# EndoLink: A Knowledge Graph-Based Platform for Crowdsourced Endonym and Place Name Collection

Janine Laura Hindermann, Sina Ahmadi

*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*

**Abstract**

This paper introduces EndoLink, a web-based crowdsourcing platform that enables users to explore, collect, and contribute missing location names through a map-based interface. The platform addresses the underrepresentation of local languages in knowledge graphs, which limits the performance of natural language processing tools for geographic information extraction, particularly for low-resource languages. We present a quantitative analysis of place name coverage in Wikidata for Iran, Turkey, and Thailand, revealing a strong bias toward official and Western languages. User feedback indicates that EndoLink improves usability and lowers the barrier to contribution. Additionally, an evaluation with large language models demonstrates their limited ability to retrieve missing labels for endonyms in Central Kurdish, highlighting dependencies on existing multilingual data. EndoLink is publicly available at https://janinelaura-hindermann.github.io/endolink.

**Keywords**

endonyms, crowdsourcing, knowledge graphs, low-resource languages, geographic information extraction

## 1. Introduction

Place names, or toponyms, are more than mere geographic labels; they carry cultural, historical, and political significance [1]. An *endonym* refers to a place name used by local communities in their native language, such as *Krung Thep* for Bangkok or *Züri* for Zurich. These local names often differ substantially from the *exonyms* used by outsiders, yet they remain crucial for preserving linguistic heritage and ensuring accurate geographic information extraction (GeoIE).

Named Entity Recognition (NER) systems, a foundational component of GeoIE pipelines, depend heavily on the quality and diversity of training data [2]. However, many languages lack sufficient labeled resources, making it challenging to build effective NER systems for these contexts [3]. This limitation is particularly problematic for location names: NER systems perform poorly on rare or previously unseen entities [4], and geographic databases exhibit well-documented biases toward Western regions, leading to the underrepresentation of populations in the Global South [5].

Knowledge graphs such as Wikidata offer a promising resource for addressing these gaps. With over 116 million items and community-driven maintenance, Wikidata provides structured, multilingual data that can enhance NER systems [6, 7]. We focus on Wikidata rather than alternatives such as OpenStreetMap because its RDF-based structure and SPARQL endpoint allow for precise, programmatic identification of missing labels per language, a query pattern that OpenStreetMap's tagging system does not natively support. However, the coverage of local place names remains highly uneven across languages. Our analysis reveals that while official languages like Persian, Turkish, and Thai achieve near-complete coverage in their respective countries, most local and minority languages have little to no representation.

To address this infrastructural gap, we present **EndoLink**, a web-based crowdsourcing platform designed to facilitate the discovery and contribution of missing location names in various languages through an interactive map-based interface. Our contributions are threefold: (1) we introduce the EndoLink platform, which allows users to visualize missing toponym labels, contribute endonyms,

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

janinelaura.hindermann@uzh.ch (J. L. Hindermann); sina.ahmadi@uzh.ch (S. Ahmadi)

https://orcid.org/0000-0001-7904-6551 (S. Ahmadi)

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

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-04-29

CEUR-WS.org/Vol-4201/paper9.pdf

and export collected data for downstream applications; (2) we present a quantitative analysis of place name coverage across languages in Wikidata for three multilingual countries; and (3) we evaluate the limitations of large language models in retrieving toponyms for low-resource languages using crowdsourced Central Kurdish endonyms.

## 2. Related Work

### 2.1. Geographic Information Extraction and Low-Resource Languages

NER is a foundational component of geographic information extraction pipelines, tasked with identifying and classifying entities such as persons, organizations, and locations in text [2]. Toponyms constitute a particularly important entity type, as their accurate extraction underpins applications in disaster management, spatial humanities, and humanitarian response. While deep learning approaches have substantially improved NER performance, these gains are concentrated on high-resource languages with abundant labeled data [8]. Systems trained on such data perform poorly on rare or previously unseen entities, a problem that disproportionately affects toponyms in underrepresented languages [4, 9, 10].

Knowledge graphs such as Wikidata have been leveraged to address data scarcity in NER, particularly for fine-grained entity recognition where entities are classified into specific subcategories (e.g., distinguishing a "city" from a generic "location") [7]. However, the multilingual coverage of these resources remains highly uneven. Recent work has examined the capabilities of large language models (LLMs) for geographic tasks [11]. While LLMs perform reasonably well on general geographic knowledge, they struggle with fine-grained local knowledge, particularly outside North America and Western Europe [12]. Belliardo et al. [5] demonstrated that geolocation extraction tools require careful tuning to avoid perpetuating geographic biases that disadvantage the Global South. This creates a circular dependency: languages underrepresented in knowledge bases remain underrepresented in model outputs, reinforcing existing data gaps rather than addressing them. Prior work has also examined label completeness in Wikidata directly, showing that coverage varies dramatically across languages, with smaller languages often lacking even basic labels for common entities [13].

### 2.2. Toponym Collection and Preservation

Traditional toponym collection has been conducted primarily by government agencies as part of topographic mapping activities. The United Nations Group of Experts on Geographical Names [14] has promoted toponym preservation since 1959, encouraging standardized gazetteers. Global resources such as GeoNames and OpenStreetMap now provide additional toponymic data through volunteer contributions, though coverage varies significantly across regions and languages [15]. However, many communities are not officially recognized and face language assimilation campaigns, particularly in the Middle East [16], where toponyms have been systematically translated or replaced following political ideologies [17, 18]. In such contexts, toponym collection is not merely a cartographic effort but a language preservation one.

Participatory approaches have emerged as alternatives to traditional field surveys. Mamontova and Klyachko [19] developed a GIS-based platform for documenting indigenous Evenki place names in Siberia, combining GIS technology with vernacular cartography to enable communities to contribute and exchange toponymic knowledge. Perdana and Ostermann [20] proposed a participatory toponym handling framework for Indonesia, highlighting challenges in coordination, standardization, and data quality when involving citizen toponymists.

### 2.3. Existing Gaps

Existing work demonstrates both the importance of multilingual toponym data for geographic information extraction and the value of participatory collection methods. However, a gap persists between these two areas. Current participatory platforms for toponym collection operate independently of established

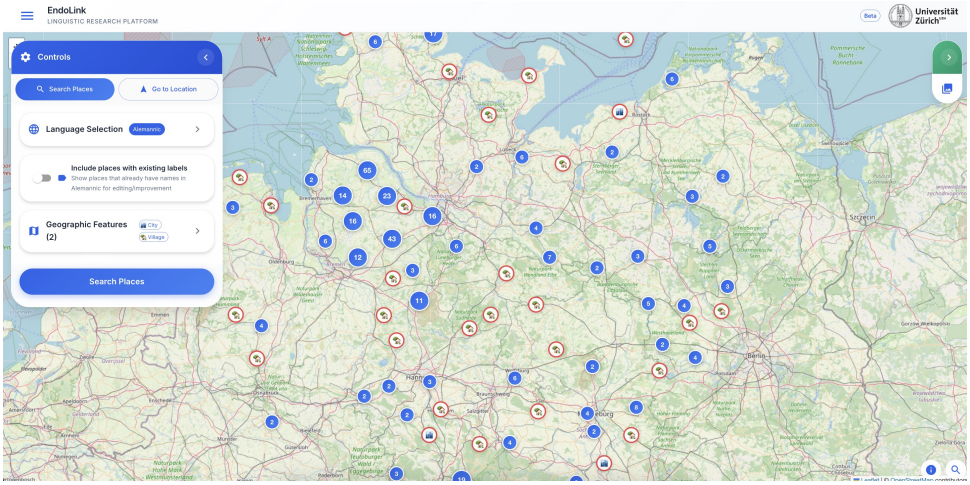


**Figure 1:** Missing city names in Alemannic in northern Germany, displayed as pins on EndoLink.

knowledge graphs, while contributing to resources like Wikidata or OpenStreetMap, directly requires either editing entries individually or writing custom SPARQL queries, both of which present substantial barriers for non-technical contributors. EndoLink addresses this gap by providing an accessible, map-based interface that enables communities to leverage the structured data infrastructure of knowledge graphs efficiently, allowing contributors to focus on providing local linguistic knowledge rather than navigating the technical complexities of knowledge graph interaction.

## 3. EndoLink

EndoLink is a front-end web application built with the JavaScript library React[1] that enables users to discover and contribute missing location names in different languages to Wikidata. The application centers on an interactive map powered by the Leaflet library[2], with the rest of the user interface built using Material-UI[3] components. The platform is hosted on GitHub Pages[4] and is publicly available.[5] The workflow consists of four steps: language selection, retrieval of missing location names, contribution of endonyms, and data export.

**Language Selection.** As the first step, users are asked to select a target language from a dropdown list populated from a locally stored copy of Wikidata's language inventory, originally queried via SPARQL from the Wikidata Query Service.[6] Each entry includes the language's Wikidata identifier, IETF language tag (`wdt:P305`), and native labels. Since the list of available languages on Wikidata does not change frequently, the application uses a locally stored copy rather than fetching it on every startup. If a user's language or dialect is not available in Wikidata, they can manually enter a custom language code.

**Retrieving Missing Location Names.** After selecting a language and navigating to a region of interest on the map, the user selects a location type (e.g., city) and triggers a SPARQL query against the Wikidata Query Service. The query, shown in Listing 1, retrieves all location entities within the highlighted circle area that lack a label in the selected language. It selects entities that are instances or subclasses of the chosen location type (e.g., `wd:Q515` for cities) and that have geographic coordinates

[1] https://react.dev
[2] https://react-leaflet.js.org
[3] https://mui.com
[4] https://pages.github.com
[5] https://janinelaura-hindermann.github.io/endolink
[6] https://query.wikidata.org

```
SELECT DISTINCT ?place ?placeLabel ?coordinates ?matchedType ?matchedTypeLabel WHERE {
  SERVICE wikibase:around {
    ?place wdt:P625 ?coordinates.
    bd:serviceParam wikibase:center
     ↪  "Point(10.17333984375000253.29805557491275)"^^geo:wktLiteral;
                  wikibase:radius "200".
  }
  VALUES ?type { wd:Q515 wd:Q532 }
  ?place wdt:P31/wdt:P279* ?type.
  BIND(?type AS ?matchedType)
  FILTER NOT EXISTS { ?place rdfs:label ?l. FILTER(LANG(?l) = "gsw") }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
} LIMIT 500
```

Listing 1: SPARQL query for retrieving cities and villages missing an Alemannic label within a map area in northern Germany.

(wdt:P625). To identify missing labels, the query uses a FILTER NOT EXISTS clause to exclude all entities that already have a label in the selected language. Results are limited to 500 entries to maintain responsiveness. Missing locations are displayed as map pins, as shown in Figure 1.

**Contributing Endonyms.** Clicking on a map pin opens a dialog where the user can enter the location name in their language. Upon saving, the pin color changes from red to green, providing visual feedback on contribution progress. The contributed names and their metadata (Wikidata ID, coordinates, language tag) are stored in a structured JSON list.

**Data Export and Upload.** Users can copy the collected data to the clipboard and paste it back into the application later, enabling interrupted workflows. The structured JSON format also makes the data reusable for downstream applications such as training NLP models or enriching custom knowledge graphs. EndoLink additionally supports uploading labels directly to Wikidata via its REST API[7]; however, this feature currently requires a local installation due to Cross-Origin Resource Sharing restrictions on the GitHub Pages-hosted version.

## 4. Results and Discussion

We evaluate EndoLink along three dimensions: a quantitative analysis of location name coverage on Wikidata for three linguistically diverse countries, Iran, Turkey, and Thailand, selected because they each have a dominant official language alongside dozens of local and minority languages; qualitative user feedback on the platform; and an exploratory test of LLM performance on underrepresented toponyms using Central Kurdish endonyms collected from Iran. All Wikidata-based analyses are based on data extracted in April 2025.

### 4.1. Distribution of Location Names on Wikidata

We analyze the coverage of location name labels across languages on Wikidata using three multilingual countries as case studies: Iran (wd:Q794), Turkey (wd:Q43), and Thailand (wd:Q869). For each country, all spoken languages were identified using the *language used* property (wdt:P2936), considering only languages with an existing IETF tag. All places of type *city* (wd:Q515), *municipality* (wd:Q15284), and *village* (wd:Q532) were extracted using the *country* property (wdt:P17). For each language, coverage

[7]Authentication process described at: https://www.wikidata.org/wiki/Wikidata:REST_API/Authentication

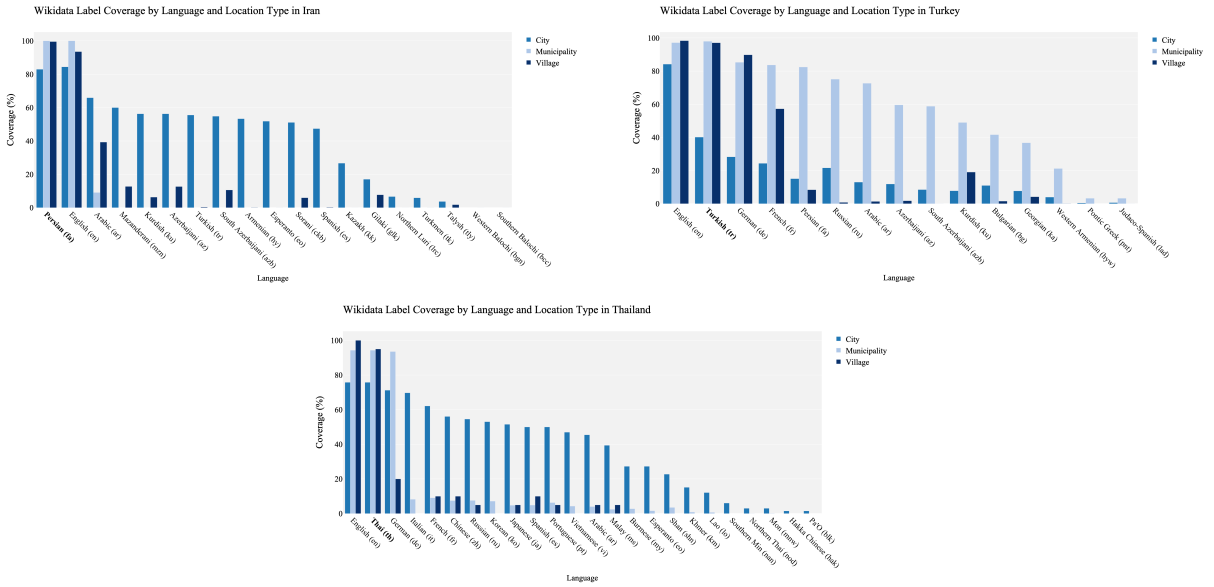


**Figure 2:** Location name label coverage across spoken languages in Iran (top-left), Turkey (top-right), and Thailand (bottom). Only languages with non-zero coverage are shown. In all three countries, official languages and English dominate while the majority of spoken languages have zero coverage and do not appear in the plots at all.

was calculated as the ratio of available labels to the total number of location entities. The results, visualized in Figure 2, reveal a consistent pattern across all three countries. Official languages (Persian, Turkish, Thai) achieve near-complete coverage, and English consistently shows very high coverage, often outperforming local languages. Several Western languages such as German, French, and Russian are disproportionately well-represented, particularly in Turkey and Thailand. In contrast, many local and minority languages have minimal or zero coverage. Regional languages such as Kurdish, Azerbaijani, Isan, and Southern Thai are largely absent from Wikidata's location labels.

These figures also expose inconsistencies in how Wikidata models location types across countries. Iran lists over 51,000 villages but only 11 municipalities, while Thailand records over 1,000 municipalities and just 20 villages. Such discrepancies complicate cross-country comparisons and reflect broader structural modelling challenges within Wikidata's location ontology (discussed further in Section 4.2).

## 4.2. User Feedback

Qualitative feedback was collected from users' interactions with EndoLink during the data collection process. Users described the interface as intuitive and engaging, with the map-based interaction being more accessible than alternatives such as manually browsing Wikidata entries or writing custom SPARQL queries. The visual feedback provided by pin color changes and the structured JSON export were valued for tracking progress and enabling data reuse in downstream applications. On the other hand, the most prominent issue was query performance: at higher zoom levels covering large geographic areas, SPARQL queries occasionally exceeded the Wikidata Query Service's 60-second execution limit [21], causing timeouts. A key contributing factor is the structural inconsistency in Wikidata's location hierarchies. For example, the city of Zurich (`wd:Q72`) is not a direct instance of "city" (`wd:Q515`) but rather an instance of "big city" and "city of Switzerland", which are subclasses at varying depths. Capturing all relevant instances therefore requires recursive SPARQL traversal via `wdt:P31/wdt:P279`, which increases computational complexity. Additionally, users noted confusion between location types (e.g., expecting municipalities but seeing villages), reflecting the cross-country modelling inconsistencies described above. Finally, Wikidata currently supports only one label per language tag, which cannot capture dialectal or regional spelling variations of the same endonym.

| Language | Exact | Accurate | Partial | Mismatch |
|---|---|---|---|---|
| English (en) | 1 (3.4%) | 10 (34.5%) | 15 (51.7%) | 4 (13.8%) |
| Persian (fa) | 4 (13.8%) | | 25 (86.2%) | |

**Table 1**
LLM evaluation on identifying endonyms. English accuracy was assessed manually considering phonetic resemblance; Persian required exact string match.

### 4.3. LLM Evaluation on Kurdish Endonyms

To test whether LLMs can compensate for missing toponym data, we conducted an exploratory evaluation using 29 Central Kurdish (Sorani) endonyms collected with the help of a local speaker familiar with villages near Sanandaj in western Iran. As the coverage analysis shows, only 5.95% of Iranian villages have a label in Central Kurdish on Wikidata, compared to 99.57% for Persian and 93.56% for English. Each endonym was given to GPT-4o with the prompt asking for the corresponding English and Persian labels.[8] Responses were compared against Wikidata labels as the gold standard.

Table 1 summarizes the results. English labels were evaluated on a four-point scale: *exact* (identical string match), *accurate* (phonetically correct but different spelling), *partial* (only parts of the label match), and *mismatch* (entirely different label). For Persian, only exact string matches were considered correct due to its standardized orthography. Overall, the LLM was confident in its responses and provided labels for all endonyms. For every endonym, it reasoned that the name in Kurdish refers to a village in the Kurdistan region of Iraq, particularly in Sulaymaniyah. This was neither requested in the prompt nor factually accurate, as the actual location of the collected places is in Iran. Many of the answers relied on phonetic transliteration and orthographic adaptation. For creating the Persian labels, the LLM often adapted Kurdish-specific characters into standard Persian characters. Similarly, the English versions were often formed by transliterating the phonetics using Latin characters, sometimes offering multiple accepted spellings.

The results indicate that the absence of the written labels in Kurdish on Wikidata hindered the LLM's ability to retrieve the corresponding labels in other languages.

## 5. Conclusion and Future Work

We presented EndoLink, a web-based crowdsourcing platform that enables users to discover, collect, and contribute missing location names to Wikidata through an interactive map-based interface. Our quantitative analysis of location name coverage in Iran, Turkey, and Thailand revealed a stark disparity: official and Western languages dominate Wikidata's toponym labels, while the vast majority of local and minority languages, including those actively spoken in these countries, have little to no representation. User feedback confirmed that EndoLink lowers the barrier to contributing multilingual location data compared to existing alternatives, though structural inconsistencies in Wikidata's location ontology pose challenges for scalability. Our exploratory LLM evaluation further demonstrated that models cannot compensate for missing toponym data, as their outputs are tightly coupled to the coverage of existing resources, reinforcing the need for upstream data collection.

The under-representation of local place names in knowledge graphs is a problem that remains largely under-explored, yet its implications extend far beyond geographic information extraction or navigation. For native speakers of marginalized communities, many of whom have experienced systematic erasure of their toponyms through state-driven renaming campaigns, seeing their place names reflected in widely used digital infrastructures is an act of recognition. Platforms like EndoLink can serve not only as data collection tools but also as instruments of cultural preservation, enabling communities to reclaim and document toponymic knowledge that might otherwise be lost. By making this process accessible to

[8]Prompt: "I have the following name written in Sorani: {name}. Could you help me get the name in English and Persian? Please only give me an answer if you are sure about the label in English and Persian."

non-technical contributors, we hope to support a broader movement toward linguistic inclusion in the digital resources that increasingly shape how the world is represented and understood. A key challenge for future deployment is building connections with language communities who can contribute; we plan to explore partnerships with existing language revitalization initiatives and diaspora networks.

For future work, we plan to further facilitate the data collection process by leveraging monolingual corpora to automatically suggest candidate toponyms that can be linked to locations on the map, reducing the manual effort required from contributors. We also envision gamification of the collection workflow to sustain engagement and scale contributions across languages and regions. Finally, implementing backend support for direct Wikidata uploads from the hosted version of EndoLink would complete the contribution pipeline and enable seamless integration into the Wikidata ecosystem.

## Acknowledgements

Sina Ahmadi gratefully thanks the support of the UZH Grant (reference number 269093).

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for assistance with code development and Claude for grammar and spelling checking. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] R. Rose-Redwood, D. Alderman, M. Azaryahu, Geographies of toponymic inscription: new directions in critical place-name studies, Progress in Human Geography 34 (2010) 453–470.

[2] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE transactions on knowledge and data engineering 34 (2020) 50–70.

[3] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, D. Klakow, A survey on recent approaches for natural language processing in low-resource scenarios, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2545–2568. URL: https://aclanthology.org/2021.naacl-main.201/. doi:10.18653/v1/2021.naacl-main.201.

[4] L. Derczynski, E. Nichols, M. van Erp, N. Limsopatham, Results of the WNUT2017 shared task on novel and emerging entity recognition, in: L. Derczynski, W. Xu, A. Ritter, T. Baldwin (Eds.), Proceedings of the 3rd Workshop on Noisy User-generated Text, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 140–147. URL: https://aclanthology.org/W17-4418/. doi:10.18653/v1/W17-4418.

[5] E. Belliardo, K. Kalimeri, Y. Mejova, Leave no place behind: improved geolocation in humanitarian documents, in: Proceedings of the 2023 ACM Conference on Information Technology for Social Good, 2023, pp. 31–39.

[6] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[7] C. Dogan, A. Dutra, A. Gara, A. Gemma, L. Shi, M. Sigamani, E. Walters, Fine-grained named entity recognition using elmo and wikidata, arXiv preprint arXiv:1904.10503 (2019).

[8] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2145–2158. URL: https://aclanthology.org/C18-1182/.

[9] D. D. Africa, S. Salhan, Y. Weiss, P. Buttery, R. D. Martinez, Meta-pretraining for zero-shot cross-lingual named entity recognition in low-resource Philippine languages, in: Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), 2025, pp. 106–127.

[10] J. Adkins, H. Collins, J. Wagner, A. Walsh, B. Davis, Named entity recognition for the Irish language, in: Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025), 2025, pp. 82–96.

[11] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, C. Guo, GPT-NER: Named entity recognition via large language models, in: Findings of the association for computational linguistics: NAACL 2025, 2025, pp. 4257–4275.

[12] T. A. Chang, C. Arnett, A. Eldesokey, A. Sadallah, A. Kashar, A. Daud, A. G. Olanihun, A. L. Mohammed, A. Praise, A. M. Sharma, et al., Global PIQA: Evaluating physical commonsense reasoning across 100+ languages and cultures, arXiv preprint arXiv:2510.24081 (2025).

[13] L.-A. Kaffee, H. ElSahar, P. Vougiouklis, C. Gravier, F. Laforest, J. Hare, E. Simperl, Mind the (language) gap: Generation of multilingual wikipedia summaries from Wikidata for articleplaceholders, in: European Semantic Web Conference, Springer, 2018, pp. 319–334.

[14] UNGEGN, Manual for the national standardization of geographical names, volume 88, United Nations Publications, 2006.

[15] E. Acheson, S. De Sabbata, R. S. Purves, A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, Computers, Environment and Urban Systems 64 (2017) 309–320.

[16] S. Ahmadi, R. Sennrich, E. Karami, A. Marani, P. Fekrazad, G. A. Baghban, H. Hadi, S. Heidari, M. Dogan, P. Asadi, et al., PARME: Parallel corpora for low-resourced Middle Eastern languages, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 30032–30053.

[17] D. Romano, H. R. H. Amin, F. F. Namdar, Arabization, Turkification, and Persianization of the Kurds and the question of genocide, in: The Palgrave Handbook of Kurdish Genocides, Springer, 2025, pp. 257–281.

[18] L. Sahakyan, The Turkification of toponyms in the Ottoman empire and the republic of Turkey, Armenian Folia Anglistika 6 (2010) 149–162.

[19] N. Mamontova, E. Klyachko, 'process toponymy': A GIS-based community-engaged approach to indigenous dynamic place naming systems and vernacular cartography, Cartographica: The International Journal for Geographic Information and Geovisualization 57 (2022) 213–225.

[20] A. P. Perdana, F. O. Ostermann, Eliciting knowledge on technical and legal aspects of participatory toponym handling, ISPRS Int. J. Geo Inf. 8 (2019) 500. URL: https://doi.org/10.3390/ijgi8110500. doi:10.3390/IJGI8110500.

[21] MediaWiki, Sparql/wikidata query service, 2024. URL: https://en.wikibooks.org/wiki/SPARQL/Wikidata_Query_Service, accessed: 2025-04-09.