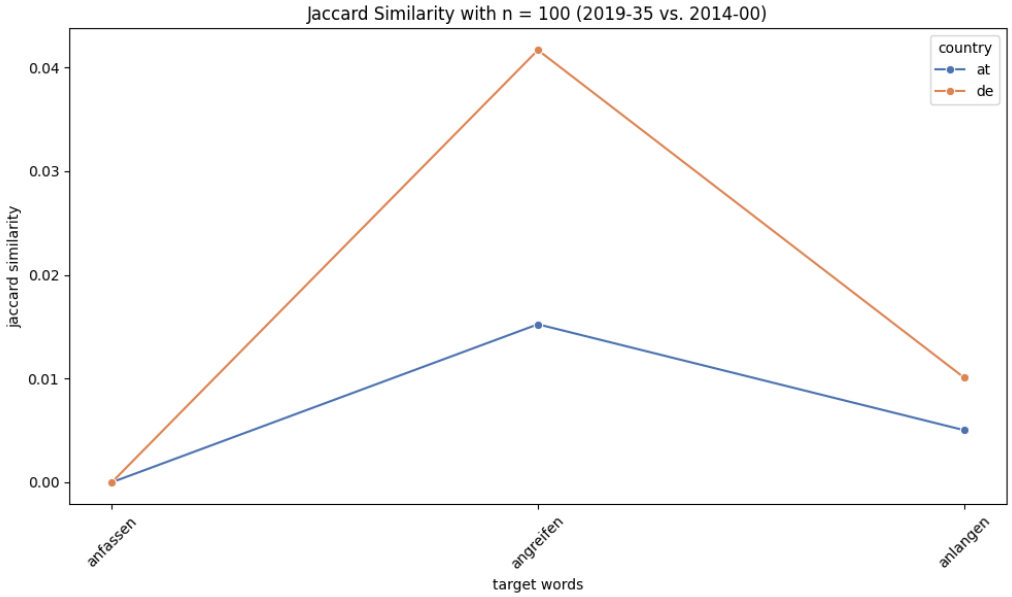

Figure 1: Comparison between 2014 and 2019

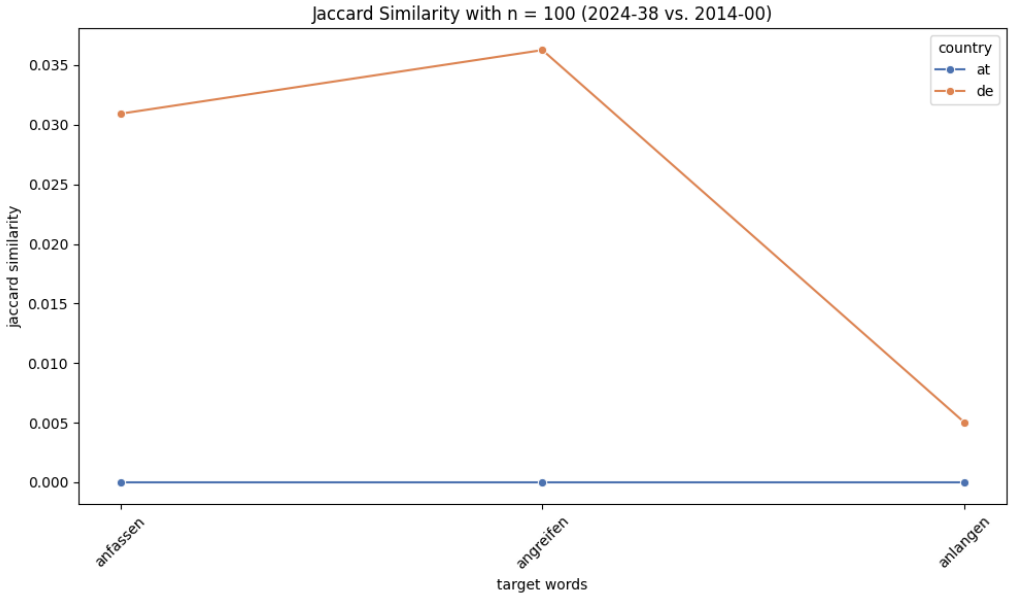

Figure 2: Comparison between 2014 and 2024

For all target words, we can observe different trends in Figure 3. The JI is calculated for word sets with varying top-level domain. The orange line indicates that the similarity between Austrian and German word sets for 'angreifen' is consistently the highest among the three target words. The similarity decreases from 2014 to 2019, but raises again in 2024. The blue line indicates a divergence in word usage related to 'anfassen' between Austria and Germany, showing that the sets became disjoint by 2019 and remained so in 2024. The green line, representing the similarity of the nearest neighbors of 'anlangen', implies that there is an overlap beginning in 2019 between Austria and Germany, which has stayed firm since then.

Table 1 illustrates that for Germany, the shared nearest neighbors for 'angreifen' are words that are usually used in the context of attacking and not of touching. This coincides with the findings

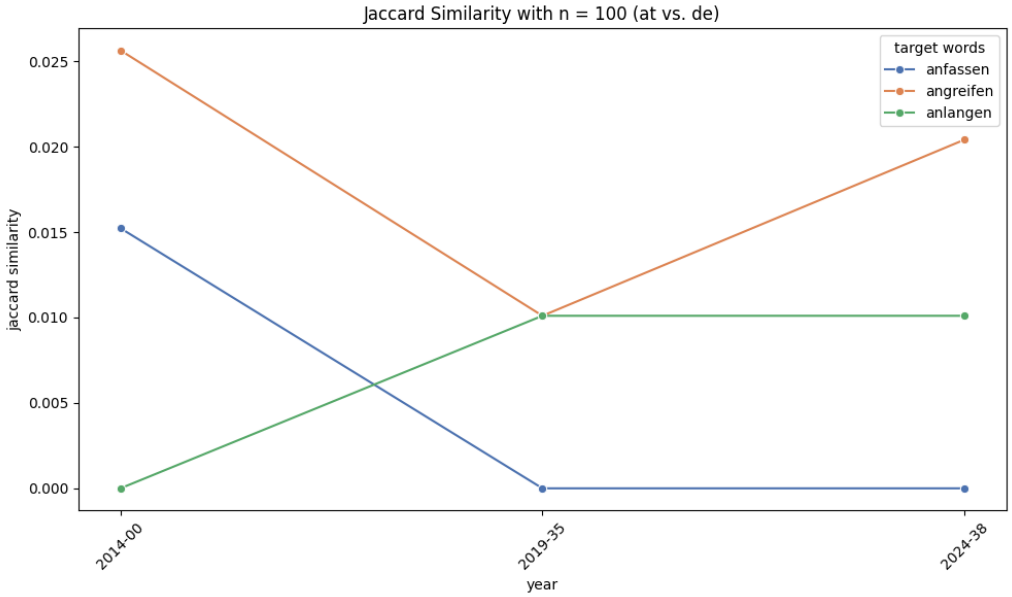

Figure 3: Comparison between Austria and Germany over time

of Ahlers and Fink (2017). The Austrian common words are neither related to a touch action, but they are also not that obviously associated with an attack, as seen in the first row of Table 1. For the target word 'anfassen', the common words are related to 'look at something', and 'anlangen' is associated with 'ankommen' for the top-level domain 'de'. These results do not match the expectations due to the lack of a 'touch' association. When comparing the word sets of both TLDs for the target word 'anfassen', we can see in the first row of Table 2 that the common words are related to body parts to touch with, like 'Finger' and 'Hand'. This is the only indication that the verb is used in the context of touching. All common words of various word sets can be found in Table 1 and Table 2 in the appendix. Furthermore, Table 3, Table 4 and Table 5 display the top 5 nearest neighbors of each of the three target words.

Challenges

While working on this project, I encountered several issues. From October 2023 on, the servers of `Common Crawl` are frequently confronted with an immense amount of download requests per second[6]. There was a time, when `Common Crawl` could not handle their requests at all, showing an error rate around 100%. If this problem occurs on a regular basis, it makes the `Common Crawl` platform unattractive for usage as a data source, due to its unreliability and instability.

Another problem I had to deal with was data storage capacity. The data is stored in a very unique way and has its own structure. However, for my task, a lot of files have to be downloaded. I am not able to filter the data by top-level-domain, since it is not stored domain-wise in its smallest unit. On an ordinary device, running this experiment is impossible without deleting WET files after each text corpus creation. There might be the option to process the data directly in the cloud, which could be seen as a further improvement.

Furthermore, since the text data is processed line per line, the data preparation takes a lot of time. I could not find a more efficient pipeline for German text data. In my opinion, the German lemmatization with `spaCy` does not work well and the original text data is not very clean. For

[6] `https://status.commoncrawl.org/`

instance, there are formatting mistakes with umlauts, which cannot be easily corrected since the umlauts are incorrectly encoded in the `Common Crawl` source files. Another typographical error is the concatenation of individual words. Addressing these issues would require a significant amount of time and effort, which exceeds the scope of this project. Consequently, the resulting word sets sometimes contain nonwords or words with spelling mistakes.

The comparison between the two varieties, Austria and Germany, is an important part of this study. The problem here is that Austria is noticeably underrepresented compared to Germany regarding the number of records `Common Crawl` provides. With the automated data collection and preparation, I am not able to create equally big text corpora for the two varieties. I have reduced the number of WET files which should be downloaded for Germany, but since one file can contain just one or over 100 relevant web pages, it is not clear in advance, how many of these files you exactly need to download to have an equal amount of text data in the end.

To answer the second research question, it is necessary to manually examine the output of the models. The downside of a manual inspection is that it can only be done for a small amount of data within a reasonable time frame.

Conclusion

In conclusion, this project aims to examine whether the usage of the German touch verbs 'anfassen', 'angreifen' and 'anlangen' has shifted over the past decade across two varieties, Austria and Germany. Using a corpus-based data science approach with data from `Common Crawl`, word embedding models for both varieties and three different time periods are created. For evaluation, the semantic similarities of the nearest neighbors are compared with the Jaccard index. While the initial goal was to analyze *how* the usage of these verbs changed over time, the limitations of the Jaccard similarity led to a revision of the first research question. The Jaccard index is not able to provide detailed context, so that is why the project focuses on determining the extent of change as reflected by the Jaccard index, rather than fully explaining how usage evolved.

The findings reveal the following key insights. First, the word sets associated with these verbs have indeed changed over time, with more pronounced changes in Austria than in Germany. Secondly, the usage of 'angreifen' shows the highest consistency between Austria and Germany, while 'anfassen' displays a clear divergence. Some word sets are clearly linked to actions of attack rather than physical touch, which was expected due to the results of Ahlers and Fink (2017).

Despite technical challenges related to the `Common Crawl` data collection and the limitations of the data pre-processing with `spaCy`, this study offers an initial exploration into the diachronic variation of touch verbs in German-speaking varieties over time. It provides a basis for future work with data from `Common Crawl`, which is the main focus of this project.

Domain-specific Lecture

The domain-specific lecture `'Introduction to Digital Humanities'`, which was selected for this project, has already been completed. It took place at the University of Vienna in winter term 2023. The lecture series was organized by Mag. Dr. Tara Andrews and provided a reasonably broad overview of the state of research and working methods in the Digital Humanities. A variety of speakers from different disciplines within the humanities discussed what digital methods bring to their own fields, and what they understand the Digital Humanities actually to be.

Literature

Ahlers, Timo and Juliane Fink (2017). "Motor-concept variation in the German verbs 'anfassen', 'angreifen', 'anlangen'. Differences between Austria, Germany and Switzerland". In: *Dialectologia et Geolinguistica* 25.1, pp. 69–91. DOI: doi:10.1515/dialect-2017-0004. URL: https://doi.org/10.1515/dialect-2017-0004.

Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: arXiv: 1301.3781 [cs.CL].

Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Appendix

Word Set 1 (Year, Country, Target Word)	Word Set 2 (Year, Country, Target Word)	Common Words
(‘2019-35’, ‘at’, ‘angreifen’)	(‘2014-00’, ‘at’, ‘angreifen’)	Hinterhalt, eingeschüchtert, versuchen
(‘2019-35’, ‘de’, ‘angreifen’)	(‘2014-00’, ‘de’, ‘angreifen’)	erschießen, Angriff, attackieren, provozieren, beleidigen, umbringen, töten, bedrohen
(‘2019-35’, ‘at’, ‘anlangen’)	(‘2014-00’, ‘at’, ‘anlangen’)	angelangt
(‘2019-35’, ‘de’, ‘anlangen’)	(‘2014-00’, ‘de’, ‘anlangen’)	ankommen, Tiefpunkt
(‘2024-38’, ‘de’, ‘anfassen’)	(‘2014-00’, ‘de’, ‘anfassen’)	bewundern, anschauen, bestaunen, ausprobieren, ganz, zugucken
(‘2024-38’, ‘de’, ‘angreifen’)	(‘2014-00’, ‘de’, ‘angreifen’)	Angriff, Gegner, provozieren, beleidigen, vorwerfen, töten, andersdenkender
(‘2024-38’, ‘de’, ‘anlangen’)	(‘2014-00’, ‘de’, ‘anlangen’)	ankommen

Table 1: Common words when comparing years

Word Set 1 (Year, Country, Target Word)	Word Set 2 (Year, Country, Target Word)	Common Words
(‘2014-00’, ‘at’, ‘anfassen’)	(‘2014-00’, ‘de’, ‘anfassen’)	Hand, Finger, Samthandschuh
(‘2014-00’, ‘at’, ‘angreifen’)	(‘2014-00’, ‘de’, ‘angreifen’)	angepöbelt, Schutzschilde, umbringen, bedrohen, bombardieren
(‘2019-35’, ‘at’, ‘angreifen’)	(‘2019-35’, ‘de’, ‘angreifen’)	beleidigen, beschimpfen
(‘2019-35’, ‘at’, ‘anlangen’)	(‘2019-35’, ‘de’, ‘anlangen’)	ankommen, angelangt
(‘2024-38’, ‘at’, ‘angreifen’)	(‘2024-38’, ‘de’, ‘angreifen’)	töten, absichtlich, Feind, zerstören
(‘2024-38’, ‘at’, ‘anlangen’)	(‘2024-38’, ‘de’, ‘anlangen’)	entschwunden, untergehen

Table 2: Common words when comparing countries

Year	Country	Top 5 of 'anfassen'
'2014-00'	'at'	schulmädchen, begrabschen, vorbeigehen, Lesson, Finger
'2014-00'	'de'	GEISTER, Samthandschuh, Reifendruck, Martinsfest, Rasierkling
'2019-35'	'at'	pinkeln, hatt, Pitsch, Sympathie, mär
'2019-35'	'de'	Heimatminister, streicheln, küssen, zusehen, sexeln
'2024-38'	'at'	surreal, Schwarzbrot, wandertauglich, Einzigartiges, Pimpinelle
'2024-38'	'de'	fotografieren, matschen, betäubend, betasten, ablichten

Table 3: Nearest neighbors of 'anfassen' by year and country

Year	Country	Top 5 of 'angreifen'
'2014-00'	'at'	nieee, gehenAssad, islami, verscherzen, planes
'2014-00'	'de'	außenstehender, Gegner, Panzerfäust, angepöbelt, sogar
'2019-35'	'at'	eindringen, diffamieren, Antideutsche, Salafist, iranisch
'2019-35'	'de'	attackieren, Eindringlinge, angeleinen, wehren, feuern
'2024-38'	'at'	Murawjow, zuhalten, Granate, beschießen, Suebe
'2024-38'	'de'	zerstören, zerstückeln, beschädigen, schädigen, loswirdn

Table 4: Nearest neighbors of 'angreifen' by year and country

Year	Country	Top 5 of 'anlangen'
'2014-00'	'at'	Maenn, ostwärts, Friedensschluss, Umrisse, angelangt
'2014-00'	'de'	erklommen, Entwickelte, zusammenfiel, Gesetzesgeber, brechen
'2019-35'	'at'	Pannonischen, Faschierte, gespannt, Laibche, Umriß
'2019-35'	'de'	Museumseingang, verkrümeln, Breidensteiner, Dolm, Klassement
'2024-38'	'at'	Untiefen, Höhepunkt, Tiefpunkt, Nidit, ganz
'2024-38'	'de'	Problemtyp, Sinistra, ankommen, Urbanik, irgendwo

Table 5: Nearest neighbors of 'anlangen' by year and country