

GEOKOKOS.ch

A Website for Collaborative Annotation of Toponyms

Prof. Dr. Martin Volk¹, Prof. Dr. Ross Purves², Dr. Simon Clematide¹, Dr. Ekaterina Egorova², Marcel Bühler¹, Janis Goldzycher¹, Lukas Meier¹, Isabel Meraner¹

¹Institute of Computational Linguistics; ²Department of Geography

Our Goal

Enrich Alpine heritage texts with semantic information:

1. recognize all text segments that represent geographical names (toponyms): mountains, mountain cabins, glaciers, places, regions, rivers, valleys, lakes;
2. link the toponyms to geographic databases (geo-referencing).

Our Corpus

The Institute of Computational Linguistics has digitized the annual publications of the Swiss Alpine Club (SAC) issued regularly since 1864 [1]. A lot of errors from automatic character recognition (OCR) have been successfully corrected in a previous Citizen Science project.

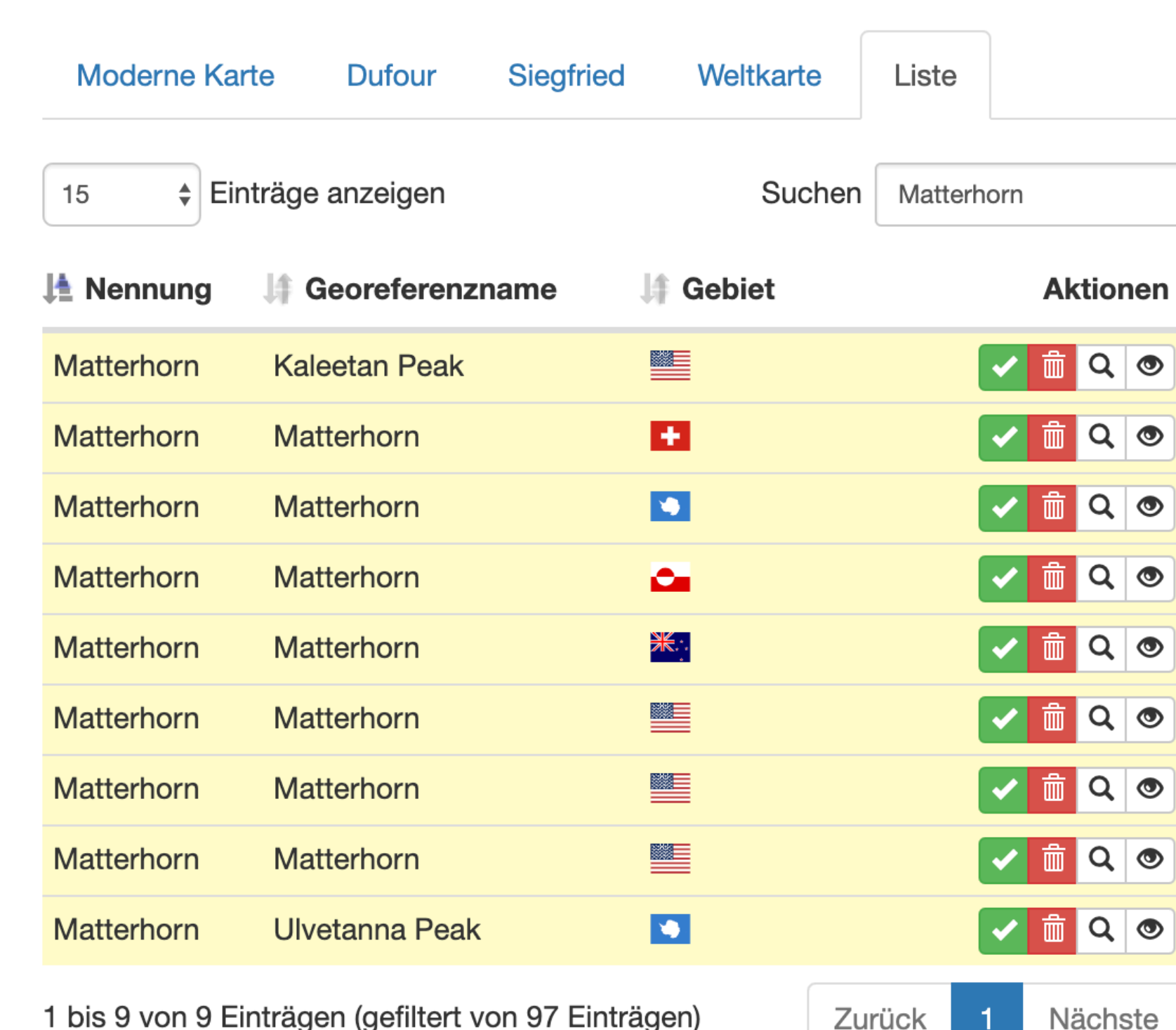
In *GeoKokos*, we want to improve the automatic annotation of toponyms.

Challenges for the Computer

1. It does not recognize all the toponyms, e.g. outdated spellings (*Viesch* vs. *Fiesch*).
2. Some toponyms also appear as normal nouns (*Jungfrau*, *Mönch*) or are part of other names (*Hotel Arosa*).
3. The same geographical name can refer to totally different places (in Switzerland, we have more than 12 peaks named *Schwarzhorn*).

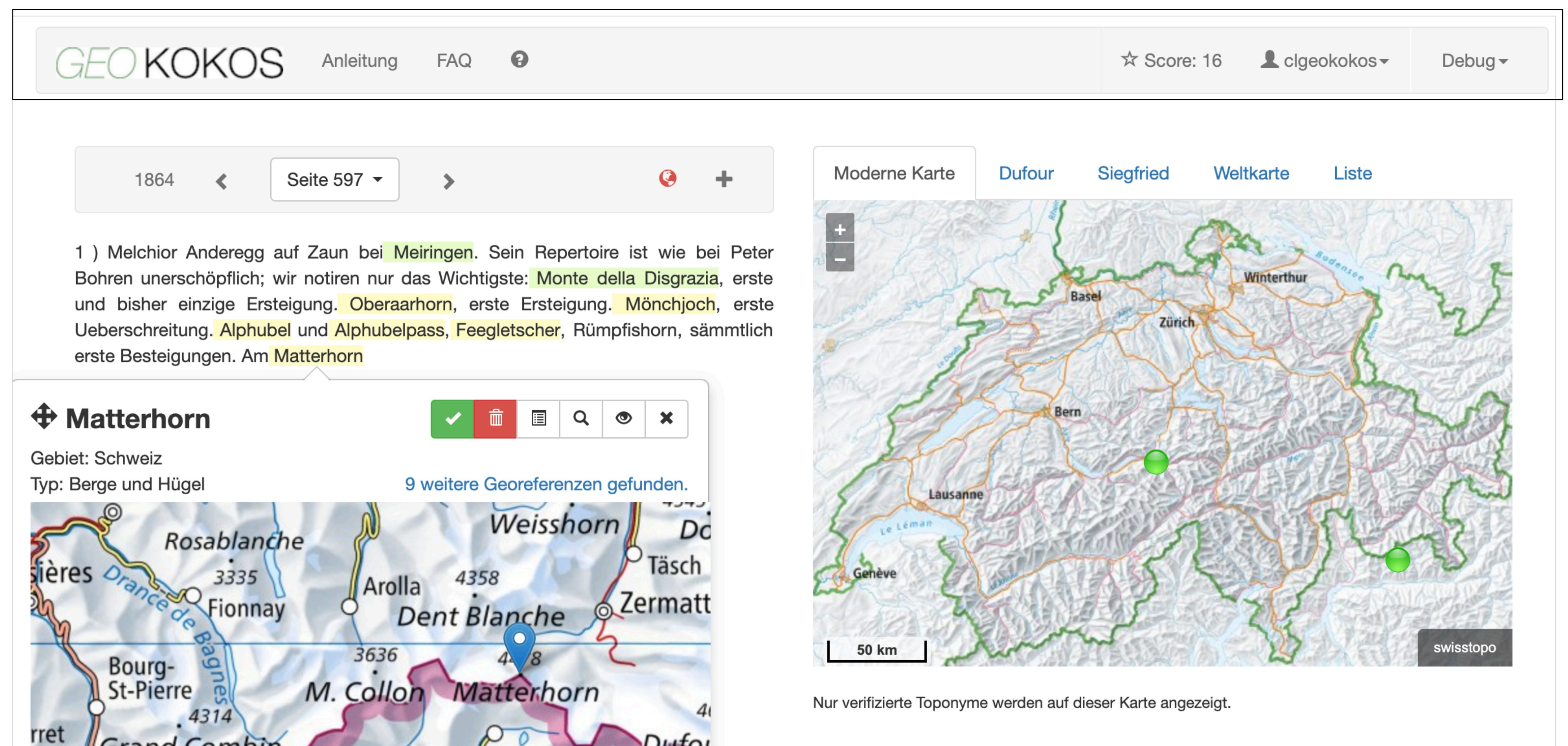
Ambiguous Toponyms

Many toponyms have multiple possible geo-references.



Nennung	Georeferenzname	Gebiet	Aktionen
Matterhorn	Kaleetan Peak	USA	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	CH	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	IT	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	FR	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	DE	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	USA	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	USA	✓ ✖ 🔍 🗑️
Matterhorn	Matterhorn	USA	✓ ✖ 🔍 🗑️
Matterhorn	Ulvetanna Peak	NO	✓ ✖ 🔍 🗑️

Screenshot of *GeoKokos* citizen science web application at <https://geokokos.ch/>. The list view enables the user to choose from multiple potential geographic referents.



Screenshot of *GeoKokos* citizen science web application.

The **left part** shows the text of a book page. The automatically recognized toponym candidates in the text are highlighted in yellow (*Oberaarhorn*, *Mönchjoch*). Verified instances are marked in green (*Meiringen*, *Monte della Disgrazia*).

The **right part** visualizes the geo-referenced toponyms on historical Swiss maps (Dufour, Siegfried) or world maps.

Citizen Scientists in the Loop

1. The recognition and geo-referencing of historical geographic names in diachronic texts requires intellectual work and often detective skills.
2. A user-friendly interface ensures that citizen scientists can efficiently verify toponym candidates and reliably geo-reference them on modern and historical maps.
3. High-quality crowd-crafted annotations allow us to iteratively improve our automatic name annotations based on deep learning methods.

Challenges for the Humans

What constitutes a toponym? [3]

1. Vague geographical terms (*Bernina-Massiv*, *Hochalpen*, *Rheinwaldgruppe*)
2. Compounds (*Matterhornfahrt*, *Maggia-Thal*)
3. Multiword expressions (*Val di Prato*)
4. Toponyms in proper nouns (*Johann v. Weissenfluh*)

→ Concise guidelines, FAQs

How to use the web application?

→ Intuitive user interface, video tutorials

Contact

Prof. Dr. Martin Volk

volk@cl.uzh.ch

<https://www.cl.uzh.ch>

Acknowledgement

UFSP Sprache und Raum

Humans and Machine Learning Working Hand in Hand

The automatically created toponym annotations based on gazetteers and pattern matching serve as a low-effort silver standard for training our initial, neural state-of-the-art model for Named Entity Recognition (NER) [2].

As a next step, the citizen scientists correct and geo-reference a small subset of the silver data. This gold standard then serves as a high-quality training material. This iterative learning setting ensures a more effective and sustainable usage of crowd-crafted annotations. Our approach leverages the learning ability of neural state-of-the-art NER for toponyms and tries to profit as early as possible from the input of the human annotators in order to reduce their correction effort.

Ultimately, this synergy produces texts annotated with precise toponyms which can be used to answer linguistic and geographic questions.

References

[1] Bubenhofer, N., Volk, M., Leuenberger, F. Wüest, D. (eds.): Text+Berg-Korpus (Release 151v01). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924, Die Alpen, Les Alpes, Le Alpi 1925-2014, The Alpine Journal 1969-2008: Institut für Computerlinguistik, Universität Zürich, 2015.

[2] Akbik, A., Blythe, D. & Vollgraf, R. (2018). Contextual String Embeddings for Sequence Labeling. *COLING 2018, 27th International Conference on Computational Linguistics*, pp. 1638–1649.

[3] Clematide, S., Egorova, E., Meraner, I., Purves, R. S. & Volk, M. (2017). Crowdsourcing toponym annotation for natural features: how hard is it? *GIScience workshop*, Melbourne, Australia.