# Geolocating News about Extreme Climate Events: A Comparative Analysis of Off-the-Shelf Tools for Toponym Identification in German

Brielen Madureira[1,2,*], Mariana Madruga de Brito[2] and Andreas Niekler[1,3]

[1]LeipzigLab - Climate Discourse, Leipzig University, Leipzig, Germany
[2]Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany
[3]Computational Humanities, Leipzig University, Leipzig, Germany

**Abstract**

Determining the geolocation of extreme climate events and disasters in texts is a common problem in climate impact and adaptation research. Named-entity recognition (NER) tools are typically used to identify a pool of toponyms that serve as candidate event locations. In this study, we conduct a comparative analysis of three off-the-shelf NER tools, namely Flair, Spacy and Stanza. We describe and quantify differences between their outputs for German news articles and evaluate them extrinsically based on three methods to determine the country where events took place. We show how their contrasts are propagated into downstream tasks and can yield distinct decisions about a document's geographical focus, which, in turn, can impact conclusions about countries' prominence in German media.

**Keywords**

NER tools, toponym identification, document geolocation, extreme climate events, German news

## 1. Introduction

> "More than 100 communities are affected and also the state's capital Porto Alegre[BRA]. This includes the Taquari[BRA]-Valley, where a minority stills speaks a German[DEU] dialect due to the migration from the German[DEU] region of Hunsrück[DEU] in the 19th century."

**Figure 1:** Excerpt from a German news article (translated by the authors)[1] reporting on an extreme climate event, annotated with toponym candidates (detected by Stanza's German NER tool) and their country. Highlighted entities must be correctly parsed to determine that Brazil, and not Germany, is the actual event's location.

The floods in South Brazil in May, 2024 were so disastrous that they found their way to international news, as shown in Figure 1. Many human readers have the needed linguistic and cognitive skills to easily realise that the event location in this example is *not* in Germany, despite the referring expressions alluding to this country. But for computational tools this distinction is not so simple: they can be misled by the presence of Germany-related terms detected as toponyms (i.e. "named entities that label a particular location" [1]). Indeed, *accurate* automation of document geolocation remains an unsolved problem, even for deep learning-based solutions [2].

Corpora of news articles have been increasingly used to analyse media coverage of climate-related events and extract information about their impacts [3, 4, 5, 6, 7]. As climate extremes are inherently place-specific, accurate and precise location information is essential for meaningful spatial analyses. This often involves determining the so-called geographical scope (or geospatial focus) of a document, which is a well-established task in the literature [8, 9, 10, 11]. It goes beyond enriching texts with

*GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands*

brielen.madureira@uni-leipzig.de (B. Madureira); mariana.brito@ufz.de (M. M. d. Brito); aniekler@informatik.uni-leipzig.de (A. Niekler)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]Source: Deutsche Welle on May 3, 2024. Available at https://www.dw.com/de/sturzregen-w%C3%BCtet-seit-tagen-im-s%C3%BCden-brasiliens/a-68988756.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-04-29

CEUR-WS.org/Vol-4201/paper1.pdf

token-based geolocation: after performing toponym identification and resolution, the metadata must be used to infer the actual location of the event of interest [12, 13]. Such a task cannot be realistically performed at scale without the aid of automated NLP tools.

The initial subtask of identifying all toponyms in the text is usually accomplished by named-entity recognition (NER) tools that treat locations as a specific entity. When this task is of secondary importance in a project, researchers often rely on off-the-shelf NER tools, taking their correctness at face value. However, such tools are not error-free and their inaccuracies can (possibly unnoticeably) are propagated to downstream decisions and results.

To investigate that, we conduct a comparative analysis of the toponyms identified by three popular off-the-shelf NER tools in a corpus of German news articles about various disasters and extreme climate events (e.g. floods, heat waves, fires and droughts). We quantify variations in their outputs and compare three prediction methods to show how these differences influence the determination of documents' geographical focus and, consequently, higher-level conclusions about the prominence of countries in German media.

## 2. Related work

NER typically subsumes toponym identification as it usually includes specific labels for locations. There is vast literature about NER tools and datasets for German texts [14, 15, 16, 17, *inter alia*]. However, the performance of off-the-shelf natural language processing (NLP) tools is not always optimal, particularly for languages other than English. Systematic evaluation is therefore essential to establish a robust and context-sensitive plan of action for each use case and domain. Previous studies, such as [18, 19, 20], assessed the performance of available tools for various NLP tasks in German. Yet, to our knowledge, recent comparative studies focusing on NER tools for German are missing from the literature.

A similar gap exists in the context of geotagging. While toponym identification has been widely evaluated, most empirical studies rely on English-language datasets [21, 22, 23, 2]. Although existing tools can correctly identify many toponyms, they also incur false positives and false negatives. In their practical guide to geoparsing evaluation, [1] argue that "off-the-shelf NER taggers are inadequate for location extraction". Common sources of error include the lack of distinction between pragmatic types of toponyms, limited handling of metonyms, entity boundary errors and case sensitivity issues [21, 1]. Moreover, German climate-related news poses its own challenges [24].

Several studies use a gold standard as a reference for evaluating toponym identification. For instance, [25] compared the performance of various NER tools for historical corpora in English to manually annotated toponyms and, more recently, [26] used a large language model to annotate German news, using the resulting dataset to train and evaluate a custom NER model. However, in real-world settings, annotated samples of comparable data (e.g. of same genre, domain and region) are not always available.

Our work builds on the existing literature but addresses a particular situation in which off-the-shelf NER tools must be selected and used without an available gold standard for evaluation. To assess their performance under these constraints, we adopt an extrinsic evaluation approach that measures how reliably the geotagged toponyms of each tool can be used for document-level geolocation and frequency-based ranking of countries in a corpus of German news.

## 3. Methods

**Task** Our task is a special case of geographical focus identification, similar to [7]: given a collection of news articles reporting on or mentioning disasters and extreme climate events, our goal is to determine the country (or countries) in which the event(s) of interest occurred. This task is a step towards data-driven analyses of the coverage of worldwide climate events in German newspapers.

**Procedure** We perform this task in a 4-step pipeline as described below. Here, we do *not* aim to propose a ground-breaking method for document-level geolocation. Instead, the main contribution

lies in the evaluation: comparing NER tools and examining whether they yield diverging downstream conclusions. For that, steps 1-4 rely on basic heuristics inspired by existing literature but without sophisticated optimisation. The deliberate use of simpler methods ought to avoid obscuring the effects of NER decisions behind more complex modelling choices.

1. **Identification** Each document in the corpus is enriched with an annotation layer for each NER tool indicating tokens they identify as candidate toponyms (i.e. labeled as `location`).
2. **Querying** Each unique toponym type identified over the whole corpus is used to build a query for a database of geographical coordinates. The top $n$ matches are retrieved with their metadata, which includes (when available) latitude, longitude and country.
3. **Resolution** Each toponym type in each document is mapped to its likely geographical coordinates. When multiple coordinates were retrieved for a toponym type, the ambiguity is resolved by taking the coordinates that are closest to the polygon formed by the unambiguous toponyms in the document. This approach is based on the "spatial minimality" heuristic used by [27] and the clustering approach by [28]. If all toponyms are ambiguous, the most probable match (i.e. the one ranked first by the database's search engine) is selected for each toponym instead.
4. **Prediction** For each tool and geographical database pair, we apply three prediction methods that get a document's toponyms and coordinates as input and map them to countries.

**Evaluation** The NER tools' performance is first assessed *bilaterally*, by intercomparison of their outputs, and then *extrinsically*, by their effect on subsequent tasks of document geolocation and frequency ranking, using human annotation as a gold standard for the documents' geographical focus. We also examine characteristics of the outputs that may explain the observed behaviour.

## 4. Experiments

This section presents an overview of the design decisions for our experiments regarding five main dimensions: data, NER tool, geographical database, country prediction method and evaluation metric. Further details can be found in the Appendix. Our implementation is available for documentation purposes at https://codeberg.org/briemadu/german-ner-geoparsing.

**Data** Our dataset is derived from an ongoing research project about climate discourse in the German media. It is comprised of 983 documents from German newspapers published between 2000 and 2024 retrieved from the `wiso-net`[2] news aggregator database using a selection of keywords related to seven hazards (heat wave, wildfire, flood, storm, drought, cold wave and landslide). The texts are in German and were annotated by humans with the type of event they discuss and the country (or countries) where the event of interest happened. The number of characters per text ranges from 178 to 11,512 with, on average, 2,468.

**Models** The analysis assesses three popular off-the-shelf tools that perform NER in German: Flair [29], using model `de-ner-large`,[3] Spacy [30], using model `de_core_news_lg`,[4] and Stanza [31], using model `de`.[5] All tools were trained with four labels (for person, location, organisation and other). In this study, only the label LOC for location is relevant. Using LLMs is also a possibility, but not the focus of this study, as they are not NER tools *by design*.

**Geographical database** The latitude and longitude coordinates with their respective country for each toponym type are retrieved from two geographical databases: Geonames and Nominatim.[6]

[2] https://www.wiso-net.de/
[3] https://flairnlp.github.io/docs/tutorial-basics/tagging-entities
[4] https://spacy.io/models/de#de_core_news_lg
[5] https://stanfordnlp.github.io/stanza/ner_models.html
[6] https://www.geonames.org/ and https://nominatim.org/, respectively.

**Prediction methods** Three methods are used to determine the geographical focus (here, the country where the discussed event occurred) of a document based on its identified toponyms:

i **Majority voting**: the most common country (or countries, if there is a tie) among the pool of toponyms' countries in a document is (are) selected as its geographical focus. This is inspired by voting methods used in other tasks [2] and the "popularity approach" by [28].
ii **Closest to centroid**: the centroid of the concave hull of the polygon formed by the document's toponyms, after removing outliers, is computed. The toponym closest to its centroid is selected as the document's geographical focus. This is similar to the approaches by [32, 33, 34].
iii **Keyword proximity**: the list of a document's toponyms is reduced to only those that co-occur in a sentence with a hazard keyword (the same used to query documents) before applying majority voting. If no sentences contain both a keyword and a toponym, we resort to the closest toponyms (in number of tokens) that comes preferably before each keyword (or after, in case none precedes it).

**Evaluation metrics** The intersection over union (IoU) metric (or Jaccard Index) is used to measure how much the set of identified toponyms per document coincide between pairs of tools, considering only exact matches (i.e. same initial and last character). Predicted countries are assessed in relation to the gold standard using the percentage of exact matches (when the exact same country or countries were predicted) and the percentage of partial matches (when at least one correct country was predicted). Finally, Kendall $\tau$ and Spearman's rank correlation coefficients [35, 36] are used to compare the resulting rankings of country frequencies in the corpus.

## 5. Analysis

The analysis is divided into four parts. We first compare the sets of toponyms identified by the three NER tools, considering a realistic situation in which gold standard toponyms are not available. Then, we examine how effective each tool's outputs are for two higher level tasks using our annotation: determining the document's geographical focus and assessing the ranked frequency of each country in the corpus, which could inform conclusions about countries' prominence in the media. We finish by discussing properties of the NER outputs that may explain some of the observed variations.

### 5.1. Toponym identification: bilateral comparisons

In total, 9,231 toponym types were identified in the corpus, among which 5,836 (63.22%) had matches in Nominatim and 3,613 (39.14%) were found in Geonames. Table 1 shows the number of toponyms and of toponym types identified by each tool over the whole corpus, as well as the percentage of types that returned at least one valid match in the geographical databases. Although Spacy identifies most toponyms and most unique types, it has the worst coverage in both databases. Flair, on the other hand, identified fewer toponym types but with a higher percentage of valid ones.

| | $n$ toponyms | $n$ types | % in Geonames | % in Nominatim |
|---|---|---|---|---|
| Flair | 11,816 | 4,639 | 63.16 | 81.20 |
| Stanza | 14,230 | 5,330 | 54.11 | 77.97 |
| Spacy | 14,367 | 6,700 | 42.75 | 63.58 |

**Table 1**
Overview of the number of toponyms identified by each tool and % of matches in geographical databases.

Figure 2 shows the distribution of intersection-over-union values computed for the sets of toponyms per document for each pair of models. In our dataset, Spacy and Stanza tended to have more disagreements, whereas Stanza and Flair tended to agree comparatively more. While the three medians are similar, the important insight is that most values fall between 0.4 and 0.8, indicating that, on average, there is considerable variation in the set of toponyms identified by different tools for the same document.

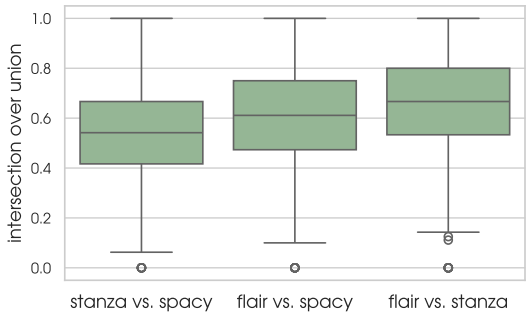


**Figure 2:** Box plots of toponyms' intersection-over-union values per document for each pair of tools.

## 5.2. Higher-level task: country prediction

Moving on to the main task of determining the geographical focus of each document, we first compare NER tools bilaterally, without the gold standard. Table 2 presents the percentage of exact agreement between the prediction of countries based on the toponyms identified by each tool. Again, Spacy and Stanza incur the most disagreements, whereas Stanza and Flair agree the most in all but one experiment.

| | Geonames | | | Nominatim | | |
|---|---|---|---|---|---|---|
| | centroid | keyword | majority | centroid | keyword | majority |
| Stanza vs. Flair | 80.47 | 87.69 | 89.01 | 72.84 | 83.01 | 82.40 |
| Spacy vs. Flair | 76.50 | 82.60 | 82.91 | 70.50 | 82.50 | 83.01 |
| Spacy vs. Stanza | 72.74 | 80.98 | 81.49 | 65.41 | 77.11 | 79.86 |

**Table 2**
Percentage of exact agreement among documents' predicted countries for each model pair and method.

These results show that the variations in the toponyms identified by each NER tool caused at least 10% (but in some cases more than 20%) of the country predictions to be different, a consistent finding across prediction methods and geographical databases.

To assess the actual performance in this task, the next step is comparing them with the gold-standard. Table 3 presents the percentage of exact and partial matches for country predictions based on each tool's outputs and method. We observe that, for our data, predictions using toponyms near hazard-related keywords led to the best overall performance in most cases, followed by majority voting. Additionally, Nominatim was more effective than Geonames in general. However, we reiterate that our aim here is not yet to find the best prediction method for solving the problem. The different methods are a means to assess the NER *tools*, and to what extent their distinction is consistent across varied experiments.

For all prediction methods and both geographical databases, Flair had the highest results, followed (in most cases) by Stanza. The main finding here is that, *ceteris paribus*, there can be a difference of up to 5 p.p. in the accuracy of the predictions which is only due to the mere choice of NER tool. Therefore, if only one tool had been used directly without careful inspection and comparison, the performance of the predictions methods in this downstream task could have suffered an unwarranted negative impact.

## 5.3. Higher-level task: ranking by prominence

In our pipeline, NER is used to guide predictions which, in turn, are supposed to be used for even higher-level conclusions. Assume we are interested in ranking countries by media prominence and consider the number of documents about a country in the corpus as a proxy for this construct. We thus order countries by how many documents were mapped to them in the gold standard and by each tool.

This part of the evaluation is conducted using results from the keyword method with Nominatim, which achieved the best metric values in determining the geographical focus so far. Table 4 shows the

| | | Geonames | | Nominatim | |
|---|---|---|---|---|---|
| | | exact | overlap | exact | overlap |
| keyword | Flair | 62.67 | 75.38 | 70.60 | 83.01 |
| | Stanza | 59.61 | 74.06 | 65.11 | 81.18 |
| | Spacy | 57.07 | 70.80 | 65.82 | 79.86 |
| majority | Flair | 60.73 | 73.14 | 65.01 | 78.33 |
| | Stanza | 58.70 | 72.23 | 61.95 | 74.47 |
| | Spacy | 57.58 | 71.31 | 62.77 | 77.62 |
| centroid | Flair | - | 63.68 | - | 66.84 |
| | Stanza | - | 62.46 | - | 61.95 |
| | Spacy | - | 60.22 | - | 62.26 |

**Table 3**
Percentage of exact and overlapping matches of predictions based on each tool in relation to the gold standard. Exact match is not presented for the centroid method because it does not predict more than one country.

Spearman rank correlation and the Kendall's $\tau$ for the rankings when considering only the countries that have at least 10 documents in the gold standard. While Spacy and Stanza's exhibited very similar correlations with the reference, Flair reached a higher correlation.[7]

| | Spacy | Stanza | Flair | | Spacy | Stanza | Flair |
|---|---|---|---|---|---|---|---|
| Spearman r | 0.753 | 0.758 | 0.833 | Kendall $\tau$ | 0.648 | 0.641 | 0.714 |

**Table 4**
Correlation between the predicted and observed number of documents per country (with at least 10 instances).

Figure 3 shows the top 14 countries ranked by the number of documents that have them as geographical focus, for NER tools and the gold standard. All tools ranked the same set of 14 countries at the top, and the two most frequent countries were also correctly determined. This is evidence that good overall agreement in identifying the most prominent countries could be reached despite the contrasts in the tools' outputs. Still, the country order varies. For instance, Switzerland (`che`) ranks much higher across all tools than it does in reality in the corpus. Stanza and Spacy lead to similar conclusions for most countries: 9 of them appear in identical positions and 3 are in adjacent positions. The main discrepancy between these two tools is that Spacy ranks France (`fra`) 3 positions higher than Stanza. Flair and Stanza, on the other hand, coincided only in 4 positions. The Kendall correlation for this portion of the ranking was 0.89 between Stanza and Spacy, 0.78 between Flair and Stanza and 0.75 between Flair and Spacy. Stanza and Spacy agreed more with each other, but Flair was superior in relation to the gold standard (correlation of 0.53, as opposed to 0.47 and 0.45 of Spacy and Stanza, respectively).

## 5.4. Error analysis

To conclude the analysis, we investigate which properties of each NER tool's outputs may have contributed to the different predictions. First, we note that average number of toponyms per document was higher for Spacy (mean 14.61, sd 12.74) and Stanza (mean 14.47, sd 12.25) than for Flair (mean 12.02, sd 10.63). Part of these extra toponyms may have contributed to divergent predictions, as two extra tokens can already change majority voting or a polygon's centroid, for example.

Looking closely, among the 9,231 toponym types identified in the corpus, 683 (7.40%) were unique to Flair, 1,276 (13.82%) were unique to Stanza and 2,847 (30.84%) were unique to Spacy. All of Flair's unique toponyms occur no more than 5 times in the corpus and may have had a lower impact on our outcomes (although this can also become a problem in larger corpora). Among the most common types

[7]This difference reduces if lower rank positions are used, but the metrics are less reliable due to the small number of instances.

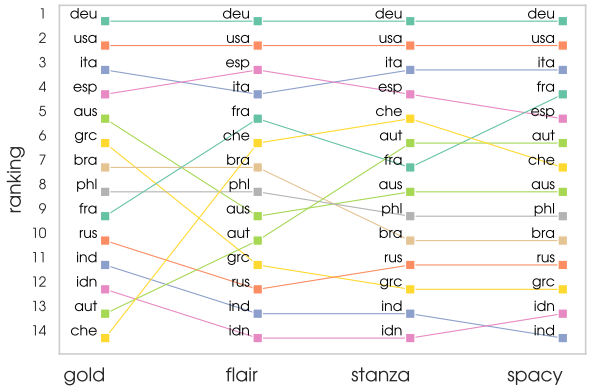


**Figure 3:** The top 14 countries with the highest frequency in the gold standard ranked using each NER tool.

identified only by Spacy, we see detached generic terms like *Stadt* (city), *Altstadt* (old town), *Innenstadt* (downtown), *Flughafen* (airport), *Provinzen* (provinces), *Gemeinde* (community) and *Rathaus* (town hall). Among the top unique toponyms identified by Stanza, apart from the word *Sonne* (sun) that appears very often, we find mostly adjectival and demonymic forms of country and city names, classified as *associative toponyms* in the taxonomy proposed by [1]. Although these forms often do not label a specific place, some of them correspond to location names in different countries, which introduces a source of noise into the methods. Associative toponyms may even map to countries not associated with the toponym itself (for instance, *italienischen*, the declined form of the adjective Italian, triggered place names like embassies in six countries–and none of them was even in Italy). It is in fact debatable to what extent associative toponyms should be included (see [1]).

There is mutual influence between geographical databases and NER outputs. On the one hand, the limited coverage of toponyms in geographical databases (Table 1) may be a sign that they help safeguard the subsequent use of tools against their invalid toponyms. But the spurious multiple matches for some underspecified place names in different countries require a more fine-grained treatment. The labelling scheme also contributes to the drawbacks: in some other languages, NER tools are trained with more location labels that enable a distinction between geopolitical entities, nationalities/affiliations, facilities and locations. German is impaired by a single label that encompasses various concepts. A possibility is to filter identified toponyms by their part-of-speech tag and syntactic role in their sentence.

## 6. Conclusion

We conducted a comparative analysis of three off-the-shelf NER tools for German applied to news articles about disasters and extreme climate events. In a setting where toponym annotation was not available, we first assessed the tools' performance by how closely they aligned with one another and then by their extrinsic performance on higher-level decisions. Flair achieved the best results in our data, in line with its superior F1 score reported by the provider. Stanza and Spacy were comparable, with some advantage to one or to the other depending on the experiment. Nevertheless, we avoid making general claims about which is inherently superior and thus did not statistically test any hypothesis in that regard. [25] has actually shown that using a majority voting ensemble of NER tools can improve toponym identification and that tools' performance may be corpus dependent.

What our experiments corroborate is that tool selection can discernibly influence downstream results. Should we blindly decide to use a single tool without comparative experiments, more than 10% of the predictions could differ, accuracy could be 5p.p. lower and the resulting ranking could be less correlated to the real observations, depending on the method. These findings provide empirical evidence that set-up decisions which may appear minor at first glance can still affect key conclusions. This underscores the need for careful evaluation and error analysis of off-the-shelf NER tools, even those widely accepted in the community, as their performance can vary in non-negligible ways across specific use cases.

## Acknowledgments

We thank Maike Reichel and Julius Hehenkamp for their valuable help in annotating the data.

## Declaration on Generative AI

The authors have not directly employed any Generative AI tools to write this paper. One of the authors uses Grammarly for grammar and spelling checks.

## Appendix: Implementation Details

**Data** Our corpus of news documents was queried using keywords related to seven types of climate extremes, similar to [37, 38, 39]. Two student assistants annotated a sample of 3,150 documents, 450 for each hazard and, among them, 18 for each year. In this study, we used only the subset of 1,008 documents identified as relevant, i.e. those that do discuss an extreme climate event. We excluded 25 documents that either do not contain any hazard-related keywords in their main text (because title and subtitle were not considered here) or have no identifiable countries (e.g. refer to Europe as a whole).

**NER** We extracted all sequences of tokens corresponding to an entity with the label LOC, together with its start and end characters. For Stanza, we loaded only the `tokenize`, `mwt` and `ner` components.

**Geographical databases** Nominatim was queried via the Python library `geopy.geocoders` and Geonames via the library `geocoder`. For each query, we cached a maximum of 20 matches with the needed metadata: latitude, longitude, country code, type and rank position in the search results. For Geonames, we additionally passed the parameters `name_equals` with the toponym string and set `fuzzy=` 1. This resulted in 34,803 instances for Nominatim and 30,836 for Geonames. Toponym matches without a known country were filtered out, so that 90.0% and 81.68% were retained respectively. For duplicates with the same name and country, we kept only the first match (in our use case, only the country is relevant, without the need for a precise geolocation within it). The final list had 11,300 toponym instances (32.47% of the initial results) for Nominatim and 10,392 (33.70%) for Geonames.

**Prediction** For majority voting, we counted and ranked the number of toponyms for each country. If there was a tie, we returned all countries that shared the top position. This method accounted for all repetitions of toponym types. In the keyword method, we first selected only toponyms that co-occur in the same sentence with one of the keywords. If there was none, we selected the toponyms closest to each keyword in number of tokens, preferably before, but after if there was none. Repeated keywords and types were also considered. For the centroid method, we first turned each coordinate into a Point object and sorted them in counter clockwise order with respect to their mean. Points were used to construct a Polygon object (we resorted back to a Line object when only two points were found or to a Point object when only one was found). Then, if there was more than one point, we computed the center and excluded all outlier points whose absolute z-score of the distance to the centroid is greater than 1. A new polygon was created with the remaining points and its centroid was computed. Finally, we selected the toponym of the point closest to the centroid. In this method, repeated toponyms were automatically aggregated into a single point in the polygon.

**Limitations** We did not account for overlapping sequences of tokens when comparing NER outputs, which could increase their agreement but would complicate merging adjacent tokens and partially overlapping sequences of characters. Furthermore, the queries to the geographical databases used only the identified tokens, in isolation, without contextual disambiguation or toponym declination. Ideally, textual context should be taken into account to resolve ambiguity. As toponyms were not lemmatised here, their German declination can also have caused mismatches in the geographical databases.

## References


[1] M. Gritta, M. T. Pilehvar, N. Collier, A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics, Language Resources and Evaluation 54 (2020) 683–712. URL: http://link.springer.com/10.1007/s10579-019-09475-3. doi:10.1007/s10579-019-09475-3.

[2] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, F. Klan, Location Reference Recognition from Texts: A Survey and Comparison, ACM Computing Surveys 56 (2024) 1–37. URL: https://dl.acm.org/doi/10.1145/3625819. doi:10.1145/3625819.

[3] D. Otto, M. Pfeiffer, M. M. de Brito, M. Gross, Fixed Amidst Change: 20 Years of Media Coverage on Carbon Capture and Storage in Germany, Sustainability 14 (2022). URL: https://www.mdpi.com/2071-1050/14/12/7342. doi:10.3390/su14127342.

[4] J. Sodoge, C. Kuhlicke, M. M. d. Brito, Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning, Weather and Climate Extremes 41 (2023) 100574. URL: https://www.sciencedirect.com/science/article/pii/S2212094723000270. doi:https://doi.org/10.1016/j.wace.2023.100574.

[5] J. H. Lochner, A. Stechemesser, L. Wenz, Climate summits and protests have a strong impact on climate change media coverage in Germany, Communications Earth & Environment 5 (2024) 279. URL: https://doi.org/10.1038/s43247-024-01434-3. doi:10.1038/s43247-024-01434-3.

[6] P. H. L. Alencar, J. Sodoge, E. Nora Paton, M. Madruga De Brito, Flash droughts and their impacts—using newspaper articles to assess the perceived consequences of rapidly emerging droughts, Environmental Research Letters 19 (2024) 074048. URL: https://iopscience.iop.org/article/10.1088/1748-9326/ad58fa. doi:10.1088/1748-9326/ad58fa.

[7] I. Kong, R. S. Purves, Analyzing Geographic Bias of Newspaper Articles Reporting Global Climate Disasters, Annals of the American Association of Geographers (2025) 1–19. URL: https://www.tandfonline.com/doi/full/10.1080/24694452.2025.2564220. doi:10.1080/24694452.2025.2564220.

[8] E. Amitay, N. Har'El, R. Sivan, A. Soffer, Web-a-where: geotagging web content, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Sheffield United Kingdom, 2004, pp. 273–280. URL: https://dl.acm.org/doi/10.1145/1008992.1009040. doi:10.1145/1008992.1009040.

[9] G. Andogah, G. Bouma, J. Nerbonne, Every document has a geographical scope, Data & Knowledge Engineering 81-82 (2012) 1–20. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169023X12000687. doi:10.1016/j.datak.2012.07.002.

[10] B. R. Monteiro, C. A. Davis, F. Fonseca, A survey on the geographic scope of textual documents, Computers & Geosciences 96 (2016) 23–34. URL: https://linkinghub.elsevier.com/retrieve/pii/S0098300416301972. doi:10.1016/j.cageo.2016.07.017.

[11] F. Melo, B. Martins, Automated Geocoding of Textual Documents: A Survey of Current Approaches, Transactions in GIS 21 (2017) 3–38. URL: https://onlinelibrary.wiley.com/doi/10.1111/tgis.12212. doi:10.1111/tgis.12212.

[12] W. Zong, D. Wu, A. Sun, E.-P. Lim, D. H.-L. Goh, On assigning place names to geography related web pages, in: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries, ACM, Denver CO USA, 2005, pp. 354–362. URL: https://dl.acm.org/doi/10.1145/1065385.1065464. doi:10.1145/1065385.1065464.

[13] S. J. Lee, H. Liu, M. D. Ward, Lost in Space: Geolocation in Event Data, Political Science Research and Methods 7 (2019) 871–888. URL: https://www.cambridge.org/core/product/identifier/S2049847018000237/type/journal_article. doi:10.1017/psrm.2018.23.

[14] D. Benikova, C. Biemann, M. Reznicek, NoSta-D named entity annotation for German: Guidelines and dataset, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 2524–2531. URL: https://aclanthology.org/L14-1251/.

[15] M. Riedl, S. Padó, A named entity recognition shootout for German, in: I. Gurevych, Y. Miyao

(Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 120–125. URL: https://aclanthology.org/P18-2020/. doi:10.18653/v1/P18-2020.

[16] K. Labusch, C. Neudecker, D. Zellhöfer, Bert for named entity recognition in contemporary and historic german, in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, German Society for Computational Linguistics & Language Technology, Erlangen, Germany, 2019, pp. 1–9. URL: https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf.

[17] E. Leitner, G. Rehm, J. Moreno-Schneider, A dataset of German legal documents for named entity recognition, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4478–4485. URL: https://aclanthology.org/2020.lrec-1.551/.

[18] K. Ortmann, A. Roussel, S. Dipper, Evaluating Off-the-Shelf NLP Tools for German, in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, German Society for Computational Linguistics & Language Technology, Erlangen, Germany, 2019, pp. 212–222. URL: https://sfb1102.uni-saarland.de/sfbunisb/uploads/2020/10/KONVENS2019_paper_55.pdf.

[19] S. Scheible, R. J. Whitt, M. Durrell, P. Bennett, Evaluating an 'off-the-shelf' POS-tagger on early Modern German text, in: K. Zervanou, P. Lendvai (Eds.), Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics, Portland, OR, USA, 2011, pp. 19–23. URL: https://aclanthology.org/W11-1503/.

[20] R. Laarmann-Quante, L. Prepens, T. Zesch, Evaluating automatic spelling correction tools on German primary school children's misspellings, in: D. Alfter, E. Volodina, T. François, P. Desmet, F. Cornillie, A. Jönsson, E. Rennes (Eds.), Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, LiU Electronic Press, Louvain-la-Neuve, Belgium, 2022, pp. 95–107. URL: https://aclanthology.org/2022.nlp4call-1.10/.

[21] M. Gritta, M. T. Pilehvar, N. Limsopatham, N. Collier, What's missing in geographical parsing?, Language Resources and Evaluation 52 (2018) 603–623. URL: http://link.springer.com/10.1007/s10579-017-9385-8. doi:10.1007/s10579-017-9385-8.

[22] J. Wang, Y. Hu, Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers, Transactions in GIS 23 (2019) 1393–1419. URL: https://onlinelibrary.wiley.com/doi/10.1111/tgis.12579. doi:10.1111/tgis.12579.

[23] Z. Liu, K. Janowicz, L. Cai, R. Zhu, G. Mai, M. Shi, Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing, AGILE: GIScience Series 3 (2022) 1–13. URL: https://agile-giss.copernicus.org/articles/3/9/2022/. doi:10.5194/agile-giss-3-9-2022.

[24] N. Doms, T. Schlachter, L. Hahn-Woernle, A Geo-Parser for German Documents with Environmental Context, in: V. Wohlgemuth, H. Kandil, A. Ramzy (Eds.), Advances and New Trends in Environmental Informatics, Springer Nature Switzerland, Cham, 2025, pp. 21–33. URL: https://link.springer.com/10.1007/978-3-031-85284-8_2. doi:10.1007/978-3-031-85284-8_2, series Title: Progress in IS.

[25] M. Won, P. Murrieta-Flores, B. Martins, Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora, Frontiers in Digital Humanities 5 (2018) 2. URL: http://journal.frontiersin.org/article/10.3389/fdigh.2018.00002/full. doi:10.3389/fdigh.2018.00002.

[26] L. Kriesch, S. Losacker, A geolocated dataset of German news articles, Scientific Data 12 (2025) 1128. URL: https://www.nature.com/articles/s41597-025-05422-w. doi:10.1038/s41597-025-05422-w.

[27] J. L. Leidner, G. Sinclair, B. Webber, Grounding spatial named entities for information extraction and question answering, in: Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, 2003, pp. 31–38. URL: https://aclanthology.org/W03-0105/.

[28] M. Badieh Habib Morgan, M. van Keulen, Named entity extraction and disambiguation: the

missing link, ESAIR '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 37–40. URL: https://doi.org/10.1145/2513204.2513217. doi:10.1145/2513204.2513217.

[29] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: W. Ammar, A. Louis, N. Mostafazadeh (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 54–59. URL: https://aclanthology.org/N19-4010/. doi:10.18653/v1/N19-4010.

[30] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). URL: https://spacy.io/. doi:10.5281/zenodo.1212303.

[31] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 101–108. URL: https://aclanthology.org/2020.acl-demos.14/. doi:10.18653/v1/2020.acl-demos.14.

[32] D. A. Smith, G. Crane, Disambiguating Geographic Names in a Historical Digital Library, in: G. Goos, J. Hartmanis, J. Van Leeuwen, P. Constantopoulos, I. T. Sølvberg (Eds.), Research and Advanced Technology for Digital Libraries, volume 2163, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 127–136. URL: http://link.springer.com/10.1007/3-540-44796-2_12. doi:10.1007/3-540-44796-2_12, series Title: Lecture Notes in Computer Science.

[33] R. C. Pasley, P. D. Clough, M. Sanderson, Geo-tagging for imprecise regions of different sizes, in: Proceedings of the 4th ACM workshop on Geographical information retrieval, ACM, Lisbon Portugal, 2007, pp. 77–82. URL: https://dl.acm.org/doi/10.1145/1316948.1316969. doi:10.1145/1316948.1316969.

[34] M. A. Radke, N. Gautam, A. Tambi, U. A. Deshpande, Z. Syed, Geotagging Text Data on the Web—A Geometrical Approach, IEEE Access 6 (2018) 30086–30099. URL: https://ieeexplore.ieee.org/document/8371593/. doi:10.1109/ACCESS.2018.2843814.

[35] C. Spearman, The Proof and Measurement of Association between Two Things, The American Journal of Psychology 15 (1904) 72. URL: https://www.jstor.org/stable/1412159?origin=crossref. doi:10.2307/1412159.

[36] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93. URL: https://doi.org/10.1093/biomet/30.1-2.81.

[37] N. Li, S. Zahra, M. Brito, C. Flynn, O. Görnerup, K. Worou, M. Kurfali, C. Meng, W. Thiery, J. Zscheischler, G. Messori, J. Nivre, Using LLMs to build a database of climate extreme impacts, in: D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, C. Christiaen, N. Kushwaha, V. Muccione, S. A. Vaghefi, M. Leippold (Eds.), Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 93–110. URL: https://aclanthology.org/2024.climatenlp-1.7/. doi:10.18653/v1/2024.climatenlp-1.7.

[38] M. Madruga de Brito, J. Sodoge, H. Kreibich, C. Kuhlicke, Comprehensive assessment of flood socioeconomic impacts through text-mining, Water Resources Research 61 (2025) e2024WR037813. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024WR037813. doi:https://doi.org/10.1029/2024WR037813.

[39] T. M. N. Carvalho, A. Niekler, C. Kuhlicke, J. Zscheischler, M. M. de Brito, Global synthesis of peer-reviewed articles reveals blind spots in climate impacts research (2025). URL: http://dx.doi.org/10.21203/rs.3.rs-6095740/v1. doi:10.21203/rs.3.rs-6095740/v1.