Universität Zürich
Geographisches Institut
Winterthurerstrasse 190
8057 Zürich

Abteilung GIS
Masterarbeit

Geographic Information Retrieval:

Identifikation der geographischen Lage von Zeitungsartikeln

Autor

Tobias Brunner
Bülachhof 2/31
8057 Zürich
tbrunner@geo.uzh.ch

Betreuungsperson

Dr. Ross S. Purves
ross.purves@geo.uzh.ch

Fakultätsvertretung

Prof. Dr. Robert Weibel
weibel@geo.uzh.ch

Abgabe: 30. Mai 2008

Abstract

Geographic Information Retrieval (GIR) can be seen as a hybrid of general information retrieval (IR) and geographic information systems (GIS). The main task is to extract and retrieve information based on spatial and thematic components identified in unstructured data. Being a novel discipline, multiple methods have been applied to this task (Ferrés 2007). In some work (e.g. Smith & Crane 2001, Leidner et al. 2003, Pouliquen et al. 2004) the authors have made use of the toponym's inherent spatial information to disambiguate ambiguous toponyms.

In this thesis, a GIR-system (lacking the ranking-component) based on GATE (a General Architecture for Text Engineering) has been developed and implemented in Java. This program has been used to extract toponyms from newspaper articles. The resulting information in combination with various gazetteers has been used to spatially compare the distribution of toponyms in newspaper articles with a random distribution. Further, both have been compared with the spatial distribution of the referents of ambiguous toponyms.

Statistical analysis (Mann-Whitney-U-Test with $p < 0.001$) shows, that neither the distribution of toponyms in newspaper articles nor the distribution of ambiguous toponyms are similar to the random sample, but that both are more spatially autocorrelated. Further, the mean distance of the referents of ambiguous toponyms is 29% smaller than the distribution of toponyms in newspaper articles. It has also been shown, that the size of the scope of a newspaper article depends on the type (e.g. village, towns, etc.) of toponyms it contains.

These results question the usability of measures based on location for the disambiguation of ambiguous toponyms. Such methods are unlikely to be successful in many cases and their use has to be restricted by hard preconditions.

Further, the investigation of a detailed gazetteer has revealed, that the relative ambiguity of a gazetteer is dependent on the depth of its hierarchy. This means that toponyms which are situated higher in the hierarchy are referentially less ambiguous than toponyms located in deeper levels of the hierarchy.

Zusammenfassung

Geographic Information Retrieval (GIR) kann als Schnittstelle von Information Retrieval (IR) und Geographic Information Systems (GIS) angesehen werden. Es befasst sich mit der Extraktion von Toponymen aus unstrukturiertem Text. Da sich dieses Gebiet erst vor kurzem etabliert hat, wurden dafür verschiedenste Methoden verwendet (Ferrés 2007). So wurde in einigen Arbeiten (z.B. Smith & Crane 2001, Leidner et al. 2003, Pouliquen et al. 2004) die räumliche Information von geographischen Orten für die Auflösung von mehrdeutigen Toponymen verwendet.

Auf der Basis von GATE (a General Architecture for Text Engineering) wurde in Java ein GIR-System ohne Ranking-Komponente entwickelt. Dieses Programm wurde verwendet, um Toponyme aus Zeitungsartikeln zu extrahieren. Anhand dieser gewonnenen Information sowie verschiedenen Gazetteers wurde die räumliche Verteilung sowohl von Toponymen in Zeitungsartikeln als auch der Referenten von mehrdeutigen Toponymen miteinander sowie auch mit einer zufälligen Verteilung verglichen.

Dabei konnte festgestellt werden, dass beide Verteilungen nicht zufällig, sondern räumlich autokorreliert sind. Die Referenten der mehrdeutigen Toponyme sind jedoch durchschnittlich um 29% weniger gestreut als die Toponyme aus Zeitungsartikeln. Diese Unterschiede konnten mit einem Mann-Whitney-U-Test bei $p < 0.001$ bestätigt werden. Weiter wurde erkannt, dass die Grösse des Scopes eines Zeitungsartikels mit der Art (z.B. Stadt, Dorf, u.s.w.) von dessen Toponymen zusammenhängt.

Diese Erkenntnisse bedeuten, dass es unwahrscheinlich ist, ein mehrdeutiges Toponym allein über die räumlichen Eigenschaften derer Referenten aufzulösen. Daher kann die Anwendung von räumlichen Disambiguationsmethoden nur sehr selten erfolgen und muss an strikte Bedingungen gebunden sein.

Weiter hat die Untersuchung eines detaillierten Gazetteers ergeben, dass sich die Ambiguität mit zunehmender Hierarchietiefe vergrössert was bedeutet, dass hierarchisch höher gelegene Ortschaften weniger mehrdeutig sind als hierarchisch tiefer gelegene.

Inhaltsverzeichnis

Figurenverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Kontext

Laut Schell (1999) hat 80% aller Information einen räumlichen Bezug. Da die Internetsuche zum täglichen Begleiter für viele Personen in der heutigen Informations-Gesellschaft geworden ist, muss zunehmend auch diese räumliche Komponente abfragbar werden. Internetsuchen bestehen nicht nur noch aus queries wie „Pamela Anderson“ und „Photoshop Tutorial“ sondern es wird immer mehr auch nach dem nächstgelegenen Schwimmbad, Einkaufshaus oder der Autogarage gesucht. Eine solche Abfrage ist für eine Datenbank mit räumlichen Informationen kein Problem, bei der Suche in einem Text ohne Metadaten oder einer Internetseite ist diese Information jedoch nicht explizit gegeben.

Das Geographic Information Retrieval (GIR) befasst sich unter anderem mit der Extraktion der räumlichen Komponente aus purem Text oder HTML um sie nachher in Abfragen prozessieren zu können. Eine der grössten Herausforderungen ist dabei das Auflösen (Disambiguation) der Mehrdeutigkeit von Ortsbeschrieben (Toponymen) mit mehreren möglichen Referenten wie dies bei „Pfäffikon“ (die Ortschaften im Kanton Zürich und Schwyz) der Fall ist.

Einzelne bisherige Studien (z.B. Leidner et al. 2003, Pouliquen et al. 2004, Overell & Rüger 2006) haben für die Auflösung dieser Mehrdeutigkeit räumliche Informationen verwendet. Dies ist ein einzigartiger Ansatz, da in den meisten Arbeiten nur textliche Hinweise und hierarchische Informationen verwendet wurden. Aufgrund der Aussage, von Schell (1999), ist dieser Ansatz jedoch nahe liegend.

1.2 Zielsetzung und Fragestellung

Diese Arbeit untersucht die räumliche Verteilung von Toponymen, insbesondere diejenige der mehrdeutigen Toponyme, was in den bisherigen Arbeiten versäumt wurde. Dabei wird die Verteilung von Toponymen auf folgende Hypothese getestet:

H_1: Die mehrdeutigen Toponyme sind nicht zufällig über den Raum verteilt, sondern sind räumlich konzentriert.

Parallel dazu wird ein GIR-System ohne die Ranking-Komponente für deutsche Texte auf der Basis des Frameworks GATE[1] (A General Architecture for Text Engineering) entwickelt und anhand von Südostschweizartikeln[2] überprüft. Die räumlichen Informationen dieser Texte, spezifischer die räumliche Ausdehnung von Zeitungsartikeln (Scope), werden wiederum untersucht. Diese wird mit der räumlichen Verteilung von mehrdeutigen Toponymen verglichen, wobei folgende Hypothese getestet wird:

H_2: Die räumliche Verteilung von mehrdeutigen Toponymen ist konzentrierter als die Ausdehnung des Scopes von Zeitungsartikeln.

Die Erkenntnisse dieser Untersuchungen werden schliesslich im Hinblick auf räumliche Methoden zur Auflösung von mehrdeutigen Toponymen angewandt. Weiter werden die extrahierten räumlichen Informationen graphisch dargestellt und auf zeitungsspezifische Phänomene untersucht.

1.3 Gliederung

Der erste Teil der vorliegenden Arbeit beschreibt den aktuellen Forschungsstand im Gebiet des GIR. Dabei wird das Augenmerk speziell auf Methoden zur Auflösung von mehrdeutigen Toponymen gelegt. Im darauf folgenden Teil werden die hier verwendeten Daten und Methoden vorgestellt. Im nächsten Abschnitt werden die Resultate präsentiert und statistisch ausgewertet. Diese werden danach in der Diskussion auf die in Kapitel 1.2 definierten Fragestellungen angewandt. Schliesslich wird ein Fazit dieser Arbeit gezogen und daraus folgende, neue Fragestellungen definiert.

[1] http://gate.ac.uk/
[2] http://www.suedostschweiz.ch/

2 Forschungsstand

In diesem Kapitel soll der aktuelle Forschungsstand im Gebiet des Geographic Information Retrieval zusammengefasst werden, welcher aus für diese Arbeit relevanter Literatur stammt. Nach einer kurzen Einführung in das GIR werden die beiden elementaren Schritte „Toponym Recognition“ und „Toponym Resolution“ beschrieben. Da sowohl die Gazetteer-Lookup-Methode als auch die „Named Entity Recognition“ (NER) nicht klar in diese Schritte einteilbar sind, ist ihnen ein eigenes Kapitel gewidmet. Schliesslich werden die gewonnenen Erkenntnisse kurz zusammengefasst und führen in die Forschungsfragen über.

2.1 Geographic Information Retrieval

Schon seit den Anfängen des Internets wird versucht, anhand von Metadaten die Resultate von Suchabfragen zu verbessern. Dass dieses Ziel jedoch nie erreicht werden kann, ist spätestens seit der Argumentation von Corey Doctorows „MetaCrap“[3], in der er die sieben unüberwindbaren Hindernisse auf dem Weg zu perfekten Metadaten aufführt, bekannt. Die Folge von den fehlerhaften oder fehlenden Metadaten ist, dass ein grosser Anteil der verfügbaren Daten unstrukturiert und daher mit simplen Suchabfragen schwer verwertbar. Aus solchen unstrukturierten Daten Informationen zu gewinnen ist das erklärte Ziel von Information Retrieval.

Geographic Information Retrieval, kurz GIR, ist ein noch relativ junges Forschungsgebiet und entstammt dem Information Retrieval. Das Gebiet des GIR hat sich seit dessen erstmaliger Definition im Jahre 1996 von Larson ständig weiterentwickelt und sich inzwischen als eigene Disziplin aus dem allgemeinen Information Retrieval herauskristallisiert. Als Plattformen wurden dabei bisher das GeoCLEF, welches dem Cross Language Evaluation Forum untergeordnet ist (Gey et al. 2006), sowie ein GIR-Workshop, welcher alternierend an der SIGIR (Purves & Jones 2004) und der CIKM (Jones & Purves 2005) stattfindet, genützt.

GIR wurde erstmals von Larson (1996, S. 82) wie folgt definiert:

„Geographic information retrieval (GIR) is concerned with providing access to geo-referenced information sources“.

[3] http://www.well.com/~doctorow/metacrap.htm

Dies ist eine sehr allgemeine Definition, welche die Hauptstossrichtungen von GIR nicht per Definition mit einschliesst. Daher haben Purves & Jones (2006, S. 375) diese Definition von GIR zehn Jahre später revidiert aber vor allem auch spezifiziert:

„…we define GIR as the provision of facilities to retrieve and relevance rank documents or other resources from an unstructured or partially structured collection on the basis of queries specifying both theme and geographic scope".

Der Hauptunterschied der beiden Definitionen liegt darin, dass Purves & Jones (2006) GIR von Geographic Data Retrieval unterscheiden. Während man die Suche nach geographischen Daten in strukturierten Archiven wie Datenbanken dem Gebiet des Geographic Data Retrieval unterordnen kann, so ist dies bei GIR nicht der Fall. GIR befasst sich also mit un- oder nur teilweise strukturierten Daten. Es kann somit als eine Grundvoraussetzung für Geographic Data Retrieval angesehen werden. Die Herausforderungen, welche die Definition von Purves & Jones (2006) aufwirft, lassen sich in folgende drei Kategorien zusammenfassen.

1. Die Detektion und eindeutige Zuweisung von Toponymen zu geographischen Orten.
2. Die thematische und geographische Indexierung sowie deren Kombination.
3. Die Möglichkeit zu bieten, sinnvolle Abfragen mit geographischem Bezug zu formulieren und zu prozessieren.
4. Die Gütigkeit der Resultate zu ermitteln (Ranking).

Diese Arbeit befasst sich vor allem mit dem ersten Schritt, welcher die grundlegende Voraussetzung für GIR darstellt. Dieser wurde von Larson (1996) in zwei Schritte aufgeteilt, nämlich das „Geo-Parsing" und das „Geo-Coding". Diese Begriffe werden jedoch nur selten verwendet und es gibt daher mehrere, sich überlappende Definitionen, welche diese beiden Schritte oftmals in zusätzliche Unterschritte unterteilen. Auch bei der Benennung dieser herrscht, wie anhand der Tabelle 1 zu erkennen ist, kein Konsens. Für die Definition der Begriffe Geo-NonGeo-Disambiguation und Geo-Geo-Disambiguation sei auf die Kapitel 2.2.1 sowie 2.3.1 verwiesen.

Autor	Erkennung mögl. Toponyme	Geo-NonGeo-Disambiguation	Geo-Geo-Disamb.
Leidner 2008	Toponym Recognition	Toponym Recognition	Toponym Resolution
Ferrés 2007	-	Toponym Resolution	Toponym Resolution
Overell & Rüger 2006	Information Extraction	Disambiguation	Disambiguation
Pouliquen 2006	Geo-parsing	Geo-coding	Geo-coding
Li et al. 2002	-	-	Location Normalization
Larson 1996	Geo-parsing	Geo-parsing	Geo-coding

Tabelle 1: Verschiedene Definitionen der Schritte „Toponym Recognition“ und „Toponym Resolution“

Diese Unterschritte wurden meist aufgrund der Implementierung voneinander abgegrenzt. So werden in dieser Arbeit der Definition von Leidner (2008) entsprechend die Begriffe „Toponym Recognition“ und „Toponym Resolution“ verwendet. Aufgrund der Art der Implementierung wird die Toponym Recognition zudem in die Unterschritte „Gazetteer-Lookup“ (Erkennung von möglichen Toponymen im Text) und „Geo-NonGeo-Disambiguation“ aufgeteilt.

2.2 Toponym Recognition

In diesem ersten Schritt werden Toponyme in unstrukturierten Daten erkannt. Dieser Schritt wird oft auch Named Entity Recognition genannt, wobei im Hinblick auf diese Arbeit schliesslich nur die Erkennung von Toponymen und nicht auch die Erkennung von Personen- und Organisationsnamen relevant ist.

Bei der „Toponym Recognition“ stellen sich folgende Herausforderungen:

- Erkennung von Namen, welche als Toponyme verwendet werden
- Erkennung von wirklichen Toponymen in diesen Namen
- Unterscheidung von metonymisch verwendeten Toponymen

2.2.1 Problem der Geo-NonGeo-Ambiguität

Amitay et al. (2004, S. 273) verwendeten den Begriff Geo-NonGeo-Ambiguität als erste und haben ihn folgendermassen definiert:

"A geo/non-geo ambiguity occurs when a place name also has a non-geographic meaning, such as a person name (e.g., Berlin) or a common word (Turkey)".

Wenn also Toponyme auch für die Bezeichnung von Entitäten die keine Ortschaften sind verwendet werden können, ist das Toponym in Bezug auf Geo-NonGeo-Ambiguität mehrdeutig. Leidner (2008) nennt diese Art von Ambiguität auch morpho-syntaktische-Ambiguität. Ein sehr bekanntes Beispiel von Geo-NonGeo-Ambiguität ist das Toponym „Paris", welches zugleich auch ein Vorname sein kann („Paris Hilton").

2.2.2 Metonymische Verwendung von Toponymen

Von metonymischem Gebrauch eines Wortes wird nach Lakoff & Johnson (1980) dann gesprochen, „wenn eine Entität verwendet wird, um eine andere, mit ihr verbundene, Entität zu referenzieren". Es ist damit also nicht die literarische (bei Toponymen die geographische) Bedeutung des Wortes gemeint, sondern eine damit Verwandte (Leveling & Hartrumpf 2006). Das wohl berühmteste Beispiel hierfür ist, wenn man die US-amerikanische Regierung mit dem Toponym „Washington" referenziert. Markert & Hahn (2002) haben deutsche Magazine untersucht und dabei herausgefunden, dass 17% aller Äusserungen metonymisch verwendet werden können. Laut Leveling & Hartrumpf (2006) kann dies auch auf Toponyme übertragen werden. Sie haben einen deutschen Korpus untersucht und ebenfalls bei 17% aller Toponyme eine Metonymie festgestellt.

Leveling & Veiel (2006) entfernten in ihrer Arbeit alle Annotationen von Toponymen, falls diese auch metonymisch verwendet werden können. Dies hat zur Folge, dass z.B. die Ortschaft „Washington" nie gefunden werden kann. Trotzdem wird dadurch eine höhere Mean Average Precision (MAP, siehe Kapitel 2.7) erreicht. Dabei hängt das Resultat immer damit zusammen, ob metonymisch verwendete Toponyme als Toponyme betrachtet werden oder nicht. In den meisten Publikationen herrscht in dieser Hinsicht kein Konsens. Pouliquen et al. (2006) zum Beispiel weisen dem französischen Wort „Les Parisiens", welches die Einwohner von Paris beschreibt, das Toponym „Paris" zu (siehe Kapitel 2.2.3). Markert & Nissim (2002, S. 4) hingegen werten die Verwendung vom Toponym „Albania" im Folgenden als metonymisch und nicht als Toponym:

„The G-24 group expressed readiness to provide Albania with food aid".

Markert & Nissim (2002) unterscheiden 3 verschiedene Fälle von Toponym-spezifischer Metonymie:

- place-for-people („Die *Schweiz* trinkt zu viel Alkohol")
- place-for-event („Der *Irak* scheint ein zweites *Vietnam* zu werden")
- place-for-product („Ich hätte gerne einen *Montepulciano*")

Ob man metonymisch verwendete Toponyme nun von den eigentlichen Toponymen differenzieren soll oder nicht, ist, wie auch der Vergleich der Arbeiten Markert & Nissim (2002) und Pouliquen et al. (2006) veranschaulicht, nicht klar (Leidner 2008). Nach Kilgarriff & Rosenzweig (2000) ist die Erkennung von Metonymen nicht nur für Computer, sondern auch für Menschen nicht trivial. So sind eindeutige Richtlinien sowie Trainings notwendig, damit Personen das gleiche Verständnis von Metonymie erhalten und folglich die selben Annotationen vornehmen.

2.2.3 Toponym Normalisierung

In Kapitel 2.3.1 wird Geo-Geo-Ambiguität besprochen, was die Mehrdeutigkeit im Hinblick auf die Referenten ist. Oftmals sind aber für einen geographischen Ort mehrere Toponyme gültig. So können einerseits mehrere sprachspezifische Toponyme wie „Genf", „Genève" und „Geneva" (Purves et al. 2007), andererseits auch solche innerhalb einer Sprache („New York City", „NYC" und „Big Apple") alle auf den selben Referenten weisen. Auch können Toponyme mit der Zeit variieren, so wurde zum Beispiel „New York" früher „New Amsterdam" genannt (Ferrés 2007). Die Gleichsetzung dieser Toponyme wird oft als „Normalisierung" beschrieben (Leidner 2008). Ein Ausnahme bilden Li et al. (2002), welche den Begriff Toponym Normalisierung für die Geo-Geo-Disambiguation verwenden, damit jedoch die Einzigen sind.

2.2.4 Lösungsansätze für die Toponym Recognition

Der erste Ansatz zur Erkennung von Toponymen ist die Verwendung eines so genannten Gazetteers (Register mit Ortsnamen). Dabei wird im Text nach Einträgen dieses Registers gesucht und diese im Text als Toponyme markiert. Dieser Ansatz wird Gazetteer Lookup genannt (Kapitel 2.4). Als Resultat eines Gazetteer Lookups

erhält man Namen, welche als Toponyme verwendet werden. Um aus diesen Namen Toponyme zu extrahieren, können entweder handgefertigte Regeln (Kapitel 2.5.2) oder an Trainingsdaten trainierte Algorithmen verwendet werden (Kapitel 2.5.1). Dieselben Methoden werden verwendet, um metonymisch verwendete Toponyme (Kapitel 2.2.2) auszuschliessen.
Im Hinblick auf GIR hat der Gazetteer Lookup den pragmatischen Vorteil, dass allen gefundenen Toponyme auch ein räumlicher Bezug gegeben werden kann. Weiter ist die Genauigkeit – ein guter Gazetteer vorausgesetzt – sehr gut (Mikheev et al. 1999, Clough 2005) und es wird kein bereits annotierter Korpus benötigt.

Ein zweiter Ansatz nennt sich Named Entity Recognition und basiert auf Regeln, welche wiederum von Hand oder per Machine-Learning ermittelt werden können. Die NER bewältigt alle Herausforderungen der Toponym Recognition in einem Schritt. Es erfolgt also bereits eine Geo-NonGeo-Disambiguation sowie eine Unterscheidung zwischen Toponymen und deren metonymischer Verwendung.
NER hat gegenüber einem simplen Gazetteer Lookup im Hinblick auf GIR die Vorteile, dass auch Toponyme erkannt werden können, welche nicht im Gazetteer vorkommen und es können Orts- von Personen- und Organisationsnamen unterschieden werden (Mikheev et al. 1999).

Abschliessend zur Toponym Recognition muss darauf hingewiesen werden, dass sich die Erkennung von Toponymen in der Deutschen Sprache von derjenigen in der Englischen Sprache stark unterscheidet. Der Hauptunterschied ist, dass im Deutschen alle Nomen gross geschrieben werden, im Englischen jedoch nur Named Entities (NE, Rössler 2004). Dies macht die Named Entity Recognition im Englischen erheblich leichter, da man praktisch alle gross geschriebenen Wörter als NE auffassen kann (Amitay et al. 2004). Im Deutschen hingegen ist die Grossschreibung von Wörtern kein Indiz dafür, dass es sich um Named Entities handelt.

2.3 Toponym Resolution

Bei der Toponym Recognition erhält man Toponyme, welchen nun bei der Toponym Resolution ein eindeutiger räumlicher Bezug gegeben wird. Dieser räumliche Bezug entsteht meist indem jedem Toponym ein Koordinatenpaar zugewiesen wird. Im

Folgenden wird die Geo-Geo-Ambiguität, die Haupt-Herausforderung der Toponym Resolution, beschrieben sowie verschiedene Lösungsansätze vorgestellt.

2.3.1 Problem der Geo-Geo-Ambiguität

Die in der Toponym Recognition gefundenen Toponyme können immer noch mehrdeutig sein, diesmal jedoch in Bezug auf den Referenten. Dabei wird oft von Geo-Geo-Ambiguität gesprochen (Amitay et al. 2004). Diese ist klar von der Geo-NonGeo-Ambiguität zu unterscheiden, da sie erst im Schritt der „Toponym Resolution“ aufgelöst werden kann. Ein typisches Beispiel hierfür ist „Aesch“ in der Schweiz. Dieses Toponym hat sowohl in den Kantonen Zürich, Baselland als auch Luzern einen Referenten.

Der Anteil der mehrdeutigen Toponyme wurde bereits in verschiedenen Gazetteers untersucht: Purves et. al (2007) zufolge sind in Grossbritannien etwa 10% aller Ortsnamen mehrdeutig. Smith & Mann (2003) sagen, dass 17% aller europäischen Orte referenziell mehrdeutig sind und es in den USA sogar 57% sein sollen. Li et al. (2003) untersuchten "The Tipster Gazetteer"[4] und fanden darin einen ähnlichen prozentualen Anteil von 18%. Es kann also davon ausgegangen werden, dass je nach Datengrundlage zwischen 10 und 60 Prozent aller Toponyme referenziell mehrdeutig sind.

Für die Auflösung dieser Geo-Geo-Ambiguität gibt es verschiedene Methoden (Heuristiken) die sich gegenseitig nicht ausschliessen, sondern fast immer in Kombination miteinander verwendet werden. Hier soll ein Überblick über elementare, bisher verwendete Heuristiken gegeben werden.

2.3.2 Auflösung durch hierarchischen Kontext

Oftmals verfügen Gazetteers über hierarchische Informationen. Diese können sich auf die Landeszugehörigkeit beschränken (Geonames[5]), oder auch alle administrativen Einheiten mit einschliessen (World Gazetteer[6]). Hierarchische Methoden machen sich das Auftreten von hierarchisch höher liegenden Toponymen im Text zur Auflösung von Geo-Geo-Ambiguität zu Nutze.

---

[4] http://crl.nmsu.edu/cgi-bin/Tools/CLR/clrcat
[5] http://www.geonames.org
[6] http://www.world-gazetteer.com

Hauptmann & Olligschlaeger (1999) suchen in der gegebenen Reihenfolge nach Toponymen mit der Hierarchiestufe „State“, „Country“ und „Continent“. Alle Referenten eines mehrdeutigen Toponyms welche nicht in dieser administrativen Einheit liegen werden weggelassen.

Pouliquen et al. (2006) suchen zuerst im Dokument nach Toponymen mit eindeutigen Referenten. Diesen Schritt nennen sie „shallow parsing“. In einem zweiten Schritt („deep parsing“) wird das Dokument nochmals durchsucht, und nur Toponyme annotiert, welche in denselben Ländern liegen wie diejenigen des ersten Schrittes.

Li et al. (2003) verwenden einen Graphen, welcher der administrativen Hierarchie des verwendeten Gazetteers entspricht. Dieser enthält alle möglichen Referenten der mehrdeutigen Toponyme sowie alle eindeutigen Toponyme. Es wird nun versucht, das Gewicht des Graphen zu maximieren, wobei die Kanten ein höheres Gewicht bekommen, wenn die Knoten innerhalb der gleichen administrativen Einheit liegen.

Amitay et al. (2004) verfolgen den gleichen Ansatz, verwenden jedoch keine Graphen, sondern lösen die mehrdeutigen Toponyme so auf, dass sie in der tiefstmöglichen, gleichen Hierarchiestufe wie ein eindeutiges Toponym des Dokuments zu liegen kommen.

2.3.3 Auflösung durch Reduktion des Gazetteers

Pouliquen et al. (2004) schränken ihren Gazetteer aufgrund der Population der einzelnen Orte ein, um die Ambiguität dessen zu verringern. Hiermit kann nicht nur die Auflösung der Geo-Geo-Ambiguität erleichtert werden, sondern auch diejenige der Geo-NonGeo-Ambiguität. Der Vorteil davon ist, dass die Precision (Kapitel 2.7) erhöht werden kann. Der Recall (Kapitel 2.7) leidet jedoch unter dieser Massnahme (Kornai & Thompson 2005). Obwohl Krupka & Hausman (1998) bei der Reduktion ihres Gazetteers von ca. 110'000 Einträgen auf ca. 9'000 Einträge nur eine kleine Einbusse des F_1-Wertes (Kapitel 2.7), von 0.92 auf 0.89, hinnehmen mussten, wird die Reduktion von Gazetteers im Allgemeinen mehr als Simplifizierung des Problems denn als Lösung davon betrachtet (Leidner 2008).

2.3.4 Auflösung durch beschreibenden Kontext

Vielfach sind Toponyme durch beschreibende Wörter umgeben, welche sie eindeutig auflösbar machen. So kann der Referent des Toponyms „Zürich“ sowohl die Stadt als

auch der Kanton sein. Falls jedoch „Kanton Zürich“ steht, so ist das Toponym „Zürich“ klar dem Referenten „Kanton“ zuweisbar.

Rauch et al. (2003) geben jedem Toponym aufgrund eines Trainingskorpus eine Wahrscheinlichkeit, mit welcher dieses im Text gemeint ist. Danach wird aufgrund von im Kontext erwähnten Wörtern die Wahrscheinlichkeit dieses Toponyms positiv oder negativ verändert. Dasselbe Prinzip wird von Li et al. (2003) verwendet, wobei die initiale Wahrscheinlichkeit aufgrund der Angaben des TIPSTER-Gazetteers (Harman 1992) gewählt wird.

2.3.5 Verwendung von „Default Referents“

In einigen Fällen gibt es für ein Toponym zwar mehrere Referenten, jedoch ist mit sehr hoher Wahrscheinlichkeit immer der Selbe gemeint. So gibt es laut Leidner (2008) im GNS-Gazetteer[7] 980 Referenten für San Francisco. Trotzdem ist höchst selten ein anderer Referent als die Millionenstadt San Francisco im US-Bundesstaat Kalifornien gemeint. Aus diesem Grund wird oftmals ein Referent bestimmt, welcher mit der höchsten Wahrscheinlichkeit der Richtige ist. Diese Default-Referenten werden oftmals anhand ihrer Population bestimmt (Ferrés 2007).

Der TIPSTER-Gazetteer enthält für einen Teil der Einträge bereits Salienz-Werte. Diese wurden entweder von der Populationsgrösse abgeleitet oder aufgrund menschlicher Intuition festgelegt (Leidner 2008). Da sie aber lediglich für einen kleinen Teil des Gazetteers vorhanden sind, haben Li et al. (2002) für alle mehrdeutigen Toponyme eine Yahoo[8]-Abfrage gemacht, und dabei das Auftreten des Toponyms mit anderen, eindeutigen Toponymen statistisch erfasst. Aufgrund dieser Werte wurde dann für die Hälfte der mehrdeutigen Toponyme ein Default-Referent festgelegt. Die restlichen Toponyme konnten den festgelegten Schwellwert nicht erreichen, und wurden aus dem Gazetteer gelöscht. Laut Li et al. (2002) wird dadurch die System-Performance nur marginal beeinflusst, da diese Toponyme auch in Texten äusserst selten vorkommen.

Default Referents werden in den meisten Arbeiten verwendet. Oftmals wird diese Methode angewandt (z.B. Amitay et al. 2004, Rauch et al. 2003, Pouliquen et al. 2004), wenn keine anderen Toponym Resolution-Methoden greifen, oder trotz dieser noch mehrere Referenten vorhanden sind.

[7] http://earth-info.nga.mil/gns/html/index.html
[8] http://www.yahoo.com

2.3.6 Auflösung durch „Co-occurrence“

Der Leser eines Textes kann ein Toponym vielfach aufgrund des textlichen Kontexts auflösen. Diese Methode versucht, aufgrund von Beispieltexten, für jedes Toponym Wörter zu finden, welche oft in Kombination mit dem Toponym vorkommen (Clough 2005). So wird in vielen Texten über den verstorbenen Musiker Kurt Cobain die Ortschaft „Aberdeen“ erwähnt wobei man aber in Wikipedia[9] 22 Ortschaften mit dem Namen Aberdeen findet. Im Zusammenhang mit Kurt Cobain ist jedoch immer diejenige im US-Bundesstaat Washington gemeint, auch wenn es nicht die Grösste oder Bekannteste ist. In diesem Falle ist das gemeinsame Auftreten (Co-occurrence) des Toponyms „Aberdeen“ mit dem Personennamen „Kurt Cobain“ der disambiguierende Faktor.

In der bisher einzigen Arbeit, welche sich mit Co-occurrences befasst, durchsuchten Overell & Rüger (2006) Artikel in Wikipedia[10] nach Toponymen und erstellten für jedes Toponym eine Liste mit Co-occurrences. Sie beschränkten sich aber auf Toponyme, welche zusammen auftreten, weshalb diese Methode besser mit „Toponym-Co-occurrence“ bezeichnet werden sollte. Die besten Resultate (F_1-Wert von 0.80) erhalten sie, indem sie diese Co-occurrences nur unter gewissen hierarchischen und geometrischen Bedingungen (minimales konvexes Polygon) verwenden.

Die Ausweitung der Co-occurrences auf Nicht-Toponyme wurde bisher nur vorgeschlagen (Clough 2005), aber noch nie verwendet.

2.3.7 Verwendung des „One Referent per Discourse“-Prinzips

In der allgemeinen linguistischen Analyse müssen oft Wörter disambiguiert werden, da sie mehrere Bedeutungen haben. So kann das Wort „Tor“ sowohl eine grosse Türe, das „Fussballtor“ (welches wiederum die Aktion oder das physische Tor bedeuten kann) als auch einen Narr beschreiben. Dieser Umstand wurde von Gale et al. (1992) untersucht, und sie sind zum Schluss gekommen, dass in 98% der Fälle nur eine Bedeutung des Wortes im gleichen Text vorkommt. Es kann also davon ausgegangen werden, dass ein Wort in einem Text immer die gleiche Bedeutung hat.

Diese Erkenntnis wird auch bei der Toponym Resolution oftmals genutzt (z.B. Amitay et al. 2004, Hauptmann & Olligschlaeger 1999, Schilder et al. 2004).

[9] http://de.wikipedia.org/wiki/Aberdeen_%28Begriffsklärung%29
[10] http://www.wikipedia.org

2.3.8 Auflösung durch räumliche Informationen

Diese Methode versucht die Toponyme aufgrund ihrer räumlichen Information aufzulösen. Es ist die Einzige, welche sich die räumliche Information der Toponyme in einem Gazetteer zu Nutze macht. Drei verschiedene Ansätze solcher räumlicher Methoden wurden bisher verwendet.

Rauch et al. (2003) sagen, dass es eine starke räumliche Korrelation zwischen geographischen Referenzen in textlicher Nähe gibt. Daher gewichten sie jeden möglichen Referenten mit dem Kehrwert der räumlichen Distanz zum im Text nächstgelegenen, eindeutigen Toponym. So würde in der Textpassage „...Reichenau bei Bonaduz...“ die räumliche Distanz zum eindeutigen „Bonaduz“ Aufschluss darüber geben, dass es sich bei „Reichenau“ um die Koordinate (750312/188006) handelt. Dieser Ansatz wurde auch von Pouliquen et al. (2004, 2006) verwendet und wird oftmals als „Distanzminimierung“ bezeichnet.

Smith & Crane (2001) berechnen für jedes Dokument den Zentroid aller möglichen Referenten des mehrdeutigen sowie der eindeutigen Toponyme, und löschen nachher all diejenigen wieder, welche mehr als zwei Standardabweichungen von diesem Zentroid entfernt liegen. Falls immer noch mehrere Referenten für ein Toponym vorhanden sind, wird dieses anhand der Distanz zum erneut berechneten Zentroiden sowie der Distanz zum nächsten eindeutigen Toponym aufgelöst.

Leidner et al. (2003) minimieren in ihrem Ansatz die Fläche des, alle Toponyme umschliessenden, konvexen Polygons. Dieselbe Heuristik wird auch von Overell & Rüger (2006) verwendet.

Diese räumlichen Methoden basieren jedoch auf Annahmen, welche nie überprüft wurden. Deshalb werden diese Annahmen in Kapitel 2.8.1 ausgearbeitet und kritisch analysiert.

2.4 Gazetteer Lookup

Der älteste Ansatz um Toponyme in Texten zu finden ist per „Gazetteer Lookup“ (Jones et al. 2001). Dies bedeutet, dass Einträge von Listen im Text gesucht und danach im Text entsprechend annotiert werden. Im Falle von mehreren übereinstimmenden Einträgen im Gazetteer, wie zum Beispiel bei „Zürich Oerlikon“, welches sowohl auf Gazetteer-Einträge wie „Zürich“ als auch auf „Oerlikon“ sowie

„Zürich Oerlikon“ zutreffen würde, wird der jeweils längste übereinstimmende (in diesem Fall „Zürich Oerlikon“) Eintrag gewählt (Woodruff & Plaunt 1994).

Als eines der Hauptprobleme des GIR haben Cucchiarelli et al. (1998) und Fonseca et al. (2002) die geringe Verfügbarkeit von grossen Gazetteers erkannt. Daraufhin haben Mikheev et al. (1999) jedoch mit ihrer Untersuchung zum Nutzen von NER ohne Gazetteers gezeigt, dass dies nur auf das Erkennen von Personen- und Organisationsnamen zutrifft. Auch mit einem kleinen Gazetteer, welcher in ihrem Falle aus der Prozessierung von nur 30 Artikeln gewonnen wurde, konnten sehr gute Resultate (Precision-Werte von 0.9 und Recall-Werte von 0.85) erreicht werden, womit die Ergebnisse von Krupka & Hausman (1998) (Kapitel 2.3.3) bestätigt werden. Auch Kornai & Thompson (2005) kommen in ihrem, entsprechend betitelten Paper „Size doesn’t matter“ zum Schluss, dass die Grösse des Gazetteers nicht entscheidend für den Erfolg eines GIR-Systems ist. Allein die Grösse der meisten Gazetteers lässt vermuten, dass ein Grossteil der Einträge nur höchst selten in den Untersuchungsdaten zu finden ist. Hier gilt also das Sprichwort „Qualität vor Quantität“.

Grundsätzlich können Gazetteers auf zwei verschiedene Wege zusammengestellt werden. Zum Einen kann man sie manuell (Mikheev et al. 1999), zum Anderen automatisch, aufgrund eines bereits annotierten Korpus (Stevenson & Gaizauskas 1999) erstellen. Um häufig auftretende Fehler zu vermeiden, können kritische Einträge (wie z.B. „Zug“) aus dem Gazetteer gelöscht werden, oder in einer separaten Liste (Stopword-Liste) gespeichert werden. Die Einträge der Stopword-Liste bekommen dann beim Lookup eine höhere Priorität und verhindern dadurch den Lookup im Gazetteer (Pouliquen et al. 2004, Amitay et al. 2004). Die Gazetteers können aber auch zusätzliche Informationen wie Typ des Orts, dessen geographische Koordinaten (Hill 2000) oder hierarchische Informationen enthalten. Die Definition der Inhalte eines Gazetteers wird oftmals als „Geographic Ontology“ bezeichnet (Jones et al. 2001). Ferrés (2007, S. 44) nennt seine Ontologie „Geographical Knowledge Bases“ (GKB) und definiert sie wie folgt:

„Geographical Knowledge Bases can be defined as geospatial dictionaries of geographic names with some relationships among place names. Usually these places can be political and administrative areas, natural features, and man-made structures”.

Durch die erweiterten Informationen können solche GKB auch für die Schritte „Toponym Recognition“ (Kapitel 2.2) und „Toponym Resolution“ (Kapitel 2.3) verwendet werden.

Gazetteers werden sowohl von kommerziellen Firmen, als auch auf OpenSource-Basis angeboten. Der wohl meist verwendete, frei verfügbare und globale Gazetteer ist der WorldGazetteer[11] (Ferrés 2007). Weitere sind Geonames[12], Falling Rain Global Gazetteer[13] und GeoNet Names Server[14] (GNS). Ein bekannter privater, globaler Gazetteer ist der Getty Thesaurus of Geographic Names[15] (TGN). Er wurde aus den Einträgen anderer Gazetteers im Hinblick auf Kunst und Architektur erstellt wodurch die meisten Einträge zusätzlich sehr gut dokumentiert sind. Ein weiterer Gazetteer wäre der Tipster Gazetteer, welcher zusätzlich auch noch einen Salienzwert für ein Toponym enthält (Li et al. 2003). Der Tipster Gazetteer wird jedoch nicht mehr erneuert.

	Anzahl Einträge	**Hierarchie**	**Feature Types**	**Population**
WorldGazetteer	250'000	✓	✓	✓
Geonames	8 Mio	✓	✓	✓
GNS	5.5 Mio	✓	✓	✓ (klassifiziert)
TGN	1.1 Mio	✓	✓	x
Falling Rain	2.9 Mio	✓	✓	✓
Tipster	160'000	✓	✓	✓

Tabelle 2: Vergleich der Eigenschaften verschiedener Gazetteers.

Oftmals werden die Informationen aus mehreren Gazetteers zu einem neuen kombiniert (Amitay et al. 2004, Clough 2005), wodurch eine grössere Abdeckung erreicht werden kann.

2.5 „Named Entity Recognition (NER)“

Dieser zweite Ansatz versucht, aufgrund des linguistischen Kontexts Toponyme zu erkennen. Dabei werden Regeln angewandt, welche zum Einen von Hand erstellt

[11] http://www.world-gazetteer.com
[12] http://www.geonames.org
[13] http://www.fallingrain.com/world
[14] http://earth-info.nga.mil/gns/html
[15] http://www.getty.edu/research/conducting_research/vocabularies/tgn

werden können, oder aber anhand von Machinelearning-Tools erhoben werden. Nach Kornai & Thompson (2005) können Menschen auch ohne spezifisches Wissen über Toponyme (wie einem Gazetteer) diese in einem Text erkennen. Daher sollte es auch für Computer möglich sein, Toponyme ohne einen Gazetteer zu erkennen.

2.5.1 Machine-Learning Tools

Machinelearning-Tools wie zum Beispiel Alias-I LingPipe[16] analysieren einen bereits annotierten Korpus statistisch, weshalb sie bereits eine grosse Datenmenge benötigen, welche manuell annotiert sein muss (Overell & Rüger 2006). Dabei gibt es die beiden Grundprinzipien des verborgenen Markow Modell (HMM) (Zheng & Sue 2002) und der maximalen Entropie (Borthwick et al. 1998). Das verborgene Markow Modell versucht die verborgenen Zustände einer Prozesskette aufgrund von beobachtbaren Ausgabesymbolen sowie Übergangswahrscheinlichkeiten zwischen den verborgenen Zuständen aufzudecken (Lawrence 1990). Das Prinzip der maximalen Entropie versucht, aus allen modellierten und stochastisch signifikanten Zusammenhängen von Eingabevariablen eines Trainingssatzes mit deren Ausgangsvariablen denjenigen Zusammenhang auszuwählen welcher die bedingte Entropie maximiert (Bender et al. 2003). Sowohl das verborgene Markow Modell als auch das Prinzip der maximalen Entropie sind in LingPipe implementiert. LingPipe erkennt aufgrund dieser Prinzipien Named Entities und klassifiziert sie in die drei Gruppen „Persons“, „Organizations“ und „Locations“.

Bischoff et al. (2006) haben LingPipe sowohl für englische als auch für deutsche Texte verwendet, wobei sich ihre Arbeit auf Query-Expansion fokussierte. Als Modell für die englische NER wurde das Standardmodell von LingPipe verwendet. Das deutsche Modell wurde an einem Korpus der Frankfurter Zeitung mit 36 Mio. Wortformen (Strötgen et al. 2005) trainiert. Die Evaluation ihres GIR-Systems bezog sich auf spezifische Abfragen und nicht auf die Toponym-Erkennung. Daher sind die Ergebnisse für diese Arbeit nicht relevant.
Schockaert et al. (2006) haben LingPipe verwendet, um anhand von Regeln aus englischen Webseiten unscharfe Regionen zu geo-codieren. Sie haben sich für diesen Ansatz entschieden, da es im Englischen relativ simpel ist, Regeln für Named Entities

[16] http://www.alias-i.com/lingpipe

zu generieren und LingPipe bereits ein Standardmodell zur Verfügung stellt. Ihr Geocoding wurde jedoch zusätzlich in Kombination mit einem Gazetteer Lookup gemacht und es wurden keine Angaben zur Genauigkeit ihres GIR-Systems gemacht.

2.5.2 Handgefertigte Regeln

Ausser durch Machine-Learning können sprachliche Regeln auch manuell erstellt werden. Ein Programm, welches manuell erstellte NER-Regeln unterstützt, ist GATE (Cunnigham et al. 2002). Im Unterschied zu LingPipe werden hier mit dem „semantic tagger“ Named Entities anhand von manuell erstellen Regeln erkannt. Diese Regeln werden in JAPE-rules (Java Annotations Pattern Engine) implementiert, wobei GATE standardmässig ein Grundset an Regeln mitliefert. Weitere Informationen zu GATE, vor allem zu dessen Gazetteer-Funktionen, werden in Kapitel 3.3.2 aufgeführt.

Auch Mikheev et al. (1999) haben versucht Texte ohne Gazetteers und nur mit Regeln zu geo-codieren. Dabei haben sie nur die Regeln, welche sie zusammen mit den Gazetteers angewandt haben, verwendet. Bei Personennamen und Organisationsnamen waren die Resultate nicht viel schlechter als bei der Verwendung von Regeln in Kombination mit Gazetteers. Für Toponyme waren die Ergebnisse jedoch ernüchternd. Sie erhielten einen Recall-Wert von 0.46 und einen Precision-Wert von 0.59. Als Begründung liefern sie die relativ einfachen Regeln, welche für Toponyme angewandt wurden. Weiter war es das Ziel des Papers, die Nützlichkeit von kleinen Gazetteers zu evaluieren, und die Resultate sind daher nicht mit anderen NER-Arbeiten zu vergleichen.

Bilhaut et al. (2003) haben für GIR-Zwecke 160 handgefertigte Regeln für die französische Sprache erstellt. Sie nutzten aber auch die Informationen zu den administrativen Einheiten eines Gazetteers (GeoNet Names Server)[17] und machten keine Angaben zu der Evaluation ihrer Regeln.

2.6 Kombination von Gazetteer Lookup mit Regeln

Mikheev et al. (1999) haben gezeigt, dass nur durch Gazetteer Lookup und ohne jegliche Regeln Precision-Werte von bis zu 0.94 und Recall-Werte von bis zu 0.84 erreicht werden können. In Kombination mit manuell erstellten Regeln können diese

[17] http://www.nima.mil/gns/html/

Werte auf 0.95 respektive 0.94 gehoben werden. Allgemein ist die Arbeit von Mikheev et al. (1999) die Einzige, welche einen isolierten Gazetteer Lookup ohne jegliche Regeln verwendet und überprüft. Gazetteer Lookup wird praktisch immer in Kombination mit Regeln verwendet, um Toponyme von anderen Named Entities zu unterscheiden (Clough 2005).

2.7 GIR-Benchmarks

Um die Güte eines GIR Systems zu messen, werden die üblichen Benchmarks aus dem IR genommen. Diese wurden von Kent et al. (1955) erstmals spezifiziert und sind eigentlich für die Bewertung von Ranking-Algorithmen gedacht. Sie lassen sich jedoch ohne weiteres auf die Erkennung von Toponymen anwenden.
Es sind dies Precision (P) und Recall (R). Diese Benchmarks werden wie folgt berechnet:

$$P = \frac{Anzahl\ korrekte\ Toponyme}{Anzahl\ gefundene\ Toponyme}$$

$$R = \frac{Anzahl\ gefundene\ Toponyme}{Anzahl\ vorhandene\ Toponyme}$$

Die Precision misst also die wie viele der gefundenen Toponyme korrekt sind (man sucht nach „False Positives“) während der Recall misst, wie viele aller im Text vorkommenden Toponyme man gefunden hat. Aufgrund dieser Definition existiert ein Tradeoff zwischen diesen beiden Werten. So könnte man in einem Dokument alle Wörter als Toponyme markieren und hätte einen Recall-Wert von 1 wodurch der Precision Wert jedoch sehr tief wäre. Umgekehrt wäre der Precision Wert 1, wenn man nur ein Toponym erkennt, dieses jedoch richtig ist.

Aufgrund dieser Tatsache hat Van Rijsbergen (1979) den F-Wert definiert, welcher den gewichteten Mittelwert von Precision und Recall darstellt, wobei β das Gewicht ist:

$$F_\beta = \frac{\left(\beta^2 + 1\right) * P * R}{\beta^2 * P + R}$$

Meist wird dabei ein Gewicht von 1 genommen und damit der harmonische Mittelwert, F_1 genannt, angegeben (Van Rijsbergen 1979):

$$F_1 = \frac{2 * P * R}{P + R}$$

Ein weiterer Benchmark, welcher für das Information Retrieval relevant ist, ist die Mean Average Precision (MAP). Dieser ist der Mittelwert der Average Precision für eine Gruppe von Anfragen. Die Average Precision wird wie folgt definiert:

„The mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved“ (Buckley & Voorhees 2000, S. 2).

Da die MAP Anfragen (queries) erfordert, wird diese nicht bei den Unterschritten „Toponym Recognition“ und „Toponym Resolution“ verwendet, sondern nur für das gesamte (Geographic) Information Retrieval.

2.8 Schlussfolgerungen und Forschungsfragen

Geographic Information Retrieval ist ein relativ junges Forschungsfeld welches dem allgemeinen Information Retrieval entstammt. Die verschiedenen Herausforderungen (Toponym Recognition, Toponym Resolution), welche sich dabei stellen, wurden aus der allgemeinen IR-Perspektive (z.B. NER, Machine-Learning), aber auch aus neuen, geographischen Perspektiven angegangen. Da dieses Gebiet aber noch jung ist, haben sich bisher keine „State of the Art“-Methoden herauskristallisiert. Diese Arbeit soll die Prämissen, welche für die bisher angewandten räumlichen Methoden unbewusst getroffen wurden aufdecken und untersuchen.

2.8.1 Prämissen für die räumliche Disambiguation

Wie bereits in Kapitel 2.3.8 erwähnt, können mehrdeutige (Geo-Geo-Ambiguität) Toponyme aufgrund deren räumlicher Verteilung disambiguiert werden. Im Folgenden werden die Prämissen, welche für diese Ansätze notwendig sind, analysiert um eine spätere Beurteilung der Anwendbarkeit dieser Ansätze zu ermöglichen.

Alle bisherigen Implementierungen von räumlichen Disambiguationsmethoden werden nicht isoliert verwendet, sondern in Kombination mit Anderen. Daher geben die Evaluationen dieser Algorithmen keinen Aufschluss über deren Anwendbarkeit. Die isolierte Anwendung dieser Methoden würde sicherlich schlechtere Resultate liefern, wäre aber im Hinblick auf die Nutzbarkeit dieser Methoden sehr informativ.

Smith & Crane (2001) lösen mehrdeutige Toponyme auf, indem sie ein Raster (Netz) mit einer Zellengrösse von je einem Grad auf den Globus projizieren. Alle möglichen Referenten der im Text vorkommenden Toponyme werden nun auf dieses Raster interpoliert und davon der Zentroid berechnet. Im nächsten Schritt werden alle Referenten, welche mehr als zwei Standardabweichungen vom Zentroid entfernt liegen, als mögliche Referenten gelöscht. Die Methode soll laut Smith & Crane (2001) mehrdeutige Toponyme wie „Spain“ (das Land und die Ortschaft im US-Bundesstaat Tennesse) auflösen.
Mit ihrer Methode gehen Smith & Crane (2001) davon aus, dass innerhalb einer 1x1-Grad-Zelle nur ein Referent pro Toponym vorkommt. Falls dies nicht der Fall ist, kann ihre Methode nicht greifen. Es werden auch keine Angaben über die Verteilung der Toponyme eines Textes im Raum gemacht. Die Wahl der Zellengrösse wurde daher nicht im Hinblick auf die Methode, sondern mehr im Hinblick auf die Komplexität des Algorithmus gemacht. Auch das von ihnen gelieferte Beispiel mit „Spain“ rechtfertigt die Methode nur teilweise, da dieses Beispiel einfacher mit anderen Methoden (z.B. Hierarchie) gelöst werden könnte. Trotzdem lässt dieses Beispiel darauf schliessen, dass die Methode besser für weit auseinander liegende Referenten von Toponymen geeignet ist.
Die Toponym Resolution von Rauch et al. (2003) basiert auf „Confidence“. Es wird also jedem Referenten eine „Vertrauenswürdigkeit“ (Wahrscheinlichkeit), dass er der Richtige ist, zugewiesen. Diese geht von einem Initialwert aus und wird dann von

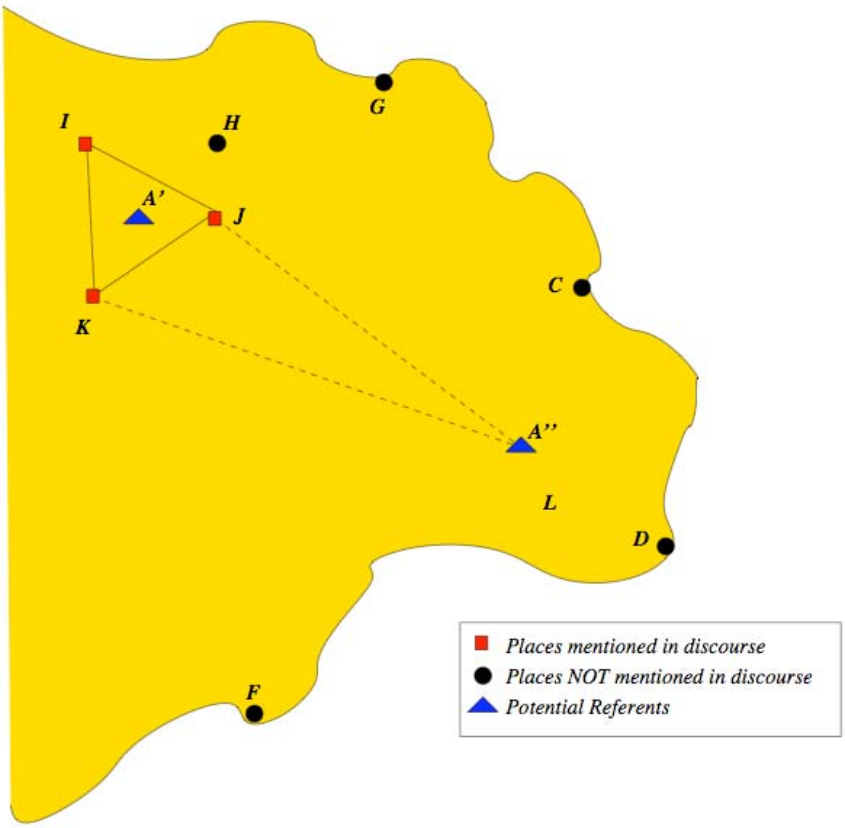
verschiedenen Heuristiken beeinflusst. Dies sind der lokale Textkontext, der räumliche Distanzfaktor sowie ein Populationsfaktor. Wie stark der geographische Faktor die Vertrauenswürdigkeit beeinflusst ist nicht bekannt.

Der geographische Faktor ist eine Funktion der geographischen Distanz zu den anderen im Dokument vorkommenden Toponymen, sowie deren textlicher Nähe. Rauch et al. (2003) weisen darauf hin, dass sie vor allem auch die Distanz zu hierarchisch höheren Toponymen verwenden (zum Beispiel „Chur“ und „Graubünden“). Dies kann zwar als räumliche Distanz gemessen werden, könnte aber durch Bool'sche Geometrieregeln wie „enthalten in“ oder einfach nur hierarchisch gerade so gut gelöst werden.

Dasselbe Prinzip wird von Pouliquen et al. (2004) verwendet, wobei hier Auskunft über die Gewichtung der räumlichen Methode im Hinblick auf die gesamte Toponym Resolution gemacht wird: Sie haben aufgrund von empirischen Versuchen festgestellt, dass Distanzen von weniger als 200 km signifikant sind, und gewichten mit diesen Erkenntnissen die Arcus-Cotangens-Formel (Bronstein et al. 1999).

Auch die von Leidner et al. (2003) verwendete „MINIMALITY“ Methode kombiniert den räumlichen Faktor mit anderen Methoden (linguistischer Kontext sowie Landeshierarchien). Der räumliche Faktor wird so gemessen, dass um alle eindeutigen Referenten sowie um jeweils einen möglichen Referenten eines mehrdeutigen Toponyms im Dokument ein „Minimum Bounding Rectangle“ (MBR) gelegt wird. Die Fläche dieses MBR wird nun minimiert. Leidner et al. (2003) schlagen vor, dass anstatt des MBR ein „Minimum Bounding Polygon“ (MBP) verwendet werden könnte, sie aber das MBR aufgrund der sonst zu hohen algorithmischen Komplexität wählen. Falls mehrere der möglichen Referenten eines Toponyms innerhalb des MBRs liegen, wird derjenige genommen, welcher im Gazetteer weiter oben steht, wodurch diese Notlösung einer zufälligen Auswahl entspricht.

Ihre Methode widerspricht der Aussage von Rauch et al. (2003), welche davon ausgehen, dass die textliche und räumliche Distanz von Toponymen korrelieren. So würden in Figur 1, unabhängig von der räumlichen Dimension, nur 16.7% der, nach Rauch et al. (2003) möglichen, Referenten in Betracht gezogen, falls ein Referent innerhalb des MBP der eindeutigen Toponyme liegt.



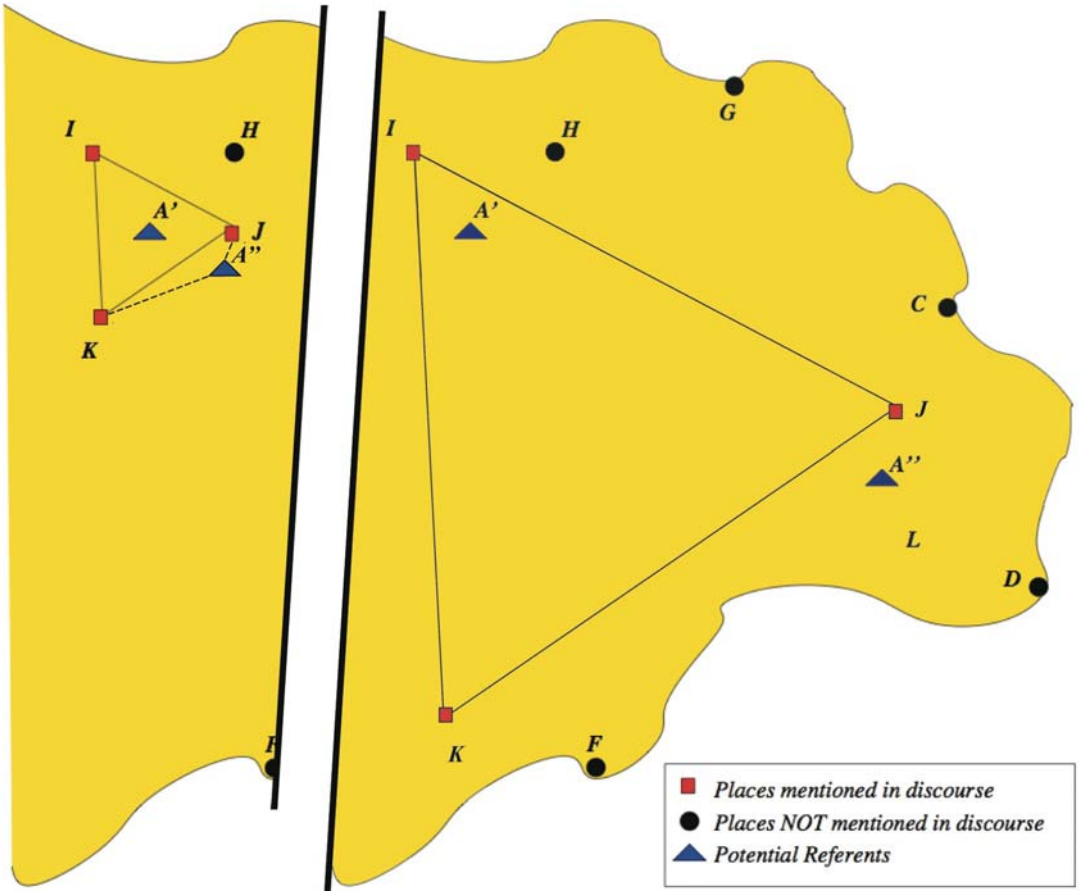
Figur 1: Visualisierung der angenommenen Verteilung von Toponymen und deren Referenten von Rauch et al. (2003) und Leidner et al. (2003), aus Leidner et al. (2003).

Darauf, dass die räumliche Verteilung der Toponyme von Relevanz für diese Methode sein könnte, weisen auch Leidner et al. (2003, S. 2) hin:

„*Probably the smaller the span, the more often this heuristic will be valid*".

Dementsprechend wird auch in Figur 1 eine relativ konzentrierte Verteilung der Toponyme gezeigt. Wie gross die durchschnittliche Spannweite (span) der im Text vorkommenden Toponyme ist, wird jedoch nicht untersucht oder angegeben. Dass nicht nur die räumliche Verteilung der aufgelösten Toponyme, sondern auch die der möglichen Referenten diese Methode beeinflussen, ist in Figur 1 unter anderem durch das Fehlen des Massstabs zu erkennen. In der von Leidner et al. (2007) gegebenen, optimalen Voraussetzung ist der eine mögliche Referent weit von den eindeutigen Toponymen entfernt, während der Andere in der Nähe (hier sogar in der Mitte des MBR / MBP) liegt. Die Variation sowohl der Dimensionen der möglichen Referenten (blaue Triangel) als auch der eindeutigen Toponyme (rote Quadrate) hat dasselbe

Szenario zur Folge, welches durch die MINIMALITY-Methode nicht aufgelöst werden kann (Figur 2).



Figur 2: Modifizierte räumliche Verteilung der möglichen Referenten von Toponymen, angepasst von Leidner et al. (2003).

Ähnlich wie Leidner et al. (2003) haben auch Overell & Rüger (2006) die Fläche des MBR minimiert. Sie erweitern jedoch die Methode von Leidner et al. (2003) dahingehend, dass auch mehrere, im MBR liegende, Referenten auf räumliche Weise disambiguiert werden können. Dabei wird in der angesprochenen Situation der Referent genommen, welcher am nächsten zum Zentrum des MBR liegt. Die Gültigkeit dieser Methode ist, wie bei der Methode von Leidner et al. (2003), nicht mit der Beobachtung von Rauch et al. (2003) übereinstimmend.

2.8.2 Forschungsfragen

Abschliessend kann festgestellt werden, dass sowohl die räumliche Verteilung der möglichen Referenten, als auch die Verteilung der eindeutig identifizierten Toponyme

eines Dokumentes ausschlaggebend für den Erfolg dieser Methoden sind. Dieser Umstand kann in die drei folgenden Forschungsfragen aufgebrochen werden:

Forschungsfrage 1:

Wie sind die Referenten von mehrdeutigen Toponymen räumlich verteilt?

Forschungsfrage 2:

Wie sind die in einem Zeitungsartikel erwähnten Ortschaften über den Raum verteilt?

Forschungsfrage 3:

Unter welchen Bedingungen lässt sich die Geo-Geo-Ambiguität von Toponymen durch räumliche Algorithmen auflösen?

3 Methodik

3.1 Daten

Für die Durchführung dieser Arbeit ist die Verwendung einer Vielzahl von Daten nötig. So müssen Gazetteers, Stopword-Listen sowie weitere Lookup-Listen aber auch ein Korpus als Beispieldatensatz vorhanden sein. Die meisten Daten (v.A. die Gazetteers) müssen zudem bereinigt werden. Für die Toponym Resolution wurden aus SwissNames (siehe Kapitel 3.1.1) und Geonames (siehe Kapitel 3.1.2) je eine komplette Datenbank erstellt. Aufgrund der Objektorientierung sowie der guten Einbindung in Java wurde die „db4objects“[18] gewählt.

3.1.1 SwissNames-Gazetteer

Swisstopo bietet alle Namen der 1:25’000 Landkarte auch separat unter dem Produkt SwissNames[19] an. Dieser Datensatz enthält 155’571 georeferenzierte Toponyme und ist damit der Umfangreichste und Ausführlichste über die Schweiz. Oftmals wurde die Qualität der Gazetteers als entscheidendes Kriterium für erfolgreiches GIR erwähnt (Mikheev et al. 1999, Kornai & Thompson 2005). Die Qualität (als kommerzielles Produkt, welches vom Staat unterstützt wird) von SwissNames ist neben der simplen Verfügbarkeit ein weiteres Kriterium welches für SwissNames als Gazetteer spricht. Zusätzlich ist die Vollständigkeit von SwissNames sehr gut und vor allem über den gesamten beschriebenen Raum (die Schweiz) gleich. Dies erlaubt höchste Flexibilität und beste Vergleichbarkeit.
Ein Nachteil, welcher SwissNames mit sich bringt ist, dass die Toponyme in der offiziellen Sprache der jeweiligen Region gespeichert sind. Somit werden französischsprachige Toponyme in SwissNames nur erkannt, wenn die französische und die deutsche Form die Selbe sind oder im Text die französische Form verwendet wird. Dasselbe gilt für die Toponyme der italienischsprachigen Schweiz.

In SwissNames haben Namensobjekte eine Vielzahl von Attributen. Dabei ist im Hinblick auf diese Arbeit vor Allem die Objektart (Objectvalue) wichtig. Diese

[18] http://www.db4o.com
[19] http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/downloads/landscape/toponymy.html

unterscheidet nicht nur den Typ des Objekts, sondern gibt bei Siedlungen, durch einen zusätzlichen Buchstaben, auch deren Einwohnerzahl mit an (Tabelle 3).

Attribute
X-Coord
Y-Coord
Gemnr
Gemname
Kanton
Objectid
Ojectorigin
Objectvalue
Altitude
Yearofchange

Objectvalue	Beschreibung
HGemeinde	Gemeinde > 50'000 Einwohner
GGemeinde	Gemeinde 10'000 – 50'000 Einwohner
MGemeinde	Gemeinde 2’000 – 10'000 Einwohner
KGemeinde	Gemeinde < 2’000 Einwohner
GOrtschaft	Grosse Ortschaft (> 2’000 Einwohner)
MOrtschaft	Mittlere Ortschaft (< 2’000 Einwohner)
KOrtschaft	Kleine Ortschaft (50 – 100 Einwohner)
Weiler	Weiler (< 50 Einwohner)
Streusiedl	Streusiedlung
Einzelhaus	Einzelhaus

Tabelle 3: SwissNames-Attribute sowie die verschiedenen Siedlungs-Objektarten (Objectvalue), Quelle: http://www.swisstopo.ch, Zugriff 20.08.07.

3.1.2 Geonames Gazetteer

Da SwissNames nur Informationen zu Schweizer Toponymen liefert, muss für die weltweiten Toponyme eine zusätzliche Datenquelle verwendet werden.

Geonames[20] ist ein frei verfügbarer (unter der Creative Commons Attribution 3.0-Lizenz[21]) Gazetteer, welcher über 8 Millionen Toponyme beinhaltet. Diese beschreiben 6.5 Millionen verschiedene geographische Merkmale, wovon 2.2 Millionen Siedlungen beschreiben. Damit ist Geonames der umfangreichste verfügbare Gazetteer. Die Merkmale werden in 9 verschiedene Klassen (Tabelle 4) eingeteilt, welche wiederum in 645 Unterklassen aufgeteilt sind.

[20] http://www.geonames.org
[21] http://creativecommons.org/licenses/by/3.0/

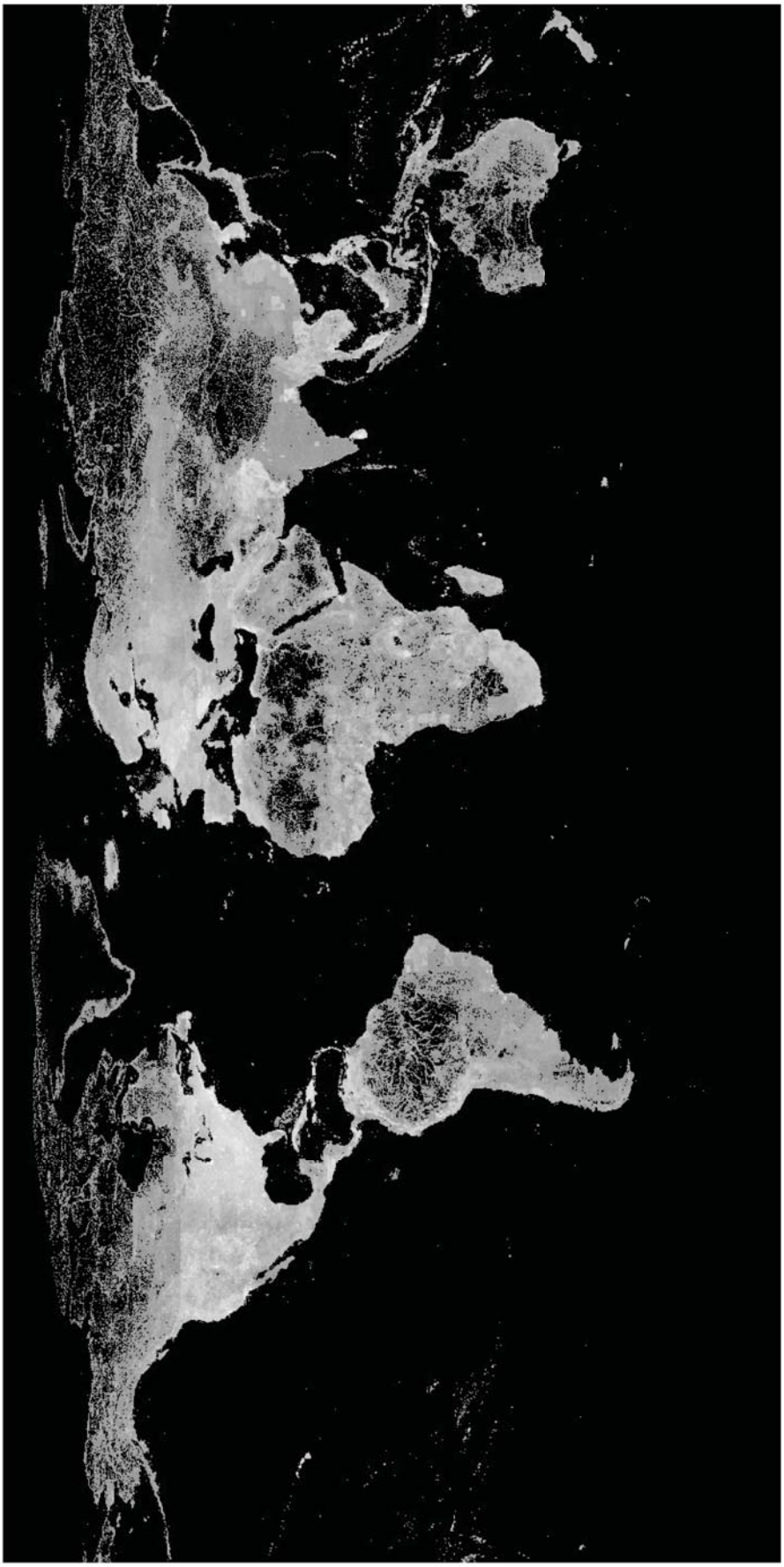
Label	Beschreibung
A	Administrative Merkmale (Länder, Staaten, Regionen)
H	Hydrologische Merkmale (Flüsse, Seen)
L	Flächige Merkmale (Parks, Gegenden)
P	Bevölkerte Merkmale (Städte, Dörfer)
R	Transportwege (Strassen, Eisenbahnschienen)
S	Einzelne Merkmale (Punkte, Gebäude, Bauernhöfe)
T	Terrain-Merkmale (Berge, Hügel)
U	Unterwasser-Merkmale
V	Vegetations-Merkmale (Wälder)

Quelle: http://www.geonames.org/export/codes.html, Zugriff: 18.02.2008

Tabelle 4: Neun Hauptkategorien von Geonames, Quelle: http://www.geonames.org/export/codes.html, Zugriff: 18.02.2008.

Die gesamten Daten können als Export der Datenbank (täglich neu) herunter geladen werden. Weiter werden verschiedene Webservices angeboten, welche Zugriff auf die Datenbank ermöglichen. Die Koordinaten sind im WGS84 (World Geodetic System 1984) gespeichert. Die wichtigsten Quellen für Geonames sind der, in Kapitel bereits erwähnte, Gazetteer GNS, die National Geospatial-Intelligence Agency[22] (NGA) sowie Wikipedia[23]. Die Einträge können aber von den Benützern anhand eines Wiki-Interfaces editiert, korrigiert sowie neu erstellt werden. Die unterschiedlichen originalen Quellen, sowie die von Benützern erstellten Einträge haben zur Folge, dass die Dichte der Merkmale pro Land nicht mit der Realität korreliert (Figur 3). So hat Bosnien-Herzegowina die grösste Dichte an Toponymen, während die Länder mit der grössten Bevölkerungsdichte (wie Indien oder China) viel weniger Einträge haben.
Geonames wird von vielen bekannten Unternehmen verwendet, so z.B. von ESRI, Microsoft, BBC, Adidas und Nike.

[22] http://gnswww.nga.mil/geonames/GNS/index.jsp
[23] http://www.wikipedia.com



Figur 3: Dichtekarte der Geonames-Einträge, Quelle: http://geonames.wordpress.com/2006/12/07/geonames-feature-density-map, Zugriff: 18.02.2008

Der prozentuale Anteil an mehrdeutigen Toponymen in Geonames ist sehr schwer zu berechnen. Zum Einen ist die enorme Grösse der Datenbank (800+ Megabyte) ein technisches Hindernis, da sehr hohe rechnerische Leistung verfügbar sein muss. Zum Anderen sind in Geonames, durch die Möglichkeit der Benutzer Geonames-Einträge zu erstellen, sehr viele Einträge mehrfach vorhanden. Die Qualität der Datenbank ist dadurch zwar nicht gefährdet, denn es gibt pro Toponym mehrere richtige Referenten, welche sich nur in der Beschreibung unterscheiden. Dies führt jedoch dazu, dass die Anzahl Referenten pro Toponym künstlich erhöht wird, was die Auflösung dieser Geo-Geo-Ambiguität kompliziert.

3.1.3 Lookup Listen

Der GATE-spezifische Lookup (siehe Kapitel 3.3.1) verlangt Listen im UTF-8 Format, weshalb die Gazetteers, welche als db40-Datenbanken gespeichert sind, in Listen umgewandelt werden müssen. Aufgrund der hohen Geo-Geo-Ambiguität in SwissNames (siehe Kapitel 4.2) wurden dabei nur die Objektklassen (siehe Tabelle 3) HGemeine, GGemeine, MGemeinde, KGemeinde, GOrtschaft, MOrtschaft und KOrtschaft in die SwissNames-Lookup Liste (*sn.lst*) extrahiert. Dies resultierte in 7’033 unterschiedlichen Toponymen.

Da in dieser Arbeit Zeitungsartikel deutscher Sprache analysiert werden, sind in erster Linie deutschsprachige Toponyme von Nutzen. Geonames bietet eine Liste von Toponymen an, deren Sprache explizit gekennzeichnet ist. Dieser Liste[24] wurden alle deutschen Toponyme entnommen, was in 10'632 Toponymen resultierte (*geonames.lst*). Deshalb wurden alle Städte mit einer grösseren Population als 15'000[25] in eine weitere Liste kompiliert (*largestcities.lst*), welche 20’977 Toponyme enthält.

Zusätzlich zu den Toponymen wird im Lookup-Schritt dieser Arbeit wenn möglich bereits Geo-NonGeo-Ambiguität erkannt. Dies geschieht aufgrund von Listen welche Toponyme beinhalten, welche auch Vornamen (*ambvornamen.lst*) oder Nomen (*nomen.lst*) sein können.

[24] http://download.geonames.org/export/dump/alternateNames.zip
[25] http://download.geonames.org/export/dump/cities15000.zip

Für die *ambvornamen.lst*-Datei wurden alle Knaben- und Mädchennamen vom Online-Dienst Namenfinder[26] mit einer Java-Anwendung extrahiert. Diese wurden mit den Toponymen der Datenbanken (SwissNames und Geonames) verglichen. Falls sie gefunden wurden, wurden sie der *ambvornamen.lst*-Datei angehängt. Dies ergab 193 Toponyme, welche auch Vornamen sein können.
Um Toponyme zu finden, welche auch „normale“ Nomen sein können (d.h. keine Ortschaften oder Vornamen), wurden die 10'000 häufigsten Wörter im Deutschen[27] wiederum mit den SwissNames- und Geonames-Toponymen verglichen und entsprechend in die *nomen.lst*-Datei geschrieben.

Gewisse weitere Toponyme sind nicht in SwissNames oder Geonames enthalten, und wurden deshalb manuell nachgetragen. Alle Länder[28] wurden in die *laender.lst*-Datei geschrieben, die eindeutigen Kantonsnamen in die *kantone.lst*-Datei und die mehrdeutigen Kantonsnamen (z.B. Zug oder Zürich) in die *ambkantone.lst*-Datei.

Die Toponym Resolution macht sich zudem weitere Annotationen zunutze. Dies sind Vornamen[29], welche nicht in SwissNames oder Geonames vorkommen (*vornamen.lst*), Titel wie „Herr“ oder „Fr.“ (*titel.lst*), die Abkürzungen der Schweizer Kantonsnamen (*kantonsabbrev.lst*), Firmennamen[30] (*firmen.lst*), Sportclubs (*sport.lst*), Ortstypen (*loctype.lst*), einer klassischen Stopword-Liste (*stopwords.lst*) sowie einem Bindestrich.
Die Listen der Titel, Sportclubs und Ortstypen wurden selbst und teilweise aufgrund von Wikipedia-Anfragen erstellt.
Die Stopword-Liste wurde erstellt, indem in einigen Dokumenten, unter anderem der elektronischen Ausgabe der Bücher-Serie „Der dunkle Turm“ von Steven King sowie einer Auswahl an Webartikeln der NZZ, nach den Toponymen der Gazetteers gesucht wurde. Dabei wurden vom Autor 87 Toponyme erkannt, welche für ihn klar keine Ortschaftsnamen sondern andere Namen waren.

[26] http://www.namenfinder.de
[27] http://wortschatz.uni-leipzig.de
[28] http://de.wikipedia.org/wiki/ISO-3166-1-Kodierliste
[29] http://www.namenfinder.de
[30] http://zefix.admin.ch/

3.1.4 Text-Korpus

Auf Anfrage stellte die Südostschweiz die gesamten Artikel ihrer Zeitung des Jahres 2006 zur Verfügung. Diese sind im XML-Format gespeichert und enthalten zusätzlich zum Text auch Metadaten wie das Datum oder den Autor. Insgesamt umfasst diese Kollektion etwas mehr als 32'000 einzelne Artikel.

3.2 Berechnung der Ambiguität in SwissNames

Um die Relevanz der Geo-Geo-Disambiguation für SwissNames zu beurteilen wurden diese Daten in einer Java-Applikation statistisch ausgewertet. Diese macht sich die Objektorientierung von Java zunutze und speichert somit die gesamten Informationen pro Ortsangabe in einem Objekt. Um die Häufigkeit der einzelnen Toponyme zu berechnen, müssen alle diese Objekte miteinander verglichen werden. Dies hat eine Komplexität von quadratischer Ordnung zur Folge. Daher werden die Objekte zuerst alphabetisch nach dem Namen sortiert, wobei der Merge-Sort-Algorithmus (Knuth 2000) zwecks seiner tiefen Komplexität angewendet wird. Die Häufigkeit der sortierten Objekte lässt sich nun in linearer Ordnung berechnen.

3.3 Toponym Recognition

Sowohl NER als auch Gazetteer Lookups sind im Hinblick auf die Toponym Recognition in der Regel sehr erfolgsversprechend, sind aber an verschiedene Voraussetzungen gebunden. So muss für die NER ein bereits annotierter Korpus vorhanden sein, um das Modell daran zu trainieren. Für einen Gazetteer Lookup benötigt man einen Gazetteer. Diese Grundvoraussetzungen legen die Verwendung des Gazetteer Lookups nahe, da das Geographische Institut der Universität Zürich eine Lizenz für SwissNames besitzt und kein bereits annotierter Korpus vorhanden ist. Weiter sind online viele Gazetteers vorhanden, welche sowohl weltweit, als auch kostenlos verfügbar sind (siehe auch Kapitel 2.4). Für die Verwendung eines Gazetteer Lookups spricht weiter, dass er speziell im Hinblick auf die Toponym Recognition sehr gute Resultate erreichen lässt (Mikheev et al. 1999).

3.3.1 Verwendung eines Gazetteer Lookups

Es gibt verschiedene Möglichkeiten einen Gazetteer Lookup zu implementieren. Die Einfachste ist, ein Programm mit dem OpenSource-Tool FLEX[31] (Fast Lexical Analyzer) zu erstellen. FLEX erkennt Muster in Texten und ermöglicht somit die Suche nach den Einträgen einer Liste im Text. Die Aufwändigste aber auch Flexibelste wäre ein eigenes Programm zu schreiben. Ein initiales Experiment, welches sowohl für die Toponym Recognition als auch die Toponym Resolution dieselbe Datenbank verwendete, führte zum Ergebnis, dass die Prozessierungsgeschwindigkeit zu langsam ist. Dieses Test-Programm wurde in Java in Kombination mit einer db4o-Datenbank erstellt. Der Vorteil eines solchen Programms wäre, dass es kompakter als eine Kombination aus mehreren Lösungen ist.

Einen Mittelweg zwischen einer kompletten Eigenentwicklung und FLEX bietet das von der Natural Language Processing Group der Universität von Sheffield entwickelte GATE an. Von den Entwicklern wird dieses wie folgt beschrieben:

> *„…a framework and graphical development environment which enables users to develop and deploy language engineering components and resources in a robust fashion“* (Cunningham et al. 2002, S. 1).

Folgende fünf Punkte charakterisieren GATE (Cunningham et al. 2002):

- Die verschiedenen Programm-Schichten wie Datenspeicherung, Visualisierung, Prozessierungskomponenten, Datenstrukturen und Algorithmen werden klar voneinander getrennt
- Automatisches Überwachen der Leistung der Sprachprozessierung.
- Integration von offenen Standards wie Java und XML und damit höchstmögliche Portabilität.
- Ein Grundset von NLP-Komponenten, welche weiterentwickelt aber auch ersetzt werden können.

[31] http://www.gnu.org/software/flex/

Das Herzstück von GATE ist ANNIE (A Nearly New IE-System). ANNIE ist ein Paket aus mehreren Prozessierungs-Ressourcen, welche miteinander kombiniert werden können. Für diese Arbeit sind die beiden Module „Gazetteer“ und „Semantic Tagger“ relevant.

Der Gazetteer sucht im Text nach den Wörtern in einer oder mehreren Listen. Diese Listen werden in einer Index-Datei referenziert, wo auch die gewünschte Annotation pro Liste gemacht wird (z.B. „Location“ für die Toponym-Liste). Der Lookup erfolgt, indem die Listen in endliche Automaten (Gill, 1962) umgewandelt werden. Diese bestehen aus Zuständen, Übergängen und Aktionen und werden dann auf den Text angewendet (Cunningham et al. 2007).

Der Semantic Tagger funktioniert mit handgeschriebenen Regeln, welche gewisse Muster erkennen und diesen eine Annotation zuweisen. Diese Regeln werden in JAPE geschrieben, was eine speziell für GATE entwickelte Sprache ist, welche auf der CPSL-Sprache von Appelt (1996) basiert. Die Muster können sowohl geschriebenen Text als auch Annotationen beinhalten, so kann z.B. eine „Location“-Annotation, vor welcher die beiden Grossbuchstaben „FC“ stehen, neu als „Fussballclub“ annotiert werden. Die in JAPE geschriebenen Regeln werden sequentiell abgearbeitet und die Priorisierung dieser Regeln kann nach den folgenden Prinzipien geschehen (Cunningham et al. 2007):

- „brill“: Alle Regeln werden angewandt. Somit können auch mehrere Annotationen für ein Wort gemacht werden.
- „all“: Auch hier werden alle Regeln angewandt, aber nachdem eine verwendet wurde, wird erst ab dem nächsten Wort bzw. der nächsten Annotation weiter gesucht.
- „first“: Das zuerst gefundene Muster wird erkannt und annotiert.
- „once“: Falls ein Muster erkannt wird, wird dieses annotiert und der Semantic Tagger beendet. Diese Regel eignet sich nicht für mehrere Annotationen in einem Text.
- „appelt“: Nur eine Regel kann pro Wort/Annotation angewandt werden. Diese wird aufgrund von drei Regeln bestimmt:
    1. Die Regel, welche auf die längste Region zutrifft wird angewendet.
    2. Wenn mehrere Regeln zutreffen, wird diejenige mit der höchsten Priorität angewendet.

3. Wenn mehrere Regeln die gleiche Priorität haben, wird die zuerst in der JAPE-Datei aufgeführte Annotation genommen.

GATE wurde bisher nicht nur für allgemeine NLP, sonder auch spezifisch für GIR-Zwecke verwendet. Zong et al. (2005) haben mit ANNIE geographische Metadaten aus Webseiten extrahiert, wobei der Standard-Gazetteer um den Faktor 10 erweitert wurde. Clough (2005) hat im Zuge des SPIRIT-Projekts[32] aufgrund von GATE-Komponenten (siehe Kapitel 3.3.2) Metadaten aus Webseiten extrahiert. Beide Arbeiten wurden auf Dokumente englischer Sprache geprüft, wobei SPIRIT grundsätzlich auch mehrsprachige Dokumente prozessieren können sollte.

3.3.2 Beschrieb der Geotagger-Klasse

GATE kann sowohl als GUI als auch als Framework verwendet werden. Basierend auf Beispielcode (CookBook.java und CorpusSaver.java) wurde von Clough (2005) eine Wrapperapplikation für GATE geschrieben, welche von Pasley et al. (2007) für die Erkennung von umgangssprachlichen Regionen angepasst wurde. Dabei wurde der Standard-Gazetteer von GATE durch den HashGazetter[33] von OntoText[34] ersetzt (Clough 2005). Dieser wandelt die Listen nicht in Automaten, sondern in HashMaps[35] um. Dadurch kann durchschnittlich 75% an Arbeitsspeicher gespart werden und auch die Geschwindigkeit wird um den Faktor 3 erhöht (Cunningham et al. 2007). Dieser Code wurde freundlicherweise zur Verfügung gestellt und für diese Arbeit angepasst. Das originale Geotagger.java liest Dateien aus einem angegebenen Ordner ein, führt die Gazetteer-Funktion von GATE (mit dem HashGazetteer) aus und schreibt die annotierten Dateien im HTML-Format wieder aus. Für die Anwendung auf den Südostschweiz-Korpus wurde der Code an das XML-Format angepasst. Weiter wurden nebst der „Location“-Annotation mehrere zusätzliche Annotationen hinzugefügt, welche in einem späteren Schritt sowohl für die Auflösung von Geo-NonGeo-Ambiguität als auch für die Toponym Resolution verwendet werden. Diese beiden GIR-spezifischen Schritte werden von Geotagger.java aufgerufen, werden jedoch durch komplett neuen Code ausgeführt (siehe Kapitel 3.3.7). Die von

[32] http://www.geo-spirit.org/
[33] http://www.gate.ac.uk/releases/gate-4.0-build2752-ALL/doc/javadoc/com/ontotext/gate/gazetteer/package-summary.html
[34] http://www.ontotext.com
[35] http://java.sun.com/j2se/1.5.0/docs/api/java/util/HashMap.html

Geotagger.java gemachten Annotationen sind in der *lists.def*-Datei indexiert und werden nachher durch JAPE-Regeln (*gaz_transfer.jape*) in die entsprechenden Annotationen (auch „Tags“ genannt) umgewandelt. Die Priorisierung folgt nach dem „Appelt“-Prinzip, was mehrere Annotationen verhindert. Somit ist sie durch die Kombination der Prioritätsstufen in *gaz_transfer.jape* und die Reihenfolge in *lists.def* definiert.

Als Ausgabedatei von Geotagger.java erhält man das ursprüngliche XML-File, in welchem die entsprechenden Wörter durch die in Kapitel 3.1.3 aufgeführten Annotationen gekennzeichnet sind.

3.3.3 Identifizierung von Metonymen

Wie in Kapitel 2.2.2 bereits besprochen, wird mit metonymisch verwendeten Toponymen unterschiedlich umgegangen. Da der Fokus dieser Arbeit nicht auf der Metonymie-Auflösung liegt, wird Metonymie aufgrund folgender Überlegungen nicht betrachtet:

In der Literatur herrscht keine eindeutige Auffassung über die Definition von metonymisch verwendeten Toponymen. So fassen Pouliquen et al. (2006) die Personen aus Paris („Les Parisiens“) als die Ortschaft Paris auf, während Markert & Nissim (2002) jedes eigentliche Toponym nicht als solches klassieren, falls es durch die Wendung „die Personen in ... (dem original verwendeten Toponym)“ ersetzt werden kann. Diese beiden Auffassungen widersprechen sich zu 100%.
Wenn Markert & Nissim (2002) das Toponym durch ein weiteres Toponym (und die kontextuelle Ergänzung „Die Personen in ...“) ersetzen, geht das Toponym und dadurch auch der geographische Bezug nicht verloren, auch wenn es eigentlich metonymisch verwendet wurde. Da in dieser Arbeit die geographische Aussage und nicht die linguistische Korrektheit untersucht wird, werden solche, nach Markert & Nissim (2002) metonymischen Toponyme, als echte Toponyme aufgefasst. Dies gilt sowohl für die Toponym Recognition / Resolution als auch für die manuelle Annotation des Vergleichsdatensatzes

3.3.4 Normalisierung von Toponymen

Da ein einzelner Ort durch verschiedene Toponyme referenziert werden kann (z.B. „New York City“, „Big Apple“, „NYC“, siehe Kapitel 2.2.3), werden diese Toponyme manchmal durch das originale Toponym („New York City“) ersetzt (Leidner 2008). Dieses Problem kann aber auch gelöst werden, indem diese Toponyme sowohl in die Gazetteers als auch in die Datenbank (welche für die Toponym Resolution verwendet wird) zusätzlich aufgenommen werden.

Um die Datenbank nicht manuell erweitern zu müssen, werden solche Toponyme in die „*largestcities.lst*“-Datei geschrieben. Diese, beim Gazetteer-Lookup als „City“ getaggten Toponyme werden nicht in der lokal gespeicherten Datenbank, sondern im, von Geonames zur Verfügung gestellten, Webservice nachgeschaut.

Von Pouliquen et al. (2006) werden bei der Toponym Normalisierung Adjektive mit dem Toponym gleichgesetzt. So würde das Wort „Zürcher“ durch das Toponym „Zürich“ ersetzt. Um den Aufwand, welche diese Implementierung mit sich bringen würde zu verhindern, werden solche Adjektive nicht als Toponyme aufgefasst. Ein weiterer Grund dafür ist, dass es im Deutschen für verschiedene Orte verschiedene Adjektiv-Formen gibt. So wird aus „Zürich“ „Zürcher“, aus „La Chaux-de-Fonds“ „La Chaux-de-Fonnier“, aus „Lugano“ „Luganesi“ aus „Nice“ „Nicoise“ und aus „Mailand“ „Mailänder“. Diese verschiedenen Formen hätten zur Folge, dass entweder jedem Toponym manuell das entsprechende Adjektiv zugewiesen werden müsste oder es für jedes Toponym mindestens 5 verschiedene Adjektive gäbe, wovon jeweils nur eines richtig wäre. Die erste Variante ist sehr zeitaufwändig und die Zweite erhöht sowohl die Gazetteergrösse (um den Faktor 5) als auch den Prozessierungsaufwand. Zudem würde die Geo-NonGeo-Ambiguität mit der zweiten Methode erhöht und somit der Prozessierungsaufwand nochmals steigen. Weiter ist nicht klar, ob als Adjektive verwendete Toponyme auch als Toponyme verstanden werden sollen, so ist daran zu zweifeln, ob „Schweizer Käse“ mit dem Toponym „Schweiz“ gleichzusetzen ist.

Aufgrund dieser Feststellungen, welche stark am Nutzen von Adjektivformen zweifeln lassen, werden sie in dieser Arbeit nicht durch die entsprechenden Toponyme ersetzt.

3.3.5 Verwendung des „One Referent per Discourse“-Prinzips

Da laut Gale et al. (1992) in 98% der Fälle nur ein Sinn pro Wort in einem Text verwendet wird und diese Heuristik von praktisch allen bisherigen GIR-Untersuchungen verwendet wurde, wird sie auch hier angewandt. Dies sowohl für die Unterscheidung von Geo-NonGeo-Ambiguität, als auch von Geo-Geo-Ambiguität.

3.3.6 Verwendung von handgefertigten Regeln

Beim Gazetteer Lookup werden zwischen Toponymen, welche klar Toponyme sind und solchen, welche auch andere Wörter sein können, unterschieden. Mögliche Kandidaten für Toponyme sind neben den „Location“-, „GLocation“-, „Land“-, „Kanton“- und „AmbKanton“-Tags auch die Tags „Ambvorname“ und „Nomen“.
Um zu prüfen, ob diese Tags Toponyme sind, werden folgende handgefertigten Regeln angewendet:

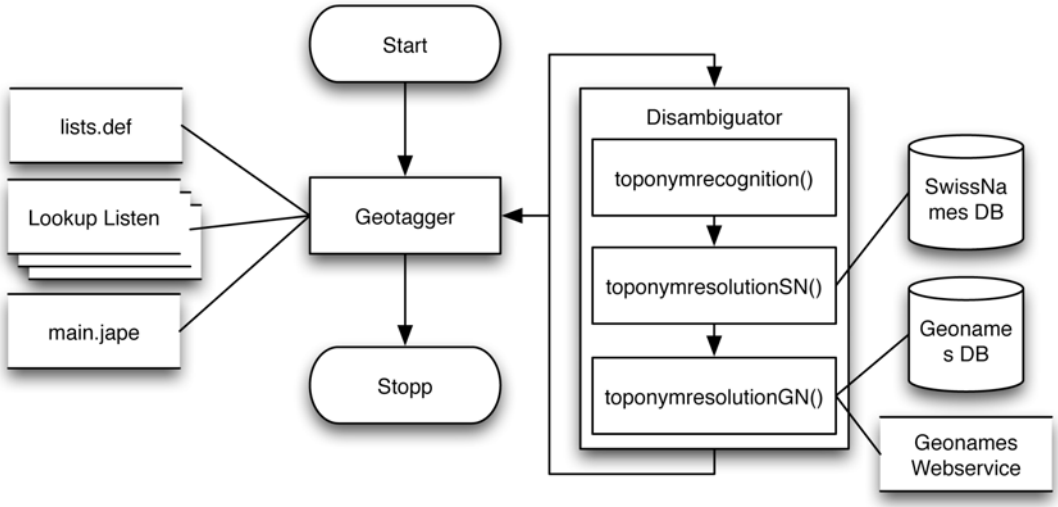
- <Ambvorname> <“weiteres Tag“> => kein Toponym, sondern Name
- <Nomen> sind nur dann Toponyme, wenn im Text einmal die Typangabe gemacht wurde (ORPD) oder wenn vorher „in“ steht.

Es werden aber auch die eindeutigen Toponyme darauf untersucht, ob sie nicht als Personennamen verwendet werden. Dies wird gemacht, da Toponyme oft auch als Nachnamen verwendet werden können, so z.B. „George Washington“. Die Regel dafür ist, dass die Toponyme nicht als solche annotiert werden, falls ein Vorname vorangestellt ist.

3.3.7 Implementierung der Toponym Recognition in der Disambiguator-Klasse

Wie in Kapitel 3.3.2 bereits beschrieben, wurde die von Paul Clough und Robert Pasley programmierte Wrapperklasse (Geotagger.java) verwendet um damit die Gazetteer-Funktion von GATE auszuführen. Diese Klasse wurde nur minimal und den Beschreibungen aus Kapitel 3.3.2 entsprechend angepasst. Für genauere Informationen sei auf Kapitel 3.3.2 sowie Clough (2005) verwiesen.
Da jedoch nur die Gazetteerfunktion von GATE verwendet wurde, müssen für die Toponym Recognition noch die einzelnen Tags in Toponyme und „nicht-Toponyme“ unterschieden werden. Dazu wird aus Geotagger.java das XML-File an die Disambiguator.java-Klasse weitergegeben. Diese wurde vollständig neu entwickelt und beinhaltet neben der Unterscheidung besagter Toponyme auch die komplette

Toponym Resolution. Die Geotagger.java-Klasse übergibt der *toponymrecognition()*-Methode eine XML-Datei, welche die Annotationen aus allen Lookup-Listen (in *lists.def* definiert) enthält.



Figur 4: Einbindung der Toponym Recognition und Toponym Resolution in Geotagger.java

Um aus den Tags die Toponyme zu extrahieren wird das XML-Dokument gemäss Figur 5 durchsucht. Die einzelnen Wörter werden jeweils in generischen Vector-Objekten der Klasse String gespeichert und die darauf angewandten Regeln sind fest eingebaut („hard-coded“). Somit werden in der Geo-NonGeo-Disambiguation sowohl Toponyme, welche auch Personennamen sein können, als auch Toponyme, welche keine Named Entities beschreiben speziell unterschieden.

Als Return-Wert der *toponymrecognition()*-Methode erhält man wiederum eine XML-Datei, welche nebst dem ursprünglichen Text aus <Location>-, <GLocation>- (für Toponyme aus Geonames) sowie <Kantonabbrev>- und <Loctype>-Tags besteht.

1.1.1 Beispiel zur Toponym Recognition

Fiktiver Beispieltext

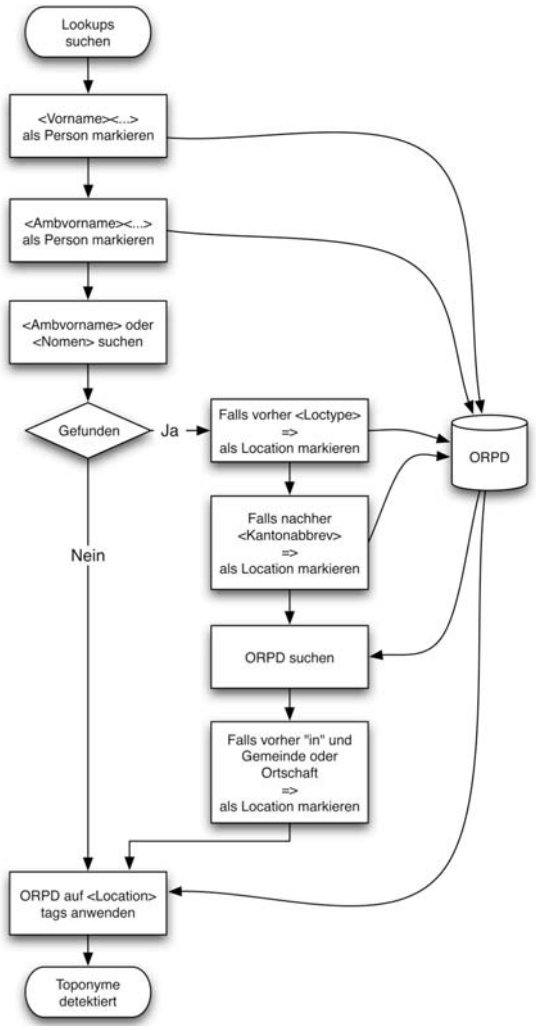
„Eduard Caminada aus holte sich dieses Jahr den Titel im Kirschsteinspucken. Von Anfang an dominierte er den Wettkampf, welcher in der Stadt Zug stattfand. Caminada gilt interkantonal als...“

Annotationen nach Gazetteer Lookup
„<Vorname>Eduard</Vorname> <Location>Caminada</Location> holte sich... ..den Wettkampf in der <Loctype>Stadt</Loctype> <Nomen>Zug</Nomen> stattfand. <Location>Caminada</Location> ...“

Vorgehen nach Flussdiagramm

1. Aufgrund des Vornamens „Eduard“ wird das Toponym „Caminada“ als Nachnamen aufgefasst und in der ORPD-Datenbank gespeichert.
2. Es wird versucht, die Geo-NonGeo-Ambiguität der <Nomen>-Annotation von „Zug“ anhand eines Toponymbeschriebs aufzulösen. Dieser wird in „Stadt“ gefunden, und damit „Zug“ als Toponym aufgefasst.
3. Nun wird das One Referent per Discourse-Prinzip auf die restlichen Annotationen (welche Toponyme sein könnten, nämlich „Location“, „Nomen“, „Ambvorname“, „Ambkanton“, „Kanton“ und „City“) angewandt. Dadurch wird auch das andere „Caminada“ wiederum nicht als Toponym aufgefasst.

Annotationen nach Toponym Recognition
„... in der <Loctype>Stadt</Loctype> <Location>Zug</Location> ...“



Figur 5: Flussdiagramm der Toponym Recognition

3.4 Toponym Resolution

Für die Toponym Resolution gibt es unzählige Methoden, wovon die meisten in Kapitel 2.3 vorgestellt wurden. Aufgrund der zur Verfügung stehenden Daten, müssen diese Methoden auf ihre Eignung überprüft werden.

Die Daten in SwissNames, sowie auch diejenigen aus Geonames, enthalten hierarchische Informationen. Weiter sind in beiden Gazetteers Populationsdaten (diejenigen in SwissNames sind diskretisiert) vorhanden. Diese beiden Gazetteer-Eigenschaften bieten sich für die Toponym Resolution an. Da die Resultate Aufschluss über die Anwendbarkeit von räumlichen Toponym Resolution-Methoden geben sollen, fallen diese als mögliche Kandidaten weg.

3.4.1 Verwendung von hierarchischer Information

In SwissNames ist zu jedem Toponym dessen übergeordnete Gemeinde sowie Kanton gespeichert. Aufgrund der Konsistenz der Daten in SwissNames kann diese administrative Hierarchie als vollständig betrachtet werden. In Geonames sind zwar hierarchische Informationen abgespeichert, diese variieren jedoch von Natur aus von Land zu Land. Daher ist die einzige konsistente Hierarchiestufe diejenige der Länder.

Hierarchische Disambiguationsstrategien wurden oft und mit guten Ergebnissen angewendet (Kapitel 2.3.2). Hier wird das Auftreten von hierarchisch höher liegenden Toponymen verwendet.

3.4.2 Verwendung des „One Referent Per Discourse“-Prinzips

Das schon in Kapitel 3.3.5 beschriebene Prinzip des „One Referent Per Discourse“ von Gale et al. (1992) auch für die Toponym Resolution verwendet.

3.4.3 Verwendung von Toponym-Typen

Linguistisch kann ein mehrdeutiges Toponym durch die Typangabe aufgelöst werden. So wird das Toponym „Luzern“, welches den Kanton beschreiben soll, oftmals von dieser Typangabe begleitet („Kanton Luzern“). Dasselbe kann auch für die Stadt Luzern gemacht werden. Um diese linguistische Hilfe zu nutzen, werden Typen von Toponymen in den fünf vorangehenden Worten gesucht. Diese Typen umfassen die Wörter See, Fluss, Bach, Stadt, Ortschaft und Gemeinde. Damit kann zwischen Gewässern und Ortschaften unterschieden werden, aber auch zwischen Ortschaften und Gemeinden. Als Städte bezeichnete Toponyme werden in dieser Arbeit nur unter

den SwissNames-Typen „*Gemeinde“ gesucht, da dem Autor keine Stadt bekannt ist, welche nicht auch eine Gemeinde ist. Beim Gazetteer Lookup werden Kantone, welche auch Ortschaften sein können speziell gekennzeichnet (<Ambkanton>). Diese sind in einer separaten Liste gespeichert.

3.4.4 Verwendung von „Default Referents“

Default Sense ist eine häufig verwendete Heuristik und soll hier, wie in den meisten bisherigen Arbeiten, als letzte Chance für die Geo-Geo-Disambiguation verwendet werden. Dabei ist sie jedoch nicht zu unterschätzen, denn in vielen Fällen, in denen keine textliche Disambiguation vorgenommen werden kann, ist davon auszugehen, dass der bekannteste Ort gemeint ist (Kapitel 2.3.5).

Um einen Default Referent zu bestimmen wurden bisher meist Populationsdaten verwendet (Ferrés 2007). Es ist jedoch anerkannt, dass der Default Referent für ein Toponym nicht nur von der Populationsgrösse abhängt. So ist das Toponym „San Bernardino“ in der Schweiz entweder mit dem Pass oder der Ortschaft im Kanton Tessin verknüpft, in den USA jedoch mit dessen grossflächigstem „County“. Dies zeigt, dass mehrere Faktoren in den Gebrauch von Default Referenten einfliessen. Da aber bisher noch keine Taxonomie dazu besteht, wird auch in dieser Arbeit primär die Populationsgrösse als Indikator für den Default Referenten verwendet. Da es sich aber beim Evaluationskorpus um die schweizer Zeitung „Südostschweiz“ handelt, werden schon bei der Toponym Recognition grundsätzlich zuerst Toponyme im SwissNames-Gazetteer gesucht, bevor die weltweiten Toponyme überhaupt erst in Betracht gezogen werden.

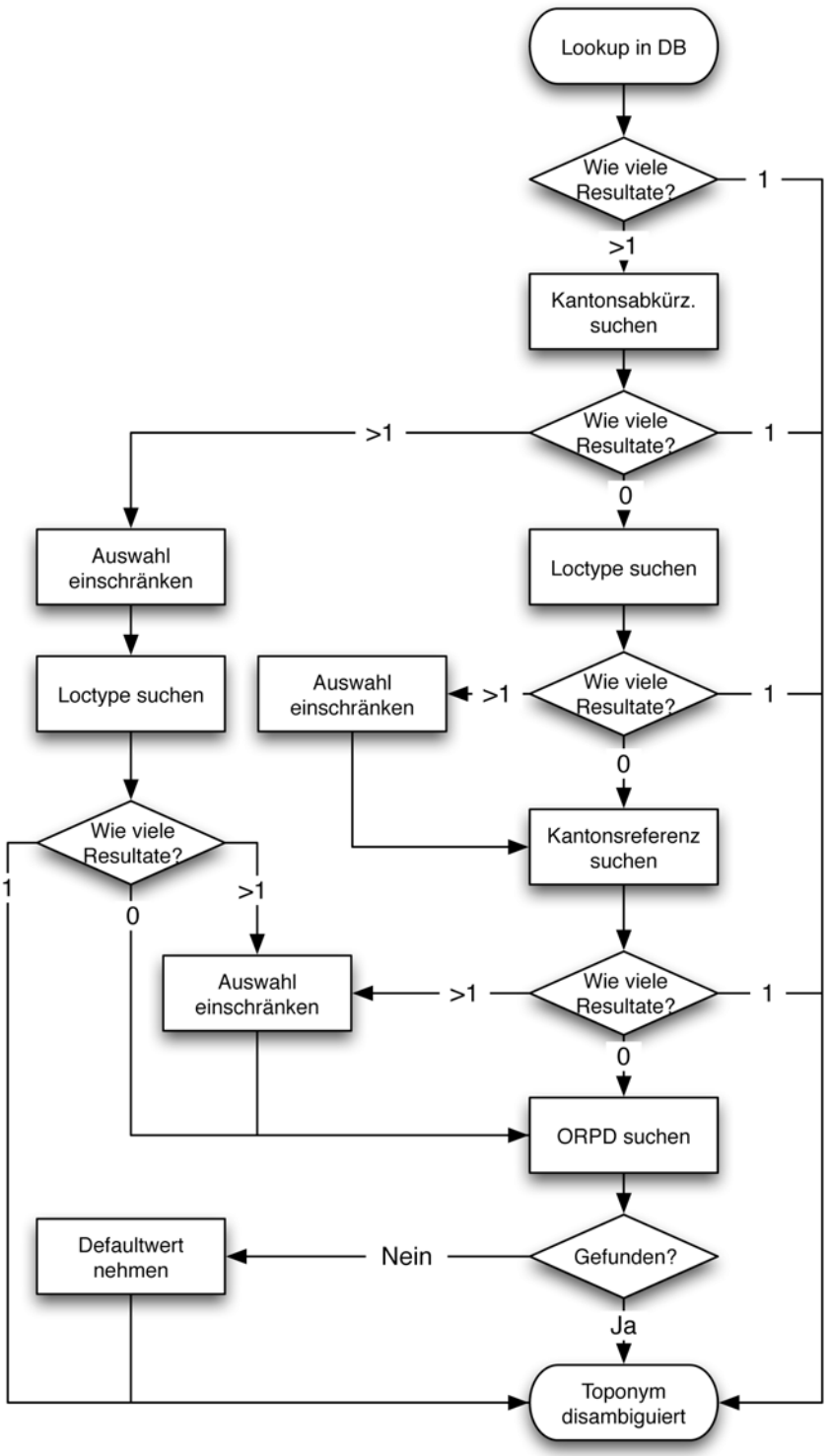
Da in SwissNames keine direkten Populationszahlen zur Verfügung stehen, werden die Typdeklarationen aus Tabelle 3 verwendet. Dabei werden zuerst die Gemeinden betrachtet und erst nachher die Ortschaften.

In Geonames sind Populationsdaten vorhanden, jedoch sind sie über den Webservice nicht abrufbar. Bei Toponymen, welche in der lokalen Datenbank gefunden werden, ist diese Information vorhanden und wird für die Berechnung des Default Referenten verwendet. Fall der Webservice verwendet wird, wird der von Geonames zuerst angegebene Referent verwendet. Damit wird auf die Indexierung von Geonames zurückgegriffen, wobei dessen Prinzip nicht publiziert ist.

3.4.5 Implementierung der Toponym Resolution in der Disambiguator-Klasse

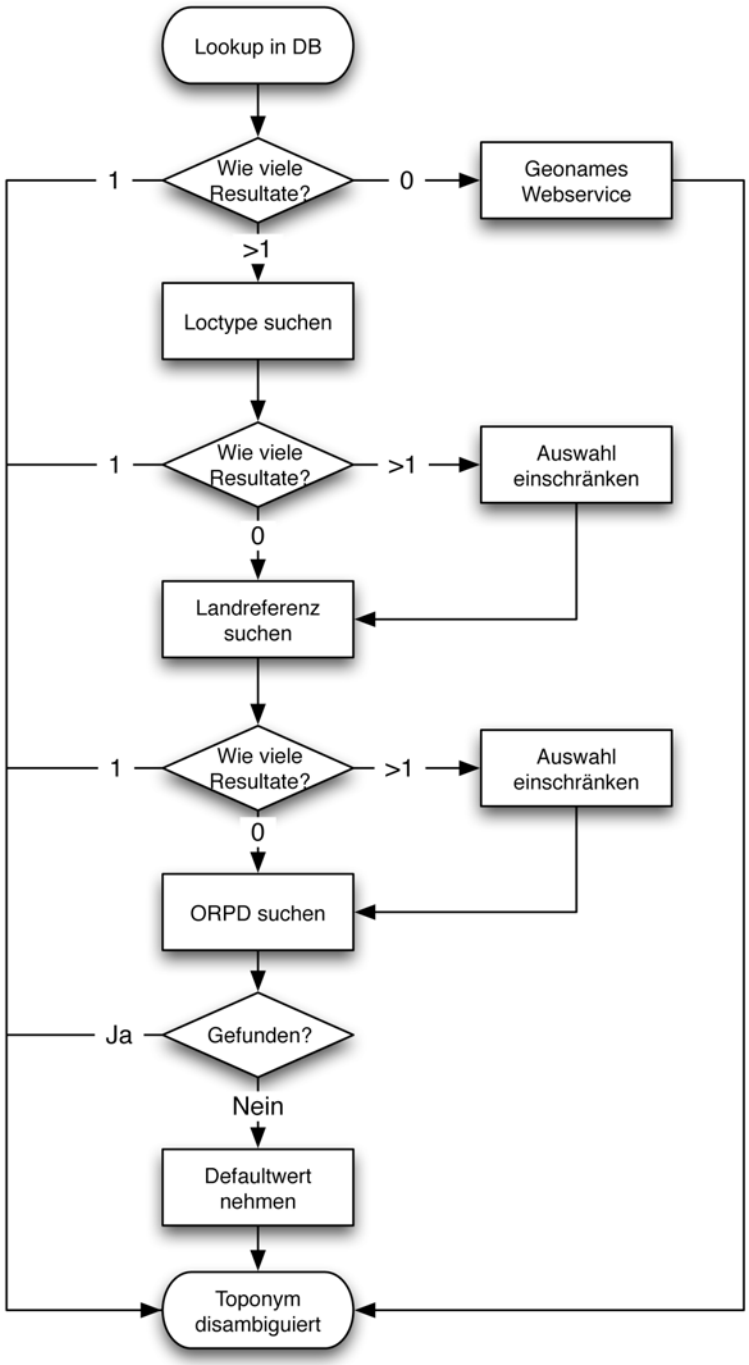
Da die zwei verwendeten Gazetteers nicht einheitlich in Aufbau und Struktur sind, wurden für die in SwissNames und Geonames gefundenen Toponyme separate Methoden geschrieben (Figur 4, 6 & 7). Dies hat den Vorteil, dass Eigenschaften wie Populationsgrösse (Geonames) oder Hierarchiestufen (SwissNames), welche nicht in beiden Gazetteers vorkommen, trotzdem genützt werden können. Beim Gazetteer-Lookup werden die Toponyme entsprechend der Gazetteer-Quelle gekennzeichnet, so sind Toponyme, welche Einträge in Swissnames beschreiben als <Location> gekennzeichnet während Toponyme aus Geonames als <GLocation> gekennzeichnet sind. In der *toponymresolutionSN()* werden also nur <Location>-Tags auf Geo-Geo-Ambiguität untersucht und in der *toponymresolutionGN()* nur <GLocation>-Tags.
Wie schon in der *toponymrecognition()*-Methode wird auch für die Toponym Resolution ausgiebig von generischen Vector-Objekten als Datenstruktur Gebrauch gemacht.

Bevor die SwissNames-Annotationen der Figur 6 zufolge untersucht werden, wird das Dokument auf Kantone untersucht. Dies um nachher eine Hierarchie-Heuristik anwenden zu können.
Wie in Figur 6 dargestellt, wird zuerst die Schweiz-übliche Disambiguation anhand des Kantonkürzels versucht. Falls diese mehr als ein Resultat zurückliefert (z.B. Sils (GR)), wird die Hierarchie-Heuristik, da diese dem Kantonkürzel inhärent ist, weggelassen und der „Loctype“ gesucht (Kapitel 3.4.3). Falls das Toponym immer noch nicht eindeutig disambiguiert ist, wird es zuerst im ORPD-Vector (Kapitel 3.4.2) gesucht und nachher der Defaultreferent (Kapitel 3.4.4) berechnet.
Falls kein Kantonkürzel gefunden wird, wird zuerst ein „Loctype“ gesucht und danach geschaut, ob bereits ein gültiger Kanton für dieses Toponym gefunden wurde, was der Hierarchie-Heuristik entspricht. Falls dies noch keinen eindeutigen Referenten hervorgebracht hat, wird wiederum zuerst im ORPD-Vector gesucht bevor schliesslich der Defaultreferent berechnet wird.



Figur 6: Flussdiagramm der Toponym Resolution für SwissNames-Toponyme

Da Geonames keine einheitliche Hierarchie für alle Länder zur Verfügung stellt, wird die Hierarchie-Heuristik bei der Toponym Resolution für <GLocation>-Tags nur auf der Ebene des Landes durchgeführt. Dabei werden wiederum zuerst alle Länder im Text gesucht und in einem Vector-Objekt abgespeichert.



Figur 7: Flussdiagramm der Toponym Resolution für Geonames-Toponyme

Im Gegensatz zu *toponymresolutionSN()* ist der Ablauf von *toponymresoltuionGN()* der Figur 7 entsprechend linear. Nachdem nach einem „Loctype“ gesucht worden ist, wird die Hierarchie-Heuristik angewandt, danach im ORPD gesucht und schliesslich der Defaultsense berechnet. Diese Heuristiken werden so lange der Reihe nach gefeuert, bis ein eindeutiger Referent für das Toponym bestimmt werden kann.

3.5 Erstellung von manuellen Vergleichsdaten

Um das programmierte GIR-System zu überprüfen, wurden zufällig (basierend auf der java.util.Random-Klasse) 100 Dokumente aus dem Südostschweiz-Korpus extrahiert. Diese wurden manuell auf Toponyme untersucht und die entsprechenden Annotationen in GATE vorgenommen. Diese Auswahl unterlag keinen Kriterien in Bezug auf die Anzahl Toponyme pro Artikel, da dies ein oft kritisierter Punkt in der Bewertung von GIR-Systemen ist (Leidner 2008).
Zur Vereinfachung der manuellen Arbeit wurde ein Such-Interface basierend auf den beiden Gazetteer-Datenbanken sowie dem Geonames-Webservice programmiert. Wie in Kapitel 3.3.3 beschrieben, werden metonymisch verwendete Toponyme als Toponyme aufgefasst. Als Adjektive verwendete Toponyme werden jedoch nicht als Toponyme klassifiziert.

3.6 Berechnung der räumlichen Verteilung der mehrdeutigen Toponyme

Um die erste Forschungsfrage zu beantworten, muss die räumliche Verteilung der einzelnen Referenten der mehrdeutigen Toponyme untersucht werden. Um dies statistisch zu untersuchen wurden nicht der Median und die Standardabweichung der Referenten genommen, sondern alle möglichen Distanzen zwischen den Referenten eines Toponyms, welche dann in einer Häufigkeitskurve dargestellt wurden. Diese Methode hat gegenüber gemittelten Beschreibungen den Vorteil, dass die häufiger auftretenden Toponyme auch dementsprechend gewichtet werden. Weiter gehen dadurch auch keine Extremalwerte verloren.

Als Vergleichsverteilung wurden zufällig Einträge aus demselben Gazetteer genommen und mit diesen dieselben Berechnungen durchgeführt. Die Gruppen der zufällig ausgewählten Toponyme haben eine Grösse von 4, wobei Tests durchgeführt wurden, welche aussagen, dass die Gruppengrösse keinen Einfluss auf die Häufigkeit der unterschiedlichen Distanzen hat.
Die Berechnungen für diese Häufigkeitsverteilungen wurden in Java programmiert und zu Visualisierungszwecken anhand von Excel Charts dargestellt. In der Java-Applikation werden alle Distanzen zwischen den Referenten eines Toponyms berechnet und gespeichert. Diese Distanzen werden dann in 10 km (oder beliebig

grosse) Intervalle diskretisiert. Schliesslich wird die Häufigkeit jeder Distanz (in den angegebenen Diskretisierungs-Schritten) berechnet und in einem, von Excel importierbaren Textfile ausgegeben.

3.7 Berechnung des Scopes eines Zeitungsartikels

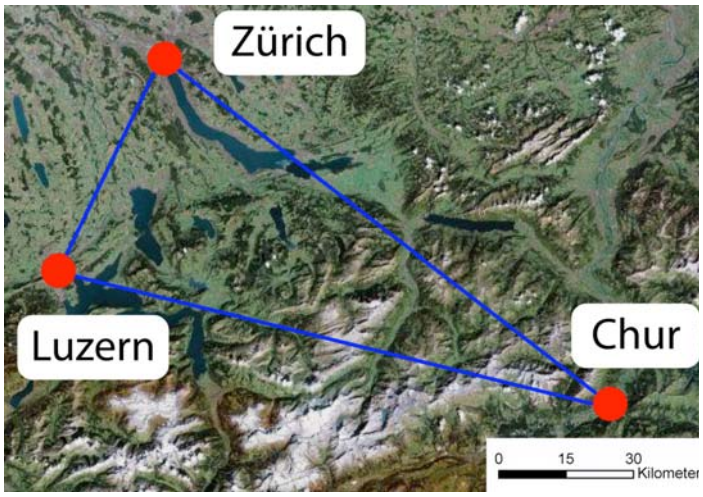
Um die Forschungsfragen 2 und 3 beantworten zu können, muss die geographische Ausdehnung von Zeitungsartikeln berechnet und dann mit der räumlichen Verteilung von mehrdeutigen Toponymen verglichen werden können. Dazu bietet sich sowohl die Analyse einzelner Artikel, aber auch die gesamtheitliche, statistische Auswertung aller Artikel an. Um eine allgemeingültige Aussage machen zu können, darf diese jedoch nicht spezifisch auf einzelne Artikel ausgerichtet sein. Daher wurde die geographische Ausdehnung (auch Scope genannt) aller Artikel berechnet. Dies wurde analog zur Berechnung der räumlichen Verteilung von mehrdeutigen Toponymen gemacht (Kapitel 3.6). Der Scope eines Zeitungsartikels ist also als

„die Summe aller Distanzen zwischen den, im selben Artikel auftretenden Toponymen“

definiert. Figur 8 illustriert das Prinzip dieses Scopes. Dabei wurden mehrfach vorkommende Toponyme nur einmal verwendet, um Distanzen von 0 m zu vermeiden. Es kann jedoch argumentiert werden, dass die mehrfach vorkommenden Toponyme auch dementsprechend gewichtet werden müssen. Im Kontext der Forschungsfrage 3 ist eine solche Gewichtung jedoch nicht sinnvoll, da sie für die bisher angewandten räumlichen Disambiguationsstrategien schlecht verwendet werden kann. Die einzige Methode, welche eine Gewichtung der Toponyme vornimmt, ist die von Smith & Crane (2001). Den Argumenten von Kapitel 2.8.1 zufolge ist diese Methode jedoch wenig sinnvoll und nur auf internationale Ambiguität verwendbar.

Bei der Prozessierung der Zeitungsartikel wurden nur SwissNames-Einträge genommen, damit die Vergleichbarkeit mit den Ergebnissen der Ambiguitäts-Analyse von SwissNames gewährleistet ist. Weiter ist die räumliche Verteilung der Toponyme in Geonames zu stark auf die Region um Zürich konzentriert (siehe auch Figur 24), was die Ergebnisse zusätzlich beeinflussen würde.

Toponyme	Distanz
Chur-Luzern	96.2 km
Chur-Zürich	95.2 km
Luzern-Zürich	41.6 km



Figur 8: Beispielillustration des Scopes eines Zeitungsartikels welcher die Toponyme „Chur“, „Luzern“ sowie „Zürich“ beinhaltet.

3.8 Visualisierung der Resultate

Um die räumliche Komponente der erkannten Toponyme aus den Artikeln der Südostschweiz weiter zu verwenden, wurden verschiedene Karten in ESRI[36] ArcGIS erstellt. Beim räumlichen Bezug der Toponyme handelt es sich um einzelne Koordinaten, was Einfluss auf die Darstellungsmöglichkeiten hat. So enthält die Südostschweiz des Jahres 2006 knapp 1.5 Millionen Toponyme, weshalb eine simple Punkt-Darstellung sinnlos wäre. Da die einzelnen Toponyme keine Gewichtung haben, bietet sich die Kernel-Density-Methode an. Diese berechnet für jede Rasterzelle der Karte die Dichte der Punkte innerhalb eines bestimmten Suchradius, wobei diese aufgrund einer Gauss'schen Verteilung gewichtet werden (O'Sullivan & Unwin 2003). ArcGis 9.2 bietet neben dem Import von ASCII-Daten (*ASCII to Raster3D*) auch die Berechnung von Dichteoberflächen (*Kernel Density*) sowie deren Visualisierung an.

Auch für die Visualisierung der Einträge in den Gazetteers eignet sich eine Darstellung als Dichteoberfläche analog zu Figur 3. Damit kann eine schnelle Aussage über die Verteilung der Toponyme in den Gazetteers und deren Auswirkung auf weitere räumliche Auswertungen gemacht werden.

[36] http://www.esri.com

3.9 Berechnung von Box-Plots

Um den Zusammenhang zwischen räumlicher Ausdehnung eines Zeitungsartikels und einzelnen, darin enthaltenen Toponymen herzustellen, eignet sich die Berechnung eines Boxplots für die einzelnen Toponyme. Dabei wird die Distanz eines Toponyms zu allen anderen, im Text vorkommenden Toponymen berechnet und nachher die Werte Median, Maximum und Minimum sowie das erste und letzte Quartil ermittelt (Emerson & Strenio 1983). Die Daten wurden mit Java-Code aus den Südostschweizartikeln extrahiert und die Boxplots anhand des Statistikprogrammes SPSS[37] berechnet und dargestellt.

[37] http://www.spss.com

4 Resultate und Interpretationen

4.1 Evaluation des Geotaggers

GATE bietet neben den Möglichkeiten zur Spracherkennung auch die Möglichkeit, die erhaltenen Resultate anhand eines manuell erstellten Vergleichsdatensatzes (Kapitel 3.5) zu überprüfen. Dieses, AnnotationDIFF genannte, Modul berechnet die IR-üblichen Werte Precision, Recall und F-Wert (Kapitel 2.7).

Der Vergleichsdatensatz enthielt 817 Toponyme, wovon 761 korrekt erkannt wurden und auch der korrekte Referent zugewiesen werden konnte. Bei 9 Toponymen, welche zwar erkannt werden konnten, würde ein falscher Referent zugewiesen, womit der Fehler in der Toponym-Resolution zu suchen ist. 47 Toponyme wurden gar nicht erkannt, und 105 Mal wurde fälschlicherweise ein Toponym erkannt. Diese Ergebnisse resultieren in folgenden Benchmark-Werten:

Recall: 0,9303
Precision: 0,8777
F-measure: 0,9033

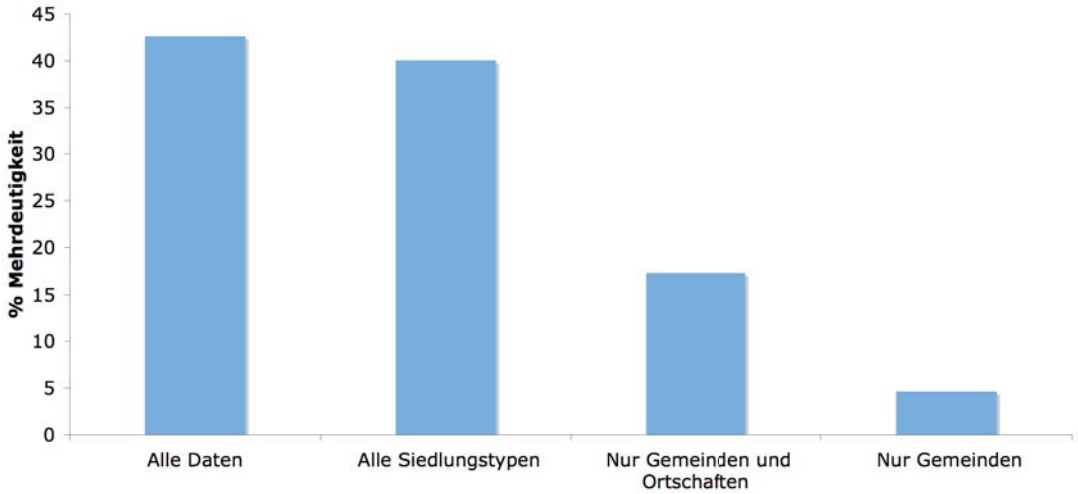
Dabei ist ein leicht höherer Recall als Precision zu beobachten. Viele der 105 „false positives“ könnten durch umfangreichere *nomen.lst*- und *ambvornamen.lst*-Listen verhindert werden. So ist z.B. „Füssen“ eine Ortschaft im deutschen Bundesland Bayern und „Caminada“ - ein typischer Bündner Nachname – auch eine Ortschaft im Tessin.

Die nicht gefundenen Toponyme waren meist gar nicht im Gazetteer vorhanden (z.B. Grossbritannien, England, Lake Louise). Manche Toponyme wurden auch aufgrund der *nomen.lst*- oder *ambvornamen.lst*-Listen nicht erkannt. Das prominenteste Opfer der *ambvornamen.lst*-Liste ist sicherlich „Paris“ (aufgrund des Namens „Paris Hilton“), welches im Vergleichsdatensatz nie erkannt wurde. Dabei ist darauf hinzuweisen, dass diese Listen keine Stopword-Listen sind, sondern nur als Indikatoren für Geo-NonGeo-Ambiguität verwendet werden. Im Falle von Paris war also nie ein klarer Hinweis vorhanden, welcher darauf hingedeutet hätte, dass es sich um ein Toponym handelt („Stadt“, “in“, vergleiche Kapitel 3.3.7). Bei den 9 Toponymen, welche einen falschen Referenten erhielten handelt es sich um 6

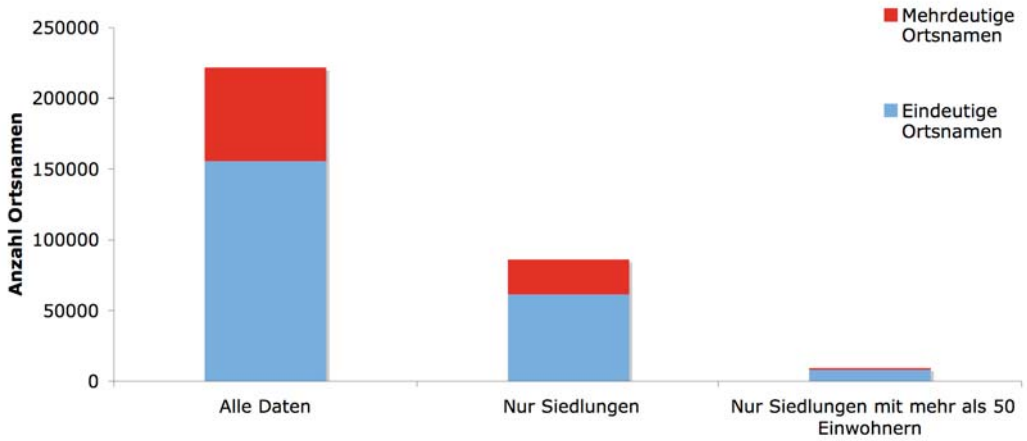
verschiedene Toponyme. Gossau wurde Dreimal sowie Rheinau und Landquart je Einmal dem falschen Referenten zugewiesen. Im Falle der Toponyme Aarhus, Aspen und Turin, wurde jeweils fälschlicherweise ein schweizer Referent dem Toponym zugewiesen.

4.2 Ambiguität in SwissNames

Die Daten von SwissNames wurden dem Prinzip von Kapitel 3.2 zufolge untersucht, wobei 42.62% aller Toponyme mehrdeutig waren. Da bisher jedoch praktisch nur weltweite Gazetteers mit einer geringeren Abdeckung untersucht wurden, wurden aus den Daten verschiedene Hierarchiestufen (siehe Figur 9) extrahiert und untersucht. Dabei wurde festgestellt, dass die Ambiguität von der Hierarchietiefe abhängt. So sank der Anteil an mehrdeutigen Toponymen bei der Reduktion von SwissNames auf Siedlungen mit mehr als 50 Einwohnern auf 17%. Diese Reduktion hat jedoch zur Folge, dass die Grösse des Gazetteers um 95% verkleinert wurde (Figur 10). Die Reduktion auf nur Gemeinden hatte zur Folge, dass nur noch 5% des Gazetteers mehrdeutig waren.

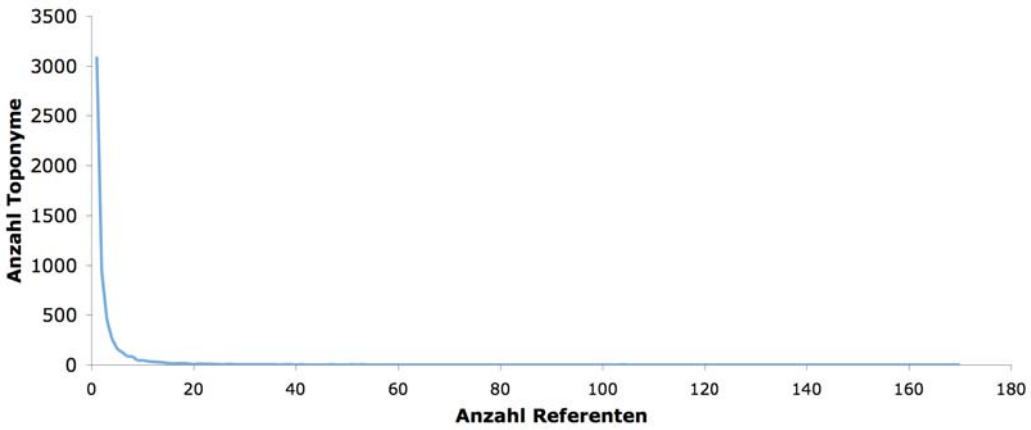


Figur 9: Prozentualer Anteil an mehrdeutigen Toponymen in SwissNames aufgeteilt auf verschiedene Hierarchiestufen.



Figur 10: Ambiguitätsgrad der Toponyme in SwissNames in Bezug zur Detailtreue.

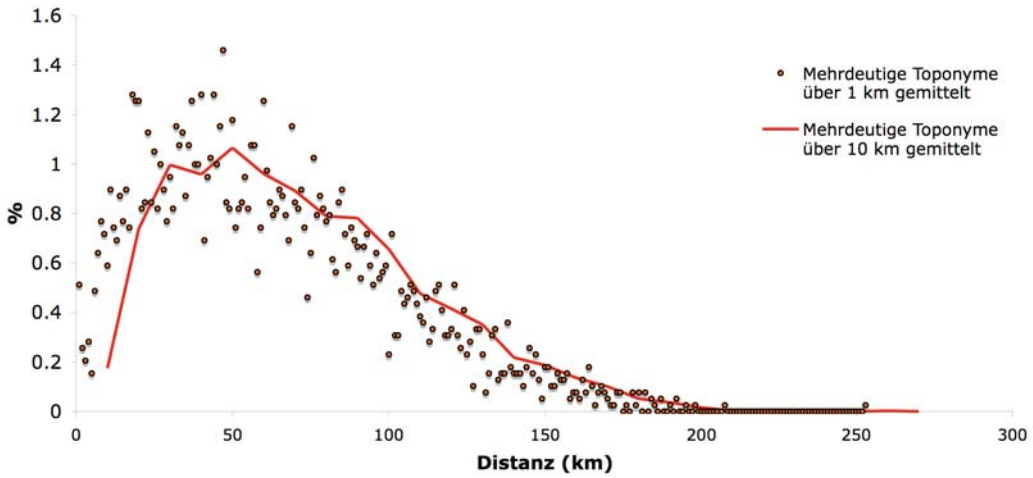
Eine Untersuchung in SwissNames ergab die folgende Grafik (Figur 11), welche die Anzahl Toponyme in Bezug zu deren Ambiguitätsgrad (Anzahl Referenten) darstellt. Dieses Verhältnis scheint umgekehrt-exponenziell zu fallen.



Figur 11: Anzahl Toponyme im Verhältnis zu deren Ambiguitätsgrad.

4.3 Räumliche Verteilung von mehrdeutigen Toponymen

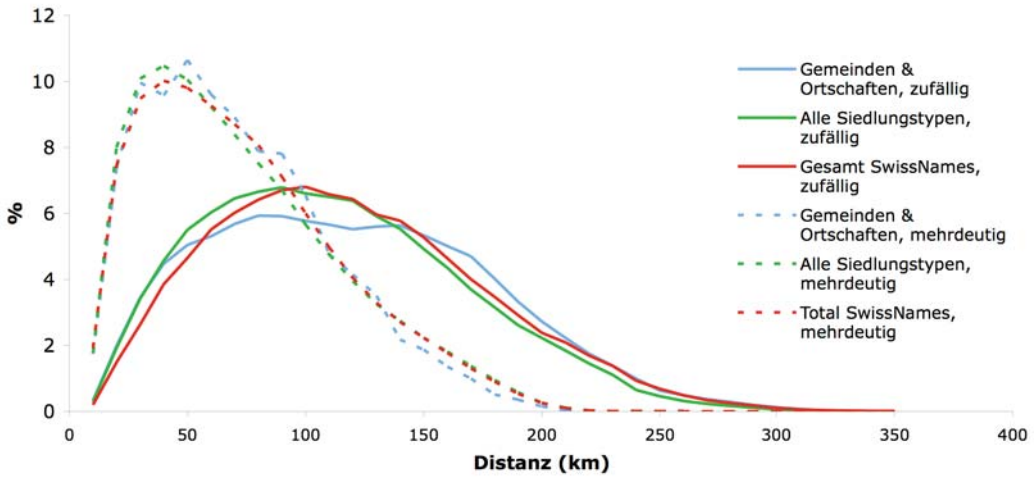
Die Verteilung der zufällig ausgewählten Toponyme ist unter anderem in Figur 13 dargestellt und verteilt sich um einen Median von 56 km. Dass diese Verteilung nicht einer Normalverteilung gleich ist (der Median ist einiges kleiner als die Hälfte des Maximums, siehe Tabelle 5), ist unter anderem (siehe Kapitel 5.2) mit der räumlichen Form der Schweiz zu begründen.



Figur 12: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen Toponymen in SwissNames (Gemeinden und Ortschaften).

Figur 12 zeigt, wie sich die Distanzen zwischen den mehrdeutigen Gemeinden- und Ortschaftsnamen verteilen. Auffällig ist dabei, dass es einige Referenten des selben Toponyms gibt, welche näher als ein km voneinander entfernt sind. Von diesen total 26 Ortschaften ist bei zwölf Ortschaften bei der Erstellung in SwissNames der Eintrag aus der 1:100'000- sowie auch derjenige aus der 1:25'000 Landeskarte aufgenommen worden. Die kleine Ortschaft „Oberdorf" wurde einfach zwei Mal aufgenommen und bei den Restlichen handelt es sich um Gemeinden, welche über Kantonsgrenzen hinausgehen (z.B. Erlinsbach (AG / SO)) oder um Ortschaften, welche gleich benannt sind wie benachbarte Gemeinden (z.B. Mur (VD / Vully (FR)).

Dieselben Berechnungen wurden dann mit allen Einträgen sowie allen Siedlungstypen von SwissNames gemacht. Diese unterscheiden sich zwar in der absoluten Anzahl der Einträge, beim relativen Vergleich ist die Verteilung dieser jedoch praktisch gleich wie die der Ortschafts- und Gemeindenamen (siehe Figur 13). Bei allen unterscheidet sich die Zufallsverteilung also stark von der Verteilung der mehrdeutigen Toponyme.



Figur 13: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten Toponymen aus SwissNames, unterteilt nach verschiedenen Toponymtypen.

Auch die, in Tabelle 5 dargestellten, statistischen Kennzahlen zur räumlichen Verteilung von Toponymen zeigen, dass sich die verschiedenen Hierarchiestufen „Sämtliche Daten“, „Siedlungstypen“ und „Gemeinden und Ortschaften“ kaum unterscheiden. Nebst der absoluten Anzahl Einträge ist einzige das Maximum der Gemeinden und Ortschaften anders als das der anderen beiden Hierarchiestufen. Eine mögliche Begründung hierfür ist, dass sich die Ortschaftsnamen zwischen den Sprachregionen stärker unterscheiden als Bezeichnungen von Weilern, Hotels oder Bauernhöfen.

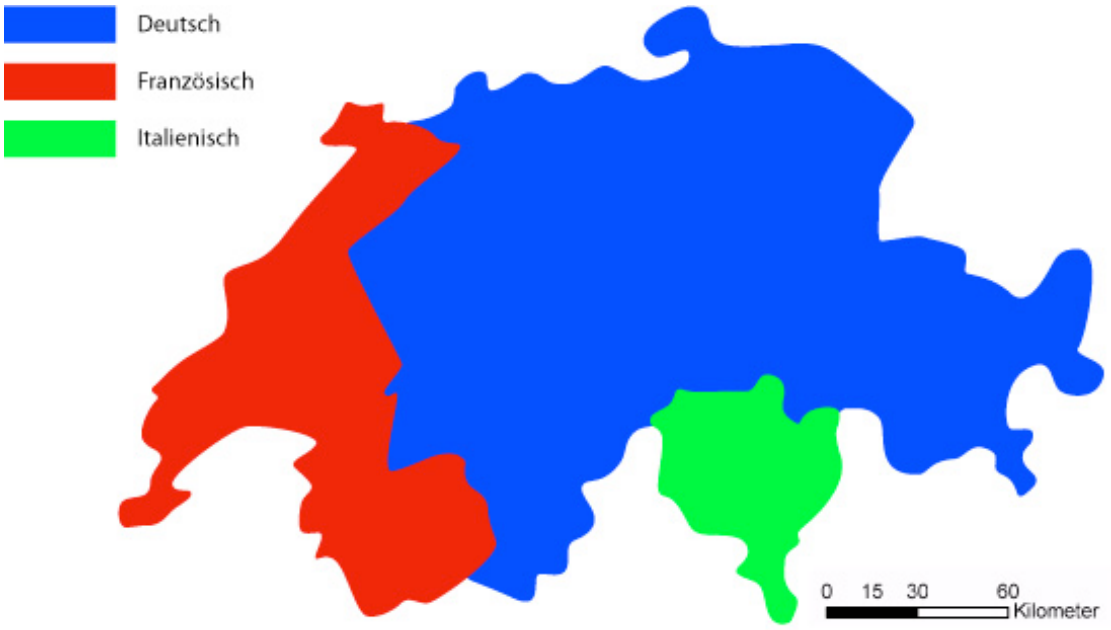
Mehrdeutige Ortsnamen	**Total**	**Min.**	**Max.**	**Median**	**Durchschnitt**
Alle Daten	736696	0 km	312 km	57 km	64 km
Alle Siedlungstypen	22920	0 km	311 km	56 km	64 km
Gem. und Ortsch.	3905	0 km	252 km	56 km	62 km

Vergleichsdaten (Stichprobe)	**Total**	**Min.**	**Max.**	**Median**	**Durchschnitt**
Alle Daten	150000	0 km	346 km	103 km	109 km
Alle Siedlungstypen	150000	0 km	343 km	98 km	103 km
Gem. und Ortsch.	150000	0 km	343 km	106 km	110 km

Tabelle 5: Statistische Angaben zur räumlichen Verteilung von Toponymen in Swissnames.

Eine mögliche Erklärung für den Unterschied zwischen der Zufallsverteilung und der Verteilung der Referenten von mehrdeutigen Toponymen wäre die räumlich

abgegrenzte Mehrsprachigkeit der Schweiz (siehe Figur 14). Da sich die Zufallsverteilung über die ganze Schweiz erstreckt, die mehrdeutigen Toponyme aber sprachbedingt innerhalb einer Sprachregion liegen können, wäre diese Diskrepanz durch die Sprachregionen erklärbar.



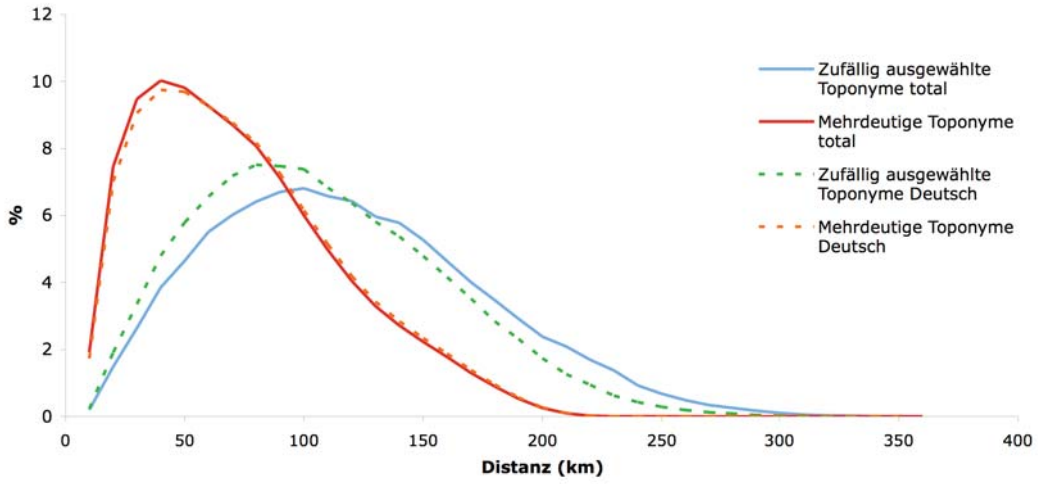
Figur 14: Sprachregionen in der Schweiz.

Um dies zu verifizieren wurde die bereits durchgeführte Statistik für die einzelnen Sprachregionen separat durchgeführt. Dabei wurden aufgrund des Schweizer Weltatlas (Spiess 2006) folgende Zuweisungen gemacht: Der italienischsprachigen Schweiz wurde der Kanton Tessin, der französischsprachigen Schweiz die Kantone Genf, Jura, Waadt, Neuchâtel, das westliche Wallis, der südwestliche Teil des Kantons Freiburg sowie der Berner Jura zugewiesen. Die restliche Schweiz wurde dem deutschsprachigen Raum zugewiesen. Dabei wurden alle SwissNames-Einträge berücksichtigt, also keine Einschränkung auf Ortschaften oder Gemeinden gemacht.

In allen drei Sprachregionen sind die mehrdeutigen Toponyme wiederum näher beieinander liegend als die zufälligen Toponyme. Die gesamten Einträge in SwissNames, aber auch die Sprachregionspezifischen wurden auf folgende Hypothese getestet:

µ0 : Verteilung der mehrdeutigen Ortsnamen = Verteilung der zufällig ausgewählten Ortsnamen

Diese konnte für alle Sprachregionen aber auch für die gesamten SwissNames-Einträge in einem Mann-Whitney-U-Test (Mann & Whitney 1947) bei einer Stichprobengrösse von je 100 Samples pro Gruppe klar abgelehnt werden. Dies war bei den deutschsprachigen sowie den gesamten SwissNames-Einträgen mit $p < 0.001$, bei den französischsprachigen Einträgen mit $p < 0.002$ und den Italienischsprachigen mit $p < 0.007$ der Fall. Für genauere Angaben zu den Tests sei auf den Anhang (8.1.1, 8.1.2, 8.1.3, 8.1.4) verwiesen.

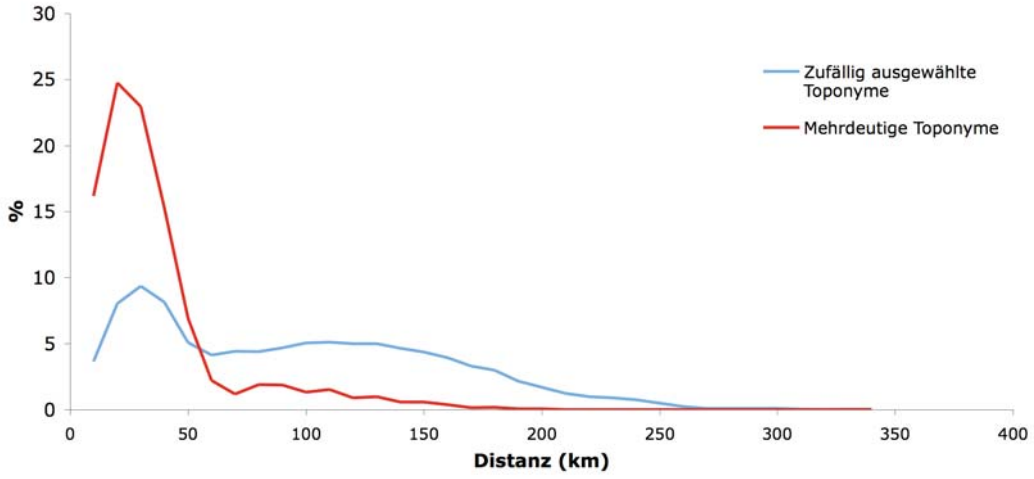


Figur 15: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten Toponymen aus der deutschsprachigen Region in SwissNames.

Um eine allgemeinere Aussage über die räumliche Verteilung von Toponymen machen zu können, müssen weitere Datenquellen sowie auch Daten über andere Länder betrachtet werden.

Um einen anderen Gazetteer zu untersuchen, wurden alle Schweizer Toponyme aus Geonames extrahiert und nach denselben Methoden untersucht. Trotz der Unterschiede zu der Verteilung in SwissNames ist das Ergebnis im Hinblick auf die erste Forschungsfrage vergleichbar. Bei den mehrdeutigen Toponymen ist die Konzentration der Distanzen im niedrigen Bereich wiederum viel stärker ausgeprägt als bei den Distanzen der Zufallsverteilung. Dies wird auch durch einen statistischen

Test mit einem Signifikanzniveau von 0.001 bei einer Stichprobengrösse von je 100 Samples pro Gruppe (Anhang 8.1.6) bestätigt.

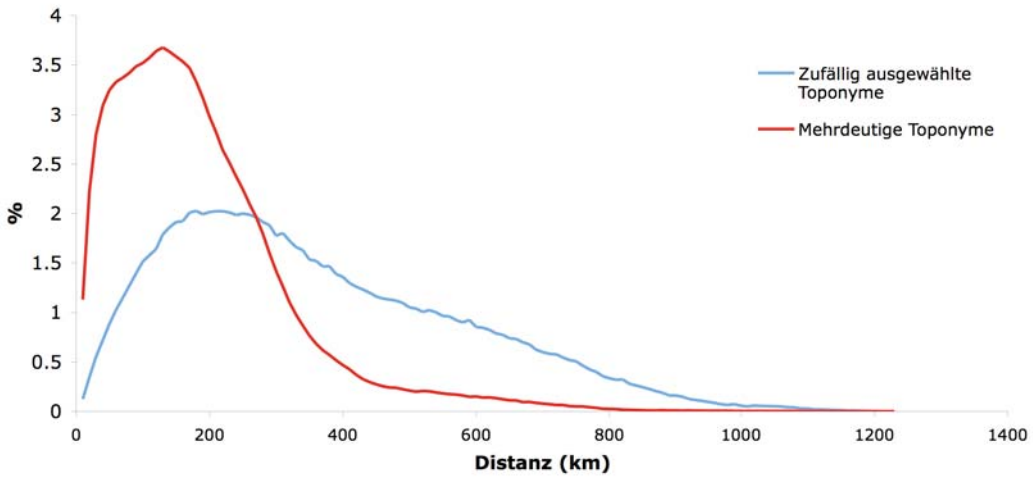


Figur 16: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten Schweizer Toponymen aus Geonames.

Um neben SwissNames eine weitere offizielle Datenquelle auszuwerten, wurden die Toponyme des 1:50'000 Scale Gazetteers[38] des britischen Ordnance Survey verwendet. Dieser Datensatz beschreibt alle Toponyme, welche auf der offiziellen 1:50'000 Karte Grossbritanniens zu finden sind. Es wurden zum Einen wieder alle Toponyme des Gazetteers ausgewertet, zusätzlich aber in einem zweiten Schritt die einzelnen Länder Schottland, Wales und England separat angeschaut (Figur 18, Anhang 8.2.1, 8.2.2).

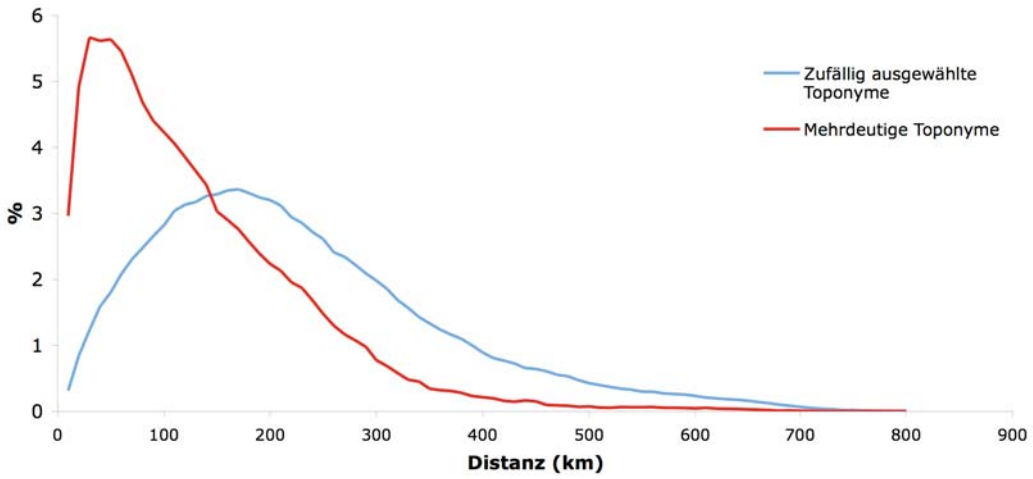
Wie schon bei SwissNames sind auch in den Ordnance Survey-Daten die mehrdeutigen Toponyme näher beieinander als die zufällig Ausgewählten (Figur 17). Am deutlichsten tritt dieses Phänomen auf, wenn nur die schottischen Toponyme betrachtet werden (Figur 20). Ein statistischer Test bei dem die gesamten Einträge im 1:50'000 Scale Gazetteer betrachtet wurden ergab einen signifikanten Unterschied bei einer Stichprobengrösse von je 100 Samples und $p < 0.001$ (Anhang 8.1.5).

[38] http://www.ordnancesurvey.co.uk/oswebsite/products/50kgazetteer/

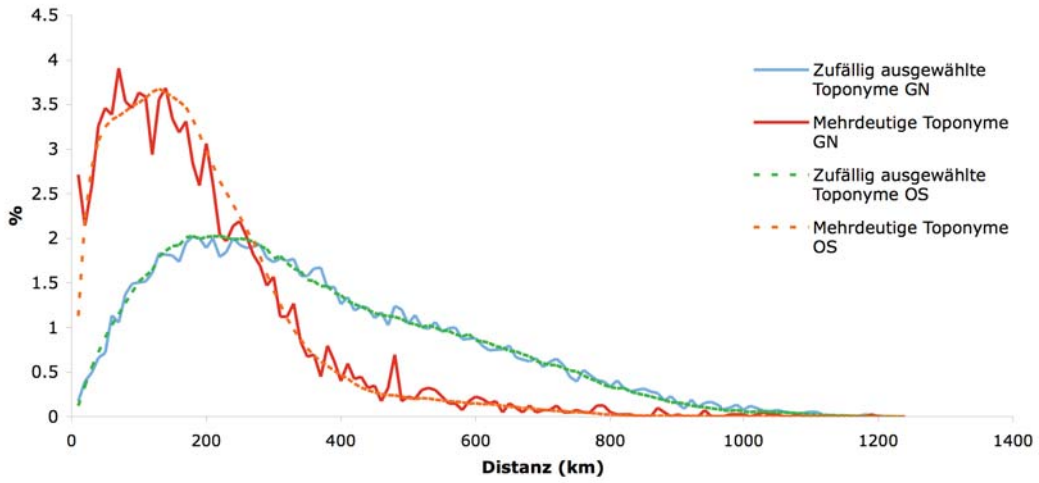


Figur 17: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten Toponymen aus dem 1:50'000 Scale Gazetteer des Ordnance Surveys.

Der Vergleich der offiziellen Datenquelle mit dem frei verfügbaren Geonames wurde auch für Grossbritannien gemacht. Auch hier sind die mehrdeutigen Toponyme näher beieinander liegend als die zufällig Ausgewählten und die Verteilung ist zudem praktisch identisch mit derjenigen des Ordnance Survey. Wiederum war ein statistischer Test mit je 100 Samples und $p < 0.001$ signifikant (Anhang 8.1.7).

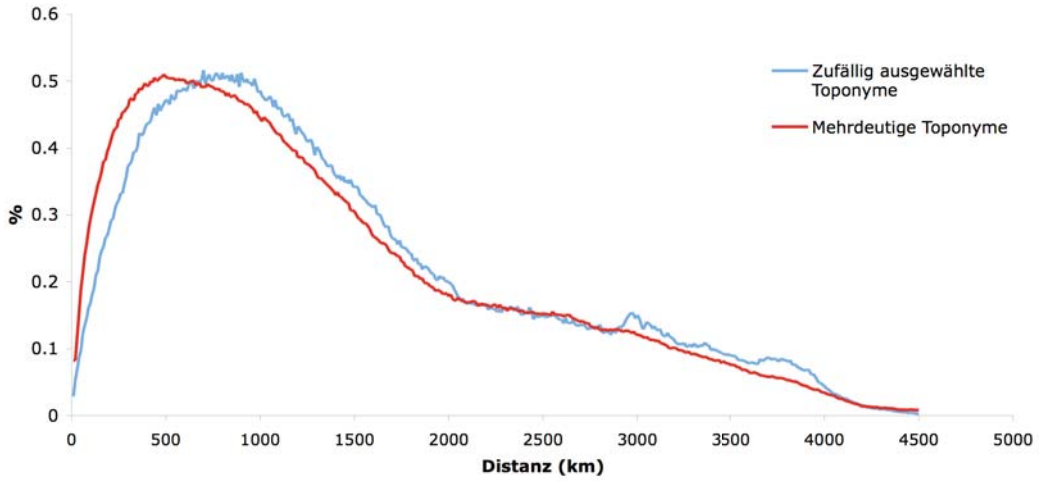


Figur 18: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten, Schottischen Toponymen aus dem 1:50'000 Scale Gazetteer des Ordnance Surveys.



Figur 19: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten, Britischen Toponymen aus Geonames sowie aus dem 1:50'000 Scale Gazetteer des Ordnance Surveys.

Um die Untersuchung auch auf weitere Länder auszuweiten, wurden die US-amerikanischen Toponyme aus Geonames untersucht. Hier unterscheidet sich die Verteilung der Distanzen von mehrdeutigen Toponymen weniger von derjenigen der zufällig Ausgewählten. Trotzdem ist ein Unterschied zwischen den beiden Verteilungen zu erkennen (Figur 20) welcher mit den bisherigen Resultaten vergleichbar ist. Um den Unterschied zwischen diesen beiden Verteilungen statistisch zu bestätigen musste allerdings die Stichprobengrösse bei einem Mann-Whitney-U-Test auf 1000 Samples angehoben werden. Der Unterschied wurde dann aber mit $p < 0.001$ signifikant (Anhang 8.1.8).



Figur 20: Häufigkeitsverteilung der Distanzen zwischen mehrdeutigen sowie zufällig ausgewählten, US-amerikanischen Toponyme aus Geonames.

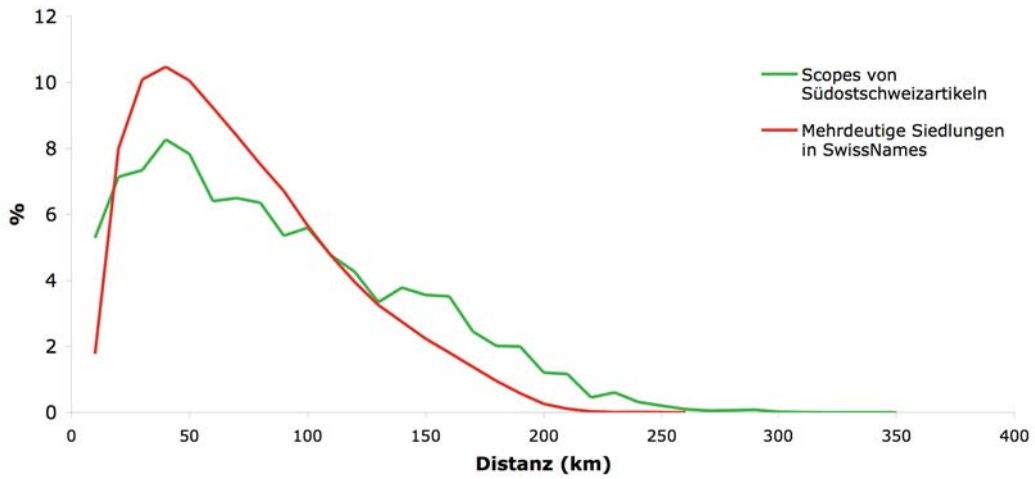
4.4 Vergleich des Scopes von Zeitungsartikeln mit der räumlichen Verteilung von mehrdeutigen Toponymen

Die Zeitungsartikel der Südostschweiz wurden nach den in den Kapiteln 3.3 und 3.4 vorgestellten Methoden prozessiert, wobei an dieser Stelle nochmals darauf hingewiesen wird, dass für den Vergleich nur Toponyme aus dem SwissNames-Gazetteer verwendet wurden.

Der Durchschnitt der Distanzen (Tabelle 6) zwischen den Toponymen in einem Zeitungsartikel (Kapitel 3.7) ist um 29.3% grösser als die durchschnittliche Distanz zwischen mehrdeutigen Toponymen und auch der Median ist um 24.5% grösser.

	Min.	Max.	Median	Durchschnitt
Mehrdeutige Toponyme (G. & O.)	0 km	252 km	56 km	62 km
Scope der Südostschweizartikel	0 km	341 km	70 km	80 km

Tabelle 6: Statistische Kennzahlen der Scopes von Zeitungsartikeln im Vergleich mit der räumlichen Verteilung von mehrdeutigen Toponymen.

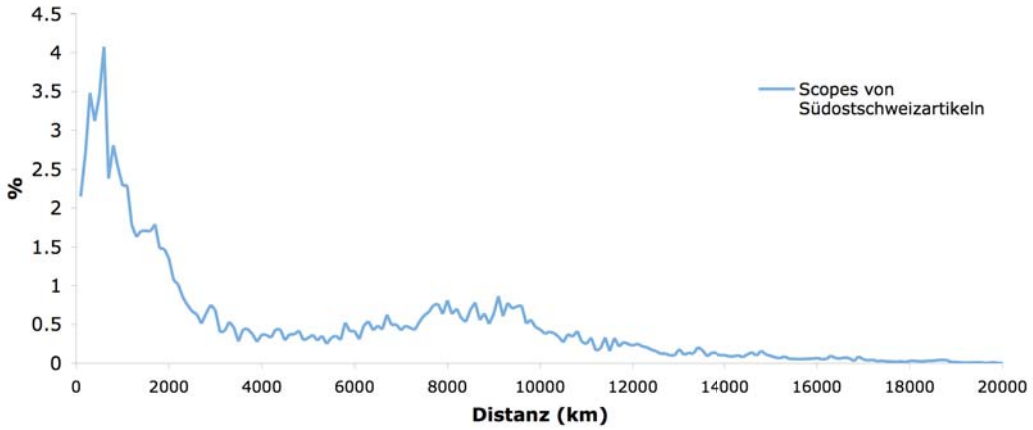


Figur 21: Vergleich der Häufigkeitskurven von Distanzen der Scopes von Südostschweizartikeln und mehrdeutigen (Siedlungs-) Toponymen aus SwissNames.

Dieser Umstand ist auch in Figur 21 zu erkennen, wo sich die beiden Verteilungen eigentlich relativ ähnlich sind. Der Unterschied ist, dass die mehrdeutigen Toponyme öfters nah (unter 100 km) beieinander liegen als die Toponyme in den Zeitungsartikeln. Der unterschiedliche Median und Durchschnitt, sowie Figur 21 lassen darauf schliessen, dass die Toponyme innerhalb eines Zeitungsartikels nicht näher beieinander liegen, als die Referenten von mehrdeutigen Toponymen, sondern

weiter voneinander entfernt sind. Um diese Ergebnisse statistisch zu untermauern wurde ein Mann-Whitney-U-Test mit einer Stichprobengrösse von je 100 Samples durchgeführt. Dieser war bei $p < 0.001$ signifikant (Anhang 8.1.9).
Die Minimalwerte von 0 km in Tabelle 6 sind eigentlich 5.2 m (Mehrdeutige Toponyme, siehe Kapitel 4.2) respektive 1.2 m (Südostschweizartikel).

Da die Kombination von SwissNames und Geonames aufgrund der Konzentration der Toponyme in den Zeitungsartikeln auf die Schweiz (und SwissNames) zu einer wenig aussagekräftigen Grafik führt, wurden die Geonames-Toponyme separat behandelt. Figur 22 zeigt die räumliche Verteilung der in Zeitungsartikeln der Südostschweiz erwähnten Toponyme aus Geonames. Dabei ist ein klares Maximum um ca. 600 km erkennbar. Dies ist in etwa das 20-Fache des Maximums der Schweizer Ortschaften welches bei ca. 40 km liegt. Ein zweiter Peak erscheint zwischen 7’000 und 10'000 km. Dieser lässt einerseits durch die Kugelform der Erde erklären. Der Umfang der Erde ist bei einer kreisförmigen Entfernung von einem Punkt aus dann maximal, wenn diese Entfernung einen Viertel des Gesamtumfangs, also etwa 10'000 km, ist. Zudem ist die räumliche Anordnung der einzelnen Länder so, dass viele globalen, im Korpus erkannten Toponyme zwischen 7'000 und 10'000 km voneinander entfernt sind (Figur 22).



Figur 22: Räumliche Verteilung der in einem Zeitungsartikel erwähnten Ortschaften (aus Geonames).

Die statistische Auswertung der Grösse der Scopes von Zeitungsartikeln hat folgende Kennwerte ergeben:

	SwissNames & Geonames	Geonames
Min	0.001 km	0.221 km
Max	19939 km	19939 km
Median	210 km	2499 km
Durchschnitt	2330 km	4497 km

Tabelle 7: Statistische Kennwerte zur Grösse der Scopes von Zeitungsartikeln.

Das wohl erstaunlichste Merkmal in Tabelle 7 ist, dass der Median bei der Berücksichtigung beider Gazetteers um den Faktor 10 kleiner ist als der Mittelwert. Dies ist darin begründet, dass mehr als die Hälfte (53.5%) der erkannten Toponyme aus SwissNames stammen und somit innerhalb der Grenzen der Schweiz liegen, was bei der Grösse der Schweiz im Vergleich zur gesamten Erde eine enorme Konzentration darstellt. Bei der Einschränkung der Toponyme auf Einträge aus Geonames wurde die Diskrepanz zwischen Mittelwert und Median viel kleiner.

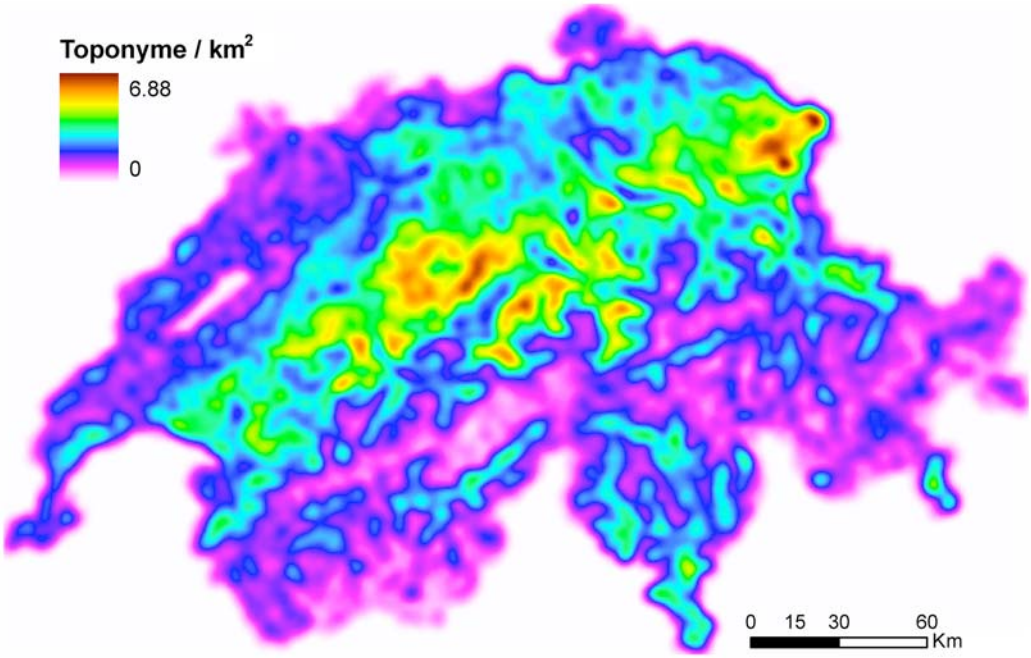
Eine mögliche Fehlerquelle der Berechnung der Scopes ist das dem Information Retrieval anhaftende Problem, dass der hier verwendete Geotagger, wie alle bisherigen und wohl auch zukünftigen, nicht fehlerfrei ist. Diese Fehler, vor allem bei der Precision, da dadurch falsche Distanzen gemessen werden, während bei nicht gefundenen Toponymen oder Referenten einfach gewisse Distanzen fehlen, können sich durchaus auf die Berechnung des Scopes auswirken. Weiter ist zu berücksichtigen, dass sich die Scopes spezifisch auf Südostschweiz-Artikel aus dem Jahr 2006 beziehen, wodurch sie sowohl an einen räumlichen Ort als auch an eine Zeitspanne gebunden sind. Daher können diese Scopes nicht unbesehen auf andere Medien übertragen werden.

4.5 Visualisierung der räumlichen Verteilung von Toponymen

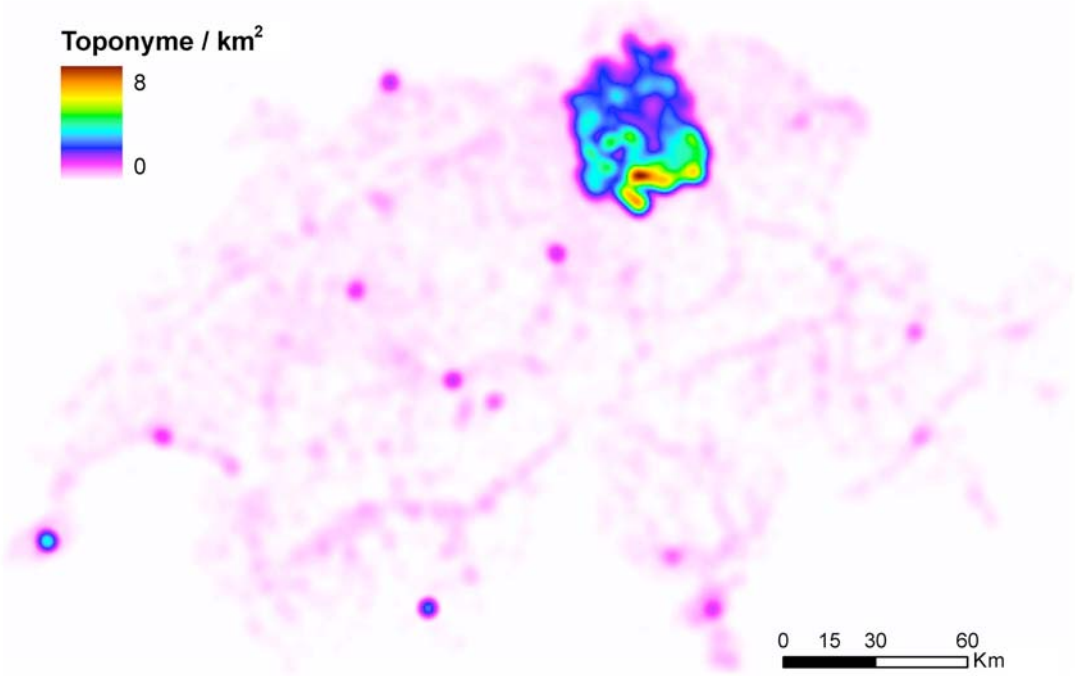
4.5.1 Visualisierung der räumlichen Verteilung von Toponymen in Gazetteers anhand von Dichteoberflächen

Um die vorhandenen Gazetteers sowie die berechneten Scopes visuell darzustellen, wurden sie entsprechend dem Kapitel 3.8 prozessiert. In Figur 23 wird eine Dichteoberfläche aller Einträge in SwissNames dargestellt. Dabei ist zu erkennen, dass sie relativ gleichmässig über den Raum verteilt sind. Die höchste berechnete Dichte war 6.88 Punkte pro km^2, welche in zwei Regionen im Bereich der Nordostschweiz festgestellt wurde. Ansonsten ist die Dichte vor allem in den alpinen

Regionen sowie dem Jura geringer, nachmals sogar gegen null Punkte pro km^2. Auf den Flächen der Seen hat es keine Einträge in SwissNames, wobei sich dies aufgrund des Suchradius von 5 km nur bei grösseren Seen, wie dem Genfer-, Boden- oder Neuenburgersee bemerkbar macht.



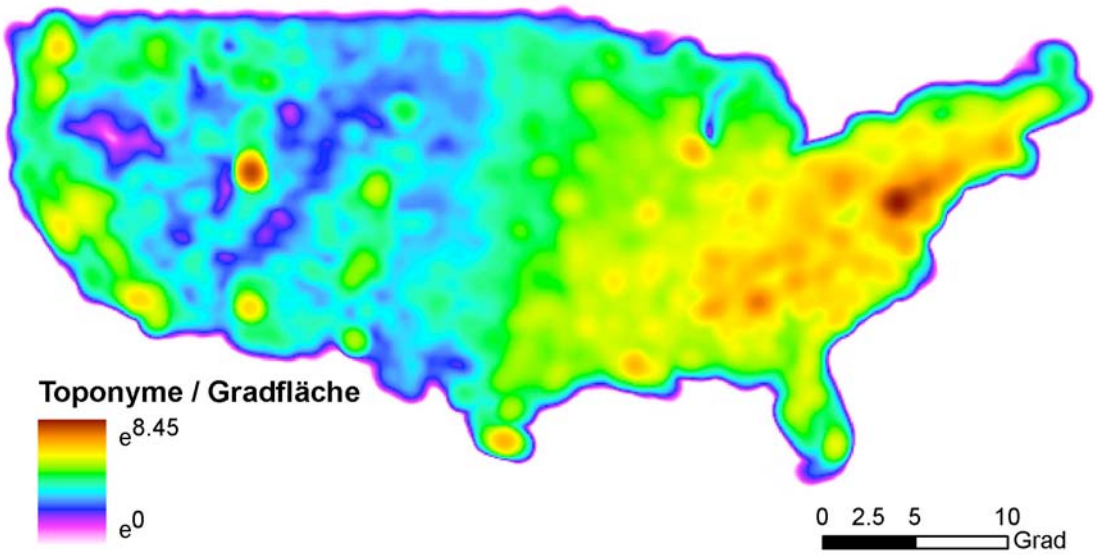
Figur 23: Dichteoberfläche der Einträge in SwissNames (Suchradius 5 km).



Figur 24: Dichteoberfläche aller Schweizer Einträge in Geonames (Suchradius 5 km).

Die räumliche Verteilung der Einträge in Geonames ist in Figur 24 dargestellt. Hier ist ein frappanter Unterschied zwischen den verschiedenen Regionen festzustellen. So ist die Punktedichte in der Region um Zürich sehr hoch, während die Abdeckung der restlichen Schweiz im Bereich von unter zwei Punkten pro km^2 ist. Diese Verteilung ist sehr unterschiedlich und rührt wohl davon, dass Geonames kein staatliches Produkt ist und auf User-Einträgen basiert. So sind neben der Region um Zürich vor allem bekannte oder touristische Ortschaften wie Genf, Basel, Bern, Lausanne, Lugano, Zermatt oder Davos in Geonames vorhanden.

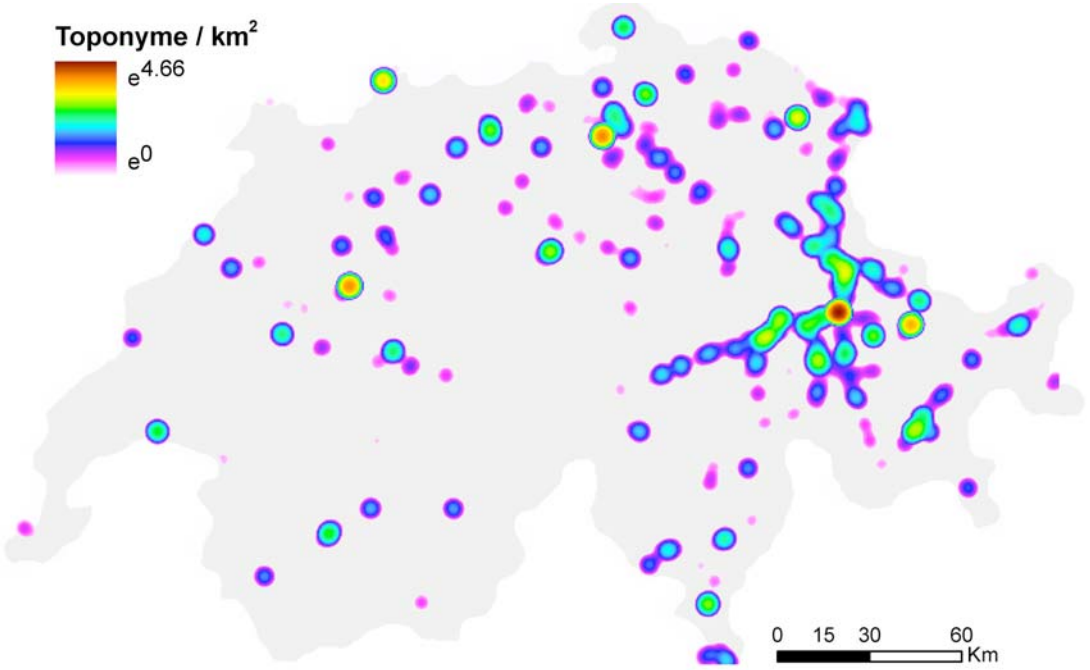
Eine weitere Dichteoberfläche wurde für die US-amerikanischen Einträge in Geonames berechnet. Diese, in Figur 25 dargestellt, ist viel unregelmässiger als diejenige der Schweiz. Wegen der grossen Streuung der Werte mussten die Dichtewerte logarithmisiert werden, um eine visuelle Aussage machen zu können. Der Allgemeine Trend der räumlichen Verteilung der US-amerikanischen Einträge ist, dass der Osten eine viel höhere Dichte an Einträgen als der westen aufweist. Vor allem im Bereich um New York ist die Dichte mit einem Wert von 4675 Punkten pro Gradeinheit. Ein zweites Maximum ist um die Stadt Salt Lake City im US-Bundesstaat Utah zu erkennen. Dies wohl aufgrund der olympischen Spiele im Jahr 2002 sowie der vielen und bekannten Skigebiete wie Alta oder Park City.



Figur 25: Dichteoberfläche der kontinentalen (ohne Alaska, Hawaii, Puerto Rico) US-amerikanischen Einträge in Geonames (Suchradius 0.1 Grad).

4.5.2 Visualisierung der räumlichen Verteilung von Toponymen in Artikeln der Südostschweiz anhand von Dichteoberflächen

Analog zum vorherigen Kapitel wurden die, in der Südostschweiz gefundenen Toponyme visualisiert. Figur 26 zeigt eine Dichtoberfläche dieser Toponyme, jedoch auf SwissNames-Einträge eingeschränkt. Trotz ihrem überregionalen Charakter und einer relativ grossen Auflage von ca. 125'000[39] ist die Südostschweiz innerhalb der Schweiz nur schon durch ihren Namen auf ein spezifisches Gebiet (Graubünden und St. Galler Rheintal) gerichtet. Dies tritt in einem so ausgeprägten Masse auf, dass die Skala logarithmisiert werden musste, um auch andere Ortschaften als „Chur" erkennbar zu machen. Nebst Einträgen in Graubünden sowie dem St. Galler-Rheintal sind auch die meisten grösseren Städte erkennbar. Auffällig ist die Abnahme der Punkte in der Westschweiz. Dies rührt, neben des lokalen Charakters der Zeitung vor allem daher, dass die Toponyme in SwissNames in der jeweiligen Hauptsprache der Region gespeichert sind. So wurden Lausanne und Lugano, welche auf Französisch respektive Italienisch gleich lauten wie auf Deutsch, oft erkannt, das grössere und vor allem politisch wichtigere Genf jedoch nur selten.

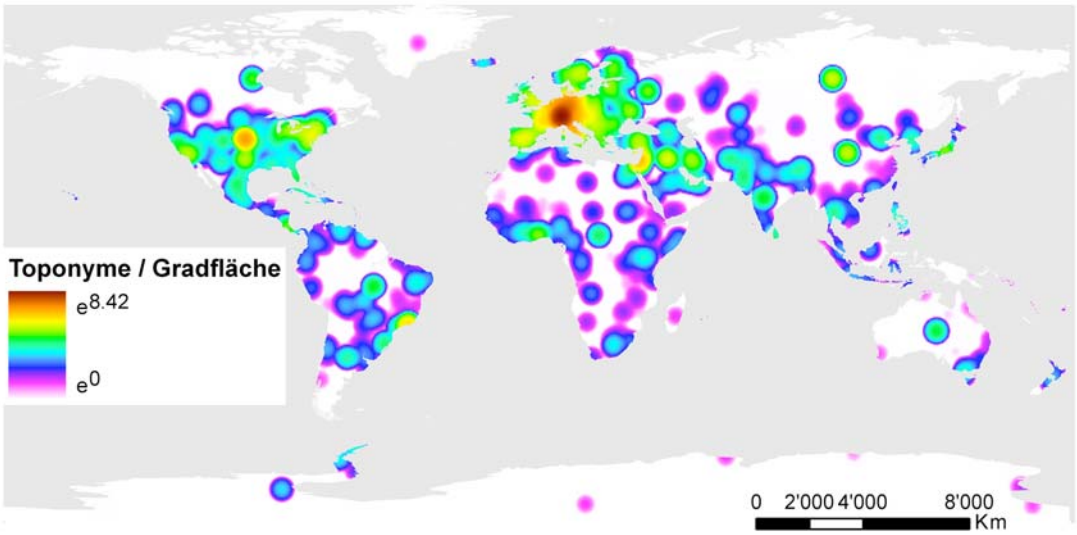


Figur 26: Dichteoberfläche der in Südostschweizartikeln erkannten Toponyme aus SwissNames (Suchradius 5 km).

[39] http://www.wemf.ch/

Auf die gleiche Art und Weise wurden die gesamten, in der Südostschweiz erkannten Toponyme, dargestellt (Figur 27). Diese Grafik veranschaulicht sehr gut, welches die politischen „Brandherde“ des Jahres 2006 waren. So ist der, in Zeitungen omnipräsente Palästina-Konflikt auch anhand der Toponyme zu erkennen. Auch über den Atomstreit zwischen Pakistan und Indien, sowie den Irak-Krieg ist im Jahr 2006 sehr oft berichtet worden. Auch die Länder China, Indien, Russland, Brasilien sowie die USA sind gut auf der Karte zu erkennen. Interessant ist, dass das in Afrika gelegene Togo relativ oft vorkommt, was daher rührt, dass die Schweizer Fussballnationalmannschaft and er Weltmeisterschaft 2006 in Deutschland gegen Togo gespielt hat. Interessant ist, dass ausser in den USA vor allem die einzelnen Länder, und nicht in diesen Ländern gelegene Städte erkannt wurden. Die Spitze, welche in der Mitte der Länder zu erkennen ist, ist darin zu begründen, dass Geonames den einzelnen Länder nicht Vektor-Geometrien sondern Punkte zuweist. Vor allem aber in den USA sind einzelne Regionen wie die Hauptstadt Washington, D.C. sowie die Regionen um New York sowie San Francisco zu erkennen.

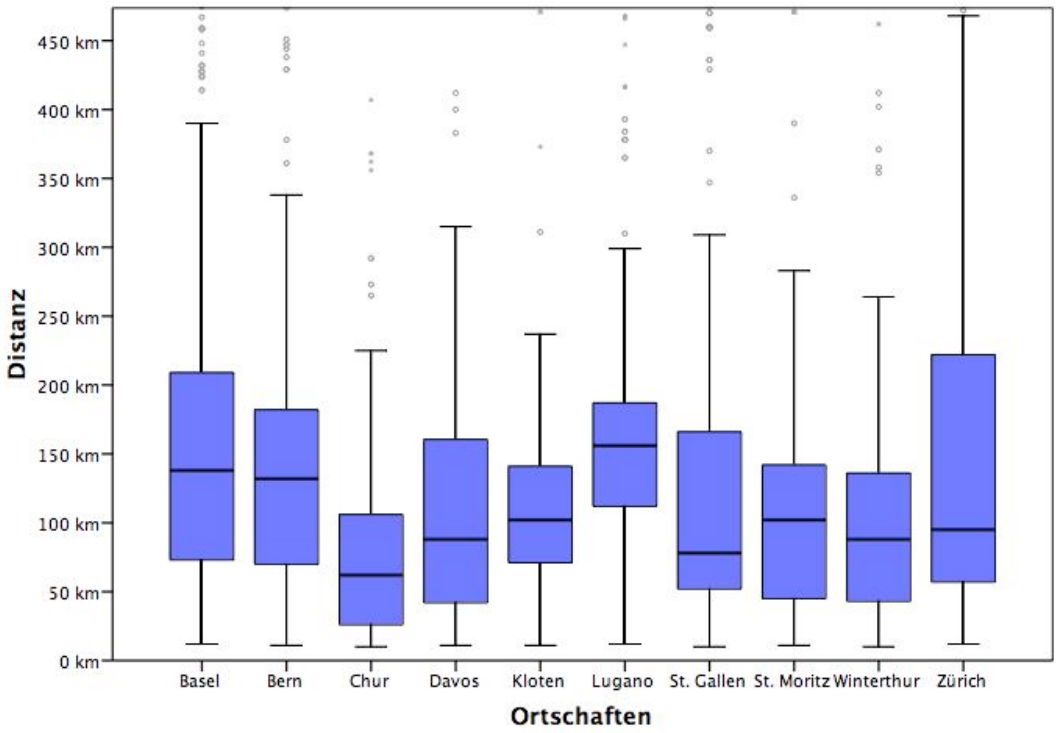
Im Allgemeinen lässt sich aber sagen, dass in der Südostschweiz neben den internationalen Konflikten vor allem Ortschaften in Europa berichtet wird.



Figur 27: Dichteoberfläche der total in Südostschweizartikeln erkannten Toponyme (Suchradius 1 Grad).

4.5.3 Visualisierung der räumlichen Verteilung von Toponymen in Artikeln der Südostschweiz anhand von Boxplots

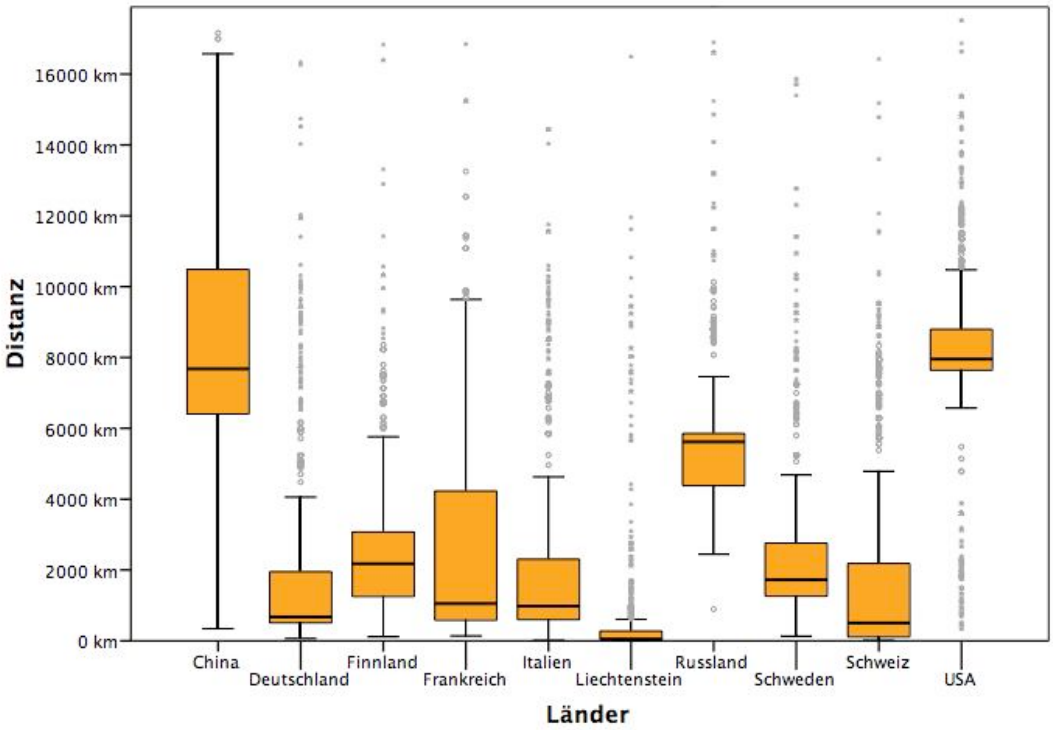
Nach der in Kapitel 3.9 beschriebenen Methodik wurden anhand der Toponyme, welche aus den Südostschweizartikeln extrahiert werden konnten Boxplots erstellt. Figur 28 zeigt die Distanzen zwischen den Toponymen der Artikel, in welchen die zehn häufigsten Schweizer Ortschaften vorkamen. Interessant ist, dass die Ortschaft mit dem höchsten Median (Lugano) weder das höchste 75%-Perzentil, noch den höchsten, nicht extremen Wert hat. Aufgrund von Figur 28 kann gesagt werden, dass Zürich zusammen mit Toponymen, die relativ weit entfernt sind, in Zeitungsartikeln der Südostschweiz vorkommt. Anderes scheint für Chur zu gelten, welches vorwiegend in räumlich konzentrierteren Artikeln vorzukommen scheint, da sämtliche Werte, also sowohl die nicht extremen-, die 25%- und 75%-Perzentilwerte aber auch der Median die Tiefsten im Vergleich zu den restlichen, in Figur 28 gezeigten Ortschaften sind.



Figur 28: Boxplot der zehn häufigsten Ortschaften in Südostschweizartikeln.

Figur 29 zeigt die zehn häufigsten Ländernamen aus Artikeln der Südostschweiz. Im Gegensatz zu den Schweizer Ortschaftsnamen aus Figur 28 sind die Boxplots teilweise ziemlich unterschiedlich. Die Mediane scheinen meist mit der Distanz der

einzelnen Länder zur Schweiz zu korrelieren. Nicht nur die Mediane, sondern auch die Streuung der werte ist bei den einzelnen Ländern ziemlich unterschiedlich. So scheinen Artikel, welche die USA beinhalten immer in etwa die gleiche Scope-Grösse zu haben, während die Grösse der Scopes von Zeitungsartikeln welche China erwähnen ziemlich unterschiedlich sind. Die Boxplots aus Figur 29 zeigen vor allem auch die Grösse der einzelnen Länder. So sind die Mediane von China, der USA sowie Russland um ein vielfaches höher als die der übrigen Länder. Im Boxplot der USA lässt sich zudem ein spezielles Phänomen entdecken. Es manifestiert sich nämlich zum Einen die Grösse des Landes, welche durch die Extremalwerte im Bereich von 0 bis 6000 km zum Ausdruck kommt, zum Anderen die Distanz zu den politischen Brandherden in Israel/Palästina und dem Irak, in welchen die USA medial omnipräsent sind.



Figur 29: Boxplot der zehn häufigsten Länder in Südostschweizartikeln.

5 Diskussion

Nachfolgend sollen zuerst die Resultate aus Kapitel 4 diskutiert werden, welche dann als Grundlage für die Beantwortung der Forschungsfragen, welche in Kapitel 2.8.2 gestellt wurden, dienen. Dabei liefern die ersten beiden Forschungsfragen wiederum die Grundlage für die Dritte, welche räumliche Disambiguationsstrategien im Kontext der Resultate dieser Arbeit kritisch analysiert und grundlegende Prinzipien für zukünftige Anwendungen definiert.

Anschliessend werden weitere Auswertungen der hier generierten Daten diskutiert.

5.1 Diskussion des hier verwendeten Geotaggers

Im Gegensatz zu den meisten anderen GIR-Systemen wurden in dieser Arbeit deutsche Artikel prozessiert. Dies gilt im Allgemeinen als komplexer als die Prozessierung von englischsprachigen Texten (Rössler 2004).

Die erreichten Precision-, Recall- sowie F_1-Werte sind jedoch wie immer mit Bezug auf die Definition von Toponymen (Kapitel 3.3.3) zu beurteilen. Diese Definition kann die Benchmarks stark beeinflussen, da sie damit auch die Komplexität der Aufgabe für den Geotagger definiert. Der hier verwendete Geotagger.java in Kombination mit Disambiguator.java könnte zum Beispiel keine als Adjektive verwendeten Toponyme erkennen. Dies fliesst jedoch aufgrund der Definition von Toponymen (Kapitel 3.3.3) in dieser Arbeit nicht in die Benchmarks mit. Ein weiterer Umstand, welcher das GIR erleichtert ist, dass hier Texte einer Zeitung prozessiert wurden und keine Internetseiten. Wenn auch der Text in Zeitungsartikeln nicht strukturiert ist, so ist das Format, in welchem die Texte gespeichert sind einheitlich. Weiter wird nicht versucht, falsch geschriebene Toponyme zu erkennen. Dies ist insofern nicht nötig, da davon ausgegangen werden kann, dass Zeitungsartikel ein hoher Grad an orthographischer Korrektheit aufweisen.

Trotz dieser, dem Gebiet des GIR innewohnenden, Varianz wird an dieser Stelle ein kurzer Vergleich dieser Arbeit mit verschiedenen GIR-Systemen angestellt.

Autor	F_1	R	P	N	Sprache
Smith & Crane 2001	0.96	0.99	0.93	4T.	gr
Mikheev et al. 1999	0.95	0.95	0.94	100 T.	en
Zheng & Su 2002	0.94	0.95	0.94	124 KB	en
Pouliquen et al. 2004	0.94	0.90	0.98	48 T.	en
Stevenson & Gaizauskas 1999	0.94	0.92	0.97	1 Mio W.	en
Kornai & Thompson 2005	0.91	0.91	0.91	100'000 W.	en
Brunner 2008	***0.90***	***0.93***	***0.88***	***100 T.***	***de***
Lee & Lee 2005	0.86	0.83	0.90	107 T.	en
Pouliquen et al. 2006	0.84	0.78	0.91	162 T.	en
Smith & Crane 2001	0.81	0.89	0.74	4 T.	en
Overell & Rüger 2006	0.80	0.80	0.80	1000 T.	en
Pouliquen et al. 2006	0.73	0.68	0.80	162 T.	de
Clough 2005	0.72	0.78	0.70	130 T.	en
Leidner 2008	0.59	0.56	0.63	946 T.	en

Tabelle 8: Vergleich von verschiedenen GIR-Systemen mit dieser Arbeit, sortiert nach F_1-Wert (N in Texten, Kilobytes oder Wörter).

Wie der Tabelle 8 zu entnehmen ist, sind die hier erreichten Werte mit praktisch allen bisherigen Arbeiten zu vergleichen. Sie setzt sich sogar stark von der einzigen anderen Arbeit, welche deutsche Texte prozessiert (Pouliquen et al. 2006) ab. Das Ergebnis von Smith & Crane (2001) ist jedoch mit Vorsicht zu geniessen, da es sich um die Prozessierung von spezifisch geographischen Texten in Griechisch, welche auch nur über ein bestimmtes Gebiet in Griechenland sind, handelt. Dieselben Heuristiken wurden von Smith & Crane (2001) auch auf englische Texte angewendet, dabei konnte aber nur ein F_1-Wert von 0.81 erreicht werden. Dies zeigt zugleich, dass der Erfolg eines GIR-Systems nicht nur von den Heuristiken, sondern auch von der Sprache sowie vom Inhalt der Texte abhängt. Diese Problem ist schon seit einiger Zeit bekannt, weshalb immer öfters versucht wird, einen „Gold-Standard-Korpus" zu erstellen (Leidner 2008), welcher die Vergleichbarkeit von verschiedenen Systemen gewährleistet.

5.2 Ambiguität in Gazetteers

Dass die Wahrscheinlichkeit für Ambiguität nicht bei allen Toponymen gleich ist, ist nur schon aufgrund der Wortlänge anzunehmen. Leidner (2008) zu Folge weisen gewisse Toponyme bis zu 1'600 (San José) verschiedenen Referenten auf. Dieser Ambiguitätsgrad fällt aber in seinem Ranking der Toponyme mit den meisten Referenten schon nach sieben Toponymen auf unter 1'000 (San Francisco mit 980) und nach 22 Toponymen auf unter 500. Somit scheinen nur wenige Toponyme einen

exorbitanten Grad an Ambiguität aufzuweisen. Zum selben Ergebnis kam die Untersuchung der mehrdeutigen Toponyme in SwissNames (Figur 11). Während es in SwissNames über 3000 Toponyme gibt, welche zwei Referenten haben, so sind es nur noch gut 900, welche drei und knapp 450 welche vier Referenten haben. Dieses Verhältnis scheint umgekehrt exponentiell zu sinken und könnte durch eine Zipf-Verteilung beschrieben werden (Clough, pers. comm.).

Um einen Anhaltspunkt für die Wichtigkeit der Geo-Geo-Disambiguation zu erhalten, kann der Ambiguitätsgrad des zur Verfügung stehenden Gazetteers berechnet werden. Dies wurde schon oft und für verschiedene Gazetteers gemacht. So sind in Grossbritannien Purves et. al (2007) zufolge etwa 10% aller Ortsnamen mehrdeutig. Smith & Crane (2001) haben „The Getty Thesaurus of Geographic Names“ innerkontinental untersucht und sagen, dass 17% aller europäischen Toponyme referenziell mehrdeutig sind, womit Europa der Kontinent mit dem niedrigsten Anteil an Geo-Geo-Ambiguität sein soll. Der grösste Anteil an Ambiguität hat Smith & Crane (2001) zufolge Nord- und Zentralamerika, wo es sogar 57% sein sollen. Auch Li et al. (2002) untersuchten den Ambiguitätsgrad von Toponymen und fanden in "The Tipster Gazetteer"[40] weltweit 18% mehrdeutige Toponyme. Verglichen mit Smith & Crane (2001) ist dies ein relativ tiefer Wert, wobei „The Getty Names Gazetteer“ verglichen mit „The Tipster Gazetteer“ viel grösser und somit auch detaillierter ist. Laut Kornai & Thompson (2005) enthält "The Tipster Gazetteer" nur schon in dessen detailliertester Region, den Vereinigten Staaten, weniger als 7% der Ortsangaben anderer Gazetteers Dies lässt darauf schliessen, dass die Ambiguität mit der Detailtreue eines Gazetteers zusammenhängt.

Die Untersuchung von SwissNames im Rahmen dieser Arbeit hat ergeben, dass 43% aller Toponyme in SwissNames mehrdeutig sind. Dies ist nicht mit den 17% der europäischen Toponyme von Smith & Crane (2001) zu vergleichen. Der Hauptunterschied ist aber auch hier in der Detailliertheit des Gazetteers zu suchen. Bisher wurden nur Gazetteers untersucht, welche eine relativ geringe Raumdichte an Toponymen aufweisen. SwissNames, als staatlich gefördertes Produkt, ist wohl einer der detailliertesten Gazetteers überhaupt, was diese Untersuchung stark von den

[40] http://crl.nmsu.edu/cgi-bin/Tools/CLR/clrcat

Bisherigen unterscheidet. Die verschiedenen Attribute von SwissNames wurden in dieser Arbeit genutzt und die Ambiguität in Abhängigkeit von der Art der Toponyme berechnet (Figur 9). Während die Einschränkung der Toponyme auf nur Siedlungstypen kaum einen Einfluss auf die Ambiguität hatte, so sank diese erst bei der Reduktion auf Siedlungstypen mit mehr als fünfzig Einwohnern. Dadurch konnte die Ambiguität auf 17% verringert werden, was mit den Untersuchungen von Smith & Crane (2001) und Li et al. (2002) übereinstimmt und näher bei den geschätzten 10% von Purves et. al (2007) liegt. Wie aber in Figur 10 gezeigt wird, wird durch diese Reduktion der Umfang des Gazetteers um 95% verringert.

Diese Erkenntnis widerspiegelt das in Kapitel 2.3.3 vorgestellte Problem der Reduktion von Gazetteers. Den Vorteil bezüglich des geringeren Ambiguitätsgrades, welchen man durch die Reduktion des Gazetteers erhält, geht Hand in Hand mit dem Verlust an Ressourcen und damit der grösstmöglichen Abdeckung. Die Herausforderung bei der Erstellung eines Gazetteers ist also, den Anteil an Ambiguität möglichst zu minimieren, während der Umfang dessen möglichst erhalten bleibt. Eine Methode zur Reduktion von Gazetteers ist, dass man sie an Trainingskorpi überprüft, und dann nur die für diesen Korpus relevanten Toponyme extrahiert (Mikheev et al. 1999), wodurch ein Machine-learning-Ansatz miteinbezogen wird. Die implizite Aussage von Mikheev at al. (1999) ist jedoch vielmehr, dass ein Gazetteer im Bezug auf die zu prozessierenden Texte erstellt werden muss. So wäre ein Gazetteer mit Informationen über Wanderwege in den Anden für das Geocodieren von Schweizer Zeitungsartikel höchst selten nützlich, bei der Prozessierung von Reiseberichten könnte dieser aber durchaus von Nutzen sein. Die adaptive Wahl des Gazetteers könnte bei der Implementierung eines Korpus-unabhängigen GIR-Systems also ein entscheidender Erfolgsfaktor sein.

5.3 Wie sind die Referenten von mehrdeutigen Toponymen räumlich verteilt?

Dieses Unterkapitel widmet sich der gleichlautenden, in Kapitel 2.8.2 gestellten Forschungsfrage.

Wie in Kapitel 4.3 aufgeführt, haben Untersuchungen im Rahmen dieser Arbeit ergeben, dass die Referenten von mehrdeutigen Toponymen nicht zufällig über den Raum verteilt sind, sondern bedeutend näher beieinander liegen. Dabei ist diese

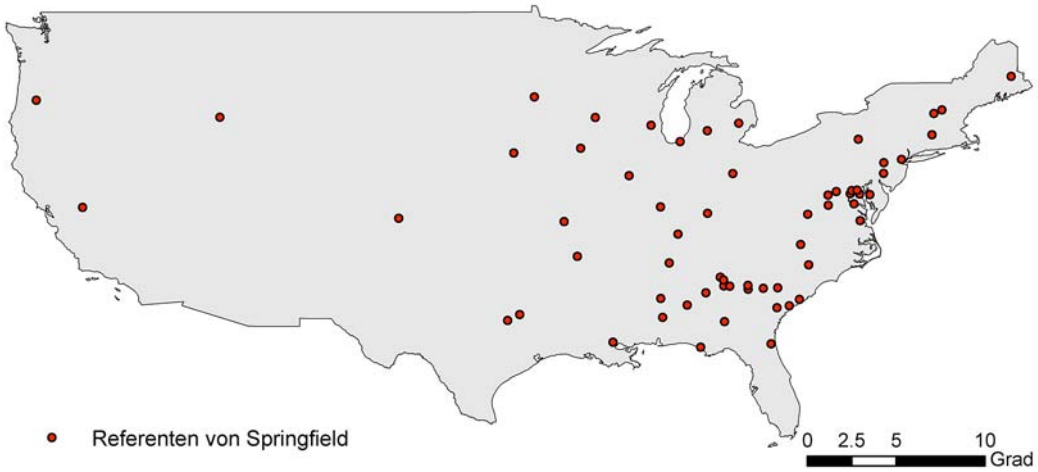
Beobachtung nicht nur für die Schweiz und SwissNames gültig, sondern auch für alle anderen untersuchten Länder (Grossbritannien und die USA) und für alle anderen untersuchten Gazetteers (Geonames und Ordnance Survey). Die Stärke dieses Unterschieds scheint jedoch sowohl von den Gazetteers als auch von den Ländern abhängig zu sein.
SwissNames ist wohl für die Schweiz als der kompletteste Gazetteer anzuschauen, daher kann dessen Auswertung als am aussagekräftigsten angesehen werden. Die Untersuchung von SwissNames ergab eine Zufallsverteilung, welche stark einer Normalverteilung gleicht. Dies ist jedoch ein Einzelfall und kann wiederum mit der Qualität des Gazetteers erklärt werden. Die räumliche Verteilung der Toponyme in SwissNames ist relativ ebenmässig (Figur 23), während diejenige der Schweizer Toponyme in Geonames räumlich sehr konzentriert ist (Figur 24). Die räumliche Konzentration der Geonames-Einträge wirkt sich denn auch in der Verteilung der zufällig ausgewählten Toponyme aus (Figur 16) aus. Diese erreichen bei SwissNames um 100 km ein Maximum, während der Peak bei Geonames zwischen 20 und 30 km zu finden ist.

Neben der Schweiz ist auch Grossbritannien ein Land, in welchem sich die räumliche Verteilung von mehrdeutigen Toponymen stark von der räumlichen Verteilung von zufällig ausgewählten Toponymen unterscheidet. Dieser Unterschied manifestiert sich am deutlichsten bei Schottischen Toponymen (Figur 18). Laut Purves (pers. comm.) ist dieser dadurch zu erklären, dass in gewissen Regionen von Schottland Toponyme Gälisch benannt wurden. Damit ist es sehr unwahrscheinlich, dass man Referenten von Toponymen findet, welche auf beiden Seiten dieser Grenze liegen.
Im Gegensatz zu den Schweizer Toponymen ist bei den Britischen Toponymen zwischen der offiziellen (Ordnance Survey) und der frei verfügbaren (Geonames) Quelle kein Unterschied zu erkennen (Figur 19). Dies lässt auf eine sehr gute Abdeckung des britischen Gebiets von Geonames schliessen.

In den USA ist der Unterschied zwischen der räumlichen Verteilung von mehrdeutigen Toponymen und zufällig ausgewählten Toponymen am schwächsten ausgeprägt (Figur 20). Hierfür liegen keine offensichtlichen Gründe auf der Hand, eine mögliche Erklärung könnte jedoch sein, dass der geringe Unterschied in den USA von der Besiedlungsart des Landes herrührt. Weiter ist die Uniformität der USA

im Hinblick auf die Sprache sowie die schnelle, räumlich parallel verlaufende Besiedlung wohl einzigartig. Dasselbe gilt für das kulturelle Gedankengut. So sind im Gegensatz zur Schweiz kaum regional spezifische Toponymmuster wie z.B. ...kon in der Region Zürich-Aargau erkennbar. Diese regionalen Muster erhöhen die Wahrscheinlichkeit für kurze Distanzen zwischen gleichen Ortschaftsnamen ungemein. Ein zusätzlicher Punkt ist, dass die Toponyme in Geonames verglichen mit einem offiziellen Gazetteer nicht sehr regelmässig über den Raum verteilt sind (Figur 25).

Die USA werden in vielen Arbeiten als das Land mit der höchsten Ambiguität bei Toponymen bezeichnet (Smith & Crane 2001). So werden auch meist US-amerikanische Toponyme wie „Springfield“ als Beispiele für Geo-Geo-Ambiguität verwendet (z.B. Markowetz et al. 2004, Kornai & Thompson 2005). Beim Beispiel von Springfield handelt es sich nach Amitay et al. (2004) um 63 Referenten welche sich über 34 US-Staaten verteilen. Dies lässt darauf schliessen, dass die Referenten von mehrdeutigen Toponymen in den USA weiter verstreut sind, obwohl es im Bundesstaat Alambama vier Referenten für Springfield gibt. In Geonames hat das Toponym „Springfield“ auch 63 US-amerikanische Referenten, deren räumliche Anordnung in Figur 30 dargestellt sind. Auf den ersten Blick sehen die Referenten sehr wohl räumlich konzentriert aus, wenn man aber die Dichte der US-amerikanischen Toponyme in Geonames betrachtet (Figur 25), so wird dies gleich wieder relativiert. In beiden Figuren ist die Dichte an Toponymen im Osten viel höher als im (mittleren) Westen. Dieses Beispiel untermauert also die Erkenntnis aus Figur 20, dass in den USA die mehrdeutigen Toponyme räumlich stärker einer zufälligen Verteilung gleichen als in der Schweiz und Grossbritannien.



Figur 30: Die 63 Referenten des Toponyms „Springfield“ in den USA.

Der hohe Ambiguitätsgrad von Toponymen bei praktisch sämtlichen Toponymtypen (siehe Kapitel 4.2) sowie die Distanz zwischen den möglichen Referenten zeigen, dass Geo-Geo-Ambguität räumlich sehr konzentriert vorkommt und dies bei der Definition von Disambiguationsmethoden berücksichtigt werden muss.
Durch diese Erkenntnis wird die erste Hypothese, welche in Kapitel 1.2 gestellt wurde, bestätigt:

H_1: Die mehrdeutigen Toponyme sind nicht zufällig über den Raum verteilt, sondern sind räumlich konzentriert.

Tobler (1970, S. 236) stellte bei der Untersuchung des wirtschaftlichen Wachstums einer urbanen Region folgendes, inzwischen als das „first law of geography“ bekanntes, fest:

„Everything is related to everything else, but near things are more related than distant things“.

Die erste Erkenntnis dieser Arbeit ist, dass Tobler’s „first law of geography“ auch auf die räumliche Verteilung von Toponymen zutrifft. So sind die Referenten von mehrdeutigen Toponymen, welche im orthographischen Sinne die stärkstmögliche Beziehung zueinander aufweisen, näher beieinander liegend als zufällig ausgewählte Referenten.
Wie Sui (2004) in seinem Paper über Tobler’s „first law of geography“ erkannte, schränkt das Ignorieren von Tobler’s „first law of geography“ die geographische Vorstellung ein. Diese ist im Anbetracht der räumlichen Toponym Resolution grundlegend.

Die Erkenntnis, dass die Referenten von mehrdeutigen Toponymen räumlich konzentriert vorkommen schmälert die Erfolgschancen der Disambiguationsstrategie von Smith & Crane (2001) drastisch. So haben Smith & Crane (2001) alle Referenten auf ein Raster mit einer Rasterweite von einem Grad interpoliert. Dies resultiert in den Breitengraden der Schweiz in einer Zellengrösse von ca. 75 km breite auf 110 km Höhe. Damit liesse sich die Schweiz in ca. zehn Zellen einteilen, aus welchen der Zentroid berechnet werden müsste. Allein die Zellengrösse von 8250 km^2 ist für die

Disambiguation gänzlich ungeeignet. Dies, da die durchschnittliche Distanz zwischen zwei Referenten eines mehrdeutigen Toponyms bei 62 km und der Median bei 56 km liegt.

5.4 Wie sind die in einem Zeitungsartikel erwähnten Ortschaften über den Raum verteilt?

Dieses Kapitel widmet sich der gleich lautenden Forschungsfrage aus Kapitel 2.8.2.
Die Analyse der Scopes von Südostschweizartikeln hat ergeben, dass vor allem über Schweizer Ortschaften berichtet wird. Dabei hatten die Scopes der Artikel einen Median von 70 km (Tabelle 6), wenn nur Schweizer Toponyme in Betracht gezogen wurden. Dies ist deutlich kleiner als eine Zufallsverteilung über die selben Toponyme mit einem Median von 106 km. Dasselbe gilt für die weltweiten Artikel mit einer starken Konzentration und einem Median von 210 km (Tabelle 7). Wenn man nur die Toponyme aus Geonames betrachtet (Figur 22), so sind sie wiederum räumlich konzentriert. Es lässt sich also auch auf die Verteilung der, in einem Zeitungsartikel erwähnten Ortschaften Tobler's „first law of geography" anwenden, was bedeutet dass sie räumlich konzentriert sind.
Wie in den beiden Boxplots (Figur 28 und Figur 29) zu erkennen ist, so haben in der Südostschweiz Artikel, welche Länder-Toponyme beinhalten tendenziell ein grösseres Scope als Artikel, welche (kleinere) Schweizer Ortschaften beinhalten (siehe Kapitel 4.5.3). Dieses Phänomen ist auch bei der Grösse der Länder beobachtbar, wobei hier auch die Distanz zum Veröffentlichungsort der Zeitung eine Rolle spielen könnte.

Sowohl Figur 26 als auch Figur 27 lassen herleiten, dass auch ein relativ kleiner, qualitativ hochwertiger Gazetteer sehr sinnvoll sein kann. Wie schon Krupka & Hausman (1998), Mikheev et al (1999) und Kornai & Thompson (2005) festgestellt haben, sind grosse Gazetteers für die meisten Anwendungen nicht notwendig. Dies widerspricht zwar den Aussagen von Cucchiarelli et al. (1998) und Fonseca et al. (2002), wird jedoch auch durch die Resultate dieser Arbeit bestätigt. Von den, in den Gazetteers enthaltenen, 101'594 Toponymen wurden nur 5'263 unterschiedliche (Referenten) erkannt, was bedeutet, dass 94.82% aller Toponyme der Gazetteers nie gefunden wurden. Dies entspricht in etwa den Resultaten von Krupka & Hausman (1998), welche ihren Gazetteer um 90% reduziert haben ohne dass das GIR sichtlich

darunter gelitten hat. Die Reduktion des Gazetteers kann jedoch nicht willkürlich geschehen, sondern es müssen die Einträge gelöscht werden, welche nicht in den zu untersuchenden Daten sind. Um nicht an Genauigkeit zu verlieren, müssten die Daten daher zwei Mal prozessiert werden. Die Alternative wäre die von Krupka & Hausman (1998), welche ihren Gazetteer hierarchisch gesäubert haben. Dies wäre im Kontext dieser Arbeit jedoch gefährlich (Kapitel 5.2), da durch den regionalen Charakter der Südostschweiz oft auch kleinere Ortschaften in Artikeln erwähnt werden.

Es ist anzufügen, dass es wie immer auf das Anwendungsgebiet des GIR ankommt, wie gross und komplett ein Gazetteer sein muss. So werden in dieser Arbeit nur Südostschweizartikel durchsucht, welche inhaltlich meist an ein räumliches Gebiet gebunden sind. Wenn das GIR jedoch auf unterschiedliche Daten angewandt wird, so wird ein grosser Gazetteer wohl notwendig sein. Mit einen grossen Gazetteer ist man demnach auf der sicheren Seite, wenn auch die Prozessierungszeit darunter leiden kann.

Abschliessend kann der Titel von Kornai & Thompson's (2005) Arbeit bestätigt werden:

Size doesn't matter

5.5 Lässt sich die Geo-Geo-Ambiguität von Toponymen durch räumliche Algorithmen auflösen?

Bisherige Ansätze zur räumlichen Disambiguation von mehrdeutigen Toponymen wurden in den Kapiteln 2.3.8 und 2.8.1 beschrieben. Dabei wurde bei sämtlichen Methoden weder die räumliche Verteilung der Referenten von mehrdeutigen Toponymen, noch das Scope der Dokumente, welche mit geographischen Referenzen verknüpft weden sollen, untersucht.

Leidner (2008) geht davon aus, dass sich nicht mehrere Referenten innerhalb des Scopes eines Dokumentes befinden (siehe Kapitel 2.8.1). Falls dies der Fall wäre, nimmt seine Methode als „Notlösung" den erstbesten Referenten.

In dieser Arbeit wurde gezeigt, dass zumindest im Falle der hier verwendeten Daten (SwissNames und Südostschweizartikel) keinesfalls davon ausgegangen werden kann, dass die Referenten von mehrdeutigen Toponymen ausserhalb des Scopes eines

Zeitungsartikels liegen. Die in Kapitel 1.2 formulierte zweite Hypothese kann somit bestätigt werden:

H2: Die räumliche Verteilung von mehrdeutigen Toponymen ist konzentrierter als die Ausdehnung des Scopes von Zeitungsartikeln.

Im Kontext der Auswertungen dieser Arbeit würde Leidners „Notlösung“ öfters angewendet werden, als dass sein MINIMALITY-Prinzip greifen könnte.
Wie schon Gould (1979) bemerkte, so kann das unbeabsichtigte Ignorieren von Tobler's Theorie die Forschungsmöglichkeiten gefährden. Zu dieser Erkenntnis, nämlich dass ihre Methode an räumliche Bedingungen geknüpft ist, gelangen auch Leidner et al. (2003, S. 2) und relativieren damit auch die Gültigkeit ihrer Methode:

„*Probably the smaller the span, the more often this heuristic will be valid*“.

Der Miteinbezug dieses *span*, in dieser Arbeit Scope genannt, in die räumliche Disambiguationsmethode würde somit die Validität der Methode verifizieren. Dabei muss der Scope in Bezug zur räumlichen Verteilung der Referenten des zu disambiguierenden Toponyms gestellt werden. Nebst dieser fallspezifischen Methode könnte man auch einen Schwellwert festlegen, welcher aufgrund der räumlichen Verteilung der Referenten von mehrdeutigen Toponymen bestimmt wurde. Pouliquen et al. (2006) haben für ihre Daten einen Schwellwert von 200 km bestimmt, wobei die Methodik, mit welcher dieser ermittelt wurde nicht dokumentiert ist. Aufgrund der Auswertungen der SwissNames-, Geonames- und Ordnance Survey- Gazetteers ist dieser Wert nicht nachvollziehbar, da sämtliche Mittelwerte der mehrdeutigen Toponyme klar unter 200 km liegen.

Sowohl Rauch et al. (2003) als auch Pouliquen et al. (2004, 2006) kombinieren die räumliche Distanz zwischen Toponymen mit der textlichen Distanz. Sie gehen daher von einer Korrelation dieser beiden Distanzen aus, welche jedoch nicht untersucht wurde. Trotzdem stellen Pouliquen et al. (2006) eine Precision für eine isolierte Anwendung ihrer „minimum kilometric distance“-Methode von 0.687 fest.

6 Schlussfolgerung

6.1 Erreichtes

In dieser Masterarbeit wurde der aktuelle Forschungsstand in Bezug auf die Toponym Recognition sowie die Toponym Resolution aufgezeigt, und speziell räumliche Disambiguationsmethoden kritisch diskutiert. Der Gazetteer „SwissNames“ wurde statistisch im Hinblick auf den Anteil an Ambiguität untersucht und mit bisherigen statistischen Untersuchungen verglichen. Zudem wurde die räumliche Verteilung von mehrdeutigen Toponymen aus verschiedenen Gazetteers extrahiert und statistisch mit einem Mann-Whitney-U-Test auf die Ähnlichkeit mit einer zufälligen Verteilung getestet.

Ein GIR für deutsche Texte wurde in Java, basierend auf GATE programmiert. Dabei wurden db4o-Datenbanken mit eingebunden sowie auch der Webservice von Geonames verwendet. Das programmierte GIR-System wurde anhand von Benchmarks getestet und mit bisherigen GIR-Systemen verglichen. Nach der Anwendung dieses GIR-Systems auf Artikel der Südostschweiz wurde die räumliche Ausdehnung (Scope) der Südostschweizartikeln des Jahres 2006 gemessen. Diese Ausdehnung wurde dann mit der räumlichen Verteilung von mehrdeutigen Toponymen verglichen und dieser Unterschied statistisch auf Signifikanz getestet. Die Ergebnisse des GIR-Systems wurden in ArcGIS 9.2 anhand von Dichteoberflächen visualisiert. Weiter wurden Boxplots der Scopes von Südostschweizartikeln in Abhängigkeit von verschiedenen Toponymen erstellt. Diese Ergebnisse wurden dann zu interpretieren versucht.

6.2 Erkenntnisse

Ein Gazetteer Lookup in Kombination mit Stopword-Listen und manuell erstellten Regeln ist eine geeignete Methode für das Geo-coding von deutschen Texten. Sowohl Precision als auch Recall lassen sich mit anderen Arbeiten vergleichen. Bei der Analyse der Ambiguität von Gazetteers konnte das von Leidner (2008) beobachtete, umgekehrt-exponentielle Verhältnis von Ambiguitätsgrad mit dessen Häufigkeit bestätigt werden. Weiter hat diese Analyse ergeben, dass mehrdeutige Toponyme nicht zufällig über den Raum verteilt sind, sondern räumlich korrelieren.

Die Scopes von Südostschweizartikeln sind durchschnittlich um 29.3% grösser als die Scopes der Referenten von mehrdeutigen Toponymen in SwissNames. Dies hat zur Folge, dass es unwahrscheinlich ist, ein mehrdeutiges Toponym allein über die räumlichen Eigenschaften derer Referenten aufzulösen. Daher muss die Anwendung von räumlichen Disambiguationsmethoden an strikte Bedingungen gebunden sein. Diese Untersuchung von Südostschweizartikeln hat weiter ergeben, dass Artikel, welche grössere Ortschaften oder Länder beinhalten tendenziell ein grösseres Scope als Artikel, welche kleinere Ortschaften beinhalten, haben.

6.3 Ausblick

Unter den Methoden, welche räumliche Informationen zur Disambiguation von Geo-Geo-Ambiguität verwenden gibt es einige (Rauch et al. 2003, Pouliquen et al. 2004), welche neben der räumlichen Nähe auch die textliche Nähe verwenden. Dabei gehen die Autoren ungeprüft davon aus, dass diese beiden Faktoren korrelieren. Diese Annahme wurde aus zeitlichen Gründen in dieser Arbeit nicht mehr überprüft. Es wird jedoch vorgeschlagen, eine Korrelationsanalyse dieser beiden Distanzen, normalisiert auf den Wert 1, vorzunehmen, um die Validität der Methoden von Pouliquen et al. 2004 sowie Rauch et al. 2003 zu überprüfen.

Weiter wäre es interessant, um für ein räumlich gebundenes Medium wie eine physische Zeitung die Distance der Referenten zum Veröffentlichungsort der Zeitung zu untersuchen. Dies aufgrund der Beobachtung, dass z.B. Typonyme wie „Glarus“ oder „San Bernardino“ sowohl in der Schweiz als auch in den USA vorkommen. Für den Schweizer Leser müssen aber nur die US-Amerikanischen Referenten textlich disambiguiert werden und für den Leser einer US-Amerikanischen Zeitung nur die Schweizer Referenten. Falls sich diese Beobachtung beweisen liesse, könnte man dadurch räumlich spezifische Default-Referenten ableiten.

Wie Leidner (2008) schon beobachtete, werden die Toponym Resolution-Methoden im Allgemeinen nicht an das jeweilige Toponym angepasst. Dadurch könnte eine sichere Methode verfälscht werden oder auch ganz verloren gehen. Zukünftige GIR-Anwendungen sollten daher eine adaptive Toponym Resolution anwenden. Diese müsste für die Referenten jedes Toponyms gewisse Kriterien überprüfen, um danach die geeignete Methode anzuwenden. So ist für die Stadt „San Francisco“ eine Default-

Referent-Methode wohl sehr präzise (Kapitel 2.3.5), da meist die Stadt im US-Bundesstaat Kalifornien gemeint ist, während sie für „Springfield“ gänzlich ungeeignet ist (Amitay et al. 2004, Figur 30).

7 Literatur

Amitay, E., Har'El, N., Sivan, R. & Soffer, A. (2004): Web-a-Where: Geotagging Web content. In: Sanderson, M., Jàrvelin, K., Allan, J. & Bruza, P. (Hrsg.) (2004): SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, 25-29 Juli, S. 273-280. ACM Press.

Appelt, D. E. (1996): The Common Pattern Specification Language. Technical report, SRI International, Artificial Intelligence Center.

Bender, O., Och, F. J. & Ney, H. (2003): Maximum entropy models for named entity recognition. In: Daelemans, W. & Osborne, M. (Hrsg.) (2003): Proceedings of CoNLL 2003, Edmonton, Alberta, Kanada, S. 148-151.

Bischoff, K., Mandl, T. & Womser-Hacker, C. (2006). Blind Relevance Feedback and Named Entity based Query Expansion for Geographic Retrieval at GeoCLEF 2006. In: Peter, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D. W., de Rijke, M. & Stempfhuber, M. (Hrsg.) (2006): Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006, Alicante, Spanien, 20-22 September, S. 946-953. Springer.

Borthwick, A., Sterling, J., Agichtein, E. & Grishman, R. (1998): NYU: Description of the MENE named entity system as used in MUC-7. In: Seventh Message Understanding Conference (MUC 7): Proceedings of a Conference held in Fairfax, Virginia, USA, 19 April-1 Mai.

Bronstein, I. N., Semendjajew, K. A., Musiol, G. & Mühlig H. (1999): Taschenbuch der Mathematik, vierte Ausgabe, Frankfurt am Main. Verlag Harri.

Buckley, C. & Voorhees, E. M. (2000); Evaluation measure stability. In: Belkin, N. J., Ingwersen, P. & Leong, M.-K. (Hrsg.) (2000): Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athen, Griechenland, 24-28 Juli, S. 33-40. ACM Press.

Clough, P. (2005): Extracting metadata for spatially-aware information retrieval on the Internet. In: Jones, C. R. & Purves, R. S. (Hrsg.) (2005): Proceedings of the ACM Workshop on Geographic Information Retrieval (GIR) held at the Conference on Information and Knowledge Management (CIKM), Bremen, Deutschland, 4. November, S. 25-30. ACM Press.

Cuchiarelli, A., Luzi, D. & Velardi, P. (1998): Automatic semantiv tagging of unknown proper names. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and Proceedings of the 17th International Conference on Computational Linguistics, Montréal, Kanada, 10-14 August, S. 286-292. ACL.

Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V. (2002): GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. & Ursu, C. (2007): Developing Language Processing Components with GATE Version 4 (a User Guide) http://gate.ac.uk/sale/tao/index.html, Zugriff 10.02.2008.

Emerson, J. D. & Strenio, J. (1983): Boxplots and Batch Comparisons. In: Hoaglin, D. C., Mosteller, F. & Tukey, J. W. (Hrsg.) (1983): Understanding Robust and Exploratory Data Analsysis. New York, USA, S. 58-96. Wiley.

Ferrés, D. D. (2007): Geographical Information Resolution and its Application to the Question Answering Systems, Universitat Politècnica de Catalunya.

Fonseca, F. T., Egenhofer, M. J., Agouris, P. & Câmara, G. (2002). Using ontologies for integrated geographic information systems. In: Transactions in Geographic Information Systems, 6 (3), S. 231-257.

Gale, W. A., Church, K. W. & Yarowsky, D. (1992): One sense per discourse. In: Proceedings of the Fourth DARPA Speech and Natural Language Workshop. Defense Advanced Research Projects Agency, Morgan Kaufmann, San Mateo, Kalifornien, USA, S. 233-237.

Gey, F., Larson, R. R., Sanderson, M., Bischoff, K., Mandl, T., Womser-Hacker, C., Santos, D., Rocha, P., Di Nunzio, G. & Ferro, N. (2006): GeoCLEF 2006: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: Peter, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D. W., de Rijke, M. & Stempfhuber, M. (Hrsg.) (2006): Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006, Alicante, Spanien, 20-22 September, S. 852-876. Springer.

Gill, A. (1962): Introduction to the Theory of Finite State Machines. McGraw-Hill, New York.

Gould, P. R. (1979): Geography 1957-1977: The Augean period. In:Annals of the Association of American Geographers, 69 (1), S. 139-151.

Harman, D. (1992): The DARPA TIPSTER project. In:. Belkin, N. J., Ingwersen, P. & Pejtersen, A. M. (1992): Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Kopenhagen, Dänemark, 21-24 Juni, S. 1-10.

Harpring, P. (1997): Proper words in proper places: The Thesaurus of Geographic Names. MDA Information, 2 (3), S. 5-12.

Hauptmann, A. G. & Olligschlaeger, A. M. (1999): Using Location Information from Speech Recognition of Television News Broadcasts. In: Robinson, T. & Renals, S. (Hrsg.) (1999): Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio. Cambridge, Grossbritannien, 19-20 April, S. 102-106.

Hill, L. L. (2000): Core elements of digital gazetteers: placenames, categories, and footprints. In: Borbinha, J. L. & Baker, T. (Hrsg.) (2000): Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), S. 280-290. Springer, London, Grossbritannien.

Jones, C. B., Alani, H. & Tudhope, D. (2001): Geographical Information Retrieval with Ontologies of Place. In: Montello, D. R. (Hrsg.) (2001): Spatial Information Theory Foundations of Geographic Information Science, COSIT 2001, 2205, S. 323-335. Springer, London, Grossbritannien.

Jones, C. B. & Purves, R. (2005): Proceedings of the ACM Workshop on Geographic Information Retrieval (GIR) held at the Conference on Information and Knowledge Management (CIKM), Bremen, Deutschland, 4. November. ACM Press.

Kent, A., Berry, M. M., Leuhrs, F. U. & Perry, J. W. (1955): Machine literature searching VIII: Operational criteria for designing information retrieval systems. American Documentation, 6 (2). S. 93-101.

Kilgarriff, A. & Rosenzweig, J. (2000): Framework and Results for English SENSEVAL. In: Computers and the Humanities, 34 (1-2). S. 15-48.

Knuth, D. E. (2000): Sorting and searching, Zweite Ausgabe. Addison-Wesley, Boston.

Kornai, A. & Thompson, B. (2005): Size doesn't matter, unveröffentlichter Entwurf, http://www.kornai.com/Drafts/size.pdf, Zugriff 10.04.2008.

Krupka, G. R. & Hausman, K. (1998): Isoquest, Inc: Description of the NetOwl extractor system as used for MUC-7. In: 7th Message Understanding Conference (MUC 7): Proceedings of a Conference held in Fairfax, Virginia, USA, 19 April-1 Mai.

Lakoff, G. & Johnson, M. (1980): Metaphors we live by. Chicago University Press.

Larson, R. R. (1996): Geographic Information Retrieval and Spatial Browsing. In: Smith, L. & Gluck, M. (Hrsg.) (1996): GIS and Libraries: Patrons, Maps and Spatial Information, University of Illinois, S. 81-124.

Lawrence, R. R. (1990): A tutorial on hidden Markov models and selected applications in speech recognition. In: Readings in speech recognition, S. 267-296. Morgan Kaufmann Publishers Inc., San Francisco, Kalifornien, USA.

Lee, S. & Lee G. G. (2005): A Bootstrapping Approach for Geographic Named Entity Annotation. In: Lee, G. G., Yamada, A., Meng, H. & Myaeng, S.-H. (Hrsg.) (2005): Information Retrieval Technology, Proceedings of the second Asia Information Retrieval Symposium, Airs 2005, Jeju, Korea, 13-15 Oktober, S. 178-189.

Leidner, J. (2005): Experiments with geo-filtering predicates for information retrieval. In: Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G. J. F., Kluck, M., Magnini, B. & de Rijke, M. (Hrsg.) (2006): Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, 21-23 September, Wien, Österreich, S. 987-996. Springer.

Leidner, J. (2008): Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Dissertation.com.

Leidner, J., Sinclair, G. & Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In: Kornai, A. & Sundheim, B. (2003): HLT-NAACL 2003 Workshop: Analysis of Geographic References, S. 31-38. Association for Computational Linguistics, Edmonton, Alberta, Kanada.

Leveling, J. & Hartrumpf, S. (2006): On metonymy recognition for geographic information retrieval. In: Purves, R. S. & Jones, C. B. (Hrsg.) (2006):

Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, USA, 10 August. ACM Press.

Leveling, J. & Veiel, D. (2006): Experiments on the Exclusion of Metonymic Location names from GIR, CLEF 2006. In: Peter, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D. W., de Rijke, M. & Stempfhuber, M. (Hrsg.) (2006): Working Notes of the Cross-Lingual Evaluation Forum (CLEF) 2006, Alicante, Spanien, 20-22 September, S. 901-904. Springer.

Li, H., Srihari, R. K., Niu, C. & Li, W. (2002): Location Normalization for Information Extraction. In: COLING 2002.

Li, H., Srihari, R. K., Niu, C. & Li, W. (2003): InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Kornai, A. & Sundheim, B. (Hrsg.) (2003): HLT-NAACL 2003 Workshop: Analysis of Geographic References, S. 39-44. Association for Computational Linguistics, Edmonton, Alberta, Kanada.

Mann, H. & Whitney, D. (1947): On a test of whether one of two variables is stochastically larger than the other. In: Annals of Mathematical Statistics 18. S. 50-60.

Markert, K. & Nissim, M. (2002): Towards a corpus annotated for metonymies: the case of location names. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), S. 1385–1392. ELRA, Paris, Frankreich.

Markowetz, A., Brinkhoff, T. & Seeger, B. (2004): Geographic Information Retrieval. 3rd International Workshop on Web Dynamics, New York, 18 Mai.

Mikheev, A., Moens, M. & Grover, C. (1999): Named Entity Recognition without Gazetteers. In: Proceedings of EACL, Bergen, Norwegen. EACL.

O'Sullivan, D. & Unwin, D. J. (2003): Geographic Information Analysis. John Wiley & Sons, Inc., Hoboken, New Jersey.

Overell, S. E. & Rüger, S. (2006): Identifying and grounding descriptions of places. In: SIGIR Workshop on Geographic Information Retrieval, S. 14–16.

Pasley, R., Clough, P. & Sanderson, M. (2007): Geo-Tagging for Imprecise Regions of Different Sizes. In: Purves, R. S. & Jones, C. (Hrsg.) (2007): Proceedings of the 4th ACM Workshop on Geographic Information Retrieval GIR'07, Lissabon, Portugal, 9 November, S. 77-82. ACM Press.

Purves, R. S., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S. & Yang, B. (2007): The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. In: Fisher, P., Gahegan, M. & Lees, B. (Hrsg.) (2007): International Journal of Geographical Information Science, 21 (7), S. 717-745. Taylor & Francis.

Purves, R. S. & Jones, C. B. (2004): Workshop on geographic information retrieval at SIGIR 2004. SIGIR Forum, 38, S, 53–56.

Purves, R. S. & Jones, C. B. (2006): Geographic Information Retrieval. In: Longley, P. (Hrsg.) (2006): Computers, Environment and Urban Systems, 30, S. 375-377. Elsevier.

Pouliquen, B., Steinberger, R., Ignat, C. & De Groeve, T. (2004): Geographical information recognition and visualization in texts written in various languages. In: Haddad, H., Omicini, A., Wainwright, R. L., Liebrock & L. M. (Hrsg.) (2004): Proceedings of the 2004 ACM Symposium on Applied Computing, S. 1051–1058. ACM Press.

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fluart, F., Zaghouani, W., Widiger, A., Forslund, A.-C. & Best, C. (2006): Geocoding multilingual texts: Recognition, disambiguation and visualisation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), S. 53–58. ELRA.

Rauch, E., Bukatin, M. & Baker, K. (2003): A confidence-based framework for disambiguating geographic terms. In: Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographical references, S. 50-54. ACL.

Rössler, M. (2004): Corpus-based Learning of Lexical Resources for German Named Entity Recognition. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), Lissabon, Portugal.

Schell, D. (1999): About Open GIS Consortium. In: Open GIS Consortium – spatial Connectivity for a Changing World. OGC Press, Wayland.

Schockaert, S., De Cock, M. & Kerre, E. E. (2006): Towards Fuzzy Spatial Reasoning in Geographic IR Systems. Workshop on geographic Information Retrieval, SIGIR, S. 34-36.

Smith, D. A. & Crane G. (2001): Disambiguating geographic names in a historical digital library. In: Research and Advanced Technology for Digital Libraries: 5th European Conference (ECDL 2001), S. 127-136.

Smith, D. A. & Mann, G. S (2003): Bootstrapping toponym classifiers. In: Kornai, A. & Sundheim, B. (Hrsg.) (2003): HLT-NAACL 2003 Workshop: Analysis of Geographic References, S. 45-49. Association for Computational Linguistics, Edmonton, Alberta, Kanada.

Spiess, E. (2006): Schweizer Weltatlas. Spiess, E. (Hrsg.) (2006): Konferenz der kantonalen Erziehungsdirektoren (EDK) Zürich. Lehrmittelverlag des Kantons Zürich.

Stevenson, M. & Gaizauskas, R. (1999): Using Corpus-derived Name Lists for Named Entity Recognition. In: Proceedings of the 6th Applied Natual Language Processing Conference and the First Meeting of the North American Chapter of the Association for Computational Linguisitcs, Seattle, Washington, USA, S. 290-295.

Strötgen, R., Mandl, T. & Schneider, R. (2005): A Fast Forward Approach to Cross-lingual Question Answering for English and German. In: Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G. J. F., Kluck, M., Magnini, B. & de Rijke, M. (Hrsg.) (2006): Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, 21-23 September, Wien, Österreich, S. 332-336. Springer.

Sui, D. Z. (2004): Tobler's First Law of Geography: A Big Idea for a Small World? In: Annals of the Association of American Geographers 94, S. 269-277.

Tobler, W. R. (1970): A computer movie simulation urban growth in the Detroit region. In: Economic Geography 46, S. 234-240.

Van Rijsbergen, C. J. (1979): Information retrieval. London, Butterworths.

Zheng, G. D. & Su, J. (2002): Named entity tagging using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, 6-12 Juli, S. 209-219. ACL.

8 Anhang

8.1 Mann-Whitney Tests

8.1.1 SwissNames (Total)

Ranks

	V2	N	Mean Rank	Sum of Ranks
SwissNames	Zufällig	100	125.35	12535.00
	Mehrdeutig	100	75.65	7565.00
	Total	200		

Test Statistics[a]

	sn
Mann-Whitney U	2515.000
Wilcoxon W	7565.000
Z	-6.072
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: V2

8.1.2 SwissNames (Deutsch)

Ranks

	V2	N	Mean Rank	Sum of Ranks
SwissNames (Deutsch)	Zufällig	100	117.83	11783.00
	Mehrdeutig	100	83.17	8317.00
	Total	200		

Test Statistics[a]

	sn_de
Mann-Whitney U	3267.000
Wilcoxon W	8317.000
Z	-4.234
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: V2

8.1.3 SwissNames (Französisch)

Ranks

	V2	N	Mean Rank	Sum of Ranks
SwissNames (Französisch)	Zufällig	100	113.07	11306.50
	Mehrdeutig	100	87.94	8793.50
	Total	200		

Test Statistics[a]

	sn_fr
Mann-Whitney U	3743.500
Wilcoxon W	8793.500
Z	-3.070
Asymp. Sig. (2-tailed)	.002

a. Grouping Variable: V2

8.1.4 SwissNames (Italienisch)

Ranks

	V2	N	Mean Rank	Sum of Ranks
SwissNames (Italienisch)	Zufällig	100	111.48	11148.00
	Mehrdeutig	100	89.52	8952.00
	Total	200		

Test Statistics[a]

	sn_it
Mann-Whitney U	3902.000
Wilcoxon W	8952.000
Z	-2.683
Asymp. Sig. (2-tailed)	.007

a. Grouping Variable: V2

8.1.5 Ordnance Survey (Grossbritannien)

Ranks

	V2	N	Mean Rank	Sum of Ranks
GB (Ordnance Survey)	Zufällig	100	130.98	13098.00
	Mehrdeutig	100	70.02	7002.00
	Total	200		

Test Statistics[a]

	os
Mann-Whitney U	1952.000
Wilcoxon W	7002.000
Z	-7.447
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: V2

8.1.6 Geonames (Schweiz)

Ranks

	V2	N	Mean Rank	Sum of Ranks
CH (Geonames)	Zufällig	100	124.30	12429.50
	Mehrdeutig	100	76.71	7670.50
	Total	200		

Test Statistics[a]

	ch_gn
Mann-Whitney U	2620.500
Wilcoxon W	7670.500
Z	-5.815
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: V2

8.1.7 Geonames (Grossbritannien)

Ranks

	V2	N	Mean Rank	Sum of Ranks
GB (Geonames)	Zufällig	100	128.42	12842.00
	Mehrdeutig	100	72.58	7258.00
	Total	200		

Test Statistics[a]

	gb_gn
Mann-Whitney U	2208.000
Wilcoxon W	7258.000
Z	-6.822
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: V2

8.1.8 Geonames (USA)

Ranks

	V2	N	Mean Rank	Sum of Ranks
USA (Geonames)	Zufällig	1000	1060.44	1060443.00
	Mehrdeutig	1000	940.56	940557.00
	Total	2000		

Test Statistics[a]

	US_GN
Mann-Whitney U	440057.000
Wilcoxon W	940557.000
Z	-4.642
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: V2

8.1.9 Südostschweizscopes verglichen mit mehrdeutigen Toponymen aus SwissNames

Ranks

	V2	N	Mean Rank	Sum of Ranks
Scopes	SO	100	118.75	11875.00
	SN	100	81.06	8025.00
	Total	200		

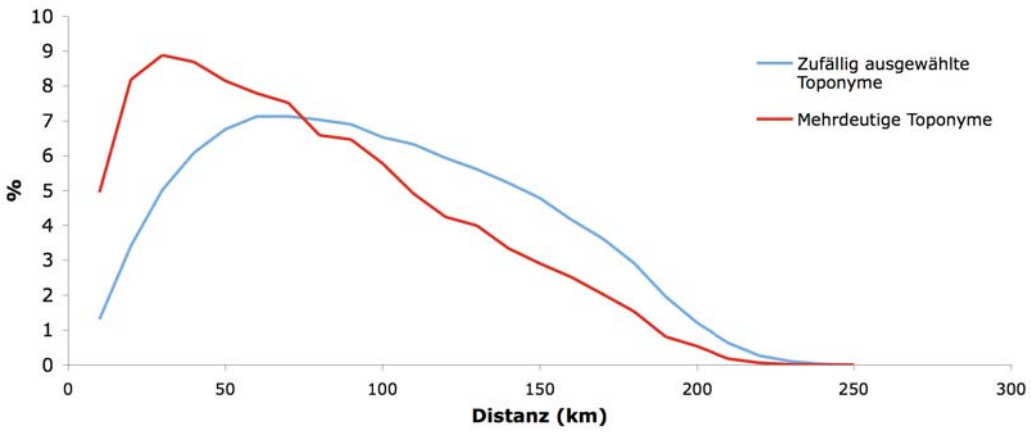
Test Statistics[a]

	V1
Mann-Whitney U	3075.000
Wilcoxon W	8025.000
Z	-4.616
Asymp. Sig. (2-tailed)	.000

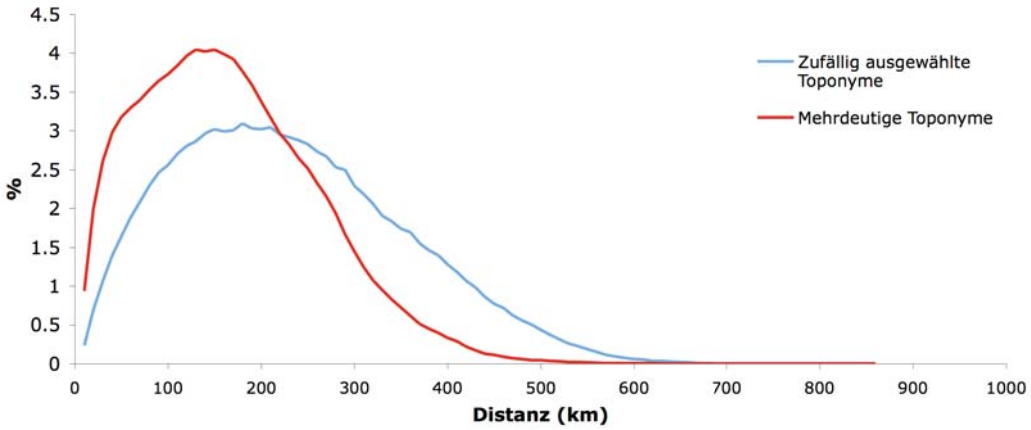
a. Grouping Variable: V2

8.2 Häufigkeitsverteilungen von mehrdeutigen Toponymen

8.2.1 Wales



8.2.2 England


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
%
0
100
200
300
400
500
600
700
800
900
1000
Distanz (km)
Zufällig ausgewählte Toponyme
Mehrdeutige Toponyme