Zur Identifikation und Verortung von Bergnamen in alpiner Literatur

Gerald HIEBEL[1], Klaus HANKE[1], Claudia POSCH[2], Gerhard RAMPL[2], Elisabeth GRUBER[2], Andrea MUSSMANN[2], Eva ZANGERLE[3]

[1] Universität Innsbruck, Arbeitsbereich Vermessung und Geoinformation ·gerald.hiebel@uibk.ac.at
[2] Universität Innsbruck, Institut für Sprachen und Literaturen, Bereich Sprachwissenschaft
[3] Universität Innsbruck, Institut für Informatik

Zusammenfassung

Berge besitzen eine besondere Anziehungskraft für Menschen. Sei es, weil sie eine kaum zu überwindende Barriere oder eine natürliche Gefahr darstellen. Die Zeitschrift des Deutschen und Österreichischen Alpenvereins (ZAV) ist für den deutschsprachigen Raum wohl als einzigartige Textquelle zu nennen. Das von der Österreichischen Akademie der Wissenschaften geförderte go!digital-Projekt „Alpenwort“ (2014-2016) digitalisierte die Alpenvereinszeitschrift (Jahrgänge 1872-1998) und erstellte ein linguistisch annotiertes Korpus. In dem interdisziplinär ausgerichtetem Folgeprojekt „Semantics for Mountaineering History“ mit Beteiligung der Vermessung und Geoinformation, der Sprachwissenschaft und der Informatik geht es um die semantische Anreicherung des Alpenwortkorpus durch die Identifizierung von Ortsnamen, Personennamen und Erstbesteigungen (z.B. die Erstbesteigung des Großvenedigers von Josef Schwab 1841). Dazu ist die Erstellung zweier Register für Orts- und Personennamen erforderlich. Die gesammelten Orts- und Personennamen werden im Anschluss im Alpenwortkorpus so weit wie möglich automatisiert identifiziert. Am Beispiel des 1990 erschienen Artikels „Skitouren in den Ötztaler Alpen“ stellen wir die eingesetzte Methodik vor und zeigen welche Herausforderungen sich bei der automatischen Identifikation und Verortung von Bergnamen und im Allgemeinen Ortsnamen ergeben.

1 Einleitung

Im Oktober 2014 startete an der Universität Innsbruck das von der Österreichischen Akademie der Wissenschaften (ÖAW) geförderte go!digital Projekt „Alpenwort. Korpus der Zeitschrift des Österreichischen Alpenvereins (1868-1998)“. Im Verlauf von drei Jahren wurde aus insgesamt 43.383 Seiten der Alpenvereinszeitschrift ein linguistisch annotiertes Korpus erstellt. Im März 2017 folgte das – ebenfalls durch ÖAW-go!digital finanzierte – Projekt „SEMOHI – Semantics for Mountaineering History“ und legte den Fokus der wissenschaftlichen Arbeiten auf die semantische Annotation des Alpenwort-Korpus. Ziel ist die Identifikation von Orts- und Personennamen innerhalb des Korpus sowie deren inhaltliche Anreicherung mit Informationen aus Orts- und Personenregistern.

Die ursprüngliche Idee zur Erstellung des Alpenwort-Korpus scheint auf den ersten Blick trivial: Berge und in diesem Zusammenhang der Alpinismus spielen eine zentrale Rolle in der Selbstwahrnehmung und –darstellung von Tiroler*innen. Im weiteren Sinne gilt dies

wohl auch für ganz Österreich. Gegen Ende des 18. Jahrhunderts zog der Alpinismus zunehmend die Aufmerksamkeit der breiten Öffentlichkeit auf sich.

Das „Semantics for Mountaineering History“ Projekt hat sich folgende Ziele gesetzt:

- Identifikation von Orts- und Personennamen im Alpenwort-Korpus
- Entwicklung eines Workflows für die automatisierte Orts- und Personennamenerkennung innerhalb dieser Quellenart
- Identifikation von Erstbesteigungen im Alpenwortkorpus
- Semantische Annotation und Repräsentation von Erstbesteigungen, Orten und Personen in maschinenlesbarer Form
- Visualisierung der in den Quellen angegebenen Orte im räumlichen und zeitlichen Kontext
- Open Access / online Dissemination

In diesem Artikel beschäftigen wir uns mit den ersten beiden Zielen in Bezug auf Orte. Herausforderungen und Lösungsstrategien der automatisierten Ortsnamenerkennung (Named Entity Recognition) und der Verknüpfung von Ortsnamen mit Einträgen in Ortsnamenregistern (Named Entity Linking) werden dargestellt. Ziel ist eine Zuordnung von Namen zu spezifischen Bergen und Orten. Über die Koordinateninformation in den Registern ist eine Verortung und Visualisierungen möglich. Darüber hinaus können räumliche Analysen durchgeführt werden. Dies erlaubt beispielsweise Abfragen zu welchen Zeiten Artikel über welche Berge oder Berggebiete geschrieben wurden, wie sich also das alpinistische Interesse im Laufe der Zeit verändert hat.

2 Quellen und Aufbereitung

2.1 Alpenwortkorpus

Die Zeitschrift des Deutschen und Österreichischen Alpenvereins wurde in einem ersten Schritt digitalisiert mit Text Encoding Initiative (TEI 2018) konformen Metadaten (z.B.: Erscheinungsjahr, Titel, Autoren, ..) versehen. Der nächste Absatz zeigt TEI Metadaten des Artikels „Skitouren in den Ötztaler Alpen: 1948 bis 1988“ mit dem ersten Absatz.

```
<text id="av_1990_114_04" decade="1990" year="1990" volume="114" articlenr="04" title="Skitouren in den Ötztaler Alpen: 1948 bis 1988" author="Rudolf Weiss" start_page="39" end_page="46">
<div> Wildspitze 1948 </div>
<div> Die harten Jahre unmittelbar nach dem Weltkrieg hatten für die bergsteigende Jugend gewisse Vorteile: Zwar gab es wenig zu beißen, zwar war die Ausrüstung kläglich, dafür aber gab es mehr Freiheit, als sich Gymnasiasten von heute vorstellen können. Man hatte einen Weltkrieg überlebt und andere Sorgen als die kleinliche Kontrolle des Schulbesuchs. Wer einigermaßen positive Schulleistungen aufzuweisen hatte, mußte keine Sanktionen befürchten, wenn er einmal ein paar Tage in gesunder Höhenluft verbringen wollte. Mein Klassenvorstand, von uns „Burschi" genannt, pflegte sich zwar eingehend nach unserer alpinen Betätigung zu erkunden — aber nicht, um dar-
```

```
aus einen Eltern-Verständigungs-Strick zu drehen, sondern um
sich Tips bezüglich der Schneeverhältnisse zu holen — er war
ebenso bergbegeistert wie wir. </div>
```

In einem nächsten Schritt wird für die Erzeugung des Textkorpus eine Paragraphen-, Satz- und Worttrennung (Tokenisierung), eine Wortartbestimmung (Part of Speech Tagging - POS) und eine orthographische und morphologische Normalisierung der Wörter durchgeführt (Lemmatisierung).

```
<corpus xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="namenannotation.xsd"
id="av_1990_114">
<text id="av_1990_114_04" decade="1990" year="1990" vol-
ume="114" articlenr="04" title="Skitouren in den Ötztaler
Alpen: 1948 bis 1988" author="Rudolf Weiss" page="39">
<div>
<s n="04-1">
<w n="04-1-1" pos="ADJA" lemma="wildspitz">Wildspitze</w>
<w n="04-1-2" pos="CARD" lemma="@card@">1948</w>
</s>
</div>
<div>
<s n="04-2">
<w n="04-2-1" pos="ART" lemma="die">Die</w>
<w n="04-2-2" pos="ADJA" lemma="hart">harten</w>
<w n="04-2-3" pos="NN" lemma="Jahr">Jahre</w>
<w n="04-2-4" pos="ADJD" lemma="unmittelbar">unmittelbar</w>
<w n="04-2-5" pos="APPR" lemma="nach">nach</w>
<w n="04-2-6" pos="ART" lemma="d">dem</w>
<w n="04-2-7" pos="NN" lemma="Weltkrieg">Weltkrieg</w>
```

Die Wortartbestimmung sollte auch Eigennamen erkennen (pos="NE"), funktioniert aber nur sehr beschränkt, wie das Beispiel der „Wildspitze" zeigt, die als Adjektiv (pos="ADJA") gekennzeichnet wird. Für die Erkennung von Ortsnamen innerhalb der alpinen Texte wird deshalb ein Ortsnamenregister verwendet, das oft auch als Gazetteer bezeichnet wird.

2.1 Ortsnamenregister - Gazetteer

Gazetteers gewinnen im Bereich der vernetzten sowie digitalen Kommunikation zunehmend an Bedeutung. Ihr Einsatz erleichtert den Zugang zu Geoinformationssystemen, die Integration ortsbezogener Daten sowie die textorientierte Suche von Rauminformation oder die räumlich orientierte Suche nach Textinformation (SHAW 2016). Für digitale Ortsnamenregister (BERMAN ET AL. 2016) gibt es unterschiedliche Konzepte bzw. ISO Standards wie beispielsweise ISO 19112 Spatial referencing by geographic identifiers (ISO 19112 2003) oder das Gazetteer Konzept der Alexandria Digital Library (ALEXANDRIA DIGITAL LIBRARY GAZETTEER 2004).

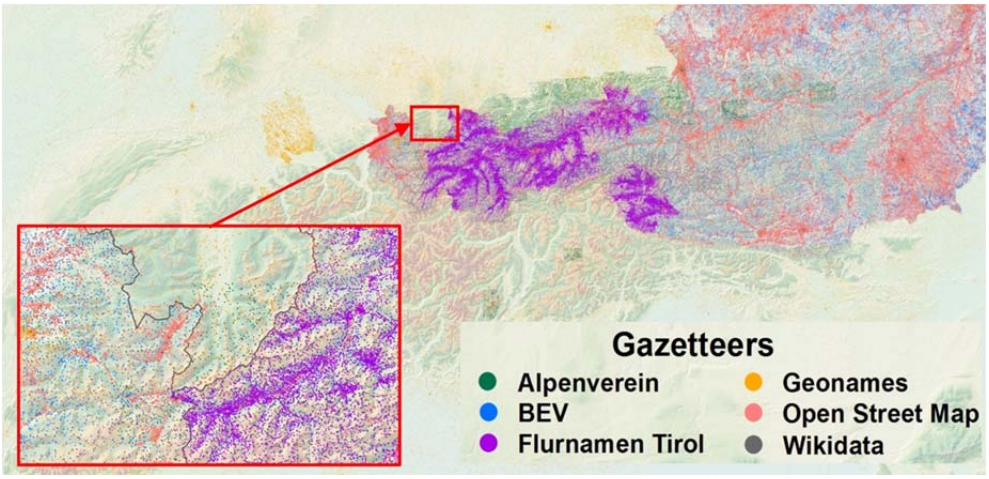

Abb. 1: Der in dem Projekt eingesetzte Gazetteer entstand aus unterschiedlichen Quellen

Elemente eines Gazetteers sind in erster Linie Namen, mögliche alternative Namen, zugehörige Identifier (ID) sowie Koordinateninformationen. Optional, aber für die vorliegende Aufgabenstellung notwendig, ist eine klare Differenzierung zwischen Namen und Referenzobjekten sowie eine Kategorisierung der enthaltenen Referenzobjekte nach definierten Ortstypen.

Um eine hohe Erkennungsrate von Ortsnamen innerhalb des Alpenwort-Korpus zu erzielen, ist der Umfang des Ortsnamensregisters von essentieller Bedeutung. Dazu wurden bereits existierende Gazetteers zusammengefasst, die unterschiedliche Herkunftsarten, Ausdehnungen und Namendichten aufweisen. Abbildung 1 zeigt die Ortsnamen folgender Quellen, die in einem Gazetteer integriert wurden:

- Österreichische Karte 1:50.000; Bundesamt für Eich- und Vermessungswesen (BEV)
- Flurnamendokumentation im Bundesland Tirol (Flurnamen Tirol 2012)
- Alpenvereinskarten
- Geonames
- Wikidata
- OpenStreetMap

Für Wikidata und OpenStreetMap wurden nur Einträge aufgenommen, die in Österreich liegen. Aus Geonames wurden bestimmte Ortsnamenkategorien wie z.B.: Siedlungen und Berge für die Alpenländer extrahiert. Jedes dieser Register hat seine eigenen Ortsnamenkategorien, die zwischen 19 (Alpenvereinskarte) und 1749 (Open Street Map) variieren. Um die verschiedenen Register gemeinsam verwenden zu können ist eine Harmonisierung der Kategorien notwendig, die innerhalb eines Thesaurus erfolgte. Die Oberkategorien dieses Thesaurus sind:

- Topographische Merkmale (z.B.: Berge oder Täler)
- Administrative Einheiten (z.B.: Länder oder Gemeinden)
- Siedlungen (z.B.: Orte oder Städte)
- Bauwerke (z.B.: Hütten oder Bahnhöfe)

- Hydrographische Elemente (z.B.: Flüsse, Seen oder Gletscher)
- Vegetation (z.B.: Wald oder Weide)
- Gebiete (z.B.: benannte kleinräumige Gebiete wie Riednamen)
- Wege (z.B.: Weitwanderwege oder Straßen)
- Aktivitäten (z.B.: Skitouren oder Klettergebiete)

3 Ortsnamenerkennung und Verknüpfung

Der Prozess der Ortsnamenerkennung und Verknüpfung zu Gazetteereinträgen gliedert sich in mehrere Schritte. Zur späteren Evaluierung ist es notwendig eine manuelle Erkennung von Ortsnamen und deren Verknüpfung zum Gazetteer in ausgewählten Artikeln durchzuführen, um so genannte Goldstandard Artikel zu erzeugen. Der nächste Schritt ist die automatisierte Eigennamenerkennung und dann die automatisierte Verknüpfung. Abschließend werden die Ergebnisse mit den manuellen Ergebnissen der Gold Standard Artikeln verglichen und der Prozess wird evaluiert und verbessert.

3.1 Manuelle Ortsnamenerkennung – Goldstandard Artikel

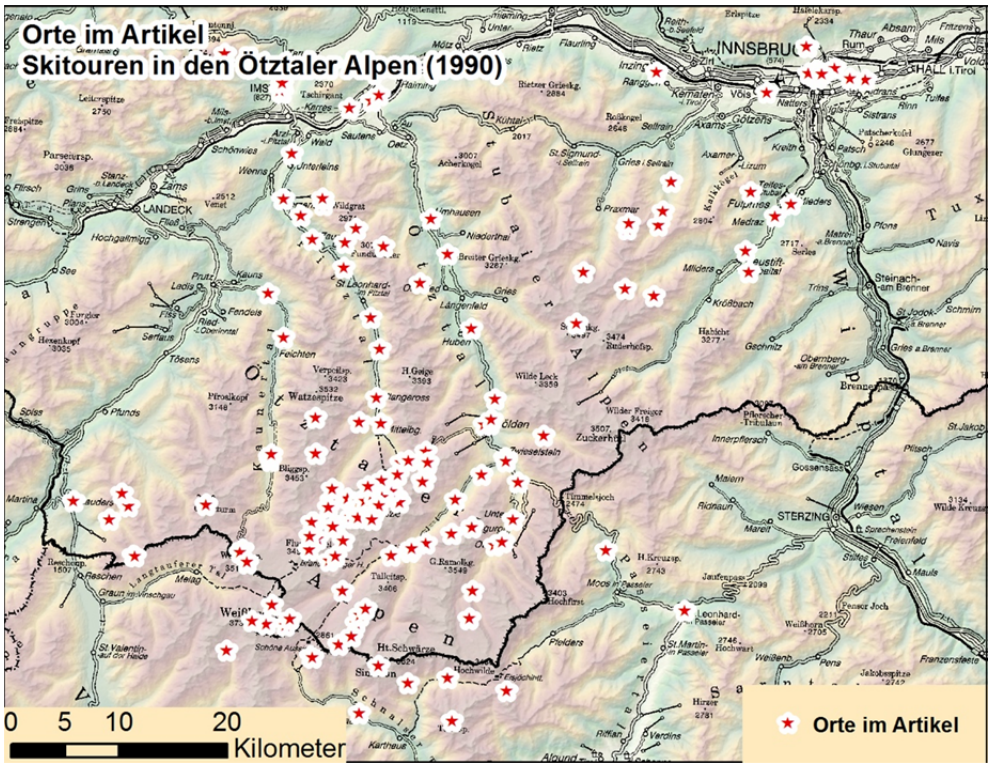

Abb. 2: Manuell identifizierte Orte in dem Artikel “Skitouren in den Ötztaler Alpen”

In diesem Zusammenhang erfolgte eine Auswahl von sieben Artikeln aus den Alpenvereinszeitschriften der Jahrgänge 1873, 1965, 1980, 1981, 1990, 1994 und 1997. Deutlich erkennbar ist, dass es sich hauptsächlich um Artikel aus der jüngeren Geschichte der Zeitschrift handelt. Diese Entscheidung ist dem Umstand geschuldet, dass sich die Schreibweise der Ortsnamen innerhalb älterer Artikel teilweise grundlegend von der aktuellen Schreibweise unterscheidet (z.B. *Thal* versus *Tal*).

Die Erstellung eines Goldstandard Artikels erfordert in einem ersten Schritt die manuelle Identifikation sowie Kennzeichnung der Ortsnamen mit Hilfe von sogenannten xml-tags innerhalb der ausgewählten Texte. Im zweiten Schritt erfolgt die manuelle Zuordnung der Ortsnamen zu Einträgen innerhalb des Gazetteers, die mit einem eindeutigen Identifikator bezeichnet sind:

```
Der <placeName gn='gn_6939049'>Roßkopf</placeName>, 2845 m,
kann dabei ohne große Mühe […]
```

Dieser Prozess führt für den 1990 verfassten Artikel „Skitouren in den Ötztaler Alpen“ zu den in Abbildung 2 dargestellten Namen und in Abbildung 3 wird der Bereich der Wildspitze vergrößert um die Dichte der manuell identifizierten Orte zu illustrieren.

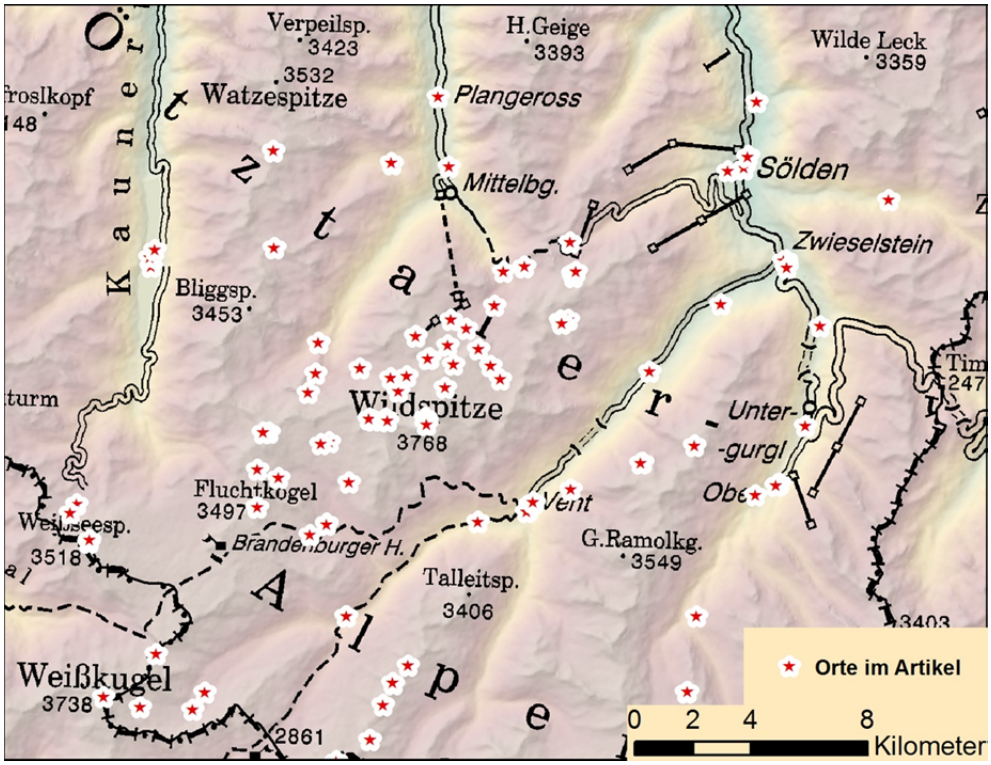

Abb. 3: Manuell identifizierten Orte im Bereich der Wildspitze

3.2 Probleme automatisierter Ortsnamenerkennung und Verknüpfung

Versucht man jetzt diesen Prozess automatisiert durchzuführen, treten eine Reihe weiterer Problemen auf. Das erste davon ist die Ortsnamenerkennung. In dem in 2.2. präsentierten integrierten Gazetteer befinden sich ca. 5,9 Millionen Namen und viele davon entsprechen Gattungsnamen (Appelativa) wie „Hütte“, „Tal“ oder „Weg“. Die Bestimmung, ob es sich bei dem im Korpus gefundenen Wort um einen Gattungsnamen oder einen Eigennamen handelt, der einen Ort bezeichnet, kann nur im lokalen Kontext entschieden werden. Das nächste Problem ist die Deklination von Ortsnamen. Im Deutschen werden Ortsnamen dekliniert und verändern damit ihre Schreibung. Was dazu führt, dass bei einem exakten Stringmatch deklinierte Ortsnamen nicht erkannt werden, oder falsche Ortsnamen erkannt werden, die zufällig der deklinierten Version des Ortes entsprechen (z.B. gibt es einen Ort „Gablers“ in Deutschland, der der deklinierten Version des Berges „Gabler“ in Österreich entspricht). Weitere Probleme der Ortsnamenerkennung sind das Weglassen eines Worts, oder Wortteils, bei der Bezeichnung eines Berges („Geiger“ statt „Großer Geiger“ , „Venediger“ statt „Großvenediger“), alternative Schreibweisen, die nicht im Gazetteer erfasst sind („...spitz“ statt „...spitze“), veraltete Schreibweisen („...thal“ statt „...tal“) oder einfach Rechtschreibfehler. Das nächste Problem der Disambiguierung tritt nach Erkennung der Ortsnamen auf. Es liegt in deren Verknüpfung zum richtigen Gazetteereintrag, also z.B. zu jenem Berg, der in diesem Artikel gemeint ist. Beispielsweise existieren in dem Gazetteer über 50 Einträge für den Roßkopf, sowohl in Deutschland, als auch in Österreich. Welcher ist der Richtige? Führt man also ein matching dieses Begriffs gegen den Gazetteer nur mit exaktem Stringmatch und dessen Deklinationen durch, ergeben sich für unseren Artikel die in Abbildung 4 dargestellten Kandidaten.

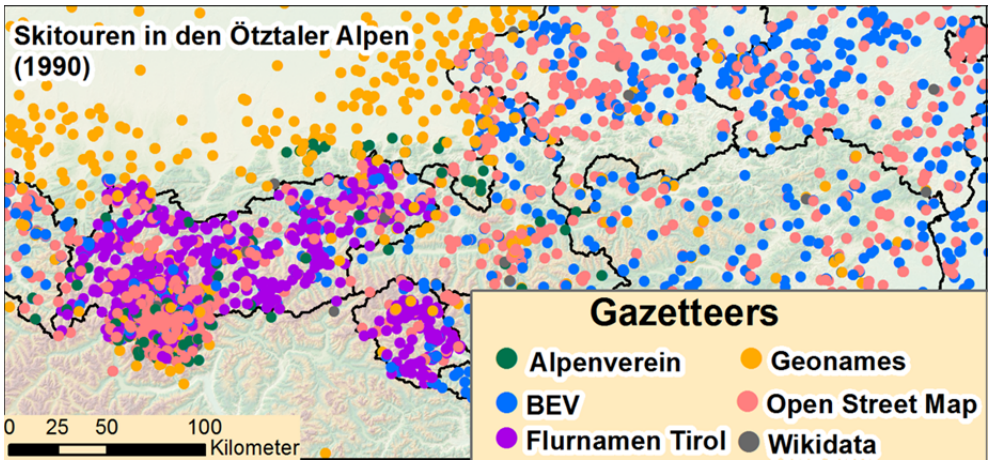

Abb. 4: Automatisch identifizierten Orte mit exaktem Stringmatch und Deklinationen

3.3 Lösungsansätze automatisierter Ortsnamenerkennung und Verknüpfung

Um die Kandidatenliste zu reduzieren, wurde ein Ansatz gewählt, der zwischen einem überregionalen Kontext und einem lokalen Kontext unterscheidet. Die zugrunde liegende

Überlegung ist, dass sich ein Artikel meist mit einem bestimmten Gebiet beschäftigt, wie in unserem Fall den Ötztaler Alpen. Innerhalb dieses Gebietes werden auch Ortsnamen mit lokaler Bedeutung verwendet. Ortsnamen außerhalb des lokalen Kontexts, müssen entweder überregionale Bedeutung haben, oder von Ortsnamen mit überregionaler Bedeutung begleitet werden um den jeweiligen Kontext herzustellen. Die Strategie ist nun erst überregional bedeutende Ortsnamen zu identifizieren, indem die Kandidaten nur aus bestimmten Ortsnamenkategorien (z.B.: Berge oder Städte) gewählt werden und Gattungsnamen ausgeschlossen werden. Dieser Ansatz führt für unseren Artikel zu den in Abbildung 5 dargestellten Namen.

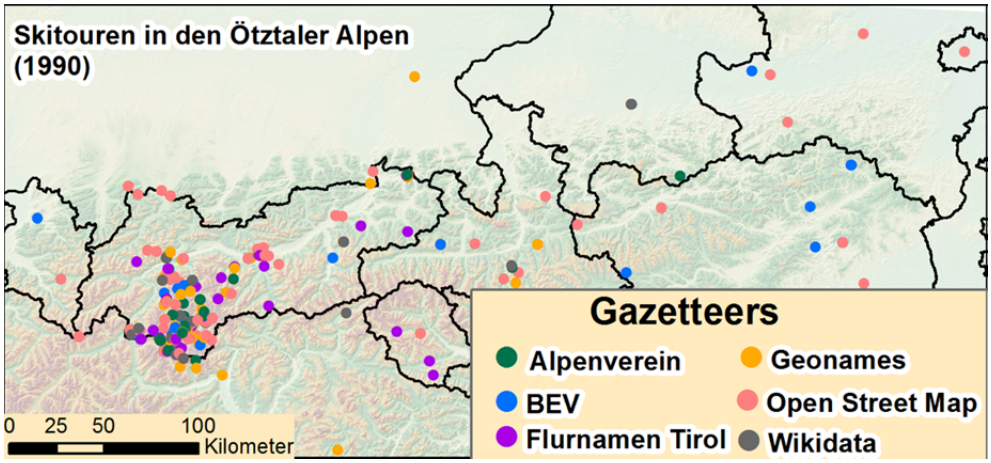

Abb. 5: Überregionale Kandidaten durch Ausschluß von Ortskategorien und Gattungsnamen

Durch eine räumliche Analyse werden Regionen mit hoher Namensdichte als „Kernregionen“ dieses Artikels definiert (Abbildung 6). Der Vergleich zu den manuell identifizierten Orten zeigt, dass durch die Beschränkungen für überregionale Kandidaten nicht alle Orte gefunden werden, die manuell identifiziert wurden. Es erfolgt nun eine Suche im lokalen Kontext der Kernregionen. Hier werden alle Ortskategorien einbezogen und auch Kandidaten, die durch Weglassen eines Wortes/Wortteiles oder alternative Schreibweisen gekennzeichnet sind. Abbildung 7 zeigt, dass über die Bestimmung lokaler Kandidaten alle manuell referenzierten Orte abgedeckt werden können. Sie zeigt auch, dass noch immer zu viele lokale Kandidaten gefunden werden. Ein weiteres noch zu lösendes Problem zeigt die Karte. Die Punkte von Open Street Map Namen ergeben Linien (z.B.: im Bereich von Sölden). Das ist darauf zurückzuführen, dass der Gazetteer in der aktuellen Version nur Punktgeometrien beinhaltet und Linien (die aus mehreren Linienelementen bestehen) aus Open Street Map in Punkte umgewandelt wurden. Flüsse und Straßen machen hier Probleme.

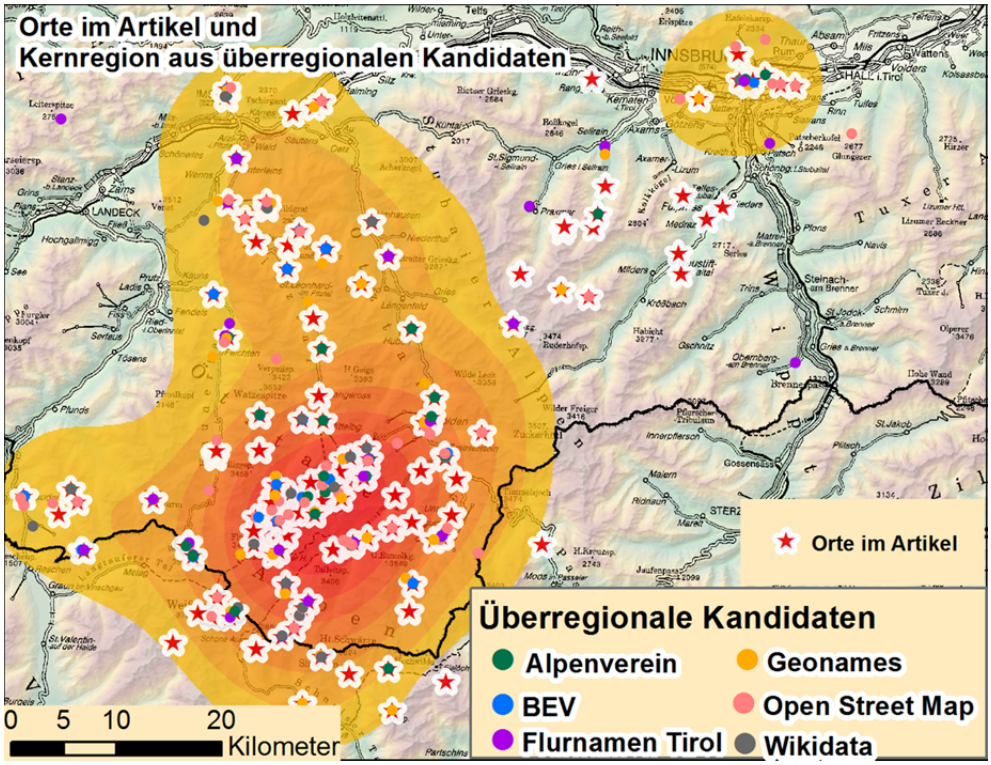

Abb. 6: Kernregion aus überregionalen Kandidaten mit Gold Standard Orten

Zusammenfassung und Ausblick

Wir haben versucht die Herausforderungen aufzuzeigen, die bei der automatischen Ortsnamenidentifikation in alpiner Literatur auftreten. Der dargestellte Lösungsansatz liefert vielversprechende Resultate. Die weitere Reduktion der lokalen Kandidaten und die Identifikation und Zusammenfassung einzelner Orte, die durch mehrere Geometrien repräsentiert werden sind Gegenstand künftiger Untersuchungen.

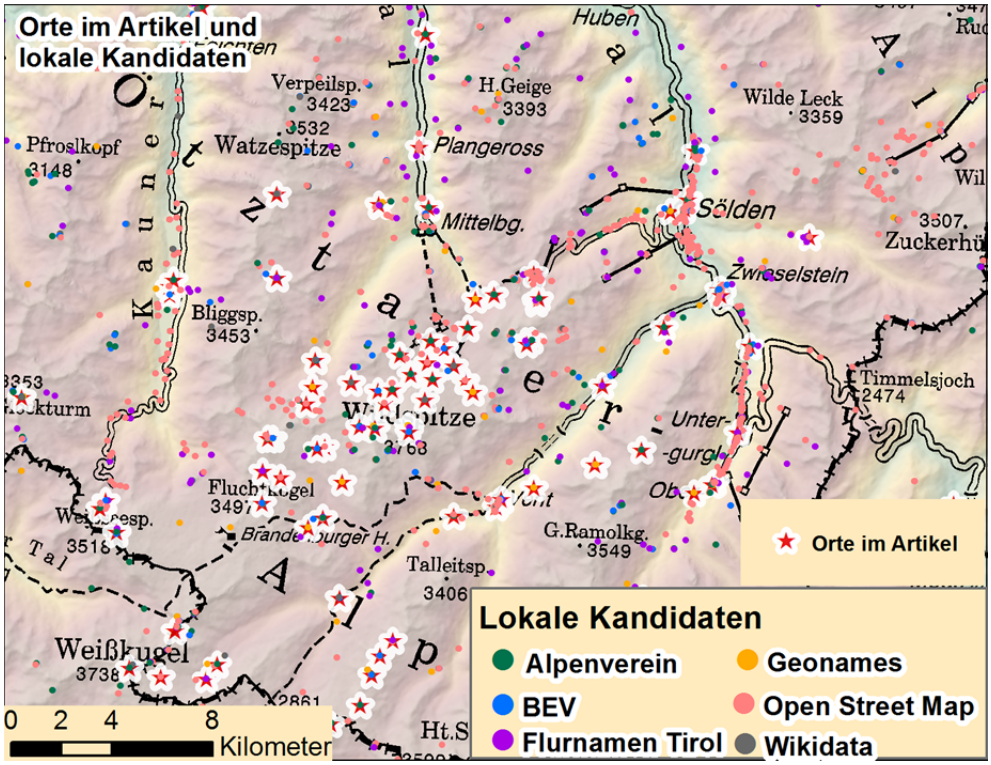

Abb. 7: Bestimmung lokaler Kandidaten innerhalb der Kernregion

Literatur

ALEXANDRIA DIGITAL LIBRARY GAZETTEER (2004). Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright: UC Regents.http://legacy.alexandria.ucsb.edu/gazetteer/, 31.10.2018

BERMAN, M. L., MOSTERN, R., SOUTHALL, H. (2016): Placing Names. Enriching and Integrating Gazetteers. Bloomington, Indianapolis: Indiana University Press.

FLURNAMEN TIROL (2012): FLURNAMENDOKUMENTATION IM BUNDESLAND TIROL http://onomastik.at/content/flurnamendokumentation-im-bundesland-tirol, 31.10.2018

ISO 19112 (2003): ISO 19112 Spatial referencing by geographic identifiers https://www.iso.org/standard/26017.html

SHAW, R. (2016): Gazetteers Enriched: A Conceptual Basis for Linking Gazetteers with Other Kinds of Information. In: Berman, Merrick Lex, Ruth Mostern, Humphrey Southall (2016): Placing Names. Enriching and Integrating Gazetteers. Bloomington, Indianapolis: Indiana University Press, 51–66.

TEI (2018) : Text Encoding Initiative, http://www.tei-c.org/, 31.10.2018