When Everything Is "Nearby": How Airbnb Listings in New York City Exaggerate Proximity

Mikael Brunila
Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada
Urban Politics & Governance Lab, School of Urban Planning, McGill University, Montréal, Canada

Priyanka Verma
Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada

Grant McKenzie
Platial Analysis Lab, Department of Geography, McGill University, Montréal, Canada

Abstract

In recent years, the emergence and rapid growth of short-term rental (STR) markets has exerted considerable influence on real estate in most large cities across the world. Central location and transit access are two primary factors associated with the prevalence and expansion of STRs, including Airbnbs. Nevertheless, perhaps due to methodological challenges, no research has addressed how location and proximity are represented in the titles and descriptions of STRs. In this paper, we introduce a new methodological pipeline to extract spatial relations from text and show that expressions of distance in STR listings can indeed be quantified and measured against real-world distances. We then comparatively analyze Airbnb reviews (written by guests) and listings (written by hosts) from New York City in order to demonstrate systematically how listings exaggerate proximity compared to reviews. Moreover, we discover spatial patterns to these differences that warrant further investigation.

2012 ACM Subject Classification Information systems → Geographic information systems; Information systems → Information extraction

Keywords and phrases spatial proximity, distance estimation, information extraction, named entity recognition, short-term rentals

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.16

Category Short Paper

Funding *Mikael Brunila*: Kone Foundation
Priyanka Verma: Social Sciences and Humanities Research Council of Canada

1 Introduction

Over the past decade, the short-term rental (STR) market has expanded in most large cities across the world. While STRs provide new economic opportunities for some, they also contribute to harmful processes such as gentrification and displacement through the removal of affordable units from the rental market [18, 1]. Airbnb in particular has become synonymous with a certain kind of gentrification, whereby conveniently located working-class and non-white neighborhoods are marketed as sites of consumption, leisure, and urban authenticity for an upwardly mobile class of white-collar professionals.

As in any market, Airbnb hosts need to communicate important information about location and other characteristics of their units to potential guests. Drawing on ideas from the interactional sociology of Ervin Goffman [6], ethnographers have likened Airbnb listings to front-stage performances whereby hosts deploy various means to manage the impressions guests will have of a listing [16]. Because real estate listings are fundamentally located *somewhere*, location figures strongly into the repertoire of *distinctions* [2] hosts can make


© Mikael Brunila, Priyanka Verma, and Grant McKenzie;
licensed under Creative Commons License CC-BY 4.0

vis-a-vis other listings. Indeed, recent research has demonstrated that location is one of the key determinants of both the average price per night and average monthly revenue of units listed on Airbnb [4]. Nevertheless, this line of inquiry has not yet been extended to the "cognitive maps" [8, 10] which translate our experience of the city into mental representations thereof and, in the context of Airbnb, which encode the relationship between residence and place for people who typically reside elsewhere. By making claims about what is "nearby", "only 10 minutes away" or "within walking distance", Airbnb hosts situate their properties within the ensemble of a city's structures and relations, including not only spatial and semiotic [10] but also ideological [8].

A wide range of work from various fields has established that our conception of what is "nearby" varies with a number of factors: larger objects tend to be considered closer than smaller ones, distances will be estimated differently depending on familiarity and activity, and so on (for an overview, see [5]). By comparing expressions of distance in listings and reviews, we can grasp how socio-economic incentives shape the production of spatial representations in discourse. While there is anecdotal evidence of the exaggeration of distance in the context of real estate advertisements [13], our paper provides a first glimpse into how these dynamics systematically unfold in a much larger dataset and in the setting of STRs. Furthermore, while others have presented models for extracting vague spatial descriptions [3, 5] as well as for assessing the linguistic distribution of concepts like "near," we provide a sociological control variable by contrasting A) listing *descriptions* with B) listing *reviews* associated with the same locations. While listing descriptions are arguably written as a profit-motivated performance for the STR market, reviewers have different motives.

In this paper, we introduce a new methodological pipeline to extract spatial relations from text and show that expressions of distance in STR listings can indeed be quantified and measured against real-world distances. With this data, we demonstrate differences in the use of terms such as "nearby" and "walking distance" across listings and reviews. We do the same for a range of toponymic categories including parks (e.g., "Central Park"), tourist attractions (e.g., "Empire State Building"), and schools (e.g., "Columbia University"). Specifically, this short paper presents preliminary work addressing the following four research questions (RQ):

RQ1 Can qualitative distance measures, such as *nearby* or *walking distance*, be quantified in STR listings?

RQ2 Do quantified distance measures in STR listings accurately reflect real-world distances?

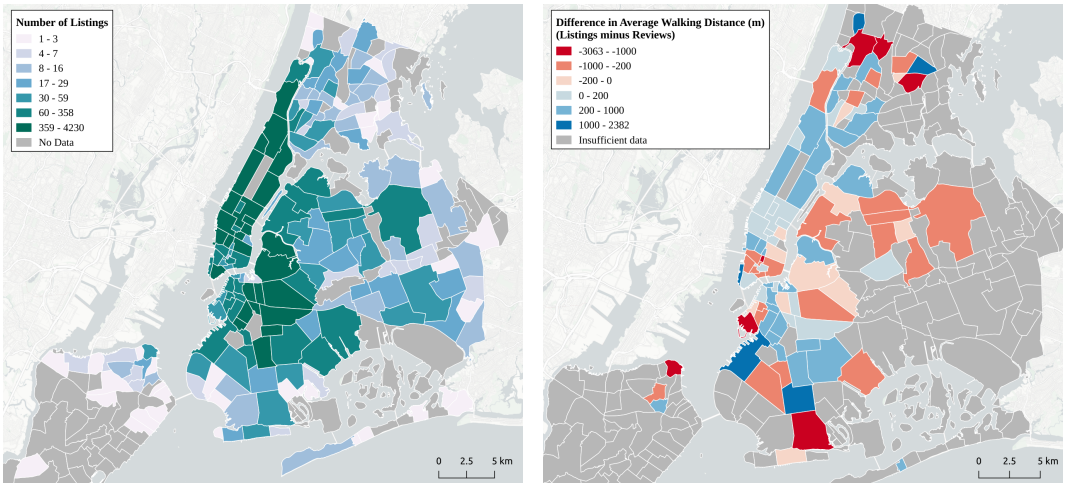
RQ3 On average, do these distances vary between listing descriptions and reviews?

RQ4 How do the above measures vary across neighborhoods in New York City (NYC)?

2 Data and Methods

The data for this paper cover all active Airbnb listings and their associated reviews for NYC in August 2019. All data were purchased from the non-profit group Inside Airbnb[1]. The data contain a total of 47,440 listings and 995,665 reviews. To make data processing more feasible, we take a sample from the latter, giving us 168,533 reviews for an average 3.55 reviews per listing (even processing this sample takes a full day). Each listing includes its title, description, and geographic coordinates. The listings are highly unevenly distributed across the city, as can be seen in Figure 1a.

[1] http://insideairbnb.com/



(a) Number of listings by neighborhood.

(b) Difference in average walking distance.

Figure 1 Cartographic representations of the Airbnb listing data. Raw count of listings per neighborhood shown in (a) and difference in average walking distance between listings and reviews shown in (b).

To extract geographical entities from the data, we manually annotated a spatially weighted random sample of 1,517 listings and 967 reviews using the annotation platform Prodigy[2] (for details on our sampling strategy, see Appendix A.1). The annotation was done by five academic annotators, including all the authors of this paper. This was effectively a named-entity recognition (NER) task, where the named entities were beyond the scope of existing general-purpose NER datasets. Annotators had 14 labels to choose between, which can be seen in Table 1 in the Appendix. For the purposes of this paper, the key labels are: (1) "Spatio-Temporal Entity" (STE) reflecting any relation between two locations, such as "15 minutes walk to" or "nearby," and (2) various toponyms ranging from tourist attractions to schools. Labels were chosen from an initial set suggested by Cadorel et al. [3] but adjusted and extended to fit our specific dataset and framework.

After annotating the data, we fit three models with DistilBERT embeddings [17]. Out of these, a model with a Conditional Random Fields (CRF) [9] classification layer performed the best, with an overall F1-score of 0.756, with a plain DistilBERT model achieving comparable results with an F1-score of 0.752. To make our work more reproducible, we use this latter model, even if it is technically slightly worse. To connect STEs with relevant toponyms, we use a combination of dependency parsing and graph partitioning: Each STE is associated with the set of toponyms that are among its immediate dependents (for all the models and other details, see the Appendix and Table 1).

To address `RQ1`, we geocoded entities using Google's Geocoding API.[3] The geocoder provides coordinates for the centroid of each entity location. This poses a challenge for larger parks such as Central Park, which expands across an area of 3.41 km^2. To address this issue, we calculated the point within the park closest to the Airbnb coordinates and used this as the final entity coordinates. We then used Open Source Routing Machine[4] to

[2] `https://prodi.gy/`
[3] `https://developers.google.com/maps/documentation/geocoding/overview`
[4] `https://project-osrm.org/`

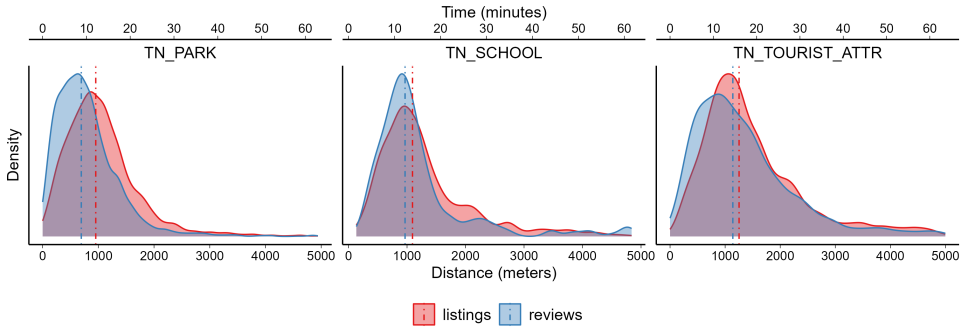


Figure 2 The distribution of distances (bottom x-axis) and walk times (top x-axis) for listings and reviews respectively. Listings consistently under-represent distance compared with reviews.

calculate the shortest walking distance between each Airbnb and entity coordinates along OpenStreetMap's pedestrian network. We use a maximum threshold value of 5,000 meters to remove any outliers in the data. We do not expect individuals to walk distances exceeding 5,000 meters since the largest STE used in our analysis is 15 minutes.

For RQ2, we generate density plots to examine how walking distances are distributed across STE groups and tags for listings and reviews. We use a secondary axis to display walk time, calculated using an average walking speed of 1.31 meters per second [14]. Differences in these distributions are assessed using a pairwise Mann-Whitney U test, a non-parametric statistical test commonly used for data that is not normally distributed [12]. We use this test to determine whether the differences in distributions are statistically significant (RQ3). Finally, for RQ4 we plot the average differences across the widely used NYC neighborhoods dataset by the non-profit BetaNYC.[5]

3 Results

Looking at Figures 2 and 3, we see that qualitative distance measures like "nearby" or "walking distance" can indeed be quantified using the methods detailed above (RQ1). By comparing these quantifications across listings and reviews, we discover that the former tend to exaggerate proximity more than the latter (RQ3). However, across both types of data, claimed walking times (5, 10, and 15 mins) were distributed widely across the actual walking times (RQ2). Walking distances from listings were on average 12 minutes when the stated distance was 5 minutes, 15 minutes for 10 minutes, and 18 minutes for 15 minutes. Walking distances from reviews, by contrast, were closer to the actual claim: 8 minutes for 5 minutes, 12 minutes for 10 minutes, and 15 minutes for 15 minutes. These differences are also reflected in how words like "near," "close," and "walking distance" are deployed on average: For listings, these are close to 15 minutes of walking, while only 10 for reviews. Furthermore, turning our attention to figure 3, words like "nearby" are always closer for parks than for schools and tourist attractions. Again, listings consistently exaggerate proximity across these three toponymic categories but the general pattern also holds: "nearby" parks are only 10 minutes away for reviews and 13 minutes away for listings, whereas schools are 12 and 15 minutes away and tourist attractions 15 and 17 minutes.

[5] https://data.beta.nyc/dataset/pediacities-nyc-neighborhoods/resource/35dd04fb-81b3-479b-a074-a27a37888ce7

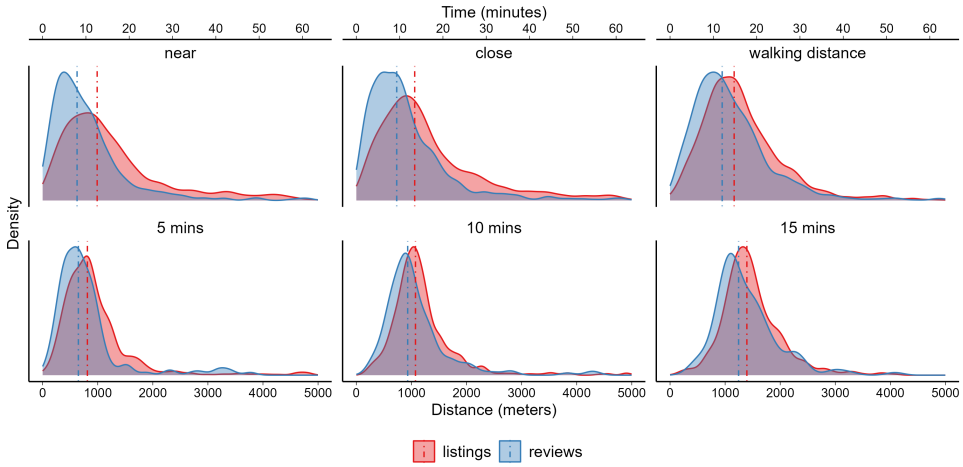


Figure 3 Walking distances and times for six different spatio-temporal entities in the data. Again, the means are consistently lower for the reviews than the listings. The differences are particularly pronounced for the vague STE qualifiers (top row).

We also found that there were statistical differences across neighborhoods (RQ4). Looking at Figure 1b, listings in Manhattan tend to exaggerate consistently compared to reviews: here, it seems, everything is "nearby." This trend is reversed only in Lower Manhattan. Outside of Manhattan, the visually distinct spatial clusters are more mixed. As we move further out of the city center, the differences become more extreme in both directions. While this is interesting to note, issues with data sparsity and outliers might be part of the explanation. Nonetheless, these patterns require further investigation. Are listings in less attractive neighborhoods more prone to exaggerate when they talk about distance? How might these patterns correlate with the cultural and economic hierarchy between different areas [11]?

4 Discussion and conclusions

In this paper, we demonstrated how spatial entities and relations can be extracted from textual descriptions of reviews and listings from Airbnb (RQ1). Claimed walking distances do not reflect real world walk-times (RQ2), but the exaggeration is more extreme in listings than reviews (RQ3). While these differences seem to be spatially clustered (RQ4), the exact nature of these clusters remains to be investigated. Although these results are preliminary, they offer a first step towards exploring the dynamics between the representation of spatial relations and place-making.

There are notable limitations to our approach. First, it remains to be seen whether our trained models would generalize well to other settings. Second, our model for extracting spatial entities and our method for parsing spatial relations are still imperfect, introducing a margin of error in the results. Third, there are sparsity issues with some of our annotated data, which is reflected in the uneven F1-scores between labels (see Table 1).

These reservations notwithstanding, we have shown how to quantify and extract vague spatial relations from text data. Moreover, we have demonstrated that there are consistent and statistically significant differences between listings and reviews – that is, between hosts and guests – in their representations of spatio-temporal relations. In this way, the results presented here open up a new vantage point to studying representations of spatial relations through geocoded text data. For example, by exploring how changes in these representations

change over time, they could be related to indices of gentrification. Furthermore, these methods could be expanded beyond the scope of Airbnb data to analyze representations of space in a number of textual contexts: short- versus long-term real estate descriptions, other forms of tourism literature, and even fictional literature.

References

1 Kyle Barron, Edward Kung, and Davide Proserpio. The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science*, 40(1):23–47, October 2020.

2 Pierre Bourdieu. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, Cambridge, MA, 1984.

3 Lucie Cadorel, Denis Overal, and Andrea G. B. Tettamanzi. Fuzzy representation of vague spatial descriptions in real estate advertisements. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising*, pages 1–4, Seattle Washington, November 2022. ACM. doi:10.1145/3557992.3565994.

4 Robbin Deboosere, Danielle Jane Kerrigan, David Wachsmuth, and Ahmed El-Geneidy. Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue. *Regional Studies, Regional Science*, 6(1):143–156, January 2019.

5 Curdin Derungs and Ross S. Purves. Mining nearness relations from an n-grams Web corpus in geographical space. *Spatial Cognition & Computation*, 16(4):301–322, October 2016.

6 E. Goffman. *The Presentation of Self in Everyday Life*. Anchor Books, New York, NY, 1959.

7 Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging, August 2015. arXiv:1508.01991 [cs]. doi:10.48550/arXiv.1508.01991.

8 F. Jameson. *Postmodernism, Or, The Cultural Logic of Late Capitalism*. Duke University Press, 1991.

9 JD. Lafferty, A. McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, June 2001. Morgan Kaufmann Publishers Inc. doi:10.5555/645530.655813.

10 Kevin Lynch. *The Image of the City*. MIT Press, Cambridge, MA, 1964.

11 David Madden. Neighborhood as Spatial Project: Making the Urban Order on the Downtown Brooklyn Waterfront. *International Journal of Urban and Regional Research*, 38(2):471–497, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-2427.12068.

12 H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, March 1947. Publisher: Institute of Mathematical Statistics. doi:10.1214/aoms/1177730491.

13 G. McKenzie and Y. Hu. The "Nearby" exaggeration in real estate. In *Proceedings of the Cognitive Scales of Spatial Information Workshop, L'Aquila, Italy*, pages 4–8, 2017.

14 Elaine M. Murtagh, Jacqueline L. Mair, Elroy Aguiar, Catrine Tudor-Locke, and Marie H. Murphy. Outdoor Walking Speeds of Apparently Healthy Adults: A Systematic Review and Meta-analysis. *Sports Medicine*, 51(1):125–141, January 2021.

15 L. A. Ramshaw and M. P. Marcus. Text Chunking Using Transformation-Based Learning. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, Text, Speech and Language Technology, pages 157–176. Springer Netherlands, Dordrecht, 1999.

16 Alexandrea Ravenelle. A return to Gemeinschaft: Digital impression management and the sharing economy. In J. Daniels, K. Gregory, and TM Cottom, editors, *Digital sociologies*, pages 27–45. Bristol University Press, 1 edition, November 2016. URL: 10.2307/j.ctt1t89cfr.

17 Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. arXiv:1910.01108 [cs]. doi:10.48550/arXiv.1910.01108.

18 David Wachsmuth and Alexander Weisler. Airbnb and the rent gap: Gentrification through the sharing economy. *Environment and Planning A: Economy and Space*, 50(6):1147–1170, September 2018. doi:10.1177/0308518X18778038.

A Appendix

Table 1 Summary of NER label frequencies in the training data, in the overall data, and performance metrics (F1, Recall, and Precision) for the DistilBERT-CRF model. The plain DistilBERT model produced similar numbers.

	Label	N Annotated (all)	N Predicted (listings)	N Predicted (reviews)	F1	Rec.	Prec.
1	TN:NEIGHBORHOOD	2265	66211	39914	0.872	0.877	0.867
2	TN:BOROUGH	1677	30014	40611	0.932	0.944	0.920
3	TN:CITY	1000	20097	46674	0.941	0.955	0.928
4	TN:STREET	552	21259	7409	0.681	0.675	0.687
5	TN:STATION	543	19828	8233	0.582	0.621	0.547
6	TN:TOURIST_ATTR	615	21619	6619	0.619	0.646	0.593
7	TN:PARK	532	19201	9548	0.893	0.941	0.850
8	TN:SCHOOL	127	3132	943	0.516	0.457	0.592
9	TN:BUSINESS	730	24059	9623	0.718	0.742	0.695
10	TN:OTHER	347	6092	3029	0.413	0.415	0.411
11	SPAT_TEMP_ENT	6643	197089	203545	0.690	0.708	0.672
12	TRANSIT	4168	126360	105646	0.787	0.806	0.768
13	GEOG_ENTITY	6663	184825	261947	0.806	0.812	0.800
14	HOST_BUILDING	915	29364	12391	0.426	0.442	0.411
	Overall	26777	769150	756132	0.756	0.771	0.742

A.1 Sampling

To sample the training data, we used the following stratified disproportionate sampling strategy:

1. Per neighborhood, all listings are included if there are 5 or fewer.
2. In neighborhoods with more listings than that, the sample for the neighborhood is 5 listings + 0.5%.
3. Each listing has 1 review sampled, but many listings have no reviews.

Sampling like this, we could ensure that all neighborhoods were represented in the training data. However, for the review data that the trained model extracted NER labels from, we used no spatial stratification, which is potentially reflected in the results. Future work should use the entire dataset of reviews or take a spatially stratified sample.

A.2 Models

We trained three different models on the annotated data: 1) DistilBERT [17] with a linear classification layer, 2) DistilBERT with a conditional random fields (CRF) [9] layer prior to the linear classifier, and 3) DistilBERT with a CRF and BiLSTM layer prior to the linear classifier [7]. For all these models, we used a 10/90 test-train split. Between the models, the

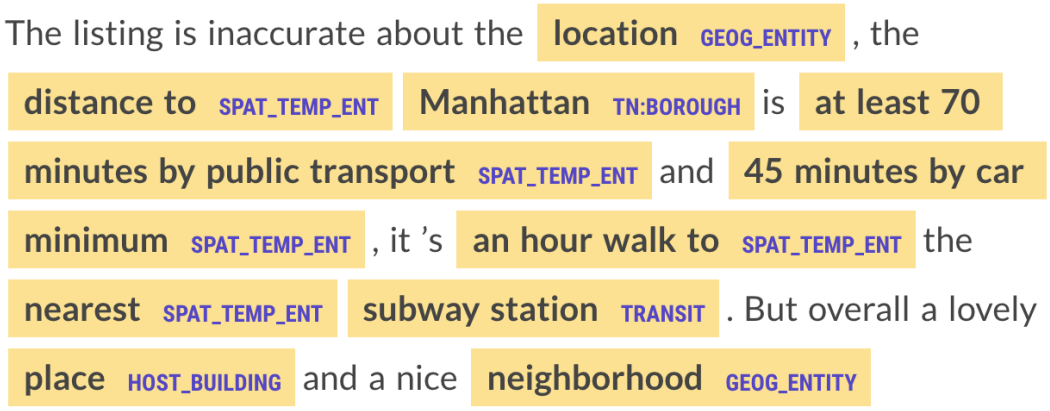


Figure 4 The annotation interface of Prodigy. This annotated review references several different types of entities related to place and spatial relations.

DistilBERT model with a CRF layer but without the BiLSTM layer performed the best, with an overall F1-score of 0.756. Almost similar results were achieved with the DistilBERT model, with a 0.752 F1-score. To keep results reproducible, all downstream tasks were performed with this model. While these F1-scores might seem on the low side, it was much higher for many of the classes in the data, as can be seen in table 1. The final models for all three architectures were trained over five epochs with a 1×10^{-4} learning rate, 1×10^{-5} weight decay, gradient clipping, and early stopping. All models were implemented in PyTorch[6] using pretrained DistilBERT models from HuggingFace[7] and using additional IOB-chunking [15].

For an example of the annotation interface and, consequently, the data that was given to the models, see figure 4.

A.3 Relationship extraction

To extract the dependencies between Spatio-Temporal Entities (STEs) and toponyms, we proceed in the following way: For each document in our corpus, we extract dependencies using the spaCy Python library[8], with entities recognized as toponyms merged into single tokens. We next identify all the dependents for all tokens for each document, using these relations to build a directed graph of each document. Given this graph, we filter for nodes that are labeled STE and remove any edges that point to this node. Next, we find the weakly connected subgraphs that remain after removing these edges, giving us a set of graphs with at most one STE node each and n nodes with other labels, including toponyms. Now, each of these other nodes is a dependent of an STE node and we can pair each toponym-labeled node with the STE of the subgraph.

[6] https://pytorch.org/
[7] https://huggingface.co/
[8] https://spacy.io/