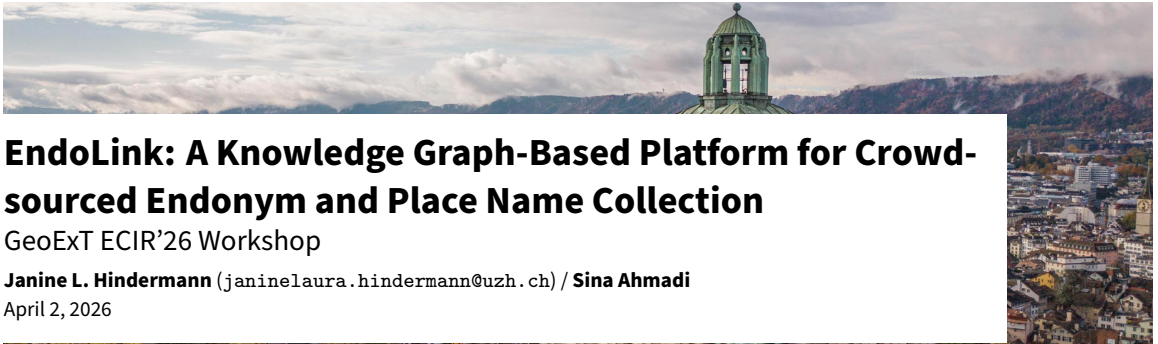
Universität Zürich[UZH]

Department of Computational Linguistics

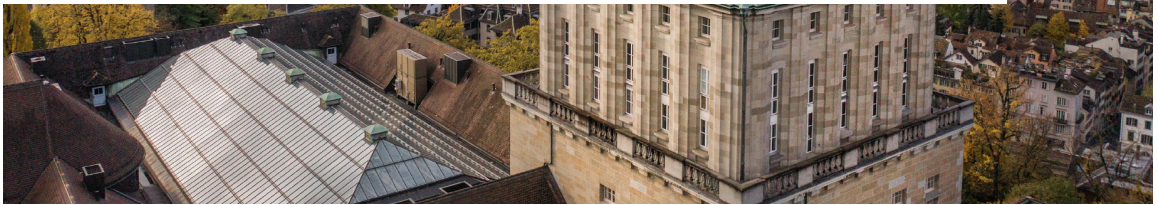


EndoLink: A Knowledge Graph-Based Platform for Crowd-sourced Endonym and Place Name Collection

GeoExT ECIR'26 Workshop

Janine L. Hindermann (janinelaura.hindermann@uzh.ch) / **Sina Ahmadi**

April 2, 2026



Introduction

Motivation and Problem

Place names (toponyms) carry cultural and political significance.

- **Endonyms**: Names used by local communities (e.g., *Züri* for Zurich).
- **Exonyms**: Names used by outsiders (e.g., *Bangkok* for *Krung Thep*).

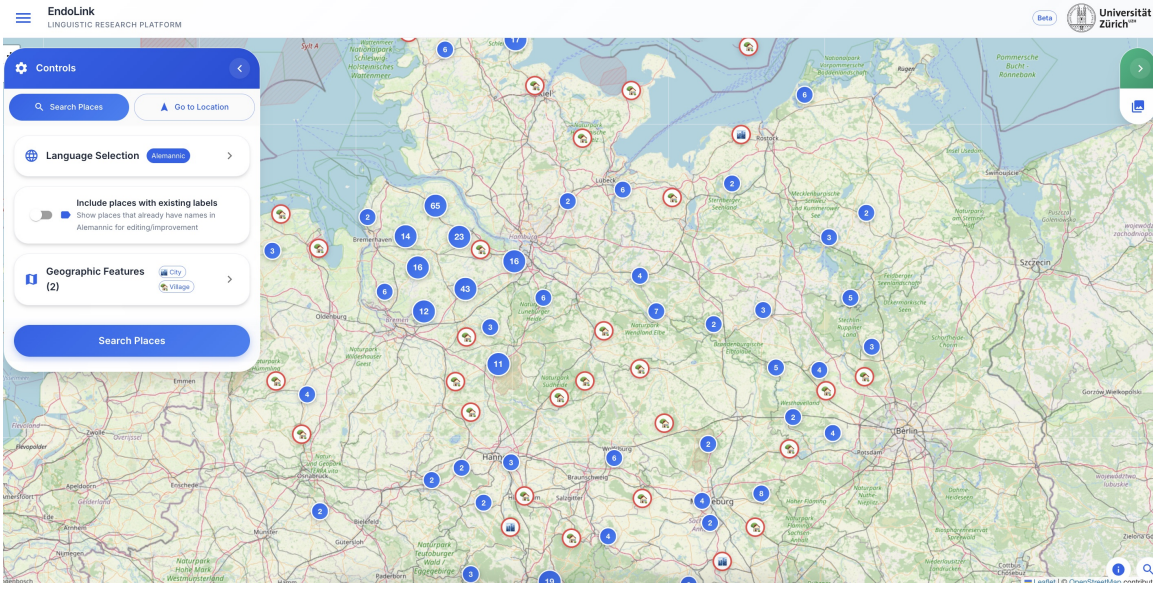
The problem:

- NER systems depend on quality training data (Li et al., 2020).
- Many languages lack labeled resources (Hedderich et al., 2021).
- Wikidata provides structured data but lacks local coverage (Vrandečić and Krötzsch, 2014).

Related Work

Gaps and Standardization

- **Extraction**: Wikidata is used for fine-grained named entity recognition (Dogan et al., 2019).
- **Standardization**: The UN promotes toponym preservation through standardized gazetteers (UNGEGN, 2006).
- **The Gap**: Current participatory platforms operate independently of KGs, while Wikidata requires high technical skills for contribution.


EndoLink
LINGUISTIC RESEARCH PLATFORM
Beta
Universität Zürich
Controls
Search Places
Go to Location
Language Selection
Alemannic
Include places with existing labels
Show places that already have names in Alemannic for editing/improvement
Geographic Features (2)
City
Village
Search Places
Bremerhaven
Oldenburg
Emmen
Zwolle
Enschede
Apeldoorn
Arnhem
Nijmegen
Hannover
Braunschweig
Salzgitter
Wolfsburg
Lübeck
Rostock
Berlin
Potsdam
Szczecin
Gorzów Wielkopolski
Leaflet | © OpenStreetMap contributors

The EndoLink Workflow

EndoLink is a React-based web application providing an interactive map interface.

- **Language Selection**: Choose from Wikidata's language inventory.
- **Retrieval**: SPARQL queries identify locations (e. g., cities) missing labels in the target language.
- **Contribution**: Users enter endonyms.
- **Upload**: Upload to Wikidata to enrich Knowledge Graph.
- **Data Export**: Outputs structured JSON for NLP training.

Demo

https://janinelaura-hindermann.github.io/endolink/

Wikidata Coverage Analysis

Iran, Turkey, and Thailand

Coverage ratio of labels for cities, villages and municipalities

- **Official Languages**: High coverage (Persian, Turkish, Thai).
- **Western Bias**: English, German, and French are disproportionately well-represented.
- **Minority Languages**: Kurdish, Azerbaijani, and Isan have minimal to zero coverage.

LLM Evaluation

We tested GPT-4o on 29 Central Kurdish (Sorani) endonyms from Iran.

- **Performance**: The LLM struggled to identify the correct entities.
- **Geographic Hallucination**: It incorrectly assumed all Kurdish names were in Iraq, even when they were in Iran.
- **Dependency**: LLMs rely on existing multilingual data.

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

Conclusion and Future Work

Closing the Data Gap

Summary:

— EndoLink lowers the barrier for non-technical users to contribute to Wikidata.
— Current knowledge graphs exhibit severe linguistic bias.
— AI models cannot yet compensate for this missing data.

Future Work:

— Gamification of the collection process.
— Direct Wikidata API integration (bypassing CORS limits).
— Suggesting candidate toponyms from monolingual corpora.

References

Cihan Dogan, Aimore Dutra, Adam Gara, Alfredo Gemma, Lei Shi, Michael Sigamani, and Ella Walters. 2019. Fine-grained named entity recognition using elmo and wikidata. *arXiv preprint arXiv:1904.10503*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

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

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.