Extrahierung von Landschaftswahrnehmungen aus Tourenberichten

---

Vorstellung einer neuen Analysemethode basierend auf Sentiment Analysis und Dependenzgrammatik

Masterarbeit GEO 511

30. September 2015

Verfasser: **Ramón Huldi** (10-731-149)

Betreuung: **Prof. Dr. Ross Purves**, **M. Sc. Flurina Wartmann**

Extrahierung von Landschaftswahrnehmungen aus Tourenberichten

Vorstellung einer neuen Analysemethode basierend auf Sentiment Analysis und Dependenzgrammatik

Masterarbeit GEO 511

Zürich, 30. September 2015

Verfasser:

Ramón Huldi

Stiegackerstrasse 6

CH – 8362 Balterswil

ramonhuldi@bluewin.ch

Matrikelnummer: 10-731-149

Betreuung:

Prof. Dr. Ross Purves

ross.purves@geo.uzh.ch

M. Sc. Flurina Wartmann

flurina.wartmann@geo.uzh.ch

Vorwort

Diese Masterarbeit stellt den letzten Schritt zum Abschluss meines Studiums an der Universität Zürich im Hauptfach Geographie dar. Im Zuge dieses Studiums habe ich meine Spezialisierung in den Geographischen Informationswissenschaften gemacht. In diesem aufstrebenden Fachgebiet erweitern neue Technologien ständig die Möglichkeiten zur Behandlung geographischer Problemstellungen. Darin war auch meine Motivation für diese Arbeit begründet. Sie stellt einen Versuch dar, Techniken aus dem Gebiet der Sentiment Analysis dafür zu nutzen, um die Wahrnehmung und Interpretation des geographischen Raumes durch den Menschen auf eine neue Art und Weise zu untersuchen.

Beim Verfassen dieser Arbeit konnte ich auf die Unterstützung einiger Personen zählen. Sie haben mir Ratschläge gegeben, für willkommene Ablenkung gesorgt oder mir den Rücken freigehalten. Im Speziellen möchte ich folgenden Personen danken:

- ***Ross Purves*** für die methodischen und inhaltlichen Ratschläge während den zahlreichen Besprechungen.
- ***Flurina Wartmann*** für die wertvollen Tipps zum Inhalt, aber auch in allen anderen Belangen rund um die Arbeit.
- ***André Bruggmann*** für zahlreiche nützliche Inputs rund um die Thematik der Sentiment Analysis.
- ***Guido Bruggmann*** für das Korrekturlesen.
- ***Remo Beerli*** für das Gegenlesen und die wertvollen Feedbacks.
- ***Adela Huldi***, ***Reto Huldi***, ***Marisa Huldi*** und ***Bruno Huldi*** für die Unterstützung und den Rückhalt während des gesamten Studiums.
- ***Patrice Frei***, ***Oliver Deseö*** und ***Jonas Hänseler*** für den regen Austausch während der Arbeit und die unterhaltsamen Stunden während der gesamten Studienzeit.
- ***Allen Umfrageteilnehmenden*** für die Bereitschaft zur Teilnahme an meiner Studie.

Zusammenfassung

In diversen Fachgebieten wie der Geographie, der Ethnophysiographie oder auch der Landschaftsbewertung ist man daran interessiert, wie Menschen Landschaften wahrnehmen und beschreiben. Diese Beschreibungen variieren sowohl mit der Landschaft selbst als auch zwischen den verschiedenen Menschen, Sprachgruppen oder Kulturen. Besonders in der Landschaftsbewertung wird dabei ein besonderes Interesse den ästhetischen Eigenschaften von Landschaften zuteil. Um zu den gewünschten Informationen zu gelangen, werden oft Befragungen oder Experimente mit Probanden und Probandinnen durchgeführt. Solche Methoden erlauben es zwar, dass Landschaftswahrnehmungen sehr differenziert aufgenommen werden können. Sie sind aber aufwändig und die räumliche und zeitliche Abdeckung ist beschränkt, weshalb man auf Verallgemeinerungen angewiesen ist.

Im Gebiet des Geographic Information Retrieval (GIR) existieren einige Arbeiten, in welchen die Interpretation und Kategorisierung des geographischen Raumes durch die Untersuchung von sogenanntem user generated content (UGC) analysiert wird. Dabei beschäftigt man sich oft damit, welche Ausdrücke von den Leuten verwendet werden, um natürliche Objekte in der Landschaft zu benennen. Wenige konzentrieren sich aber auf diejenigen Ausdrücke, welche die Eigenschaften solcher Landschaftselemente beschreiben. Dem Autor ist weiter nur eine Arbeit bekannt, in welcher die Schönheit von Landschaften mithilfe eines Ansatzes des Opinion Mining geschätzt wurde. In *Nowak* (2013) konnten dafür sowohl Textquellen als auch quantitative Bewertungen herangezogen werden, was in diesem Kontext selten der Fall ist.

Diese Arbeit untersucht eine Methodik, mit welcher aus einem nutzergenerierten Textkorpus Beschreibungswörter extrahiert werden können, welche sich auf Landschaftselemente beziehen. Dafür wird eine eigens dafür konzipierte Anwendung namens SentiTours entwickelt. Diese Anwendung erkennt mithilfe der Regeln der Dependenzgrammatik Wortbeziehungen in Texten. Als Untersuchungsgegenstand dienen die Berichte der Tourenplattform Hikr. Die extrahierten Landschaftsbeschreibungen werden auf räumliche Muster und weitere Eigenheiten untersucht. Des Weiteren wird der Versuch unternommen, mit einer Form der Sentiment Analysis die Meinungen der Leute gegenüber den beschriebenen Landschaften zu schätzen, indem die Polaritätswerte der zur Beschreibung verwendeten Adjektive in einer sogenannten Sentiment-Bibliothek nachgeschlagen werden.

Es zeigt sich, dass die Beschreibungswörter verschiedener Landschaftselemente sich unterscheiden und dass diese Unterschiede dargestellt und untersucht werden können. Des Weiteren lassen sich charakteristische Eigenschaften von verschiedenen Regionen miteinander vergleichen. Es wird ausserdem aufgezeigt, dass gewisse Landschaftselemente in verschiedenen Regionen unter-

schiedlich beschrieben werden. Zudem werden einige Erkenntnisse gewonnen über die Anwendung einer Sentiment Analysis im Kontext der Landschaftsbewertung. Dazu gehört, dass einige Ausdrücke im Zusammenhang mit Alpintouren nicht dieselbe Bedeutung haben wie im allgemeinen Sprachgebrauch. Somit wird für eine weiterführende Arbeit die Entwicklung einer an den Tourenkontext angepassten Sentiment Bibliothek vorgeschlagen.

Der Beitrag dieser Arbeit besteht darin, dass ein neues Vorgehen zur automatisierten Extrahierung von Landschaftsbeschreibungen präsentiert wird. Darauf aufbauend wird zudem ein erstes Mal untersucht, inwiefern sich Meinungen über Landschaften aus unstrukturierten Texten erkennen lassen, wenn keine quantitative Bewertung der Landschaft als Referenz vorliegt.

Inhalt

Abbildungen

Tabellen

1. Einleitung

1.1 Motivation

«*Der Unterschied zwischen Landschaft und Landschaft ist klein; doch gross ist der Unterschied zwischen den Betrachtern*» (*Emerson*, 1987).

Nicht nur der amerikanische Philosoph Ralph Waldo Emerson beschäftigte sich mit der Landschaftswahrnehmung. In diversen anderen Disziplinen möchte man wissen, wie Landschaften von den Menschen interpretiert und beschrieben werden; dies nicht zuletzt auch in ästhetischer Hinsicht. Die Kenntnisse darüber können von grosser Bedeutung sein. Zum Beispiel dann, wenn es darum geht, den Wert einer Landschaft zu ermitteln. In der Schweiz wird diese Aufgabe den Kantonen sogar gesetzlich auferlegt, wie im Raumplanungsgesetz nachzulesen ist: «*Für die Erstellung ihrer Richtpläne erarbeiten die Kantone Grundlagen, in denen sie feststellen, welche Gebiete … besonders schön, wertvoll, für die Erholung oder als natürliche Lebensgrundlage bedeutsam sind*» (*Bundesgesetz über die Raumplanung*, 1979). Bei der Ethnophysiographie handelt es sich sogar um ein eigenes Fachgebiet, in welchem untersucht wird, wie Menschen die natürliche Umwelt beschreiben. Sowohl in der Landschaftsbewertung als auch in der Ethnophysiographie werden Erkenntnisse häufig über Experimente mit Probanden und Probandinnen gewonnen. Sollen jedoch Beschreibungen einer grossen Anzahl Landschaften über verschiedene Zeiträume hinweg erfasst werden, stösst man mit dieser Methode an Grenzen.

Im Bereich des «Geographic Information Retrieval» (GIR) wurde verschiedentlich gezeigt, dass mit gewissen Anwendungen Informationen darüber gewonnen werden können, wie Menschen ihre Umwelt interpretieren. Dies geschieht häufig durch die automatisierte Analyse von Textquellen. Dabei wird zum Beispiel untersucht, welche Begriffe von den Leuten verwendet werden, um natürliche Objekte in ihrer Umwelt zu benennen (*Derungs & Purves*, 2013). Es gibt jedoch kaum Arbeiten, welche sich auf die Eigenschaften oder die Meinungen konzentrieren, welche diesen Landschaftselementen zugewiesen werden. Damit könnten jedoch interessante Rückschlüsse auf den Charakter oder die wahrgenommene Schönheit von Landschaften gemacht werden. Mit der Arbeit von *Nowak* (2013) ist dem Autor zwar ein Ansatz bekannt, in welchem versucht wurde, die Schönheit von beschriebenen Landschaften zu schätzen. Dabei konnten neben Textquellen jedoch auch quantitative Bewertungen als Referenz herangezogen werden. Eine solche Datengrundlage ist jedoch selten gegeben. Aus diesem Grund verfolgt die vorliegende Arbeit einen neuen Ansatz zur Erfassung von Landschaftsbeschreibungen einer grossen Anzahl Leute über ausgedehnte räumliche und zeitliche Skalen hinweg. Dazu stützt sie sich auf nutzergenerierte Texte im Internet.

Bei diesen Texten handelt es sich um Tourenberichte der Plattform «Hikr». Die Einträge enthalten Beschreibungen der begangenen Routen und der erlebten Landschaft. Mit computerlinguistischen Methoden sollen diese Landschaftsbeschreibungen extrahiert werden. In einem zweiten Schritt soll überprüft werden, ob die Bewertung der Schönheit der verschiedenen Landschaften mithilfe einer Variante der «Sentiment Analysis» möglich und auch sinnvoll ist.

Die Motivation dieser Arbeit besteht darin, eine Möglichkeit zu prüfen, Beschreibungen und Bewertungen von Landschaften anhand einer reinen Textquelle rechentechnisch zu untersuchen. Ein solches Vorgehen würde es ermöglichen, die Wahrnehmung der Landschaft über grosse räumliche und zeitliche Skalen, sowie über verschiedene Bevölkerungsgruppen hinweg zu erfassen. Die erhaltenen Resultate könnten in verschiedenen Disziplinen der theoretischen und angewandten Landschaftsbewertung neue Einblicke gewähren oder als Ergänzung zu bestehenden Methoden dienen.

1.2 Forschungsfragen

Das Ziel dieser Arbeit ist es, zu überprüfen, ob sich die Meinung von Leuten gegenüber Landschaften basierend auf der Extrahierung ihrer Beschreibungen ermitteln lässt. Diese Zielsetzung lässt sich in zwei verschiedene Forschungsfragen gliedern:

Forschungsfrage 1:
Inwiefern lassen sich Beschreibungen von Landschaften aus Texten rechentechnisch extrahieren?

- Welche Muster lassen sich in diesen Beschreibungen erkennen?

Forschungsfrage 2:
Inwiefern ist es möglich, mithilfe der Technik der Sentiment Analysis die Meinungen von Menschen gegenüber beschriebenen Landschaften zu ermitteln?

1.3 Untersuchungs- und Themengebiet

Abb. 1: Lage des Untersuchungsgebiets (grün eingefärbt) in Europa (*www.nouahsark.com*).

Das Untersuchungsgebiet für diese Arbeit ist die Schweiz (siehe Abbildung 1). Diese bietet mit ihren Alpinlandschaften eine interessante natürliche Grundlage für die Analyse von Landschaftsbeschreibungen. Ausserdem ist die Datenlage für die Schweiz mit 42‘286 von weltweit 71‘432

(Stand: 21.05.2015) Berichten auf Hikr grosszügig. Von diesen Texten werden nur diejenigen in deutscher Sprache analysiert.

1.4 Gliederung der Arbeit

Im zweiten Kapitel wird der Forschungsstand der für diese Arbeit relevanten Themengebiete ausgeleuchtet. Daraufhin werden die Eigenschaften des Hikr-Korpus genauer analysiert (Kapitel 3). In Kapitel 4 werden aus dieser Analyse und den Erkenntnissen aus der Literaturrecherche die Implikationen für die Methodik abgeleitet, welche dann in Kapitel 5 ausführlich erläutert wird. Anschliessend werden im Kapitel 6 die Resultate präsentiert und ausgewertet. In der Diskussion (Kapitel 7) werden die Resultate analysiert sowie allfällige Schwierigkeiten diskutiert. In der Schlussfolgerung (Kapitel 8) werden dann die Forschungsfragen beantwortet. Zum Schluss werden im Ausblick (Kapitel 9) mögliche Ansätze zukünftiger Arbeiten in diesem Gebiet diskutiert.

2. Wissenschaftlicher Hintergrund

In dieser Arbeit werden zwei sehr unterschiedliche Fachgebiete miteinander verknüpft, nämlich die Landschaftsbewertung und das «Information Retrieval» respektive die Sentiment Analysis. In diesem Kapitel werden die Hintergründe dieser Gebiete erläutert und der Forschungsstand in Bezug auf die vorliegende Arbeit wird untersucht. Dazu wird als Erstes der Landschaftsbegriff eingeführt und diskutiert, um dessen Bedeutung für diese Arbeit etwas einzugrenzen.

2.1 Der Landschaftsbegriff

Der Duden liefert folgende Beschreibung für den Begriff Landschaft: «*Hinsichtlich des äußeren Erscheinungsbildes (der Gestalt des Bodens, des Bewuchses, der Bebauung, Besiedelung o. Ä.) in bestimmter Weise geprägter Teil, Bereich der Erdoberfläche; Gebiet der Erde, das sich durch charakteristische äussere Merkmale von anderen Gegenden unterscheidet*» (*www.duden.de*). Gemäss dieser Definition handelt es sich bei einer Landschaft um einen räumlichen Ausschnitt der Erdoberfläche nicht klar bestimmter Ausdehnung, welcher visuell wahrgenommen werden kann. Gemäss *Granö* (1997) handelt es sich bei der Landschaft um diejenigen Elemente der Umwelt, welche weit vom Betrachter oder der Betrachterin entfernt liegen und vorwiegend mit dem Auge erfasst werden. Andere Autoren argumentieren, dass auch weitere Sinne zur Wahrnehmung der Landschaft beitragen können. Laut *Feld* (1996) können demnach ebenfalls Töne und Gerüche eine Rolle spielen sowie die Wahrnehmung über die Haut (z.B. Temperatur, Wind). *Mark et al.* (2011) stellen zudem das Konzept einer Unterteilung in weite und nahe Elemente in Frage. Sie kritisieren die ungenaue Formulierung und die schwierige Operationalisierbarkeit einer solchen Definition.

Untersuchungen haben gezeigt, dass sich Leute bei der Beschreibung von Landschaften meist auf Teile der Landschaft beziehen. Man spricht hierbei auch von Landschaftselementen *(Tversky & Hemenway*, 1983). Ein solches Konzept der Landschaft als Ganzes, zusammengesetzt aus einzelnen Teilen, bildet die Basis für viele empirische Untersuchungen (*Naveh & Lieberman*, 1984). Um die komplexe Natur des Landschaftskonzepts zu verstehen, ist eine solche Segmentierung notwendig. Es ist aber dennoch die Kombination dieser Elemente, welche eine Landschaft ausmacht (*Mark et al.*, 2011).

Landschaft wird in der Literatur also je nach Anwendungsgebiet und Fokus unterschiedlich definiert. *Antrop* (2005) identifiziert drei Tätigkeitsfelder, welche eigene Perspektiven, Konzepte und Methoden für den Umgang mit dem Begriff Landschaft entwickelt haben. Dazu gehören die Naturwissenschaften (mit der Landschaftsökologie als Schwerpunkt), die Humanwissenschaften und die angewandten Wissenschaften. Zu den Letzteren zählen auch die Landschaftsarchitektur und die -planung. Eine allgemeingültige und vor allem präzise Definition von Landschaft lässt sich also

nur schwer formulieren. Eine für den Kontext von Tourenberichten und somit für diese Arbeit sinnvolle Definition liefert *Palka* (1995), indem er die Landschaft charakterisiert als «the assemblage of human and natural phenomena contained within one's field of view out-of-doors» (*Palka*, 1995). Die Wahrnehmung durch andere Sinne soll jedoch nicht gänzlich ausgeschlossen werden, da auch diese zu Beschreibungen der Landschaft beitragen können. Dazu passend die Definition von *Appleton* (1980): «the environment perceived, especially visually perceived» (*Appleton*, 1980).

2.2 Landschaftsbewertung

Aus einer Landschaft lassen sich viele Werte ableiten, welche es zu erfassen gilt. Sie stellt einen wichtigen Tourismusfaktor dar und trägt auch in erheblichem Masse zur Lebensqualität der lokalen Bevölkerung bei. Diese Werte sind nicht nur ästhetischer Natur. Sie können zum Beispiel auch durch die vorhandene Biodiversität oder durch die kulturelle Bedeutung repräsentiert werden. Diese und weitere Themen können in eine Landschaftsbewertung einfliessen (*Mark et al.*, 2011). Wie oben festgehalten, liegt im Rahmen dieser Arbeit der Fokus aber auf der Bewertung derjenigen Aspekte der Landschaft, welche mit den Sinnen, vorwiegend visuell, wahrgenommen werden. Es handelt sich hierbei um ein Gebiet, in welchem es üblich ist, Erkenntnisse aus Experimenten mit Laien einfliessen zu lassen (*Dakin*, 2003). Inwiefern die Meinungen der Leute in die schlussendlichen Bewertungen einfliessen, unterscheidet sich jedoch je nach Methodik. Dies entspricht auch dem Problemfeld, welchem sich diese Arbeit widmet.

Laut *Daniel* (2001) zeichnet sich die Geschichte der Landschaftsbewertung durch einen Wettkampf zwischen Experten-basierten und Wahrnehmungs-basierten Ansätzen aus. Diese Ansätze unterscheiden sich in ihrer Konzeptualisierung und der relativen Gewichtung der subjektiven Wahrnehmung durch den Menschen gegenüber einer objektiven Bewertung der ästhetischen Qualität einer Landschaft. Dabei stellte sich heraus, dass Wahrnehmungs-basierte Bewertungen im Gegensatz zu Experten-basierten eine höhere Reliabilität aufweisen. D.h., dass anders als bei Letzteren die Variation zwischen den Landschaften grösser war als diejenige zwischen den Bewertungen derselben Landschaft durch verschiedene Leute. Dem Experten-basierten Ansatz liegt die Annahme zugrunde, dass sich die Ästhetik einer beliebigen Landschaft anhand von vordefinierten Kriterien bewerten lässt. Dabei kann es sich um das Vorhandensein von natürlichen Elementen wie Wasser oder Topographie handeln oder auch um Designparameter wie Formen, Linien und Texturen (*Dakin*, 2003). Diesen Kriterien liegen zwar auch Resultate aus Experimenten mit Laien zugrunde. Diese werden jedoch nicht für jede zu bewertende Landschaft neu ermittelt. Während des Bewertungsprozesses kommt die Rolle des Betrachters oder der Betrachterin anderweitig zum Tragen. So wird zum einen die Anzahl Aussichtspunkte ermittelt, von welchen eine Person die entsprechende Landschaft erblicken kann. Zum anderen wird für jede Landschaft ein

Sensitivitätswert berechnet. Dieser basiert auf der Anzahl Betrachter und Betrachterinnen sowie dem Kontext, unter welchem die Landschaft betrachtet wird (*Daniel*, 2001).
Dakin (2003) beschreibt mit dem sogenannten erfahrungsbezogenen Ansatz noch eine weitere Methode zur Landschaftsbewertung. Hierbei wird anstelle der Wirkung der physikalischen Elemente vielmehr die Bedeutung einer Landschaft untersucht. Diese Bedeutung ergibt sich jedoch wiederum nicht nur aus der reinen Sinneswahrnehmung, sondern auch durch die Interaktion des Menschen mit seiner Umwelt.
Die verschiedenen Ansätze verfolgen somit Methoden mit unterschiedlichen Vor- und Nachteilen. Beim Experten-basierten Ansatz entscheidet meist eine dafür geschulte Person anhand vordefinierter Kriterien über den Wert einer Landschaft (*Dakin*, 2003). Dabei werden einmal ermittelte Informationen zur Wahrnehmung der Landschaft verallgemeinert und auf andere Regionen übertragen. Beim Wahrnehmungs-basierten Ansatz hingegen fliessen die Meinungen der Probanden und Probandinnen stärker in den jeweiligen Bewertungsprozess ein. Dazu werden einer Auswahl von Leuten Landschaften (meist in Form einer Fotografie) gezeigt, worauf diese eine quantitative Bewertung abgeben müssen (*Daniel*, 2001). Beim erfahrungsbezogenen Ansatz werden Datenerhebungsmethoden angewandt, welche zudem das Element der Interaktion mit der Umwelt berücksichtigen. So werden Leute beispielsweise damit beauftragt, selbst Fotografien zu machen, Tagebücher zu schreiben und an Interviews teilzunehmen (*Dakin*, 2003). Solche Methoden sind jedoch aufwändig und es können nur wenige Teilnehmer und Teilnehmerinnen in diesen Prozess eingebunden werden. Es bestehen also verschiedene Trade-Offs zwischen dem Aufwand, dem Einbezug möglichst vieler Meinungen und der Übertragbarkeit auf andere Regionen und Situationen. Diese Lücken könnten mit einer entsprechenden Anwendung zur automatisierten Extrahierung von Landschaftsbeschreibungen aus grossen Textkorpora womöglich geschlossen oder zumindest minimiert werden. Welche Möglichkeiten bestehen, um dies zu bewerkstelligen, wird in den folgenden Kapiteln untersucht.

2.3 Sprache und Raum

Um Landschaftsbeschreibungen analysieren zu können, müssen gewisse Eigenheiten des Sprachgebrauchs im geographischen Kontext berücksichtigt werden. Die Unterteilung des Raumes in Objekte basiert auf der Kategorisierung einer in Wirklichkeit überwiegend kontinuierlichen Oberfläche (*Burenhult & Levinson*, 2008). Diese Kategorisierung ist jedoch nicht immer eindeutig. Was für den einen ein Hügel ist, ist für den anderen ein Berg. Die Definition eines Konzepts, in diesem Fall des Objekts Berg, variiert von Beobachter zu Beobachter. Dies äussert sich besonders zwischen Leuten unterschiedlicher Sprachgruppen/-kulturen. Landschaftsbeschreibungen sind demnach mit vielen Unsicherheiten behaftet (*Mark et al.*, 2010). Diese Vagheit im Sprachgebrauch ist für die Kommunikation zwischen Menschen sehr wichtig, um die Komplexität zu reduzieren (*Fisher*,

2000). Für die Verarbeitung in einem Computer stellt sie jedoch eine grosse Herausforderung dar (*Derungs*, 2014).

2.4 Geographic Information Retrieval (GIR)

Der Bewältigung solcher Herausforderungen widmet man sich im Gebiet des Information Retrieval. Im Bezug zu geographischen Fragestellungen spricht man auch von «Geographic Information Retrieval» (GIR). Es handelt sich dabei um eine Erweiterung des IR mit Methoden aus den Geographischen Informationswissenschaften (*Derungs*, 2014). In diesem Gebiet geht es unter anderem darum, aus unstrukturierten Quellen, wie Texten oder Bildsammlungen, Informationen über die Interpretation des Raumes durch die Menschen zu gewinnen.

Einige Arbeiten beschäftigen sich mit der automatischen und eindeutigen Erkennung von Toponymen (Ortsnamen) in Texten (*Habib & van Keulen*, 2012). *Scheider & Purves* (2013) schlagen weiter vor, dass man Beschreibungen in unstrukturierten Texten nutzen soll, um zu erfahren, wie Orte von Menschen lokalisiert werden. Dies unter anderem, um die Verortung von Textinhalten zu verbessern. Dazu sollen nicht nur Toponyme selbst berücksichtigt werden, sondern auch der erweiterte Kontext in Form von Ortsbeschreibungen. Eine weitere Anwendungsmöglichkeit besteht in der Untersuchung der Unterteilung von Landschaften in Entitäten. So werden in der Arbeit von *Derungs & Purves* (2013) aus einem Alpinkorpus Begriffe extrahiert, welche von Leuten verwendet werden, um Objekte in der Landschaft zu benennen. Dazu gehören Ausdrücke wie «Berg», «Tal», etc. In *Hollenstein & Purves* (2010) werden aus den zwei Fotoarchiven «Flickr» und «Geograph» nicht nur Landschaftselemente extrahiert, sondern auch Begriffe, welche auf Qualitäten der Umgebung schliessen lassen. Mit der Analyse dieser Qualitäten werden also nicht nur die durch die Nutzer und Nutzerinnen der Plattform erfolgte Einteilung des Raumes in Entitäten untersucht, sondern auch die Eigenschaften, welche dem Raum zugeschrieben werden.

Der Fokus der vorliegenden Arbeit liegt einerseits in einer solchen automatisierten Extrahierung von Beschreibungen von Landschaftselementen. Dies jedoch mit dem erweiterten Ziel, die inhärente Meinung der Leute über die Landschaft zu erhalten. Mit solchen Problemstellungen wiederum befasst man sich besonders im Gebiet der Sentiment Analysis, welche als Teilgebiet des IR betrachtet werden kann (*Pang & Lee*, 2008). Auf diese Technik wird deshalb im nächsten Kapitel ausführlich eingegangen.

2.5 Sentiment Analysis

2.5.1 Grundlegendes über die Sentiment Analysis

Das Ziel der Sentiment Analysis ist es, Meinungen aus Texten rechentechnisch zu extrahieren (*Pang & Lee*, 2008). In vielen Arbeiten geht es darum, ganze Texte automatisch den Gefühlspolaritäten «positiv» oder «negativ» zuzuordnen (*Pang et al.*, 2002). Andere verfolgen dasselbe Vorhaben auf der Ebene von Sätzen oder Satzteilen. Aus den Resultaten lässt sich dann wiederum ein Wert für einen ganzen Text berechnen (*Wilson et al.*, 2005; *Yu & Hatzivassiloglou*, 2003). Es ist auch möglich, lediglich diejenigen Meinungen zu extrahieren, welche gegenüber einem bestimmten Thema geäussert werden (*Hu & Liu*, 2004; *Hurst & Nigam*, 2003). Für gewisse Anwendungen genügt eine Einteilung in die Polaritäten «positiv» oder «negativ». Texte oder Textbestandteile enthalten jedoch meist eine Mischung aus positiven und negativen Gefühlsäusserungen. So kann die Stärke einer Gefühlspolarität auch in einer Werte-Skala angegeben werden (*Thelwall et al.*, 2010). Ein beliebter Untersuchungsgegenstand in der Sentiment Analysis sind Reviews von klassischen Bewertungsplattformen. Diese verfügen nämlich häufig über die Eigenschaft, dass die Autoren oder Autorinnen zusätzlich zum textuellen Feedback ein Rating in quantitativer Form (z.B. Sterne oder Daumen rauf/runter) abgeben können (*Pang et al.*, 2002). In der Sentiment Analysis existieren vorwiegend zwei Ansätze, welche nachfolgend erläutert werden.

2.5.2 Maschinelles Lernen

In einigen Arbeiten werden Algorithmen des maschinellen Lernens eingesetzt, um die Texte oder Textteile zu klassifizieren respektive einer Gefühlspolarität zuzuordnen (*Pang et al.*, 2002; *Pang & Lee*, 2004; *Turney*, 2001*)*. Maschinelles Lernen wird unter anderem in der Themen-basierten Textklassifikation verwendet. Die Sentiment Analysis kann als Spezialfall angesehen werden, bei welchem die positive und die negative Kategorie jeweils ein «Thema» repräsentieren. Zur Durchführung eines solchen Verfahrens müssen im Text Merkmale definiert werden, auf welchen die Klassifikationen beruhen. Dazu werden oft sogenannte «N-Gramme» erzeugt. Das heisst, der Text wird in Wortgruppen einer vordefinierten Grösse (z.B. «Unigramme», «Bigramme», etc.) zerlegt. Dabei kann auch nur eine bestimmte Wortart (z.B. Adjektive) berücksichtigt werden (*Pang & Lee*, 2004).

Wie oben erwähnt, ist bei vielen Reviews mit dem quantitativen Rating bereits ein Indikator für die Polarität der Kommentare vorhanden. Dies ermöglicht die Anwendung einer überwachten Klassifikation mit Algorithmen des maschinellen Lernens, wobei ein Teil des Korpus als Trainingsdatensatz verwendet werden kann (*Pang et al.*, 2002). Ist eine solche Referenz nicht vorhanden, besteht auch die Möglichkeit, einen Teil des Datensatzes für die Evaluation manuell zu klassifizieren (*Wilson et al.*, 2005).

2.5.3 Sentiment-Lexikon

Ein anderer Ansatz sieht die Verwendung eines sogenannten «Sentiment-Lexikons» vor. Darin ist je nach Lexikon eine unterschiedliche Anzahl Wörter mit einem dazugehörigen «Sentiment-Wert» aufgeführt. Für die Bestimmung des Polaritätswerts eines Textes oder Textteils werden diese Wörter gezählt und ihre Sentiment-Werte aufsummiert (*Yu & Hatzivassiloglou*, 2003). Ein Beispiel eines solchen Lexikons ist «SentiWordNet». Dieses enthält Stimmungswerte (positiv oder negativ zwischen 0 und 1) für 117‘659 «Synsets». Ein Synset stellt eine Gruppe von Wörtern dar, welche dieselbe inhaltliche Bedeutung haben und somit denselben Sentiment-Wert aufweisen. Zusätzlich zum Sentiment-Wert wird ein Wert für die Objektivität angegeben. Ist dieser Wert gleich 1, ist eine Wortgruppe vollständig objektiv, und es geht von den Wörtern des entsprechenden Synsets keine Stimmung aus. Ein Synset kann auch gleichzeitig negativ und positiv sein (*Baccianella et al.*, 2010). Um die Werte für die Synsets zu erhalten, wurden in einem iterativen Prozess mit jedem Synset mehrere semi-überwachte Klassifikationen durchgeführt, d.h. mit Trainingsdatensätzen, von welchen eine Teilmenge manuell bewertet wurde. Aus den Resultaten konnte dann der zutreffendste Wert entnommen werden. SentiWordNet wird in diversen Arbeiten verwendet (*Baccianella et al.*, 2010), wurde jedoch bisher nur für die englische Sprache entwickelt. Bei «BAWL-R» (Berlin Affective Word List Reloaded), der Weiterentwicklung von «BAWL» (Berlin Affective Word List), handelt es sich um eines der wenigen deutschen Sentiment-Lexika. Hier wurden die Sentiment-Werte von etwas mehr als 2‘900 Wörtern (Nomen, Verben und Adjektive) anhand eines Experiments mit Probanden und Probandinnen ermittelt. Diese mussten in einer Versuchsanordnung die Negativität respektive Positivität eines Ausdrucks auf einer Skala von -3 bis 3 angeben (*Võ et al.*, 2006). «SentiWS» (SentimentWortschatz) ist ein weiteres deutsches Sentiment-Lexikon. Es enthält Bewertungen in einer Intervallskala von -1 bis 1 von ca. 3‘500 Wörtern. Dazu gehören auch 10 Adverbien. Die Ermittlung der Werte erfolgte hierbei ebenfalls mit einem semi-überwachten Klassifikationsverfahren ähnlich demjenigen von SentiWordNet (*Remus et al.*, 2010).

2.5.4 Meinungen zu spezifischen Themen

Ein Ziel dieser Arbeit ist es, aus Tourenberichten die Meinung der Leute gegenüber der Landschaft zu extrahieren. Doch wie lassen sich diejenigen Wörter erkennen, welche sich auf ein spezifisches Thema beziehen? (*Hu & Liu*, 2004) verfolgen dabei die einfache Idee, dass Meinungswörter mit grosser Wahrscheinlichkeit nahe beim Themenbegriff lokalisiert sind, auf welchen sie sich beziehen. (*Cataldi et al.*, 2013) hingegen versuchen Wortbeziehungen anhand der Analyse der Satzstruktur zu erkennen. Dazu werden mit einem sogenannten «Dependency Parser» die syntaktischen Abhängigkeitsbeziehungen zwischen thematisch relevanten Begriffen und allfälligen Mei-

nungswörtern analysiert. Dazu werden in einem Satz die Wortbeziehungen in einem Dependenzbaum wie in Abbildung 2 (Beispielsatz: «Small and charming hotel with all the benefits of a big one.») dargestellt.

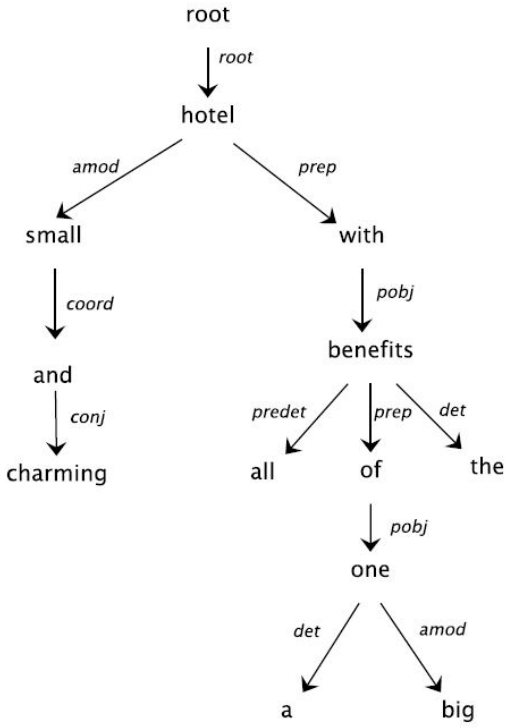


Abb. 2: Beispiel eines Dependenzbaums (*Cataldi et al.*, 2013).

Von den Meinungswörtern werden dann in einem Verfahren ähnlich demjenigen von SentiWordNet die Polaritätswerte ermittelt. Aus diesen Werten lässt sich dann die Haltung gegenüber einem Thema berechnen. Zur syntaktischen Analyse der einzelnen Sätze stützt sich diese Methodik auf die Theorie der Dependenzgrammatik. Diese wird deshalb noch etwas eingehender erläutert.

2.5.4.1 Dependenzgrammatik

Die Dependenzgrammatik beschreibt Abhängigkeitsverhältnisse sprachlicher Elemente respektive Wörter (*Jung*, 1995). *Tarvainen* (2000) bezeichnet sie auch als «Chemie der Sprache». Er vergleicht damit die Bildung von Verbindungen aus chemischen Grundstoffen mit derjenigen von grösseren Einheiten aus sprachlichen Elementen. Stehen zwei Elemente eines Satzes in einer Dependenzrelation zueinander, lässt sich vom Vorkommen des einen Elements auf das Vorkommen des anderen Elements schliessen. Dependenzrelationen können hierarchisch dargestellt werden. Dabei werden das bedingende Element als «Regens» und das bedingte Element als «Dependent» bezeichnet (*Jung*, 1995). Im Satz «Ich liebe diese Landschaft.» beispielsweise regiert das Wort «liebe» als Regens über das nominativische Element «Ich» und das akkusativische Element «diese Landschaft». Unter Zuhilfenahme der Dependenzgrammatik sollten also die allfälligen Beschrei-

bungswörter von Landschaftsbegriffen basierend auf syntaktischen Regeln ermittelt werden können. Um die Abhängigkeitsbeziehungen für eine grosse Menge an Sätzen automatisiert zu ermitteln, werden sogenannte Dependency Parser entwickelt. Ein Beispiel hierfür ist die Anwendung «ParZu», welche an der Universität Zürich von *Sennrich & Schneider* (2009) aus dem Institut für Computerlinguistik entworfen wurde.

2.5.5 Herausforderungen

Bei der Sentiment Analysis handelt es sich um ein aktives Forschungsgebiet, in welchem man noch mit einigen Herausforderungen konfrontiert ist. Der semantische Gehalt eines Ausdrucks ist nicht fixiert, sondern abhängig vom Kontext oder der Domäne, in welchem er verwendet wird (*Wilson et al.*, 2005). So hat bei der Charakterisierung eines Landschaftselements wie beispielsweise eines Felsens der Ausdruck «scharf» eine andere Bedeutung, als wenn die Klinge eines kürzlich erworbenen Küchenmessers bewertet wird. Ein Wort als solches kann auch positiv konnotiert sein, jedoch in einem negativen Kontext verwendet werden und umgekehrt. *Wilson et al.* (2005) unterscheiden dabei zwischen vorgängiger und kontextueller Polarität. Besonders deutlich wird diese Tatsache bei Negationen, welche die Polarität eines Wortes umkehren. Aber auch Adverbien können den Sentiment-Wert eines Ausdrucks modifizieren. Zudem könnten im Hikr-Korpus auch komplexere Fälle auftreten, wie ironische Äusserungen. In einigen Arbeiten werden zu Beginn einer Sentiment Analysis die neutralen Sätze oder Phrasen aus dem Textdatensatz ausgeschieden. Dann wird nur mit den restlichen eine genauere Polaritätsbestimmung durchgeführt. (*Hurst & Nigam*, 2003; *Pang & Lee*, 2004; *Wilson et al.*, 2005; *Yu & Hatzivassiloglou*, 2003). *Wilson et al.* (2005) haben dabei festgestellt, dass viele Wörter mit nicht-neutraler vorgängiger Polarität in Phrasen auftauchen, welche als neutral klassifiziert werden.

2.5.6 Sentiment Analysis und Geographie

In der Sentiment Analysis liegt der Fokus bisher stark auf Anwendungen in der Wirtschaft *(Pang & Lee*, 2008). Arbeiten mit geographischem Bezug konzentrieren sich meist auf die räumliche Verteilung von Meinungen (*Hauthal & Burghardt*, 2014). Bei *Nowak* (2013) beziehen sich jedoch auch die Meinungen selbst auf einen geographischen Inhalt, nämlich auf die Landschaft. Er untersucht textuelle Beschreibungen von Landschaftsfotografien von ganz Grossbritannien mit dem Ziel, die Meinungen bezüglich der Schönheit dieser Bilder zu schätzen. Als Referenz verfügt er jedoch zusätzlich über einen Datensatz, in welchem dieselben Bilder auf einer Punkteskala von 1-10 bewertet wurden.

2.6 User generated content

Bei den Informationen auf Hikr handelt es sich grösstenteils um Inhalte, welche von freiwilligen Nutzern und Nutzerinnen generiert wurden. Bevor der Datensatz im Kapitel 3 eingehender analysiert wird, werden deshalb noch einige Besonderheiten solcher Inhalte diskutiert.
Gemäss *Balasubramaniam* (2009) kann ein Inhalt laut OECD (Organization for Economic Cooperation and Development) dann als nutzergenerierter Inhalt («user generatet content», UGC) definiert werden, wenn er:

- öffentlich über das Internet verfügbar ist
- einen gewissen Grad an Kreativität aufweist
- ausserhalb einer professionellen Tätigkeit erstellt wurde

Handelt es sich bei den von Internetnutzern und -nutzerinnen bereitgestellten Daten um Geodaten, spricht man auch von «volunteered geographic information» (VGI) (*Flanagin & Metzger*, 2008). UGC verfügt über gewisse typische Eigenschaften, deren Kenntnis für die Arbeit mit solchen Datensätzen von Relevanz ist. Die meisten Inhalte stammen von Leuten, welche im betreffenden Gebiet kein Expertenwissen aufweisen (*Tulloch*, 2007). *Hill & Ready-Campbell* (2014) konnten jedoch aufzeigen, dass die «Weisheit der Masse» diesen Effekt kompensieren und in gewissen Fällen Einschätzungen von Experten oder Expertinnen sogar übertreffen kann. Laut *Flanagin & Metzger* (2008) hat zudem die Motivation, unter welcher ein Beitrag verfasst wird, einen entscheidenden Einfluss auf die Glaubwürdigkeit der publizierten Inhalte. Weiter ist nicht davon auszugehen, dass die Anzahl Beiträge unter den Nutzern und Nutzerinnen gleichmässig verteilt ist. Meist steuert nämlich ein kleiner Teil von Leuten eine grosse Menge an Inhalten bei, während ein grosser Teil von Leuten eine kleine Menge an Inhalten produziert. Als Faustregel kann man dabei auf das Paretoprinzip (80/20) zurückgreifen (*Ochoa & Duval*, 2008). Laut *Ochoa & Duval* (2008) und *Hollenstein & Purves* (2010) können sogar einzelne Nutzer oder Nutzerinnen eine Verzerrung in den Daten verursachen. Die Autoren kommen deshalb zum Schluss, dass zumindest die Kenntnis der Eigenschaften des zu bearbeitenden Datensatzes wichtig ist.

3. Der Hikr-Korpus

Die Touren-Website Hikr ist eine Internetplattform, auf welcher Leute in Form von Blogeinträgen über die von ihnen gemachten Touren berichten, indem sie einen Text verfassen und Fotos hochladen. Die Gründung der Seite datiert aus dem Jahr 2003. Es können jedoch auch nachträglich Berichte über Touren verfasst werden, welche bis ins Jahr 1970 zurückreichen. Je nach Aktivität kann der Bericht acht verschiedenen Kategorien zugeordnet werden: Wandern, Hochtouren, Klettern, Schneeschuhe, Klettersteig, Skitouren, Eisklettern und Mountainbike. Es können auch weitere Metainformationen angegeben werden, wie das Tour-Datum, die passierten Wegpunkte oder der Zeitbedarf. Als Pflichtangaben gelten jedoch nur die Region und ein Titel. Die publizierten Texte können Besuchern und Besucherinnen der Website bei der Planung einer anstehenden Tour als Anregung oder als Informationsquelle dienen. Hikr verfügt zurzeit (Stand 21.05.2015) über 71‘432 Berichte aus aller Welt. 42‘286 stammen aus der Schweiz. Davon sind wiederum 28‘287 von 1099 verschiedenen Autoren in deutscher Sprache verfasst worden. Diese 28‘287 Texte verteilen sich wie in Tabelle 1 gezeigt auf die verschiedenen Aktivitäten.

Aktivität	Häufigkeit
Wandern	18661
Skitouren	3839
Schneeschuhe	2083
Hochtouren	2034
Klettern	603
Mountainbike	344
Klettersteig	183
Eisklettern	32
Ohne Angabe	508

Tab. 1: Häufigkeitsverteilung der Aktivitäten in den untersuchten Hikr-Berichten.

Wie weiter unten in Abbildung 3 zu sehen ist, entspricht die Verteilung der Berichte auf die Autoren einer sogenannten «Zipf-Verteilung» (*Zipf*, 1935), wobei die Häufigkeit umgekehrt proportional zur Anzahl Berichte pro Autor steht. D.h., dass einige wenige Autoren eine grosse Menge an Beiträgen verfassen und eine grosse Anzahl Autoren nur wenige Texte beisteuern. Dies entspricht auch dem in *Ochoa & Duval* (2008) und in *Hollenstein & Purves* (2010) erwähnten Muster. Abbildung 4 zeigt zudem, wie die Längen der Berichte (Anzahl Wörter) verteilt sind. Auffällig ist hierbei

der Peak nahe bei Null. Es stellt sich heraus, dass bei einigen Berichten keine ausführliche Tourenbeschreibung gemacht wurde. So laden einige Nutzer und Nutzerinnen lediglich die von ihnen gemachten Bilder hoch.

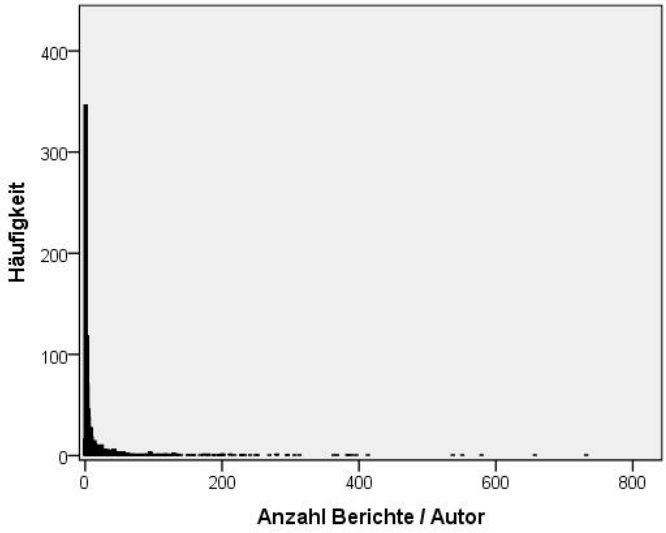


Abb. 3: Verteilung der Anzahl Beiträge auf die Autoren in den untersuchten Hikr-Berichten.

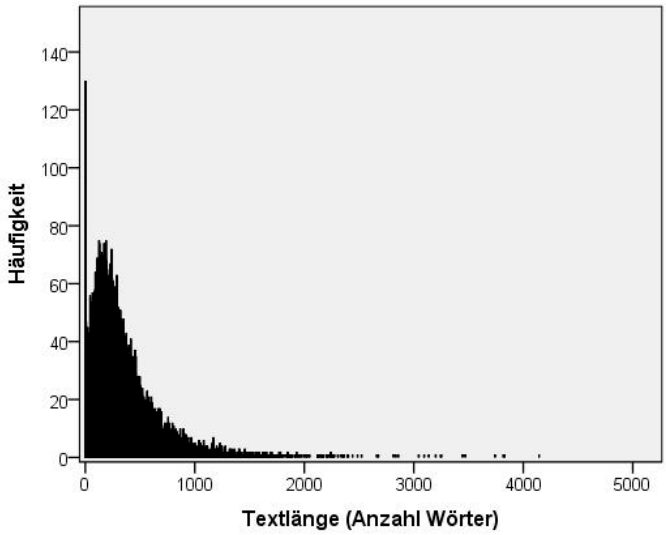


Abb. 4: Textlängen der untersuchten Hikr-Berichte.

In Abbildung 5 ist dargestellt, wie die Touren räumlich verteilt sind. Wie zu erwarten war, erzählen die meisten Texte von Unternehmungen in den alpinen Regionen. So verzeichnen diejenigen Kantone, durch welche sich die Schweizer Alpen ziehen, die grösste Anzahl Beiträge. Am meisten Berichte weisen die Kantone Bern (grösstenteils Berner Oberland), Wallis, Tessin und Graubünden auf.

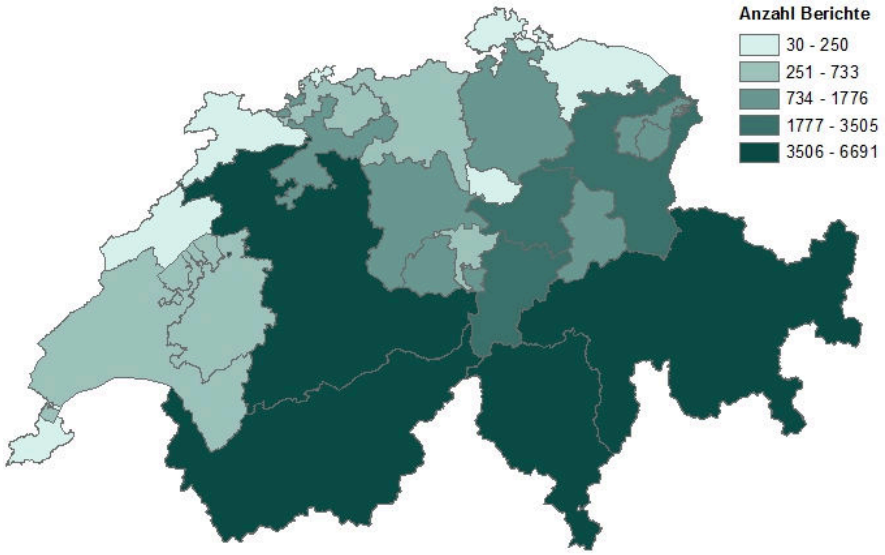


Abb. 5: Choroplethenkarte der Schweiz mit der Anzahl Berichte pro Kanton.

Abbildung 6 zeigt einen Ausschnitt aus einem Bericht auf der Tourenplattform Hikr. Im Vergleich zu vielen Arbeiten im wirtschaftlichen Kontext werden die Touren in Hikr nicht quantitativ bewertet, geschweige denn die Landschaft als solches. Des Weiteren folgen die Texte keiner vorgegebenen Struktur, wie dies teilweise bei klassischen Produktebewertungen der Fall ist.

Region:	Welt » Schweiz » Waadt » Waadtländer Alpen
Tour Datum:	11 August 2015
Hochtouren Schwierigkeit:	L
Klettern Schwierigkeit:	I (UIAA-Skala)
Wegpunkte:	Col du Pillon 1548 m (2) Tête aux Chamois 2525 m (8) Sex Rouge 2971 m (22) Le Dôme 2994 m (27) Les Diablerets - Sommet des Diablerets 3210 m (32)
Geo-Tags:	CH-VD CH-VS
Zeitbedarf:	4:30
Aufstieg:	500 m
Abstieg:	500 m
Strecke:	Scex Rouge (2971 m) - Le Dôme (2986 m) - Les Diablerets Sommet des Diablerets (3210 m)

Eine wunderbare, angenehme Bergtour bei schönem Wetter. Von der Bergstation **Scex Rouge** geht es auf der Moräne den Steinmannli entlang hinauf Richtung **Le Dôme**. Hier wartet die Schlüsselstelle, die aber mit Seilen gesichert und deshalb gut begehbar ist. Eindrücklich: Offenbar stand der Gletscher bis vor wenigen Jahren noch deutlich höher. Der Rückgang des Gletschers macht heute diesen Abstieg über einen Felsrücken überhaupt erst nötig. Auf dem **Glacier des Diablerets** angekommen heisst es anseilen und Steigeisen anschnallen. Dann geht es angenehm den Gletscher hoch auf den Gipfel **Sommet des Diablerets**. Der Gletscher war aper, die Spalten daher gut sichtbar. Nachdem wir die Aussicht genossen haben, gings auf dem gleichen Weg zurück.

Abb. 6: Beispiel eines Tourenberichts auf Hikr (*www.hikr.org*).

4. Implikationen

In diesem Kapitel wird erläutert, welche Implikationen aus der Analyse des Hikr-Datensatzes und der Literaturrecherche hervorgehen. Daraufhin werden die Forschungslücken aufgezeigt, welchen sich diese Arbeit widmen soll.

4.1 Wissenschaftlicher Hintergrund

In Anlehnung an die Arbeiten im Bereich der Wirtschaft wird für diese Arbeit die Analogie gemacht, dass auch Landschaften als Produkte angesehen werden können und deren Elemente als die Produkteigenschaften, welche es zu bewerten gilt. Die Texte auf Hikr können als eine Art Tagebucheinträge angesehen werden, in welchen man die erlebte Landschaft noch einmal Revue passieren lässt. Die Analyse solcher Tagebücher gehört auch zu den von *Dakin* (2003) vorgeschlagenen Methoden, um ein umfassenderes Bild der Landschaft aus der Sicht der Leute wiederzugeben. Die Texte auf Hikr verfügen über keine quantitativen Referenzwerte. Es besteht zwar die Möglichkeit, einen Trainingsdatensatz selbst zu erstellen, indem Probanden und Probandinnen eine Auswahl an Texten manuell bewerten, wie dies beispielsweise in *Wilson et al.* (2005) gemacht wurde. Es sollen jedoch nicht nur die Gefühlswerte der Landschaftsbeschreibungen extrahiert werden, sondern auch die dafür verwendeten Worte. Deshalb scheint die Verwendung eines Sentiment-Lexikons, in welchem der Polaritätswert von Meinungswörtern nachgeschlagen werden kann, geeigneter. Da der Sentiment-Wert sich mit der Domäne eines Textes verändern kann, wäre die Erstellung eines Domänen-spezifischen Lexikons wie in *Blitzer et al.* (2007) zwar sinnvoll. Jedoch würde dies den Zeitbedarf einer eigenen Arbeit in Anspruch nehmen und wird deshalb im Rahmen dieser Arbeit nicht umgesetzt. Um die Meinungswörter ausfindig zu machen, welche sich auf die Landschaftsbegriffe beziehen, erscheint der Ansatz von *Cataldi et al.* (2013) vielversprechend. Dabei wird durch eine syntaktische Analyse ermittelt, in welcher Beziehung die einzelnen Wörter eines Satzes zueinander stehen. Beschreibungswörter werden also nicht basierend auf Wahrscheinlichkeiten ermittelt, wie dies beispielsweise in *Hu & Liu* (2004) der Fall ist.

4.2 Untersuchungsgegenstand

Der verwendete Datensatz kann als UGC und im engeren Sinne auch als VGI bezeichnet werden. Die Texte wurden mit der Motivation verfasst, Besuchern und Besucherinnen der Website von den begangenen Touren zu berichten. Unter den Autoren und Autorinnen auf Hikr gibt es zwar Laien, ein nennenswerter Teil der Leute verfügt aber wahrscheinlich über ein fortgeschrittenes

Touren-Wissen. Aufgrund der ungleichen Verteilung der Beiträge auf die Verfasser und Verfasserinnen besteht die Möglichkeit, dass besonders in Regionen mit wenigen Berichten der Einfluss einer einzelnen Person sehr gross sein kann. Des Weiteren könnten die verschiedenen Aktivitäten aufgrund unterschiedlicher Interessen zu einer anderen Interpretation der Landschaft führen. Allgemein kann die Wahrnehmung der Landschaft durch verschiedene Faktoren beeinflusst werden. Dazu gehören zum Beispiel das Wetter, der Schwierigkeitsgrad der Tour oder der Zeitbedarf. Die Einflüsse dieser Faktoren liessen sich durch die zahlreichen verfügbaren Metainformationen untersuchen. Dies wäre hier jedoch zu weit vorgegriffen und könnte allenfalls in einer weiterführenden Arbeit berücksichtigt werden.

4.3 Forschungslücken

Im Gebiet des GIR wurden schon verschiedene Vorgehen entwickelt, um Informationen darüber zu gewinnen, wie die Menschen ihre Umgebung interpretieren. Der Fokus lag bisher aber kaum auf den Eigenschaften oder ästhetischen Qualitäten, welche der Landschaft zugesprochen werden. Gängige Methoden zur Landschaftsbewertung sind aufwändig und von den Meinungen weniger Personen abhängig. Erkenntnisse beziehen sich teilweise auf einen räumlichen und zeitlichen Ausschnitt und werden dementsprechend verallgemeinert. Mit den Techniken der Sentiment Analysis besteht die Möglichkeit, die Meinungen einer grossen Anzahl Leute zu ermitteln. Dies wurde bisher aber vor allem in Arbeiten mit wirtschaftlichem Hintergrund untersucht. Bei der Arbeit von *Nowak* (2013) handelt es sich um die einzige dem Autor bekannte Anwendung im Kontext der Landschaftsbewertung. In besagter Arbeit dienen jedoch quantitative Bewertungen der Landschaften als Referenz. Diese Situation besteht jedoch nur bei sehr wenigen Datensätzen. Damit eine solche Methodik für viele verschiedene Regionen und Textquellen übernommen werden kann, wäre es deshalb von Vorteil, wenn sie unabhängig von solchen quantitativen Referenzwerten funktioniert. Diese Arbeit setzt dort an und versucht, aus reinen Textdatensätzen aus Hikr Landschaftsbeschreibungen herauszulesen, zu analysieren und deren Gefühlspolaritäten zu bestimmen.

5. Methoden

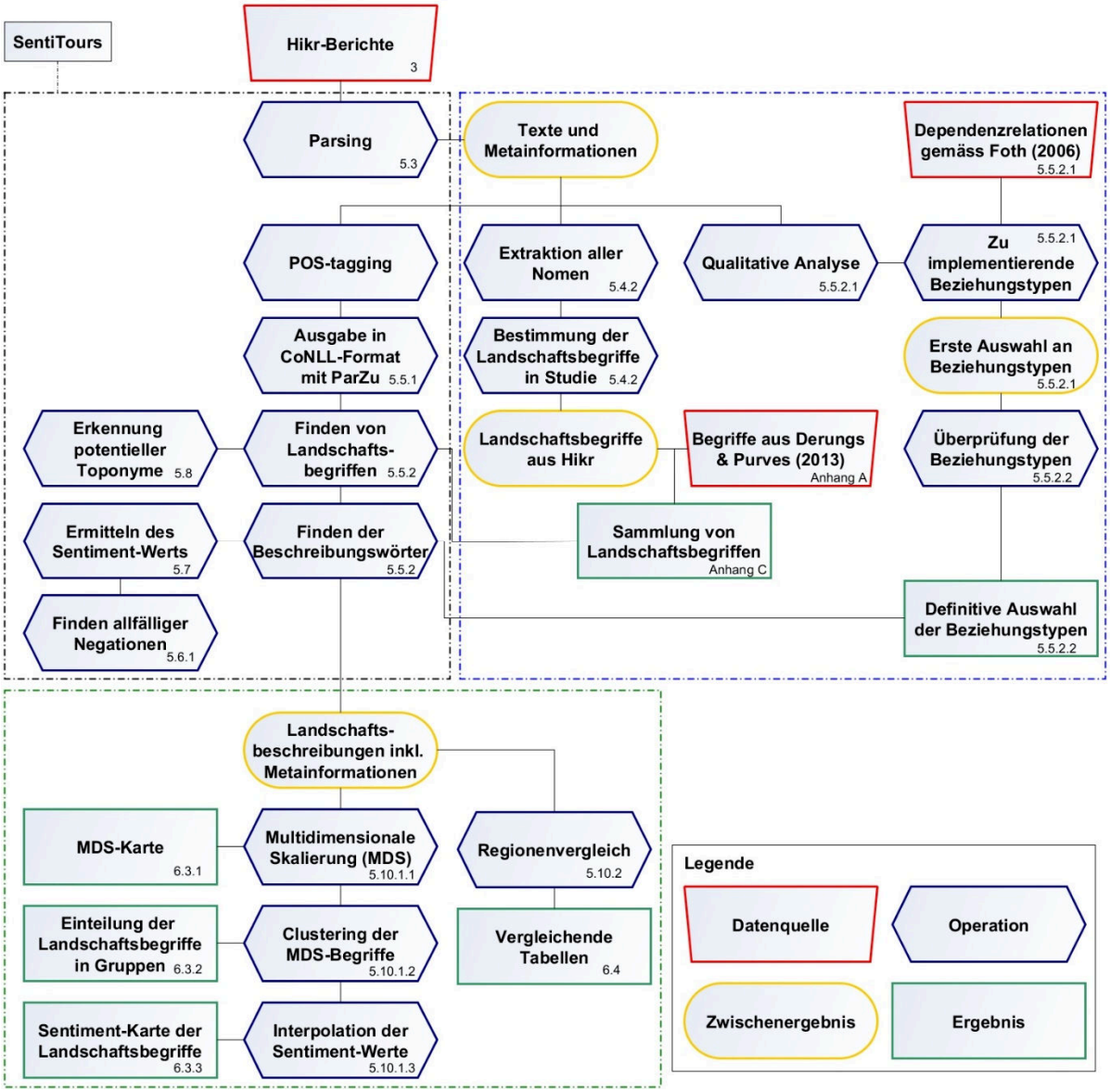


Abb. 7: Methodisches Vorgehen.

In Abbildung 7 ist das gesamte Vorgehen in einem Flussdiagramm dargestellt. Es wird eine Anwendung namens «SentiTours» entwickelt, welche Landschaftsbeschreibungen und deren inhärente Meinungen aus den Hikr-Texten herausliest. Die Kästchen der oberen rechten Hälfte der Abbildung (blau umrahmt) beschreiben Operationen und Daten, welche zur Entwicklung von SentiTours beitragen. Die Operationen, welche dann von der Anwendung oder zumindest von Teilen davon (Parsing) ausgeführt werden, sind in den Kästchen der oberen linken Hälfte aufgeführt (schwarz umrahmt). Die von SentiTours extrahierten Landschaftsbeschreibungen werden im Anschluss einigen Analysen unterzogen. Die Kästchen, welche sich darauf beziehen, sind in der unteren Hälfte der Graphik aufgeführt (grün umrandet). In einigen Kästchen ist das jeweilige Kapitel

oder der Anhang notiert, in welchem die entsprechende Operation, die Datenquelle, das Zwischenergebnis oder das Ergebnis aufgeführt oder beschrieben wird. Im Folgenden werden zuerst die Werkzeuge erwähnt, welche für die vorliegende Arbeit zum Einsatz kommen. Dann wird ausführlich auf die Entwicklung von SentiTours eingegangen. Anschliessend werden die verschiedenen Methoden präsentiert, mit welchen der durch die Anwendung erhaltene Output interpretiert und analysiert wird.

5.1 Werkzeuge

Zur Entwicklung von SentiTours wird die Programmiersprache «Python» verwendet. Als Dependency Parser wird die Anwendung ParZu aus *Sennrich & Schneider* (2009) in SentiTours eingebunden. Diese ermittelt neben den Wortarten («Part-of-Speech-Tagging») und den Grundformen («Lemmatisierung») auch die syntaktischen Abhängigkeitsbeziehungen zwischen Wörtern in einem Text. Die Programmierarbeit wird auf einer Linux-Plattform verrichtet. Dies unter anderem, weil ParZu speziell für dieses Betriebssystem entwickelt wurde. Als Sentiment-Bibliothek wird SentiWS von *Remus et al.* (2010) verwendet. Auswertungen werden einerseits ebenfalls mit Python und andererseits mit der Statistik-Software «SPSS» durchgeführt. Weiter wird für eine später in dieser Arbeit erläuterte Interpolation die Software «ArcMap/ArcGIS» benützt.

5.2 Übersicht über die Funktionsweise von SentiTours

In diesem Kapitel wird kurz zusammengefasst, wie die Anwendung SentiTours schlussendlich funktionieren soll. In den nachfolgenden Kapiteln wird dann die Umsetzung dieser Idee beschrieben.

In einem ersten Schritt sollen die html-Dateien von Hikr «geparst» (herausgelesen) und die jeweiligen Texte sowie relevante Metainformationen eingelesen werden. Dann wird in den Texten nach Landschaftsbegriffen gesucht. Wird ein Landschaftsbegriff gefunden, wird der entsprechende Satz herausgespeichert. In einem weiteren Schritt wird in den gespeicherten Sätzen geprüft, ob der darin enthaltene Landschaftsbegriff von einem anderen Wort beschrieben wird. Ist dies der Fall, handelt es sich beim entdeckten Wortpaar um eine Landschaftsbeschreibung bestehend aus Beschreibungswort und Landschaftsbegriff, welche später im Output ausgegeben werden soll. Zuerst wird diese Wortbeziehung jedoch noch auf zusätzliche Eigenschaften überprüft. So wird weiter untersucht, ob das Beschreibungswort von einer Negation modifiziert wird. Im Weiteren wird geschaut, ob es sich beim Landschaftsbegriff möglicherweise um ein Toponym handeln könnte. Der Grund dafür wird später in Kapitel 5.8 erläutert. In einem letzten Schritt wird von jedem Beschreibungswort der Sentiment-Wert, falls vorhanden, in SentiWS nachgeschlagen. Wird das Beschreibungswort im Text durch eine Negation modifiziert, wird das Vorzeichen des Sentiment-

Werts umgekehrt. Adverbien werden nicht berücksichtigt, weil in SentiWS anders als bei SentiWordNet nur für 10 verschiedene Adverbien Sentiment-Werte angegeben sind und somit nicht ermittelt werden kann, inwiefern ein allfälliges Adverb ein Beschreibungswort modifiziert.

5.3 Parsen der html-Dateien

Die html-Dateien des Hikr-Korpus (Stand: 2010) wurden vorgängig mit einem «web scraping» Verfahren extrahiert. Aus diesen müssen in einem ersten Schritt die relevanten Inhalte herausgelesen werden. Der wichtigste Inhalt ist der Text. Von den Metadaten werden folgende Informationen für die späteren Analysen als relevant erachtet, wobei in dieser Arbeit nicht alle verwendet werden:

- Region
- Tour-Datum
- Schwierigkeit/Aktivität
- Wegpunkte
- Zeitbedarf
- Autor

5.4 Ermitteln der Landschaftsbegriffe

Damit in den Texten überhaupt nach Landschaftsbegriffen gesucht werden kann, muss eine Auswahl von Begriffen festgelegt werden, welche Landschaftselemente darstellen. Als Landschaftsbegriffe werden Nomen betrachtet, welche verwendet werden, um einen Ausschnitt oder ein Element der Landschaft zu benennen. Es kann sich dabei in Anlehnung an *Palka* (1995) sowohl um ein anthropogenes oder anthropogen geprägtes (z.B. Dorf, Trampelpfad) als auch um ein natürliches Element (z.B. Grat, Fels) handeln.

5.4.1 Landschaftsbegriffe aus der Literatur

Derungs & Purves (2013) identifizierten aus einem Korpus von 10‘000 Artikeln mit Beschreibungen von Schweizer Alpinlandschaften und –aktivitäten 95 natürliche Begriffe, welche benutzt werden, um Landschaften zu kategorisieren («Text+Berg», Anhang A). Diese Ausdrücke sollten auf den Hikr-Datensatz übertragbar sein. Dies muss jedoch überprüft werden. Deshalb wurde der gesamte Hikr-Korpus nach dem Vorkommen der 95 Begriffe durchsucht. Dabei hat sich gezeigt, dass 92 der 95 Ausdrücke mindestens einmal im gesamten Korpus verwendet werden.
Mit den gefundenen Begriffen wird nun eine Häufigkeitsverteilung erstellt. Dabei wird für jedes Wort entsprechend der Anzahl Nennungen ein Rang von 1-92 vergeben. Da auch die 95 respektive

92 Begriffe aus oben genannter Quelle nach deren Häufigkeit sortiert sind, lässt sich ein Rangkorrelationstest nach Spearman durchführen. Das Ergebnis dieses Tests ist in Abbildung 8 ersichtlich.

Correlations

			Hikr	TextBerg
Spearman's rho	Hikr	Correlation Coefficient	1.000	.597**
		Sig. (2-tailed)	.	.000
		N	92	92
	TextBerg	Correlation Coefficient	.597**	1.000
		Sig. (2-tailed)	.000	.
		N	92	92

**. Correlation is significant at the 0.01 level (2-tailed).

Abb. 8: Rangkorrelation der gemeinsamen Landschaftsbegriffe in Hikr und Text+Berg.

Das Testresultat zeigt, dass die beiden Rangfolgen signifikant korrelieren, $r_s = 0.6$, $p = 0.000$. Die Reihenfolgen, basierend auf der Anzahl Nennungen, sind sich demnach ähnlich in den zwei Korpora. Die Liste mit den natürlichen Begriffen wird deshalb verwendet.

5.4.2 Landschaftsbegriffe aus dem Hikr-Korpus

Wie gezeigt, kommen nicht alle Begriffe aus *Derungs & Purves* (2013) im Hikr-Korpus vor. Es ist auch nicht bekannt, ob in Hikr womöglich weitere Landschaftsbegriffe erwähnt werden. Des Weiteren enthält die Auflistung dieser Quelle nur natürliche Objekte. Wie bereits erwähnt, werden in dieser Arbeit jedoch auch anthropogene oder anthropogen geprägte Objekte zur Landschaft gezählt.

Aus diesen Gründen werden basierend auf dem Hikr-Korpus weitere Landschaftsbegriffe gesucht. Dazu werden die 1500 häufigsten Nomen in einer Liste gespeichert. Diese Liste wird fünf verschiedenen Probanden und Probandinnen präsentiert. Es handelt sich dabei um Masterstudenten und -studentinnen mit Hauptfach Geographie. Diese müssen bei jedem Nomen entscheiden, ob es sich dabei um einen Landschaftsbegriff handelt oder nicht. Durch diese Methode kommen 181 weitere Landschaftsbegriffe zu den 92 aus *Derungs & Purves* (2013) dazu. Der Versuchsaufbau sowie die komplette Liste mit den Begriffen können Anhang B respektive C entnommen werden.

5.5 Extraktion der Beschreibungswörter

Um diejenigen Beschreibungswörter aus dem Text zu erkennen, welche sich auf die Landschaftsbegriffe beziehen, kommt die Theorie der Dependenzgrammatik zur Anwendung. Wie in Kapitel 2.5.4.1 erläutert, lassen sich die syntaktischen Beziehungen der Wörter in einem Satz mit einem Dependency Parser aufschlüsseln. Folglich lassen sich Methoden implementieren, mit welchen ausgehend von einem Landschaftsbegriff auf sein Beschreibungswort geschlossen werden kann.

5.5.1 Das CoNLL Datenformat

Mit der Anwendung ParZu lässt sich ein Output im sogennanten «CoNLL» (computational natural language learning) Datenformat generieren. Dieser Output setzt sich aus zehn verschiedenen Parametern zusammen, welche in *Buchholz & Marsi* (2006) beschrieben werden und von den Autoren auf ihrer Homepage in Abbildung 9 zusammengefasst werden.

Field number:	Field name:	Description:
1	ID	Token counter, starting at 1 for each new sentence.
2	FORM	Word form or punctuation symbol.
3	LEMMA	Lemma or stem (depending on particular data set) of word form, or an underscore if not available.
4	CPOSTAG	Coarse-grained part-of-speech tag, where tagset depends on the language.
5	POSTAG	Fine-grained part-of-speech tag, where the tagset depends on the language, or identical to the coarse-grained part-of-speech tag if not available.
6	FEATS	Unordered set of syntactic and/or morphological features (depending on the particular language), separated by a vertical bar (|), or an underscore if not available.
7	HEAD	Head of the current token, which is either a value of ID or zero ('0'). Note that depending on the original treebank annotation, there may be multiple tokens with an ID of zero.
8	DEPREL	Dependency relation to the HEAD. The set of dependency relations depends on the particular language. Note that depending on the original treebank annotation, the dependency relation may be meaningfull or simply 'ROOT'.
9	PHEAD	Projective head of current token, which is either a value of ID or zero ('0'), or an underscore if not available. Note that depending on the original treebank annotation, there may be multiple tokens an with ID of zero. The dependency structure resulting from the PHEAD column is guaranteed to be projective (but is not available for all languages), whereas the structures resulting from the HEAD column will be non-projective for some sentences of some languages (but is always available).
10	PDEPREL	Dependency relation to the PHEAD, or an underscore if not available. The set of dependency relations depends on the particular language. Note that depending on the original treebank annotation, the dependency relation may be meaningfull or simply 'ROOT'.

Abb. 9: Parameter des CoNLL Datenformats (*ilk.uvt.nl*).

Ein solcher Output wird in Abbildung 10 mit dem Beispielsatz «Wir spazierten an einem schönen Bach entlang.» illustriert.

```
1   Wir       wir       PRO    PPER   1|Pl|_|Nom          2   subj _   _
2   spazierten spazieren V    VVFIN  1|Pl|Past|_         0   root _   _
3   an        an        PREP   APPR   Dat 2   pp   _   _
4   einem     eine      ART    ART    Indef|Masc|Dat|Sg   6   det  _   _
5   schönen   schön     ADJA   ADJA   Pos|Masc|Dat|Sg|Wk| 6   attr _   _
6   Bach      Bach      N      NN     Masc|Dat|Sg         3   pn   _   _
7   entlang   entlang   PTKVZ  PTKVZ  _                   2   avz  _   _
8   .         .         $.     $.     _                   0   root _   _
```

Abb. 10: Output eines Beispielsatzes im CoNLL Datenformat.

Wie im Beispiel ersichtlich, hat der Parameter «HEAD» des Wortes «schön» die gleiche Zahl (6) wie der Parameter «ID» des Wortes «Bach». Beim Wort «Bach» handelt es sich also um den Kopf («HEAD») des Wortes «schön», welches wiederum das Attribut («attr») des Wortes «Bach» darstellt. Die zwei Ausdrücke stehen somit in einer wie in *Jung* (1995) beschriebenen Dependenzre-

lation zueinander. Basierend auf den Informationen des CoNLL-Outputs lässt sich also eine Beziehung zwischen den beiden Wörtern ermitteln. Dieser Beziehungstyp wird im Zuge dieser Arbeit als «Attributbeziehung» bezeichnet.

Es zeigt sich, dass für die Ermittlung von Beziehungen zwischen verschiedenen Wörtern die Felder «DEPREL» (Dependenzrelation) und «HEAD» von Bedeutung sind. «DEPREL» verweist auf die Art der Beziehung eines Wortes zu seinem Kopf und «HEAD» zeigt an, bei welchem Wort es sich um diesen Kopf handelt. Um diese Zusammenhänge etwas zu verdeutlichen, kann das obige Beispiel wie in Abbildung 11 auch graphisch in einem sogenannten «Dependenzbaum» dargestellt werden.

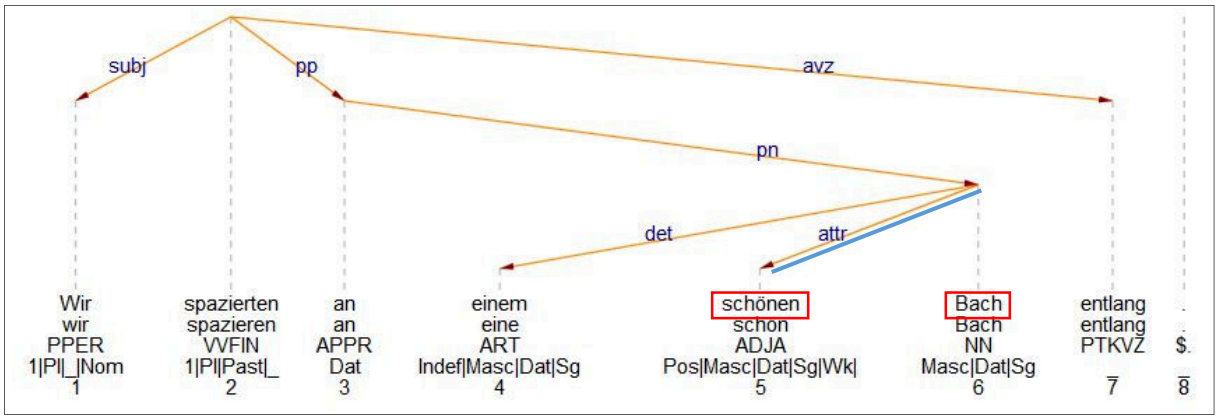


Abb. 11: Dependenzbaum des Satzes: «Wir spazierten an einem schönen Bach entlang».

In diesem Ausgabeformat wird ein Pfeil verwendet, um die Verbindung eines Wortes mit dem jeweiligen Kopf darzustellen. Wie zu sehen ist, zeigt ein Pfeil vom Kopf «Bach» zu seinem Attribut «schön». Die zu ermittelnden Beziehungen zwischen Wörtern können also jeweils als Pfade (blaue Linie) im Dependenzbaum definiert werden. Wie im nächsten Kapitel demonstriert wird, kann ein solcher Pfad auch über mehrere Wörter führen.

5.5.2 Beziehungstypen

Es gibt verschiedene Beziehungsarten, welche zwischen einem Landschaftsbegriff und einem beschreibenden Wort bestehen können. Vergleichen wir die folgenden zwei Sätze:

- «Dies ist ein schöner Bach.»
- «Dieser Bach ist schön.»

Wie bereits gesehen, handelt es sich beim ersten Satz um eine Attributbeziehung zwischen den Wörtern «schön» und «Bach». Beim zweiten Satz besteht ebenfalls eine Beziehung zwischen den Wörtern «schön» und «Bach». Es handelt es sich jedoch um einen anderen Beziehungstyp.

1	Dieser	diese	ART	PDAT	Masc|Nom|Sg	2	det	_	_
2	Bach	Bach	N	NN	Masc|Nom|Sg	3	subj	_	_
3	ist	sein	V	VAFIN	3|Sg|Pres|Ind	0	root	_	_
4	schön	schön	ADV	ADJD	Pos	3	pred	_	_
5	.	.	$.	$.	_	0	root	_	_

Abb. 12: Analyse einer Prädikatbeziehung im CoNLL Datenformat.

Wie in Abbildung 12 zu sehen ist, haben das Subjekt («subj») «Bach» und das Prädikat («pred») «schön» den gemeinsamen syntaktischen Kopf «ist». Diese «Prädikatbeziehung», wie sie fortan genannt wird, wird also auf eine andere Weise ermittelt als eine Attributbeziehung. Wie in Abbildung 13 verdeutlicht wird, besteht der Unterschied vor allem darin, dass der Dependenzpfad (blaue Verbindung) nun über ein weiteres Wort (ist) führt und der Zusammenhang zwischen dem Beschreibungswort und dem Landschaftsbegriff dadurch hergestellt wird, dass sie einen gemeinsamen Kopf haben.

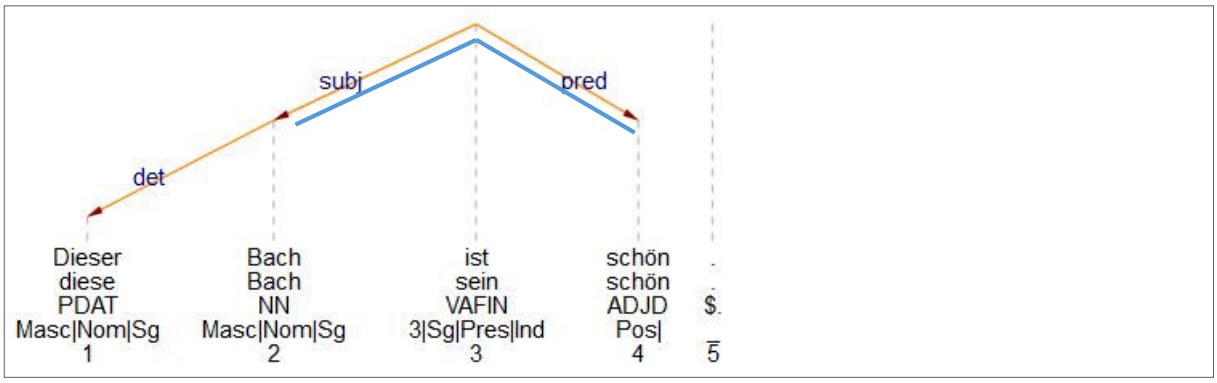


Abb. 13: Dependenzbaum des Satzes «Dieser Bach ist schön.»

Um verschiedene Beziehungstypen zu erkennen, müssen in SentiTours also verschiedene Regeln implementiert werden, nach denen, ausgehend von einem Landschaftsbegriff, nach einem allfälligen Beschreibungswort gesucht wird. Um beispielsweise das Finden einer Prädikatbeziehung zu implementieren, kann folgendermassen vorgegangen werden: Wird ein Landschaftsbegriff entdeckt, wird im CoNLL-Output des entsprechenden Satzes nach einem Wort gesucht, dessen Parameter «DEPREL» den Eintrag «pred» enthält. Dann wird geprüft, ob der Kopf des Wortes mit diesem Eintrag mit dem Kopf des Landschaftsbegriffs übereinstimmt.

Nach diesem Prinzip wird nun versucht, alle möglichen Arten von Beziehungsformen zu implementieren, welche von einem Landschaftsbegriff auf sein allfälliges Beschreibungswort schliessen lassen. Dies mit dem Bewusstsein, dass nicht nur Adjektive, sondern auch Verben und Nomen als Beschreibungswörter dienen können, wie beispielsweise in «Die Schönheit dieser Berge.» oder «Ich mag diese Berge.»

5.5.2.1 Vorläufige Auswahl der Beziehungstypen

Auf der Internetplattform «GitHub» stellen die Entwickler von ParZu eine Liste mit den wichtigsten Dependenzrelationen aus *Foth* (2006) zusammen. Wie gesehen, sind diese wichtig, um allfällige Beziehungen zwischen Wörtern herzustellen. Die aufgelisteten Abhängigkeitsbeziehungen sind jeweils mit einem Beispielsatz versehen. Zur Illustration wird in Abbildung 14 ein Ausschnitt aus dieser Auflistung gezeigt, welcher die Dependenzrelation «ATTR» erklärt.

Abb. 14: Beispiel der Dependenzrelation «ATTR» (*www.github.com*).

Es gilt nun herauszufinden, welche dieser Abhängigkeitsbeziehungen verwendet werden können, um Landschaftsbeschreibungen im Text zu erkennen. Dazu muss zuerst analysiert werden, wie Landschaftselemente im Hikr-Korpus beschrieben werden. Dies geschieht auf qualitative Weise, indem ca. 50 Texte gelesen und auf Landschaftsbeschreibungen untersucht werden.

Implikationen aus der qualitativen Analyse des Hikr-Korpus

Aus der Analyse ergibt sich, dass Landschaftselemente relativ häufig auf einfache Weise attributiv charakterisiert werden. Beispiele hierfür sind Formulierungen wie «Damals hatte es nach der Passüberschreitung noch einen *kleinen Klettersteig*.» oder «Im Aufstieg bis Biflen dann einige *nicht sehr stabile Bergbachquerungen*.» Einige Beschreibungen erfolgen durch ein angehängtes Prädikat: «Denn das *Gras ist nass* und damit recht *schlipfrig*.» Grundsätzlich werden vor allem Adjektive verwendet, um die Landschaft zu charakterisieren. Selten werden Nomen benutzt, um einen Landschaftsbegriff mit einem «Genitiven Modifikator» («GMOD») zu beschreiben: «Jedes Mal staunen wir ob der *Schönheit der Bergwelt* hier oben.» Die Erfassung dieser Arten von Landschaftsbeschreibungen gilt es somit zu implementieren. Dafür müssen die entsprechenden Dependenzrelationen vom Programm erkannt und basierend darauf muss der Dependenzpfad von einem Schlüsselwort (Landschaftsbegriff) zu seinem Komplement (Beschreibungswort) hergestellt werden können.

Der Hikr-Korpus umfasst über 70‘000 Berichte. Die Wahrscheinlichkeit ist also gross, dass noch weitere potentielle Beschreibungsformen existieren, welche es zu erkennen gilt und welche somit bei der Implementierung berücksichtigt werden müssen.

Analyse weiterer möglicher Beziehungstypen

Anstatt nur von den gefundenen Landschaftsbeschreibungen auf die Beziehungstypen zu schliessen, wird deshalb zusätzlich versucht, basierend auf den in *Foth* (2006) enthaltenen Abhängigkeitsbeziehungen mögliche Landschaftsbeschreibungen abzuleiten. Als Erklärung soll die Dependenzrelation «OBJA» dienen. Als Beispiel wird dort der Satz «Ich sehe den Mann.» angegeben. Das beschriebene Element in diesem Satz ist das Akkusativobjekt «Mann». Es wird nun die Annahme getroffen, dass ein Satz wie «Ich mag diese Berge.» durchaus einige Male auftreten könnte. Dasselbe gilt für Sätze, in denen der Landschaftsbegriff durch ein finites Verb anstatt durch ein Prädikat beschrieben wird: «Dieser Berg gefällt mir.» Anders sieht es aus, wenn es sich beim Landschaftsbegriff um ein Dativobjekt handeln sollte («OBJD»). In diesem Fall ist eine plausible Landschaftsbeschreibung kaum denkbar. Eine grammatikalisch zwar mögliche, aber dennoch unwahrscheinliche Formulierung könnte folgendermassen aussehen: «Ich huldige diesem Berg.» Es wird deshalb davon ausgegangen, dass die Erkennung eines solchen Beziehungstyps keinen Mehrwert darstellt. Noch deutlicher sieht es beispielsweise bei der Dependenzrelation «DET» aus. Wie in Abbildung 15 ersichtlich, verweist diese lediglich darauf, dass es sich bei einem Wort (z.B. «der») um das Bestimmungswort eines anderen Wortes (z.B. «Mann») handelt. Eine Verwendung dieser Dependenzrelation für die Erfassung einer Landschaftsbeschreibung ist deshalb nicht denkbar.

Abb. 15: Beispiel der Dependenzrelation «DET» (*www.github.com*).

Die aus einer ersten Auswahl resultierenden Beziehungstypen und die Dependenzrelationen, welche zur Erfassung dieser Beziehungstypen erkannt werden müssen, sind in Tabelle 2 aufgeführt. Sie sind jeweils mit einem Beispielsatz versehen. Der Landschaftsbegriff und das Beschreibungswort, mit welchem dieser in der entsprechenden Beziehung steht, sind fett markiert.

Beziehungstyp	*DEPREL*	*Beispielsatz*
Attributbeziehung	ATTR	«Wir spazierten an einem **schönen Bach** entlang.»
Prädikatbeziehung	PRED	«Dieser **Berg** ist **schön**.»
Akkusativobjekt-Beziehung	OBJA	«Ich **mag** diese **Berge**.»
Finite-Verb-Beziehung	SUBJ	«Dieser **Berg gefällt** mir.»
Genitivnomen-Beziehung	GMOD	«Wir geniessen die **Schönheit** dieser **Berge**.»

Tab. 2: Erste Auswahl zu implementierender Beziehungstypen.

Es soll dabei erwähnt sein, dass die Benennungen der Beziehungstypen für diese Arbeit so bestimmt wurden und sich nicht auf eine offizielle Nomenklatur beziehen.

Verschiedene Zeitformen

Die qualitative Analyse zeigt, dass Berichte sowohl in der Gegenwarts- als auch in der Vergangenheitsform verfasst werden. Sätze mit Landschaftsbeschreibungen stehen oft in der Zeitform Präsens («Man folgt einfach dem *kleinen Talkessel* bis zum Schibergsattel und…»), aber auch im Präteritum («Das *Gestein* war zum Teil sehr *lose*.»). Selten tauchen Beschreibungen in der Zeitform Perfekt auf («Aber dieser Klettersteig *hat* mir sehr gut *gefallen*.») Landschaftsbeschreibungen im Plusquamperfekt und solche in der Zukunftsform wurden nicht gefunden und sind deshalb wohl selten. Wird eine Beschreibung in einer anderen Zeitform notiert, können sich die syntaktischen Abhängigkeitsbeziehungen innerhalb des Satzes verändern. Welche Auswirkungen die Schreibweise der unterschiedlichen Zeitformen auf die Ermittlung der verschiedenen Beziehungstypen hat, wird im Folgenden erläutert.

Die Attributbeziehung ist nicht sensitiv gegenüber einer Veränderung der Zeitform. Das Beschreibungswort ist dem Landschaftsbegriff stets direkt vorangestellt und behält seine attributive Rolle. Ebenso ist die Ermittlung einer Genitivnomen-Beziehung unabhängig von der Zeitform des entsprechenden Satzes. Die Ermittlung einer Finite-Verb-Beziehung verändert sich hingegen mit der Zeitform. Wird der fiktive Satz «Dieser Berg gefällt mir.» in der ersten Vergangenheitsform geschrieben, wird er um ein Hilfsverb («AUX») ergänzt: «Dieser Berg *hat* mir gefallen.» Wie der Vergleich der CoNLL-Outputs in Abbildung 16 unten zeigt, verändert sich dadurch der Dependenzpfad zwischen dem Landschaftsbegriff «Berg» und dem Beschreibungswort «gefallen». Dies muss vom Programm berücksichtigt werden. Auch bei einer Prädikatbeziehung muss die Bestimmungsweise angepasst werden, wenn sie im Perfekt steht. Beim ebenfalls fiktiven Satz «Dieser Berg ist schön.» besteht das Prädikat aus dem Satzteil «ist schön». Beim Satz «Dieser Berg ist schön gewesen.» hingegen wird es um das Hilfsverb «gewesen» erweitert. Auch hier erfolgt die Ermittlung der Beziehung zwischen Beschreibungswort und Landschaftsbegriff nun über einen anderen Pfad. Zusätzlich zur bisherigen Auswahl kommen deshalb zwei neue Beziehungstypen hinzu, welche in Tabelle 3 aufgeführt sind.

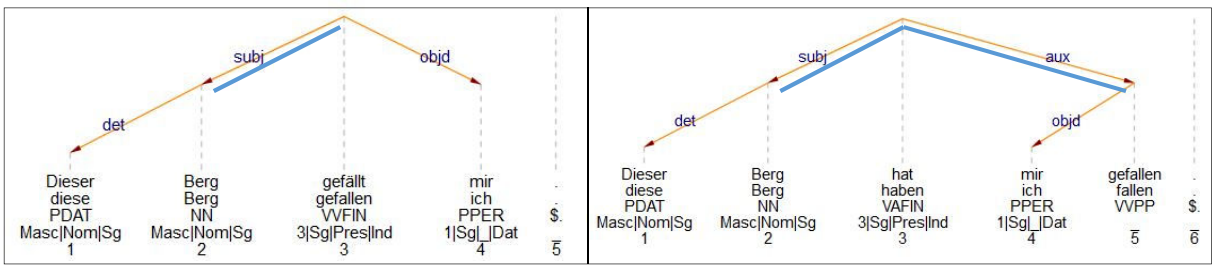


Abb. 16: Dependenzpfade einer Finite-Verb-Beziehung im Präsens und im Perfekt.

Beziehungstyp	*DEPREL*	*Beispielsatz*
Hilfsverbbeziehung	SUBJ/AUX	«Dieser **Berg** hat mir **gefallen**.»
Prädikatbeziehung Perfekt	SUBJ/AUX/PRED	«Dieser **Berg** ist **schön gewesen**.»

Tab. 3: Zusätzliche Beziehungstypen für die Berücksichtigung unterschiedlicher Zeitformen.

5.5.2.2 Definitive Auswahl der Beziehungstypen

Die Auswahl der verschiedenen Beziehungstypen erfolgte in einem ersten Schritt nicht frei von gewissen Annahmen. Es gilt nun die Sinnhaftigkeit dieser Auswahl zu überprüfen. Dazu wurden 100 zufällige Texte mit den sieben provisorisch implementierten Beziehungstypen analysiert. Darin wurden 99 Landschaftsbeschreibungen gefunden. Im Output enthalten waren der Beziehungstyp, die zwei miteinander in Beziehung stehenden Wörter sowie der Satz, von welchem diese extrahiert wurden. Anhand dieser Outputs wurde eine Statistik erstellt. In dieser kann nun überprüft werden, welchen Anteil jeder einzelne Beziehungstyp an der gesamten Anzahl an entdeckten Beziehungstypen hat. Zusätzlich wird manuell erfasst, wie viele der ermittelten Wortbeziehungen eines Beziehungstyps ein plausibles Resultat ergeben (Plausibilität). Plausibel in dem Sinne, dass das Beschreibungswort auch wirklich ein Wort ist, welches eine Eigenschaft des beschriebenen Landschaftsbegriffs charakterisiert. Das Resultat dieser Analyse ist in Abbildung 17 aufgeführt.

Attributbeziehung:		**Akkusativobjekt-Beziehung:**		**Finite-Verb-Beziehung:**	
Anteil:	51.5%	Anteil:	14.1%	Anteil:	14.1%
Plausibilität:	80.4%	Plausibilität:	14.3%	Plausibilität:	14.3%
Prädikatbeziehung:		**Genitivnomen-Beziehung:**		**Hilfsverbbeziehung:**	
Anteil:	9.1%	Anteil:	7.1%	Anteil:	4.1%
Plausibilität:	44.4%	Plausibilität:	14.3%	Plausibilität:	0.0%
Prädikatbeziehung Perfekt					
Anteil:	0.0%				
Plausibilität:	/				

Abb. 17: Überprüfung der ersten Auswahl an Beziehungstypen.

Dieser Zwischenauswertung ist zu entnehmen, dass es sich bei den meisten ermittelten Beziehungstypen um Attributbeziehungen handelt. Davon wurden 80.4 Prozent als plausibel erachtet. Aufgrund des hohen Werts erscheint es sinnvoll, diesen Beziehungstyp zu verwenden, um Landschaftsbeschreibungen im Hikr-Korpus ausfindig zu machen. Die übrigen 19.6 Prozent der erfassten Attributbeziehungen stellen entweder gar keine Landschaftsbeschreibungen dar, oder diese

werden nicht richtig erfasst Zu einem nicht plausiblen Resultat führt beispielsweise die Analyse des Ausdrucks «schön gelegene Alp». Hierbei setzt sich die Beschreibung eigentlich aus den zwei Wörtern «schön» und «gelegen» zusammen. Mit der momentan implementierten Methodik kann jedoch nur ein Beschreibungswort erfasst werden, welches in diesem Fall das Wort «gelegen» ist. Von den weiteren Beziehungstypen scheint die Prädikatbeziehung als einzige ebenfalls einen akzeptablen Plausibilitätswert aufzuweisen. Dieser ist mit 44.4 Prozent dennoch relativ tief. Die Ausschnitte aus dem Output in Abbildung 18 zeigen den Grund dafür.

Bericht: 20; Prädikatbeziehung; ***wild Landschaft;*** *Die* ***Landschaft*** *ist* ***wild****, Natur pur.*

Bericht: 13; Prädikatbeziehung; ***Gipfel Gipfel****; Das Brunegghorn ist ein schöner* ***Gipfel*** *und bietet viel Aussicht.*

Abb. 18: Plausible (grün) und nicht plausible (rot) ermittelte Prädikatbeziehung.

Der erste Ausschnitt repräsentiert ein Resultat, wie es zu erwarten wäre. Das Prädikat «wild» beschreibt den Landschaftsbegriff «Landschaft». Beim zweiten Beispiel wird jedoch der Landschaftsbegriff «Gipfel» als Beschreibungswort von sich selbst erfasst. Ein Blick in den CoNLL-Output des entsprechenden Satzes in Abbildung 19 offenbart den Grund für dieses Verhalten.

1	Das	die	ART	ART	Def|Neut|Nom|Sg	2	det	_	_			
2	Brunegghorn	Brunegghorn	N	NN	Neut|Nom|Sg	3	subj	_	_			
3	ist	sein	V	VAFIN	3|Sg|Pres|Ind	0	root	_	_			
4	ein	eine	ART	ART	Indef|Masc|Nom|Sg	6	det	_	_			
5	schöner	schön	ADJA	ADJA	Pos|Masc|Nom|Sg|St|	6	attr	_	_			
6	Gipfel	Gipfel	N	NN	Masc|Nom|Sg	3	pred	_	_			
7	und	und	KON	KON	_	3	kon	_	_			
8	bietet	bieten	V	VVFIN	_|_|Pres|_	7	cj	_	_			
9	viel	viel	ART	PIAT	Fem|_|Sg	10	det	_	_			
10	Aussicht	Aussicht	N	NN	Fem|_|Sg	8	obja	_	_			
11	.	.	$.	$.	_	0	root	_	_			

Abb. 19: CoNLL-Output einer nicht plausiblen Prädikatbeziehung.

Da der Landschaftsbegriff selbst eine prädikative Rolle einnimmt, wird er fälschlicherweise auch als Beschreibungswort erfasst. Dieser Fehler kann vermieden werden, indem die zusätzliche Bedingung implementiert wird, dass es sich beim Prädikat um ein Adjektiv handeln muss. Diese Korrektur hätte bei der aktuellen Analyse zu einer Plausibilität von 88.9 Prozent geführt. Diese Anpassung lässt sich auch auf die Prädikatbeziehung in der Zeitform Perfekt übertragen. Auch wenn dieser Beziehungstyp keinen Treffer erzielt hat, wird davon ausgegangen, dass bei einem Vorkommen die Plausibilität ähnliche Werte annehmen wird wie bei der Prädikatbeziehung in der Zeitform Präsens (eine spätere Überprüfung hat eine Plausibilität von 54 Prozent ergeben). Die beiden Beziehungstypen werden deshalb ebenfalls beibehalten.

Alle übrigen Beziehungstypen weisen tiefe Plausibilitätswerte auf. Der Blick auf die Outputs zeigt, dass es sich bei den ermittelten Wortbeziehungen zwar jeweils um die gesuchten Beziehungstypen handelt. Meist stellen die sich auf die Landschaftsbegriffe beziehenden Ausdrücke jedoch keine eigentlichen Beschreibungswörter dar. In den Abbildungen 20 bis 22 sind jeweils für jeden der übrigen Beziehungstypen ein plausibles und ein nicht plausibles Resultat aufgeführt:

Bericht: 30; Genitivnomen-Beziehung; ***Schönheit Berg****; Aber was soll ' s: es kann zu Arroganz ausarten, wenn man die* ***Schönheit*** *der* ***Berge*** *ständig nur für sich allein besitzen will.*

Bericht: 20; Genitivnomen-Beziehung; ***Seite Schlucht****; Auf der anderen* ***Seite*** *der* ***Schlucht*** *wandern wir nun wieder zurück.*

Abb. 20: Plausible und nicht plausible ermittelte Genitivnomen-Beziehung.

Bericht: 46; Akkusativobjekt-Beziehung; ***bewundern Berg****; Eine wunderbare Höhenwanderung, wie geschaffen für einen 75 +, dem die Kraft für Gipfelerstürmungen ausgegangen ist, der es aber unheimlich geniesst, die* ***Berge*** *von unten zu* ***bewundern****.*

Bericht: 17; Akkusativobjekt-Beziehung; ***erreichen Gipfel****; Nach einer kurzen Rast vor dem Abstieg zur Scharte ging es nicht mehr lange, und über grosse Blöcke und ausgesetzte Grätli* ***erreichten*** *wir den* ***Gipfel*** *des Jegihorn.*

Abb. 21: Plausible und nicht plausible ermittelte Akkusativobjekt-Beziehung.

Bericht: 19; Finite-Verb-Beziehung; ***öffnen Landschaft****; Beeindruckend ist, wie sich hier die* ***Landschaft öffnet****.*

Bericht: 7; Finite-Verb-Beziehung; ***sein Schneefeld****; Ab knapp 2300 m wird es wieder flacher, immer wieder* ***sind Schneefelder*** *und Stellen mit Felsblöcken zu queren.*

Abb. 22: Plausible und nicht plausible ermittelte Finite-Verb-Beziehung.

Die Beispiele zeigen, dass von diesen Beziehungstypen viele Ausdrücke erfasst werden, welche keine Beschreibungswörter sind. Dies ist eigentlich nicht weiter verwunderlich, da es sich anders als bei der Attributbeziehung um Verben (z.B. Akkusativobjekt-Beziehung) respektive Nomen (Genitivnomen-Beziehung) handelt.

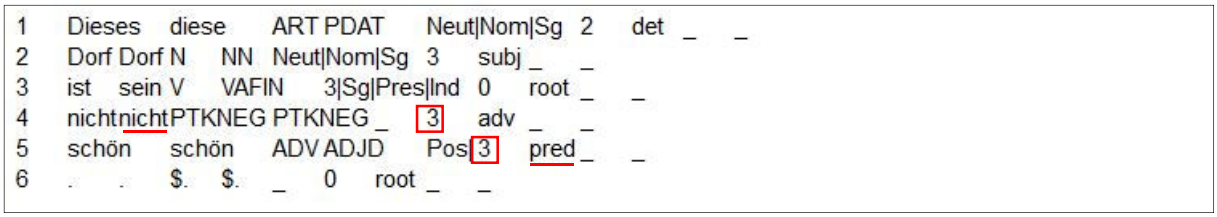
Es stellt sich die Frage, wie SentiTours zwischen beschreibenden und nicht beschreibenden Wörtern unterscheiden könnte. Eine Möglichkeit könnte die Erstellung einer Liste mit Begriffen sein, welche mit grosser Wahrscheinlichkeit verwendet werden, um Dinge zu beschreiben. Diese Problemstellung kann aus Zeitgründen im Rahmen dieser Arbeit aber nicht weiter behandelt werden. Aus diesem Grund werden diese Beziehungstypen nicht mehr weiter berücksichtigt. In SentiTours wird also die Erkennung der Attributbeziehung, Prädikatbeziehung und der Prädikatbeziehung Perfekt implementiert. Des Weiteren werden nur Adjektive als Beschreibungswörter erfasst. Wie

weiter oben gezeigt, sollte so ein grosser Teil der Landschaftsbeschreibungen mit einer hohen Plausibilität erfasst werden.

5.6 Erfassen von Negationen und Adverbien

5.6.1 Negationen

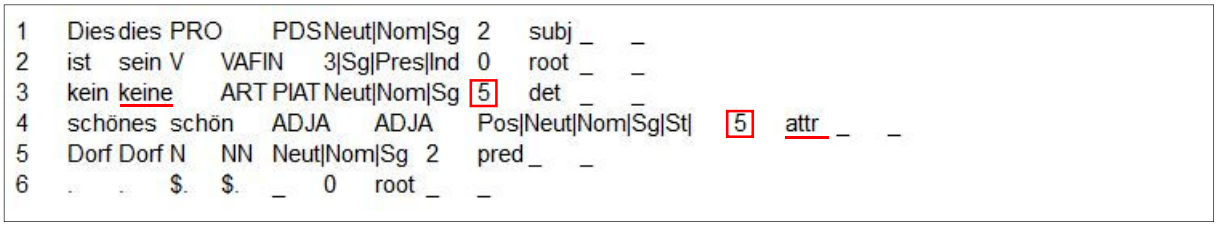
Die Negationen lassen sich nach ähnlichen Prinzipien ermitteln wie die obigen Beziehungstypen. Wie in Abbildung 23 ersichtlich, hat in einer Verneinung der Ausdruck «nicht» denselben syntaktischen Kopf wie das Adjektiv, auf welches er sich bezieht. Diese Form der Verneinung kann nur in Prädikatbeziehungen im Präsens oder Perfekt vorkommen und nicht in Attributbeziehungen.



1	Dieses	diese	ART	PDAT	Neut|Nom|Sg	2	det	_	_
2	Dorf	Dorf	N	NN	Neut|Nom|Sg	3	subj	_	_
3	ist	sein	V	VAFIN	3|Sg|Pres|Ind	0	root	_	_
4	nicht	nicht	PTKNEG	PTKNEG	_	3	adv	_	_
5	schön	schön	ADV	ADJD	Pos	3	pred	_	_
6	.	.	$.	$.	_	0	root	_	_

Abb. 23: CoNLL-Output einer Verneinung mit dem Ausdruck «nicht».

Beim Wort «kein» und seinen Deklinationen («keine», «keines» etc.) besteht dieselbe Regel (Abbildung 24). Diese Form der Verneinung kann wiederum nur in Attributbeziehungen auftreten.



1	Dies	dies	PRO	PDS	Neut|Nom|Sg	2	subj	_	_
2	ist	sein	V	VAFIN	3|Sg|Pres|Ind	0	root	_	_
3	kein	keine	ART	PIAT	Neut|Nom|Sg	5	det	_	_
4	schönes	schön	ADJA	ADJA	Pos|Neut|Nom|Sg|St|	5	attr	_	_
5	Dorf	Dorf	N	NN	Neut|Nom|Sg	2	pred	_	_
6	.	.	$.	$.	_	0	root	_	_

Abb. 24: CoNLL-Output einer Verneinung mit dem Ausdruck «kein».

Um Negationen zu erkennen, wird also zuerst nach dem Vorkommen von Verneinungswörtern gesucht. Anschliessend wird überprüft, ob sich diese auf die Beschreibungswörter der Landschaftsbegriffe beziehen oder nicht.

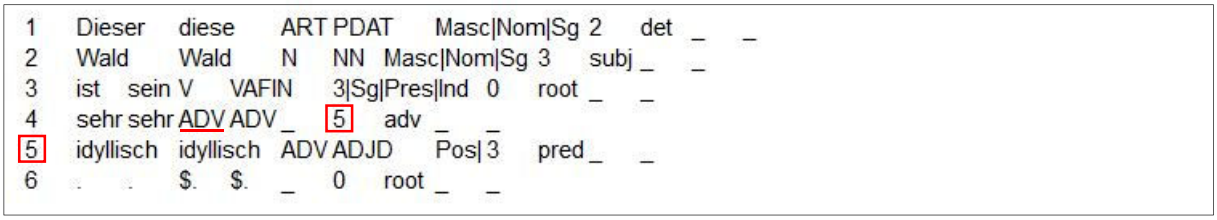
5.6.1.1 Plausibilität der identifizierten Negationen

Wie bei den Beziehungstypen gilt es auch hier zu testen, inwiefern die durch diese Methodik extrahierten Negationen plausibel sind. Dazu wird der Output einer Stichprobe von 50 zufällig ermittelten Landschaftsbeschreibungen, in welchen Negationen entdeckt wurden, manuell überprüft. Es stellt sich dabei heraus, dass in diesen 50 Fällen einmal (2 Prozent) eine Negation erfasst

wurde, wo das Beschreibungswort in Wirklichkeit nicht negiert wird. Die extrahierten Negationen sind also in den allermeisten Fällen plausibel.

5.6.2 Adverbien

Auch modifizierende Adverbien könnten mit einer ähnlichen Methodik erfasst werden. Wie in Abbildung 25 angedeutet, müsste dafür nach Adverbien gesucht werden, welche im CoNLL Ausgabeformat beim Parameter «CPOSTAG» mit «ADV» gekennzeichnet sind. Danach müsste überprüft werden, ob es sich beim Kopf dieses Adverbs um das Beschreibungswort handelt. Wie bereits erwähnt, werden die Adverbien für diese Arbeit jedoch nicht berücksichtigt.



1	Dieser	diese	ART	PDAT	Masc|Nom|Sg	2	det	_	_
2	Wald	Wald	N	NN	Masc|Nom|Sg	3	subj	_	_
3	ist	sein	V	VAFIN	3|Sg|Pres|Ind	0	root	_	_
4	sehr	sehr	ADV	ADV	_	5	adv	_	_
5	idyllisch	idyllisch	ADV	ADJD	Pos|	3	pred	_	_
6	.	.	$.	$.	_	0	root	_	_

Abb. 25: CoNLL-Output einer durch ein Adverb modifizierten Landschaftsbeschreibung.

5.7 Ermitteln des Sentiment-Werts

Für jede entdeckte Beziehung wird, falls möglich, ein Sentiment-Wert ermittelt. Dazu wird das Beschreibungswort in SentiWS nachgeschlagen. Wird das Wort gefunden, wird der entsprechende Sentiment-Wert herausgelesen und in den Output von SentiTours integriert. Der Sentiment-Wert beschreibt somit bei jedem gefundenen Beziehungstyp die Gefühlslage oder die Meinung, welche dem beschriebenen Landschaftsbegriff entgegengebracht wird. Wird das Wort im Sentiment-Lexikon nicht gefunden, wird dies im Output vermerkt. Wurde vorgängig eine Negation festgestellt, wird der Sentiment-Wert mit -1 multipliziert, da sich die Werte bei SentiWS auf einer linearen Skala von -1 bis 1 befinden.

5.8 Behandlung der Toponyme

In Kapitel 5.2 wurde erwähnt, dass Landschaftsbegriffe darauf überprüft werden, ob es sich dabei um Toponyme handelt. Der Grund dafür wird nun in diesem Kapitel erläutert.

Die qualitative Analyse des Hikr-Korpus hat nämlich gezeigt, dass Leute gelegentlich Toponyme beschreiben. Diese Toponyme enthalten in ihrem Namen häufig eine Referenz auf ein generisches Landschaftselement. So steht der Begriff «Eigernordwand» in der Realität auch wirklich für eine Nordwand. Oft wird in den Texten auch einfach nur der generische Teil des Toponyms erwähnt. In diesem Fall würde also häufig nur von der «Nordwand» gesprochen. Es gibt hingegen auch

Ortsnamen, bei denen der physische Bezug zur benannten Landschaftsform nicht besteht. Ein Beispiel hierfür ist die Gemeinde «Berg» im Kanton Thurgau. Würde im Text eine Beschreibung dieser Gemeinde entdeckt, so wäre in gewissen Fällen nicht zu unterscheiden, ob es sich dabei um einen tatsächlichen Berg handelt oder nicht. In *Derungs et al.* (2011) wird zwar darauf hingewiesen, dass die in Toponymen erwähnten generischen Bestandteile mit der erwarteten realen Repräsentation oft übereinstimmen. Dennoch bestehen einige Ausnahmen. Um abschätzen zu können, wie gross der Einfluss dieser Ausnahmen ist, sollte das Programm SentiTours fähig sein, Toponyme im Text zu erkennen. Hierzu werden die Metainformationen über die von den Leuten passierten Wegpunkte zu Hilfe genommen. Wird im Text ein Landschaftsbegriff entdeckt, wird überprüft, ob dieser Begriff auch in den Wegpunkten zu finden ist. Ist dies der Fall, wird im Output der Vermerk «Pot_Toponym» (für: potentielles Toponym) gemacht. Eine Übereinstimmung mit einem Wegpunkt wird auch dann erfasst, wenn der Landschaftsbegriff nur einen Teil eines solchen ausmacht. So würde zum Beispiel der Landschaftsbegriff «Nordwand» im Wegpunkt «Eigernordwand» gefunden und als potentielles Toponym bezeichnet. Anschliessend kann anhand einer Analyse einer Auswahl von 50 Landschaftsbeschreibungen mit potentiellen Toponymen ermittelt werden, wie oft diese auch wirklich ein Landschaftselement repräsentieren.
Von allen beschriebenen Landschaftsbegriffen wurden schliesslich 4‘354 (ca. 4 Prozent) als potentielle Toponyme gekennzeichnet. Davon wurden 50 zufällige Beschreibungen mithilfe der Sätze, aus welchen sie extrahiert wurden, analysiert. Aus dieser Auswahl konnte in zwei Fällen (4 Prozent) beobachtet werden, dass der ermittelte Landschaftsbegriff nicht mit dem erwarteten generischen Objekt übereinstimmt. Dies war beispielsweise bei der Ortschaft «Burgen» der Fall, welcher auf eine Burg schliessen lässt, in Wirklichkeit aber einen Weiler darstellt. Der Einfluss von generischen Ortsnamen ohne wirklichen Bezug auf das tatsächliche Landschaftselement ist also zu vernachlässigen.

5.9 Zusammenfassung der Funktionsweise von SentiTours

In diesem Kapitel soll die Funktionsweise der entwickelten Anwendung zur Extrahierung von Landschaftsbeschreibungen aus dem Hikr-Korpus noch einmal zusammenfassend erklärt werden.
Die Texte sowie die in Kapitel 5.3 erwähnten Zusatzinformationen werden aus den Webseiten herausgelesen. Dann werden die Texte Satz für Satz unter Einbindung der Software ParZu analysiert und für jeden Satz ein CoNLL-Output generiert. In diesen Outputs wird zuerst nach den Landschaftsbegriffen gesucht, welche, wie in Kapitel 5.4 erklärt, ermittelt wurden. Wird ein Landschaftsbegriff gefunden, werden unter Anwendung der implementierten Regeln zur Erkennung der verschiedenen Wortbeziehungen Wörter gesucht, welche sich auf diesen Landschaftsbegriff

beziehen. Zudem wird kontrolliert, ob es sich bei diesem Wort um ein Adjektiv handelt. Des Weiteren wird anhand eines Vergleichs mit den Wegpunkten ermittelt, ob es sich beim gefundenen Landschaftsbegriff um ein Toponym handeln könnte. Wird kein beschreibendes Adjektiv gefunden, wird dieser Landschaftsbegriff ignoriert und der nächste auf eine allfällige Beschreibung überprüft. Wird jedoch ein beschreibendes Adjektiv gefunden, wird überprüft, ob ein Verneinungswort («nicht» oder «kein» und seine Deklinationen) existiert, welches denselben syntaktischen Kopf hat wie dieses Adjektiv. Das Adjektiv wird dann in der Sentiment-Bibliothek SentiWS nachgeschlagen und der allfällige Polaritätswert gespeichert. Wurde vorher eine Negation entdeckt, wird der Polaritätswert mit -1 multipliziert. Die Dokumentation zu SentiTours befindet sich in Anhang D.

5.10 Auswertung des Outputs von SentiTours

Wie schliesslich der erhaltene Output für jede ermittelte Landschaftsbeschreibung aussieht, wird in Abbildung 26 an einem Beispiel illustriert.

```
Bericht: 256; Attributbeziehung; herrlich Gipfel; Negation_Nein; Pot_Toponym;
SentiWS: Ja; Sentiment: 0.4821; ;Welt Schweiz Uri; 12 Februar 2011; Hochtou-
ren; WS; 6:30; Bombo; Realp 1538 m (165) Realp - Parkplatz Furkapassstrasse
1538 m (36) Furkareuss - Brücke P.1603m 1603 m (29) Oberstafel 2221 m (38) Ro-
tondohütte SAC 2570 m (67) Witenwasserenstock - Ostgipfel 3025 m (13) Witen-
wasserenstock 3082 m (8) Tälligrat / Rottälligrat - P.2748m 2748 m (12) Stel-
liboden 2209 m (19)
```

Abb. 26: Beispiel-Output von SentiTours.

Darin enthalten sind in dieser Reihenfolge folgende Informationen: Bericht-ID; Beziehungstyp; Beschreibungswort und Landschaftsbegriff; Negationen-Status; Toponym-Status; Angabe, ob Beschreibungswort in SentiWS gefunden wurde; entsprechender Sentiment-Wert; Region; Tour-Datum; Aktivität; Schwierigkeitsgrad; Tour-Dauer; Pseudonym des Autors oder der Autorin; Wegpunkte.

Die in dieser Form extrahierten Landschaftsbeschreibungen aus dem Hikr-Korpus bilden die Grundlage für die weiterführenden Analysen, welche in den folgenden Kapiteln erläutert werden. Sie sollen aufzeigen, welche Eigenschaften und welche Aussagekraft die extrahierten Beschreibungen haben. Auch werden damit allfällige räumliche Gesetzmässigkeiten untersucht. Die Analysen sollen darüber hinaus Aufschluss geben, inwiefern die Anwendung der Technik der Sentiment Analysis im Kontext von Landschaftsbewertungen sinnvoll ist.

5.10.1 Vergleich der Landschaftselemente

Es besteht die Erwartung, dass morphologisch unterschiedliche Landschaftselemente auch unterschiedlich beschrieben werden. Diese Vermutung soll überprüft werden. Dazu werden die Landschaftsbegriffe in einem ersten Schritt mithilfe einer multidimensionalen Skalierung (MDS) entsprechend der Ähnlichkeit ihrer Beschreibungen im abstrakten Raum angeordnet. So kann anhand der Distanzen zwischen den Begriffen auf ihre Ähnlichkeit geschlossen werden. Um die Gruppierung der Begriffe erklären zu können, wird anschliessend ein Clustering durchgeführt, bei welchem einander ähnliche Landschaftsbegriffe in Gruppen zusammengefasst werden. Dann werden die 10 wichtigsten Beschreibungswörter jeder Gruppe mithilfe des sogenannten «Tf-idf-Masses» (term frequency inverse document frequency) ermittelt und miteinander verglichen. In einem letzten Schritt wird eine Interpolation basierend auf den durchschnittlichen Sentiment-Werten der im Raum angeordneten Landschaftsbegriffe durchgeführt. Dies um zu überprüfen, ob einander näher liegende Begriffe und Begriffsgruppen auch über ähnliche Sentiment-Werte verfügen.

5.10.1.1 Multidimensionale Skalierung

Um zu untersuchen, wie ähnlich die verschiedenen Landschaftsbegriffe beschrieben werden, wird die sogenannte «Kosinus-Ähnlichkeit» berechnet. Diese wird oft verwendet, um die Ähnlichkeit zwischen zwei Dokumenten respektive Dokument-Vektoren zu ermitteln. Dabei wird von jedem Begriff oder von einer bestimmten Auswahl an Begriffen in einem Dokument die Häufigkeit ermittelt. Aus den ermittelten Frequenzen kann dann ein n-dimensionaler Begriffsvektor für jedes Dokument gebildet werden. Die Ähnlichkeit zwischen zwei Dokumenten entspricht dann der Stärke des Zusammenhangs zwischen ihren Vektoren. Diese wird quantifiziert als der Kosinus des Winkels zwischen den Vektoren. Sind zwei Dokumente a und b gegeben, wird die Kosinus-Ähnlichkeit wie in Abbildung 27 berechnet.

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|}$$

Abb. 27: Formel zur Berechnung der Kosinus-Ähnlichkeit (*Huang*, 2008).

Da die Frequenz eines Terms in einem Dokument nicht negativ sein kann, resultiert ein positiver Wert zwischen 0 und 1. Ein Vorteil dieser Methode ist die Unabhängigkeit von der Dokumentenlänge. Stehen die Termfrequenzen zwischen zwei Dokumenten im selben Verhältnis und sind nur die absoluten Zahlen unterschiedlich, werden die Dokumente als identisch betrachtet (*Huang*, 2008). Treten zum Beispiel die Wörter «Haus» und «Auto» in einem Dokument ein- und zweimal und in einem anderen Dokument zwei- und viermal auf, zeigen die Vektoren (1,2) und (2,4) in die

gleiche Richtung und die Dokumente werden als gleich betrachtet. Folglich wäre der Wert für die Kosinus-Ähnlichkeit in diesem Fall 1.

Anstelle von Dokumenten werden im vorliegenden Fall die Beschreibungen der verschiedenen Landschaftsbegriffe miteinander verglichen. Es werden jedoch nur diejenigen Begriffe verglichen, von welchen mindestens 200 Beschreibungen vorhanden sind, was in 91 Fällen zutrifft. Damit kann eine repräsentative Anzahl und Verteilung der Beschreibungswörter gewährleistet werden. Ausserdem wäre die Gegenüberstellung aller Begriffe in einer MDS aus Darstellungsgründen problematisch. Für jeden dieser 91 Landschaftsbegriffe wird in einer geordneten Liste stets in derselben Reihenfolge vermerkt, wie oft jedes der gefundenen Beschreibungswörter verwendet wird, um diesen Begriff zu beschreiben. Eine solche Angabe kann auch den Wert 0 annehmen. Da der überwiegende Teil der Adjektive im gesamten Korpus nur wenige Male vorkommt, wird jeder Begriffsvektor einen grossen Anteil Nullwerte haben. Laut *Manning & Schütze (1999)* ist die Kosinus-Ähnlichkeit jedoch relativ robust in Fällen, wo Vektoren mit einer stark unterschiedlichen Anzahl an Nullwerten miteinander verglichen werden. Schlussendlich werden die Listen aller Landschaftsbegriffe miteinander verglichen. Die erhaltenen Werte für die Kosinus-Ähnlichkeit werden dann in einer Matrix dargestellt, von welcher ein Ausschnitt in Tabelle 4 aufgeführt ist.

	Berg	Restaurant	Gelände	Weglein	Weg	Gipfel
Berg	1					
Restaurant	0.123085591	1				
Gelände	0.052665286	0.096257137	1			
Weglein	0.11974234	0.204453167	0.26162471	1		
Weg	0.113653057	0.119104579	0.211573378	0.408271126	1	
Gipfel	0.762216866	0.243911948	0.071725227	0.23148897	0.161761207	1

Tab. 4: Ausschnitt der Matrix zur Erstellung der MDS.

Diese Matrix kann anschliessend verwendet werden, um anhand der darin enthaltenen Ähnlichkeitswerte die Begriffe mithilfe einer MDS im Raum anzuordnen. Wie bereits angedeutet, wird bei einer MDS die Ähnlichkeit respektive Unähnlichkeit zwischen zwei Objekten durch deren Distanz in einem niedrigdimensionalen Raum repräsentiert. Dieses Verfahren kann als Technik für explorative Datenanalysen verwendet werden. Dies um ein allfälliges Muster in Daten zu erkennen, bei denen keine explizite Theorie herbeigezogen werden kann, um deren Ausprägungen oder Zusammenhänge vorauszusagen (*Borg & Groenen*, 2005).

Mit der Statistiksoftware SPSS wird ein sogenanntes «proximity scaling» («PROXSCAL») durchgeführt. Der Parameter «proximities» wird auf «similarities» gesetzt, da es sich bei den Werten in der MDS um Ähnlichkeitswerte handelt. Die Anzahl Dimensionen wird auf zwei festgelegt.

5.10.1.2 Clustering

Clustering-Algorithmen ordnen Objekte verschiedenen Gruppen zu, auch Cluster genannt. Im Gegensatz zur Klassifikation erfolgt die Clusterbildung in einem unüberwachten Verfahren und benötigt keine Trainingsdaten (*Manning & Schütze*, 1999). Das Clustering dient hier dazu, Landschaftsbegriffe mit ähnlichen Beschreibungen in Gruppen zusammenzufassen. Als Clusteringverfahren wird das sogenannte «Ward-Verfahren» gewählt. Das Ziel dieser Methode ist es, jeweils diejenigen Objekte einem Cluster zuzuordnen, welche die Streuung (Varianz) in der entsprechenden Gruppe möglichst wenig erhöhen (*Backhaus et al.,* 2011).

Danach kann ermittelt werden, welche Beschreibungswörter respektive Adjektive für eine Gruppe von Landschaftsbegriffen typisch sind. Dazu wird für jedes Adjektiv eines Clusters das Tf-idf-Mass berechnet. Auch dieses Mass wird oft zum Vergleich von Textdokumenten herangezogen. Die Vorkommenshäufigkeit (term frequency) gibt an, wie oft ein Begriff in einem Dokument vorkommt. Da ein Wort aber möglicherweise für die Beschreibung aller Landschaftsbegriffe häufig verwendet wird, sollte ein vielfaches Vorkommen eines Begriffs nicht auch in gleichem Masse zur Relevanz beitragen. Deshalb wird die Vorkommenshäufigkeit mit der inversen Dokumenthäufigkeit (idf) normalisiert, bei welcher berücksichtigt wird, in wie vielen Dokumenten der entsprechende Begriff auftaucht (*Ramos*, 2003). Im vorliegenden Fall entspricht ein Dokument wiederum der Gesamtheit aller Beschreibungen eines Landschaftsbegriffs.

5.10.1.3 Interpolation der Sentiment-Werte

Da in der MDS die Landschaftsbegriffe nach der Ähnlichkeit ihrer Beschreibungswörter angeordnet werden, ist zu erwarten, dass Objekte eines Clusters auch ähnliche Sentiment-Werte aufweisen. Es gilt zu überprüfen inwiefern dies zutrifft. Es soll auch untersucht werden, inwiefern sich die Gefühlspolaritäten unter den verschiedenen Clustern unterscheiden.

Um die Meinungsinformationen im Raum darzustellen, wird eine Interpolation der Sentiment-Werte ähnlich wie in *Nowak* (2013) durchgeführt. Dazu wird die MDS-Graphik exportiert und in die Software ArcMap eingelesen. Dort werden die Punkte, welche die einzelnen Landschaftsbegriffe repräsentieren, digitalisiert. Bei jedem Begriff wird in der Attributtabelle das Attribut «Sentiment» mit dem entsprechenden Durchschnittswert erstellt. Basierend auf diesen Werten wird dann eine Interpolation (IDW, d^2 Funktion) durchgeführt.

5.10.2 Vergleiche im echten geographischen Raum

Um auch Untersuchungen im echten geographischen Raum anzustellen, sollen Beschreibungen in verschiedenen Regionen miteinander verglichen werden. Es wird erwartet, dass sich bei morphologisch unterschiedlichen Landschaften auch die Beschreibungen voneinander unterscheiden und dass sich gewisse für die Regionen typische Eigenschaften herauskristallisieren lassen. Um dies zu überprüfen, werden die ermittelten Daten der Regionen Safiental, Valsertal, Thurgau und Schaffhausen miteinander verglichen, und zwar in Bezug auf die erfassten Adjektive und Landschaftsbegriffe. Das Safiental und das Valsertal sind zwei benachbarte Alpentäler im Kanton Graubünden. Bei den Regionen Schaffhausen und Thurgau handelt es sich um zwei benachbarte Kantone des Schweizer Mittellandes. Die unterschiedlichen Charakteristika zwischen den beiden alpinen Landschaften und den beiden Mittelland-Kantonen werden beim Betrachten der Abbildungen 28 und 29 deutlich, welche einige auf Hikr hochgeladene Bilder der entsprechenden Regionen zeigen.

Abb. 28: Bilder aus den Gebieten Safiental (oben) und Valsertal (unten) aus Hikr (*www.hikr.org*).

Abb. 29: Bilder aus den Gebieten Thurgau (oben) und Schaffhausen (unten) aus Hikr (*www.hikr.org*).

Die vier Regionen verfügen alle über eine ähnliche Anzahl Berichte (Stand 15.06.2015: Valsertal: 112, Safiental: 128, Thurgau: 104, Schaffhausen: 99) sowie mit Ausnahme des Kantons Thurgau auch über eine relativ ähnliche Fläche (*www.bfs.admin.ch*; *www.vals.ch*; *www.safiental.ch*). Wobei der Vergleich schwierig ist, da Hikr eine eigene Regioneneinteilung anwendet.
Die Erwartung ist, dass sich die beiden alpinen Regionen und die beiden Mittellandregionen in den Beschreibungen jeweils ähnlicher sind. Ausserdem sollten zum Beispiel in den alpinen Gebieten erwartungsgemäss häufiger alpine Erscheinungen wie Berge, Gipfel oder Grate beschrieben werden. Um die Beschreibungen in den vier Regionen miteinander zu vergleichen, werden die zehn wichtigsten Beschreibungswörter und Landschaftsbegriffe jeder Region nach deren Tf-idf-Mass absteigend sortiert und einander gegenübergestellt.

6. Resultate

Insgesamt hat das entwickelte Programm SentiTours 106‘998 Beschreibungen in 22‘570 verschiedenen Tourenberichten gefunden. Nachfolgend werden die Resultate präsentiert, welche sich durch die Analyse dieser Beschreibungen mit den oben beschriebenen Methoden ergeben haben.

6.1 Deskriptive Statistik der Landschaftsbeschreibungen

Um einen Überblick über die ermittelten Landschaftsbeschreibungen zu verschaffen, werden in diesem Kapitel die beschriebenen Landschaftsbegriffe sowie die dafür verwendeten Beschreibungswörter deskriptiv untersucht.

6.1.1 Beschriebene Landschaftsbegriffe

Von den 273 Landschaftsbegriffen, nach welchen SentiTours gesucht hat, werden 271 mindestens einmal beschrieben im Hikr-Korpus. Keine Beschreibungen wurden für die Elemente «Matten», «Mittelstation» und «Thal» gefunden.

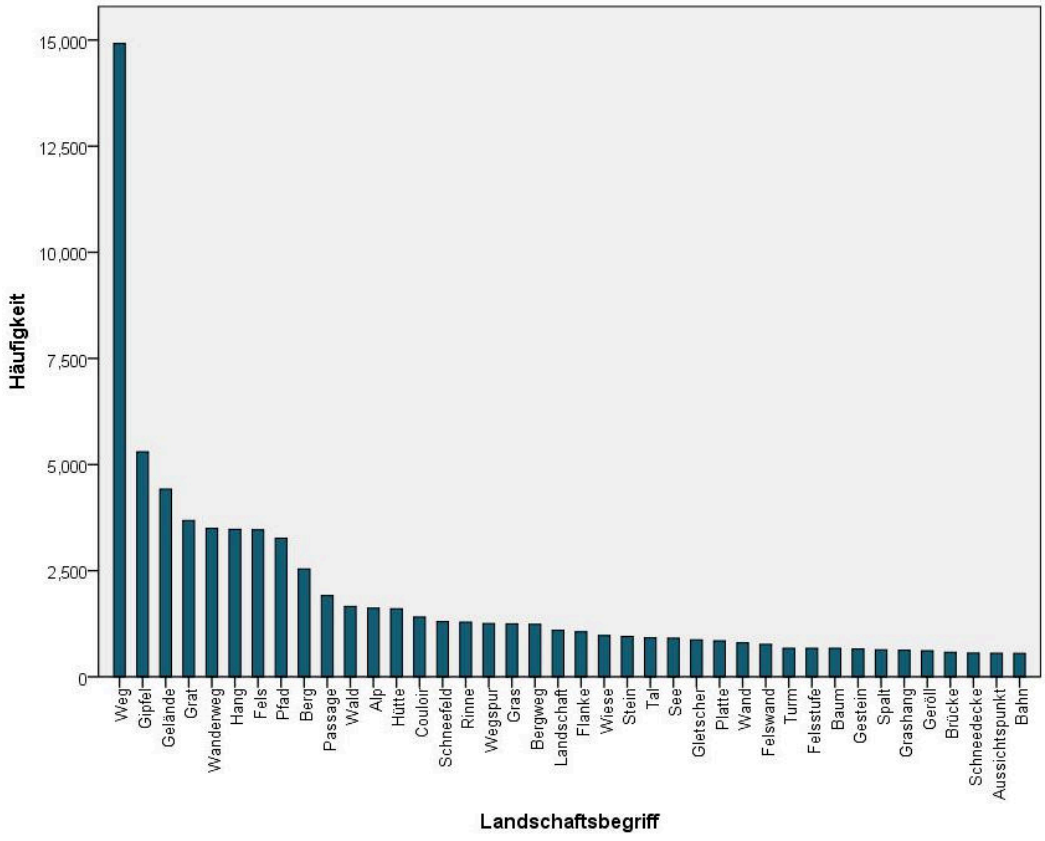


Abb. 30: Die 40 meistbeschriebenen Landschaftsbegriffe nach deren Häufigkeit sortiert.

In Abbildung 30 sind die 40 am häufigsten beschriebenen Landschaftsbegriffe absteigend nach der Anzahl Nennungen sortiert und in einem Säulendiagramm dargestellt. Die Anzahl Beschreibungen pro Landschaftsbegriff variieren beträchtlich. Das Landschaftselement «Weg» wird mit 14‘941 Mal am meisten beschrieben, während der Begriff «Küste» nur einmal in einer Beschreibung erwähnt wird. Das zweithäufigste Element «Gipfel» wird drei Mal weniger oft beschrieben wie der Begriff «Weg». Auch danach ist noch ein starker Abfall der Häufigkeiten erkennbar, bis ab dem zehnthäufigsten Begriff («Passage») die Abnahme relativ kontinuierlich geschieht.
In Abbildung 31 werden diese 40 Landschaftsbegriffe nun noch nach deren Sentiment-Werten sortiert und wiederum in einem Säulendiagramm dargestellt. Dazu wird der durchschnittliche Sentiment-Wert aller Beschreibungen des jeweiligen Begriffs berechnet.

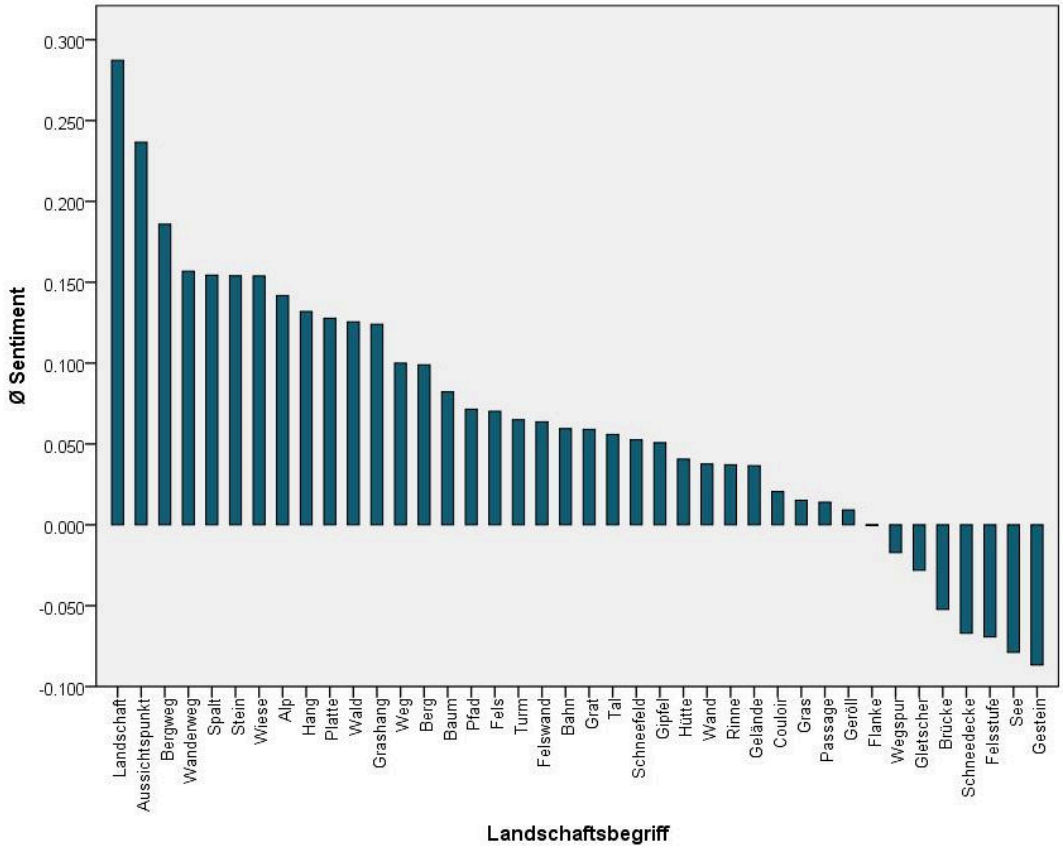


Abb. 31: Die 40 meistbeschriebenen Landschaftsbegriffe nach Sentiment-Wert sortiert.

33 der 40 am häufigsten beschriebenen Landschaftselemente werden im Durchschnitt positiv beschrieben. 7 haben einen negativen durchschnittlichen Sentiment-Wert. Die Sentiment-Werte reichen auf einer Skala von -1 bis 1 von -0.087 (Gestein) bis 0.287 (Landschaft).

6.1.2 Verwendete Beschreibungswörter (Adjektive)

Insgesamt werden 5‘772 verschiedene Adjektive verwendet, um die 271 Landschaftselemente zu beschreiben. Die höchste Frequenz weist das Wort «steil» mit 7‘396 Nennungen auf. Dabei werden 3‘099 Adjektive nur einmal verwendet, um einen Landschaftsbegriff zu beschreiben.

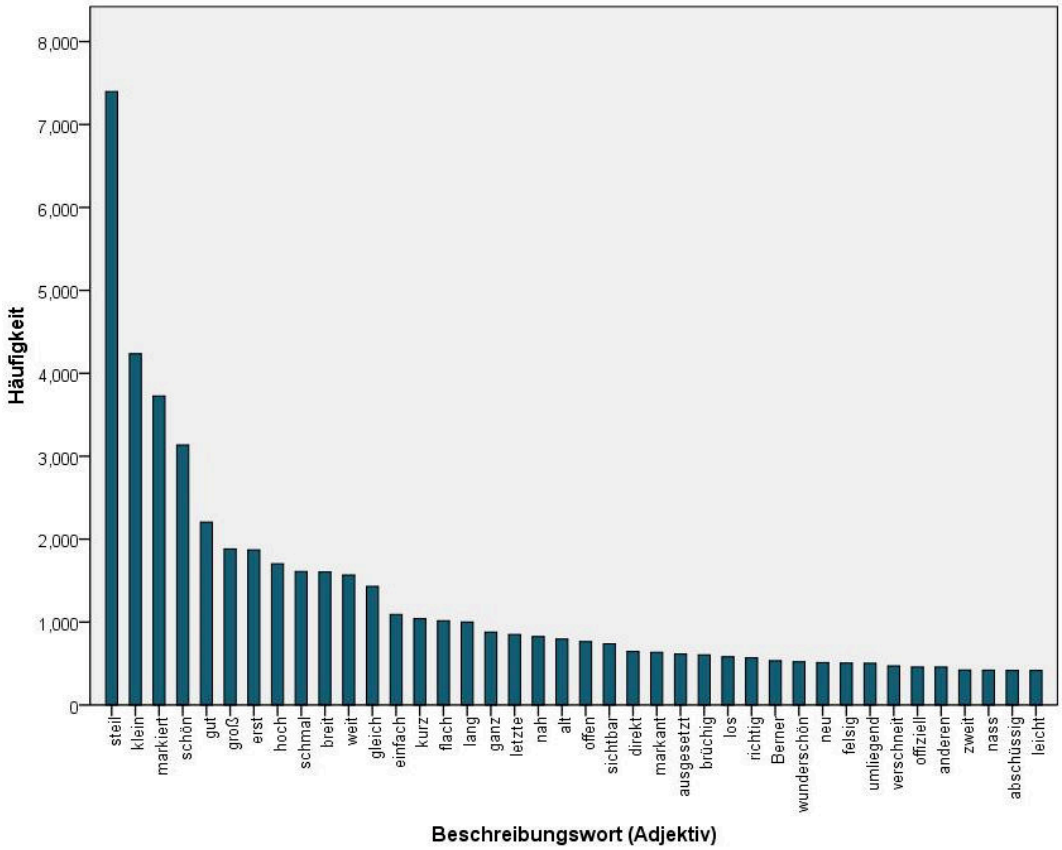


Abb. 32: Die 40 häufigsten Beschreibungswörter nach deren Häufigkeit sortiert.

In Abbildung 32 sind wiederum die 40 am häufigsten verwendeten Adjektive absteigend nach der Anzahl Nennungen sortiert und in einem Säulendiagramm dargestellt. Es zeigt sich, dass bei den Beschreibungswörtern sowie bei den Landschaftsbegriffen ebenfalls eine schnelle Abnahme der Häufigkeiten zu beobachten ist. In den Abbildungen 33 und 34 unten sind die 40 häufigsten positiven respektive negativen Adjektive jeweils absteigend nach der Häufigkeit ihres Vorkommens sortiert. Die Stärke der Polaritäten wird dabei nicht beachtet. Auch bei den positiven Beschreibungswörtern (Abbildung 33) dominieren einige wenige Ausdrücke. Insgesamt 389 Adjektiven wird durch die Sentiment-Bibliothek SentiWS eine positive Polarität zugewiesen. Bei den negativen Beschreibungswörtern (Abbildung 34) ist der Abfall der Häufigkeiten noch etwas stärker ausgeprägt. Das am meisten verwendete Adjektiv «klein» kommt mehr als viermal so häufig vor wie das zweitplatzierte Adjektiv «kurz». Die Häufigkeiten nehmen dann rasch ab und erreichen schnell tiefe Werte. Insgesamt 227 Adjektiven wird durch SentiWS eine negative Polarität zugewiesen. Es erhalten also insgesamt 616 von 5‘772 verschiedenen Adjektiven einen Gefühlswert.

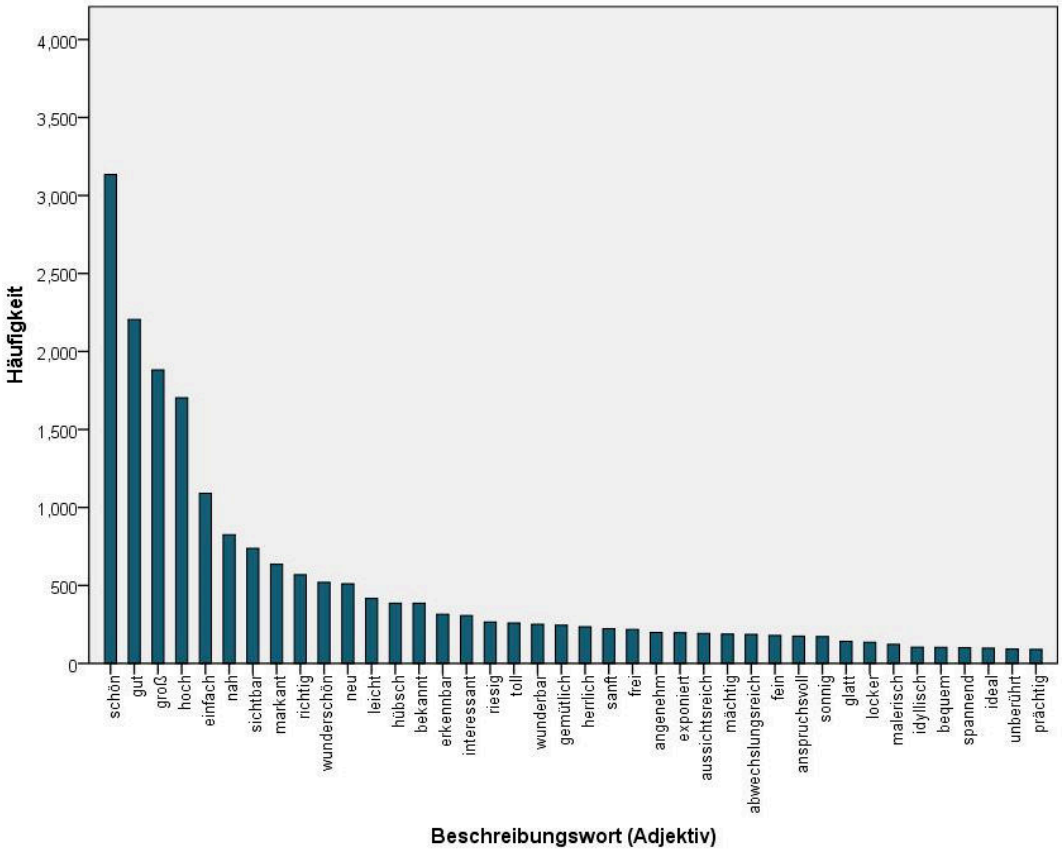


Abb. 33: Die 40 häufigsten positiven Beschreibungswörter nach deren Häufigkeit sortiert.

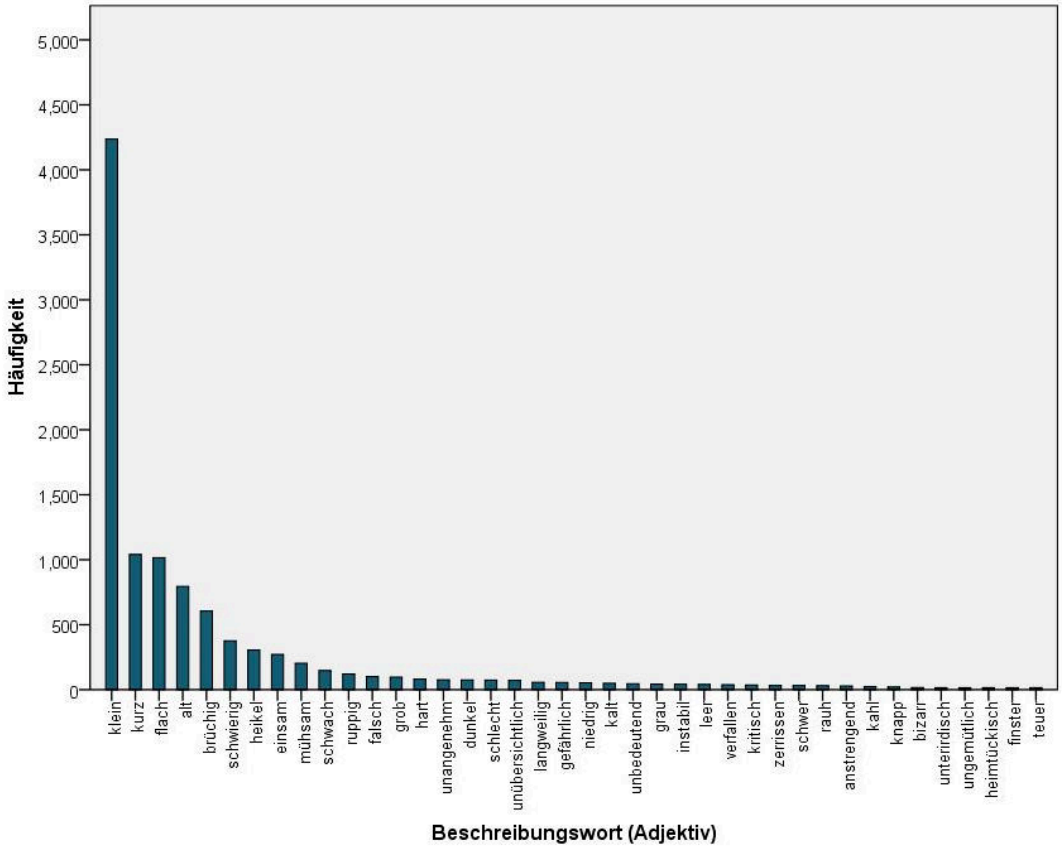


Abb. 34: Die 40 häufigsten negativen Beschreibungswörter nach deren Häufigkeit sortiert.

6.2 Vergleich der Landschaftselemente

In diesem Kapitel werden die verschiedenen Darstellungen aufgeführt, in welchen die Landschaftsbegriffe basierend auf ihren Beschreibungen miteinander verglichen werden.

6.2.1 Multidimensionale Skalierung

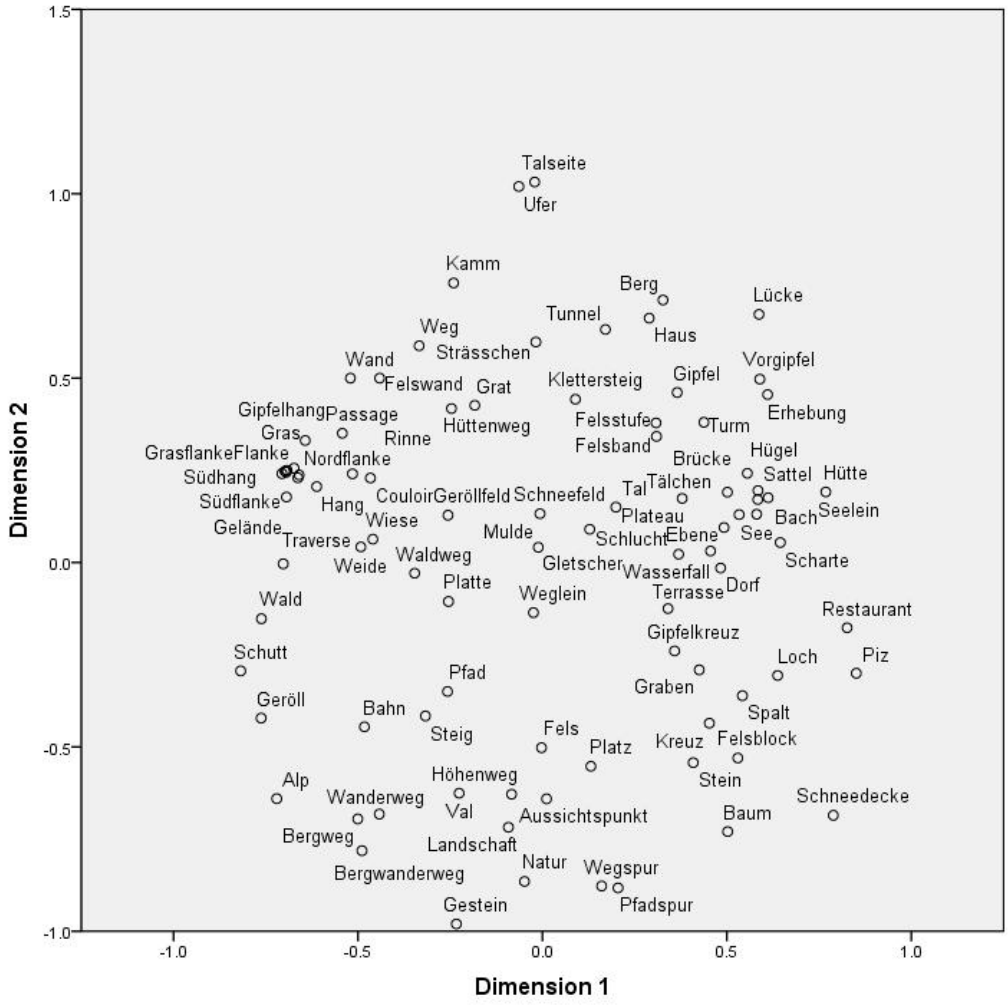


Abb. 35: MDS der Landschaftsbegriffe, basierend auf der Ähnlichkeit ihrer Beschreibungen.

In Abbildung 35 sind die 91meistbeschriebenen Landschaftsbegriffe (mind. 200 Beschreibungen) im zweidimensionalen Raum entsprechend der Ähnlichkeit ihrer Beschreibungen in einer MDS angeordnet. Die Distanzen zwischen den Begriffen beruhen auf den ermittelten Kosinus-Ähnlichkeiten. Die Stress-Masse nach Kruskal nehmen folgende Werte an: S-Stress I = 0.31, S-Stress II = 0.73, S-Stress = 0.24.

6.2.2 Clustering

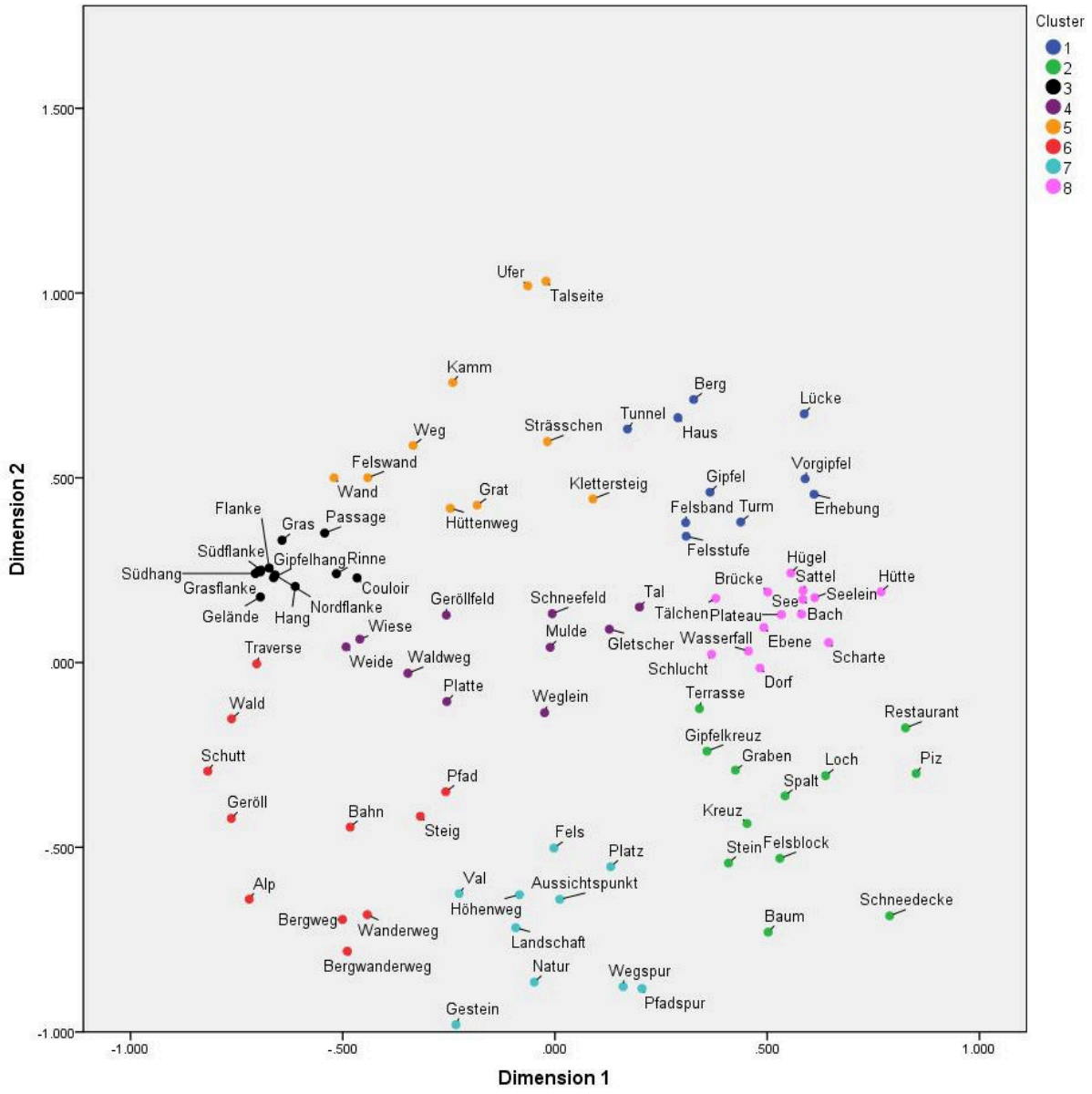


Abb. 36: Gruppierung der Landschaftsbegriffe in der MDS mittels Ward-Clustering.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**
hoch	gross	steil	steil	markiert	markiert	gut	klein
erst	umgestürzt	offen	klein	gleich	berner	schön	rosa
klein	klein	flach	gross	steil	steil	brüchig	schön
umliegend	geschlossen	abschüssig	schön	gut	gut	los	erst
schön	los	kurz	flach	breit	schmal	deutlich	weit
letzte	tief	hoch	weit	direkt	schön	fest	gross
nah	schön	felsig	erst	schmal	licht	vorhanden	nah
weit	riesig	heikel	aper	schön	offiziell	schwach	vierwaldstätter
gross	sichtbar	einfach	saftig	weit	los	wunderschön	tief
markant	umgefallt	weit	breit	einfach	dicht	griffig	rauschend

Tab. 5: Rangfolge der gemäss Tf-idf-Mass wichtigsten Begriffe pro Cluster in der MDS.

In Abbildung 36 wurden die in der MDS angeordneten Landschaftsbegriffe mithilfe eines Clustering-Verfahrens nach Ward verschiedenen Gruppen zugeteilt. Durch die Analyse der Heterogenitätsmasse wurde eine optimale Anzahl von acht Clustern ermittelt. Die verschiedenen Cluster sind farblich gekennzeichnet. In Tabelle 5 sind die wichtigsten 10 Begriffe pro Cluster gemäss Tf-idf-Mass absteigend aufgelistet. Die Adjektive in der ersten Spalte haben also jeweils die höchste, diejenigen in der letzten Spalte die tiefste Bedeutung für das entsprechende Cluster.

6.2.3 Interpolation der Sentiment-Werte

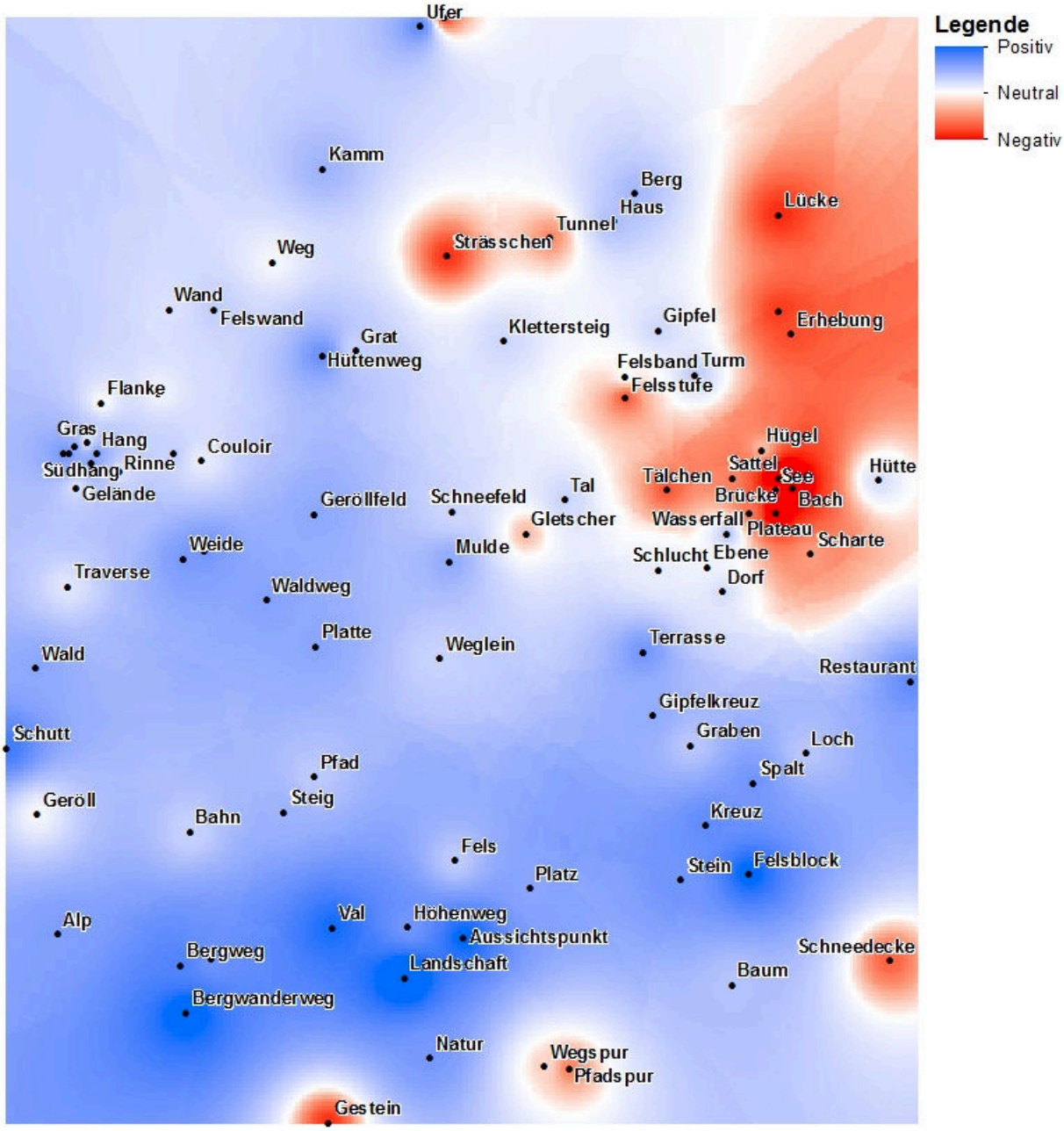


Abb. 37: Interpolation der Sentiment-Werte in Anlehnung an das Vorgehen in *Nowak* (2013).

Abbildung 37 stellt eine räumliche Interpolation (IDW, d^2 Funktion) der durchschnittlichen Sentiment-Werte der in der MDS angeordneten Landschaftsbegriffe dar. Positive Werte sind blau und negative Werte rot dargestellt. Neutrale Werte sind weiss. Der niedrigste Wert liegt bei -0.127 («Sattel») und der höchste Wert bei 0.287 («Landschaft»). Neutrale Werte sind solche nahe bei Null, da diese weder positiv noch negativ sind. Aus Darstellungsgründen sind nicht alle Punkte beschriftet.

6.3 Vergleiche im echten geographischen Raum

Vergleich der Adjektive

	Valsertal		Safiental		Thurgau		Schaffhausen
1	markiert	1	steil	1	klein	1	vorhanden
2	steil	2	lang	2	markiert	2	klein
3	klein	3	schön	3	gelb	3	gut
4	sichtbar	4	klein	4	rutschig	4	markiert
5	gross	5	breit	5	offiziell	5	hoch
6	hoch	6	windgeschützt	6	schön	6	gross
7	rotweissrot	7	begehbar	7	weit	7	schön
8	brüchig	8	schmal	8	nah	8	offen
9	gleich	9	flach	9	tannegger	9	unmarkiert
10	flach	10	gut	10	erkennbar	10	bemalt

Tab. 6: Die gemäss Tf-idf-Mass wichtigsten Adjektive pro Region.

Vergleich der Landschaftsbegriffe

	Valsertal		Safiental		Thurgau		Schaffhausen
1	Bergwanderweg	1	Hang	1	Weg	1	Weg
2	Weg	2	Weg	2	Wanderweg	2	Turm
3	Hang	3	Grat	3	Baum	3	Wanderweg
4	Gelände	4	Piz	4	Tobel	4	Fels
5	Gipfel	5	Couloir	5	Kirche	5	Hügel
6	Grat	6	Gelände	6	Hof	6	Wald
7	Kapelle	7	Mulde	7	Höhle	7	Gipfel
8	Schneefeld	8	Gipfel	8	Klettergarten	8	Pfad
9	Pfadspur	9	Passage	9	Wegspur	9	Loch
10	Gletscher	10	Tälchen	10	Gelände	10	Haus

Tab. 7: Die gemäss Tf-idf-Mass wichtigsten Landschaftsbegriffe pro Region.

In den Tabellen 6 und 7 sind jeweils pro Region die 10 Adjektive respektive Landschaftsbegriffe mit den höchsten Tf-idf-Werten absteigend sortiert. Die Anzahl gefundener Landschaftsbeschreibungen in den Regionen reicht von 129 (Schaffhausen) bis 354 (Thurgau). In den beiden anderen Regionen wurden 297 (Valsertal) und 242 (Safiental) Beschreibungen extrahiert. Dabei wurden 162 (Valsertal), 149 (Safiental), 168 (Thurgau) und 79 (Schaffhausen) verschiedene Adjektive

verwendet. Damit wurden 101 (Valsertal), 79 (Safiental), 89 (Thurgau) und 49 (Schaffhausen) verschiedene Landschaftselemente beschrieben.

7. Diskussion

Im Hinblick auf die Beantwortung der Forschungsfragen wird in diesem Kapitel zum einen die Leistung der entwickelten Anwendung SentiTours kritisch betrachtet. Zum anderen werden die Muster in den extrahierten Landschaftsbeschreibungen sowie auch die Ergebnisse der damit getätigten Auswertungen diskutiert. Die ausführliche Beantwortung der Forschungsfragen erfolgt schliesslich in der Schlussfolgerung in Kapitel 8.

7.1 Die Anwendung SentiTours

In diesem Unterkapitel wird die Anwendung SentiTours evaluiert. Um die beiden Forschungsfragen zu beantworten, stellt sich die Frage, ob die Anwendung in der Lage ist, Beschreibungen von Landschaften aus Texten rechentechnisch zu extrahieren und deren inhärente Meinungen wiederzugeben. Es soll nun beurteilt werden, inwiefern die umgesetzte Methodik diese Aufgabe erfüllt. Im Zuge dessen werden Schwierigkeiten und Limitierungen diskutiert, wodurch später Verbesserungsvorschläge abgeleitet werden können.

7.1.1 Plausibilität der extrahierten Beschreibungen

Gemäss der Analyse, welche in Kapitel 5.5.2.2 zur Evaluierung der Beziehungstypen durchgeführt wurde, ist die Plausibilität der meisten extrahierten Inhalte hoch. Werden die Plausibilitätswerte der drei Beziehungstypen, welche schlussendlich verwendet wurden (Attributbeziehung, Prädikatbeziehung, Prädikatbeziehung Perfekt) unter Berücksichtigung der Häufigkeiten der einzelnen Beziehungstypen miteinander verrechnet, beträgt die durchschnittliche Plausibilität ca. 81%. Bei vier von fünf Treffern handelt es sich bei den entdeckten Beschreibungswörtern also um diejenigen, welche den entsprechenden Landschaftsbegriff auch wirklich beschreiben. Auch Negationen werden bis auf wenige Ausnahmen korrekt erfasst. Dies spricht für die Anwendung der entwickelten Methodik zur Extraktion von Landschaftsbeschreibungen. Es muss dabei jedoch beachtet werden, dass nur diejenigen Resultate evaluiert wurden, welche zu einem Treffer geführt haben («precision»). Es wurde nicht ermittelt, wie viele Beziehungen fälschlicherweise gar nicht erfasst wurden («recall»). Dies hätte man jedoch mithilfe einer Studie schätzen können, in welcher Probanden und Probandinnen Landschaftsbeschreibungen in einer Auswahl an Texten markieren. Die Resultate der Studie hätten dann mit den Resultaten der Anwendung verglichen werden können (*Cardie*, 1997). Im Bezug zur ersten Forschungsfrage lässt sich dennoch sagen, dass sich Beschreibungen von Landschaften aus Texten extrahieren lassen. Auf die Muster und Eigenheiten,

welche man in diesen Beschreibungen erkennen kann, wird in den Kapiteln 7.2, 7.3 und 7.4 eingegangen.

7.1.2 Probleme und Limitierungen

In diesem Unterkapitel werden einige Probleme diskutiert, welche während der Entwicklung von SentiTours aufgetaucht sind. Ausserdem wird auf die Limitierungen der Anwendung eingegangen, welche sich darauf auswirken, wie akkurat die Landschaftsbeschreibungen aus den Texten erfasst werden.

7.1.2.1 Beziehungstypen

Bei der qualitativen Untersuchung der Hikr-Texte hat sich gezeigt, dass Landschaftsbegriffe in wenigen Fällen auch mit Nomen oder Verben beschrieben werden können. In einer ersten Phase wurden deshalb auch Beziehungstypen zur Erfassung solcher Beschreibungen implementiert. Wie die Analyse in Kapitel 5.5.2.2 jedoch ergeben hat, sind dadurch vor allem Begriffe erfasst worden, die sich zwar auf einen Landschaftsbegriff beziehen, aber keine eigentlichen Beschreibungswörter sind. Aus diesem Grund musste das Kriterium eingeführt werden, dass es sich bei den beschreibenden Wörtern um Adjektive handeln muss. Obwohl kein Wert für den recall ermittelt wurde, wird basierend auf der qualitativen Sichtung der Hikr-Texte die Behauptung gemacht, dass der Grossteil der Landschaftselemente mit Adjektiven charakterisiert wird. So gibt es auch in der Literatur Arbeiten aus dem Gebiet der Sentiment Analysis, welche sich nur auf Adjektive konzentrieren (*Hu & Liu*, 2004). Dennoch würde die Erfassung zusätzlicher Wortarten als Beschreibungswörter bestimmt einen Mehrwert generieren.

7.1.2.3 Kontext

Ausser der Berücksichtigung von Negationen wird mit der entwickelten Anwendung kein zusätzlicher Kontext erfasst. Daraus folgt, dass implizite Beschreibungen von Landschaftselementen nicht erkannt werden, was häufig eine Schwierigkeit von Anwendungen im Gebiet der Sentiment Analysis ist (*Cambria et al.*, 2013). Dies kann zum Beispiel dann der Fall sein, wenn sich eine Beschreibung in einem Satz auf einen Landschaftsbegriff im vorherigen Satz bezieht. Oder wenn Eigenschaften der Umgebung beschrieben werden, ohne dass dafür explizit ein Landschaftselement erwähnt wird: «Trotz Lärmschutzmauer ist der Verkehr deutlich zu hören.» Es gibt auch Fälle, wo der eigentliche Sinn einer Beschreibung verborgen bleibt oder nicht vollständig abgebildet wird. Aus folgendem Satz beispielsweise: «Mehr gibt es eigentlich nicht zu sagen, denn der Weg ist gut zu finden und er ist auf der Wanderkarte eingezeichnet.» wurde die Beschreibung «gut Weg» extrahiert. Dies bildet den Charakter der gemachten Beschreibung aber nicht vollständig ab.

In Kapitel 2.5.5 wurde erwähnt, dass Adverbien den Sentiment-Wert einer Beschreibung modifizieren können. Wie in diversen Arbeiten gezeigt, wird die Erfassung von Adverbien auch praktiziert (*Hauthal & Burghardt*, 2014). Wie demonstriert wurde, wäre die Erkennung von modifizierenden Adverbien mit der vorliegenden Methodik auch ohne Weiteres möglich gewesen. Für deutsche Texte wurde jedoch kein Sentiment-Lexikon gefunden, in welchem die Gefühlspolaritäten von Adverbien angegeben sind, wie dies z.B. in SentiWordNet der Fall ist. Eine solche oder ähnliche Auflistung der gängigsten Adverbien mit einem Wert für ihre «Modifikationsstärke» wäre jedoch notwendig. So könnte zum Beispiel unterschieden werden, dass der Ausdruck «sehr hässlich» negativer ist als der Ausdruck «etwas hässlich».

7.1.2.4 Sprache

Wie zu erwarten war, werden in den Hikr-Berichten auf Deutsch auch einige schweizerdeutsche Ausdrücke verwendet oder, wie in folgendem Beispiel ersichtlich, auch ganze Satzteile in Dialekt formuliert: «Schneefelder fast bis zur Passstrasse. Achtung: Tückische Löcher Hugi: Merci vielmol für‘s vorus go.»

Der verwendete Dependency Parser ParZu liefert ausserdem in seltenen Fällen sprachlicher Besonderheiten falsche Resultate. So wird beim Ausdruck «durch’*s Dorf*» der Buchstabe «s» für das Beschreibungswort des Begriffs «Dorf» gehalten. Ein anderer Fehler ist die Erfassung eines Anführungszeichens «„» als Beschreibungswort. Dies geschah beispielsweise bei der Analyse des folgenden Satzes: «Von meinem Ziel *„Bahn* kurz nach sieben“ habe ich mich bereits verabschiedet, kurz nach acht heißt die neue Parole stattdessen.» Diese beiden Fehler traten 204 respektive 268 Mal in den insgesamt 106‘998 Beschreibungen auf. Aus diesem Grund wurden sie aus der späteren Analyse ausgeschlossen. Es wurden auch noch weitere ähnliche Fälle entdeckt, doch handelt es sich dabei jeweils nur um eine sehr kleine Anzahl betroffener Beschreibungen (jeweils weniger als 22), so dass diese keinen Einfluss auf die Resultate haben.

7.2 Deskriptive Statistik der Landschaftsbeschreibungen

In den folgenden Unterkapiteln werden die Wortstatistiken in Form der Säulendiagramme in Kapitel 6.1 diskutiert. Damit sollen gewisse sprachliche Muster aufgedeckt werden. Der Fokus liegt dabei auf der Verteilung der jeweils 40 häufigsten und somit gewichtigsten Ausdrücke.

Die Verteilungen zeigen alle Züge einer Zipf-Verteilung. Die Häufigkeit nimmt also umgekehrt proportional mit der Rangierung der Begriffe ab. Dieser Effekt zeigt sich häufig bei Untersuchungen von Worthäufigkeiten (*Zipf*, 1935). Am stärksten ausgeprägt ist dieses Muster bei den negativen Beschreibungswörtern.

7.2.1 Beschriebene Landschaftsbegriffe

7.2.1.1 Ohne Sentiment

Es fällt auf, dass einige wenige Landschaftselemente die Beschreibungen in Hikr stark dominieren. Der Begriff «Weg» ist dabei sehr dominant. Die Tourengänger und Tourengängerinnen beschreiben somit sehr oft Wege. Dies liegt wohl daran, dass sie sich sehr oft auf solchen bewegen und die Eigenschaften dieser Wege für die jeweiligen Tätigkeiten von Bedeutung sind. Hinzu kommt, dass verwandte Begriffe wie «Wanderweg», «Pfad» oder «Bergweg» ebenfalls stark vertreten sind. Dies ist ein weiterer Hinweis darauf, dass den Eigenschaften des begangenen Untergrundes viel Beachtung geschenkt wird. Zählt man das Vorkommen der Elemente «Weg», «Wanderweg», «Pfad» und «Bergweg» zusammen, macht diese Gruppe ca. 30 Prozent der häufigsten 40 Landschaftsbegriffe aus. Diese häufigen Beschreibungen von Wegen lassen sich womöglich auch mit der von *Gibson* (1979) entwickelten Theorie der sogenannten «Affordanz» erklären. Diese besagt, dass Dinge über einen Aufforderungscharakter verfügen, mit welchem sie uns auferlegen, was wir mit ihnen tun sollen oder können. Die offensichtliche Gebrauchseigenschaft eines Weges liegt darin, dass man in begeht. Da bei denselben Touren von den Leuten wohl auch mehrheitlich dieselben Wege beschritten werden (müssen), ist es von Interesse, dass diese in einem Bericht erwähnt und beschrieben werden.

Lässt man die Wege ausser Acht, so lässt sich sagen, dass natürliche Objekte unter den 40 am häufigsten beschriebenen Begriffen stärker vertreten sind als anthropogene Elemente. Dies lässt sich wohl darauf zurückführen, dass die Leute sich vorwiegend in Gebieten aufhalten, wo sie Letzterem weniger häufig begegnen. Es ist dabei auch zu beachten, dass sich in der in Kapitel 5.4 erstellten Liste der zu suchenden Landschaftsbegriffe mehrheitlich natürliche Objekte befinden. Zudem ist zu erwähnen, dass sich einige Begriffe nicht klar einer Kategorie zuordnen lassen. Ein asphaltierter Weg lässt sich eindeutiger als anthropogenes Element bezeichnen als ein Trampelpfad.

7.2.1.2 Mit Sentiment

Das Diagramm in Abbildung 31 deutet darauf hin, dass Landschaften respektive Landschaftselemente vorwiegend positiv beschrieben werden. Dennoch lässt sich die entstandene Rangierung inhaltlich schwer nachvollziehen. Es scheint fraglich, inwiefern die Darstellung die tatsächliche Wahrnehmung der Leute wiedergibt. So befindet sich zum Beispiel das Landschaftselement «See» im negativen Bereich. Dies widerspricht Erkenntnissen in der Literatur, welche besagen, dass Gewässer einen speziell positiven Effekt auf die Wahrnehmung der Landschaft haben (*Yamashita*, 2002). Andererseits wird das Element «Spalt» im Vergleich sehr positiv bewertet. Es wäre zu erwarten gewesen, dass mit diesem Begriff eher negative Eindrücke assoziiert werden, wie im Falle von Gletscherspalten. Die Analyse der Adjektive kann auf diese Fragen womöglich Antworten liefern.

7.2.2 Verwendete Beschreibungswörter (Adjektive)

7.2.2.1 Ohne Sentiment

Bei der Häufigkeitsverteilung der verwendeten Beschreibungswörter (Abb. 32) lässt sich ein ähnliches Muster beobachten wie bei den Landschaftsbegriffen. Die Dominanz des Wortes «steil» deutet wiederum darauf hin, wie wichtig die Eigenschaften des Untergrundes für die Tourengänger und Tourengängerinnen sind. Im Gegensatz dazu gibt es eine enorme Menge an Adjektiven, welche nur wenige Male genannt werden. Darunter sind auch einige Wortkreationen, welche im normalen Sprachgebrauch nicht üblich sind. Beispiele dafür sind «trockensteingesäumt» oder «gebüschbewachst».

Es gibt auch einige Adjektive, welche nicht das äussere Erscheinungsbild oder den Zustand eines Elements beschreiben und somit im Kontext der Landschaftsbewertung weniger aussagekräftige Hinweise liefern. So wird zum Beispiel mit dem Wort «umliegend» die Position eines Landschaftselements relativ im Raum angegeben. Mit den Ausdrücken «erst», «letzte» und «zweit» wird eine zeitliche Abfolge der Elemente auf der begangenen Tour beschrieben. Auch die Beschreibung «direkt» («auf direktem Wege») liefert zumindest im Kontext dieser Arbeit keine nützlichen Informationen. Einen besonderen Fall stellt das Beschreibungswort «Berner» dar. Dieses Beschreibungswort wird oft in Verbindung mit dem Landschaftselement «Alp» verwendet. Da der Ausdruck «Berner Alp» im Gazetteer (Toponymliste) von Hikr gesucht und nicht gefunden wurde, handelt es sich hierbei anscheinend nicht um ein offizielles Toponym.

7.2.2.2 Mit Sentiment (Positiv)

Bei der angewandten Methode wird die Einteilung der Adjektive in die Polaritäten positiv und negativ durch die verwendete Sentiment-Bibliothek SentiWS bestimmt. Hierbei kann wie in *Wilson et al.* (2005) erwähnt, das Problem der Domänenabhängigkeit auftreten. Dies bedeutet, dass im vorliegenden Fall möglicherweise Adjektive in SentiWS als positiv klassifiziert wurden, welche in der Domäne der Tourenberichte normalerweise nicht als positiv angesehen werden. Die Klassifizierung der positiven Meinungswörter (Abb. 33) erscheint in den meisten Fällen durchaus sinnvoll. Bei einigen wenigen Wörtern besteht dennoch die Vermutung, dass diese im Text auch oder sogar überwiegend im negativen Kontext gebraucht werden. Dazu gehören die Adjektive «exponiert», «glatt» und «locker». Dazu wurden unter anderem folgende Beispielsätze gefunden, welche diese Vermutung bekräftigen: «Ziemlich steil in sehr *lockerem*, unstabilem Gestein.»; «Diese könnten insbesondere bei Nässe auf den *glatten* Felsen sehr nützlich sein.»

7.2.2.3 Mit Sentiment (Negativ)

Stärker ausgeprägt scheint der Effekt der Domänenabhängigkeit bei den negativen Adjektiven zu sein (Abb. 34). Die drei häufigsten Wörter «klein», «kurz» und «flach» müssen für einen Tourengänger oder eine Tourengängerin keineswegs eine negative Bedeutung haben. So ist ein kurzer und flacher Weg wohl oft etwas Erfreuliches. Diese Beschreibungswörter sind ausserdem gute Beispiele für Adjektive, welche kontextabhängig sowohl positiv als auch negativ klassifiziert werden können. Sie widerspiegeln die von *Wilson et al.* (2005) erwähnte Problematik, welche durch die Unterschiede in der vorgängigen und kontextuellen Polarität entstehen kann. Lässt man in der Auflistung die häufigsten vier Begriffe ausser Acht, erscheinen unter den restlichen als negativ klassifizierten Ausdrücken dennoch einige plausibel und unabhängig vom Kontext brauchbar. Dazu zählen Adjektive wie «brüchig», «schwierig», «heikel», «mühsam», «unübersichtlich», «langweilig» etc.

7.3 Vergleich der Landschaftselemente

In diesem Kapitel werden die Vergleiche der Landschaftsbegriffe basierend auf ihren Beschreibungen im abstrakten Raum diskutiert. Dabei wird beurteilt, ob die entstandenen räumlichen Muster nachvollziehbar sind und welche Aussagen sich daraus ableiten lassen.

7.3.1 Multidimensionale Skalierung

Die räumliche Zuordnung in der MDS in Abbildung 35 lässt sich weitestgehend nachvollziehen. Es lassen sich einige grössere und kleinere Gruppierungen erahnen, welche aufzeigen, dass Landschaftselemente mit ähnlichen Eigenschaften auch ähnlich beschrieben werden. In Abbildung 38 wurden einige Regionen markiert, in welchen Gruppierungen von Landschaftselementen erwartet werden. Im grün eingezeichneten Bereich befinden sich augenscheinlich Ausdrücke, welche häufig eine geneigte Fläche repräsentieren wie «Südhang», «Grasflanke», «Gelände» etc. Im gelb markierten Bereich sammeln sich viele Begriffe, welche eine Erhebung repräsentieren wie «Berg», «Gipfel», «Hügel» etc. Landschaftselemente, welche im Gegensatz dazu eine Vertiefung beschreiben, siedeln sich weiter unten an (rot markiert). Dazu gehören die Begriffe «Graben», «Spalt», und «Loch. Auch die Gewässer in Form der Elemente «Bach», «See», «Seelein», und «Wasserfall» gruppieren sich in der rechten Bildmitte zusammen (blau markiert).

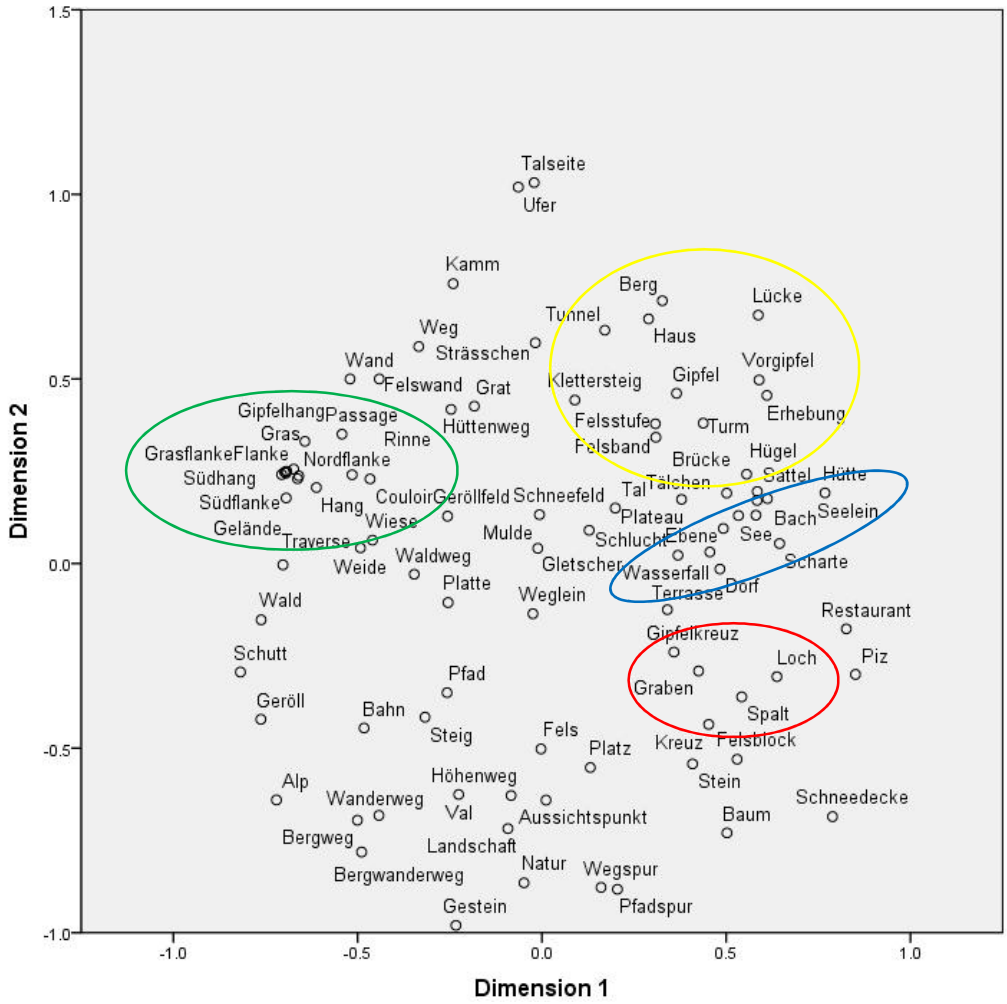


Abb. 38: Hervorhebung einiger räumlicher Muster in der erstellten MDS.

Auffällig ist zudem, dass über den gesamten Raum Landschaftselemente verteilt sind, welche verschiedene Arten von Wegen beschreiben. Im oberen Bereich der Darstellung befinden sich die Begriffe «Weg» und «Hüttenweg», in der Mitte die Begriffe «Waldweg» und «Weglein» und im unteren Bereich die Begriffe «Bergweg», «Wanderweg», «Bergwanderweg» und «Höhenweg». Um noch etwas fundierter auf die Eigenschaften eingehen zu können, welche verantwortlich sind für die räumliche Verteilung der Landschaftselemente, werden im nächsten Kapitel die durch das Ward-Clustering gebildeten Gruppen mit ihren gewichtigsten Adjektiven diskutiert.

7.3.2 Clustering

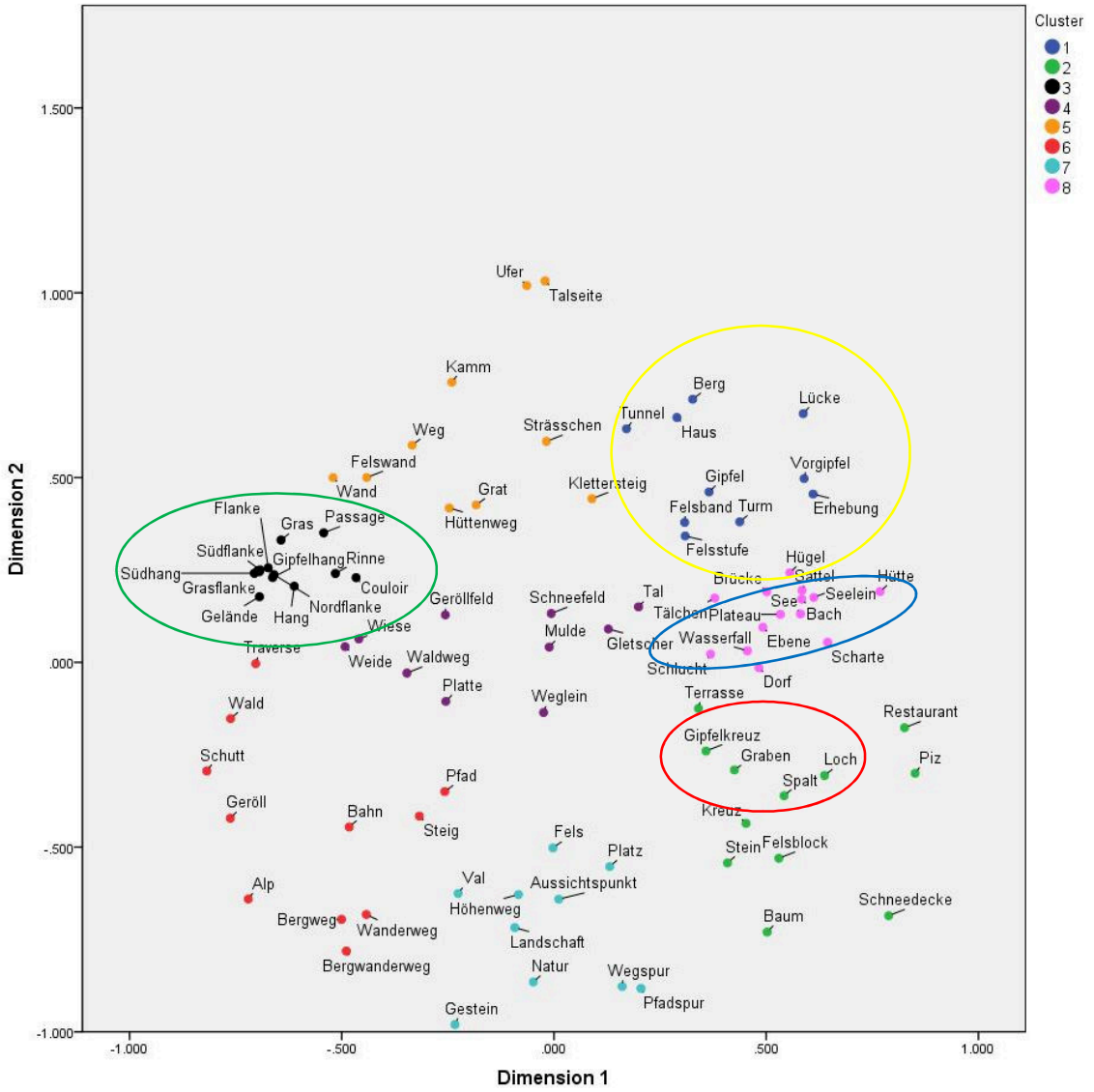


Abb. 39: Übertrag der erwarteten räumlichen Muster in die Clusterdarstellung.

Wie in Abbildung 39 verdeutlicht wird, treffen die erwarteten Gruppierungen grösstenteils zu. Im Folgenden soll mithilfe der zusätzlich ermittelten clustertypischen Adjektive diskutiert werden, welche charakteristischen Eigenschaften sich aus den Begriffsgruppierungen herauslesen lassen. Die erwähnten Beschreibungswörter werden in Tabelle 8 markiert.

Die clustertypischen Adjektive bestätigen die Vermutung, welche in Kapitel 7.3.1 gemacht wurde, dass bei der Beschreibung der Objekte in Cluster 3 vor allem die Neigung dieser Elemente von Interesse ist. Darauf verweisen Ausdrücke wie «steil», «flach» und «abschüssig» (in Tabelle 8 grau markiert). Die Cluster 5 und 6 beinhalten viele verschiedene Arten von Wegen oder wegähnlichen Elementen wie «Grat», «Pfad», «Bergwanderweg» etc. Gemeinsame Wörter wie «markiert», «schmal» oder «steil» (gelb markiert) sind wohl diesen Landschaftsbegriffen zuzuschreiben. Einige Beschreibungswörter in den jeweiligen Clustern sind ziemlich eindeutig einigen wenigen

Landschaftselementen innerhalb der Gruppe zuzuordnen. So bezieht sich das Adjektiv «rauschend» (blau markiert) in Cluster 8 mit grosser Wahrscheinlichkeit auf die Elemente «Bach» und «Wasserfall». Im Cluster 4 hingegen wird das Wort «saftig» (grün markiert) wohl ausschliesslich zur Beschreibung einer «Wiese» oder «Weide» verwendet. Dies könnte ein Hinweis darauf sein, dass eine weitere Unterteilung durch die Bildung von mehr Clustern eventuell sinnvoll wäre. Spezielle Fälle stellen die Beschreibungswörter «berner», «rosa» und «vierwaldstätter» dar (rot markiert). Wie bereits erwähnt, ist «Berner Alp» eine häufige Wortassoziation. Das Beschreibungswort «vierwaldstätter» bezieht sich auf das Toponym Vierwaldstättersee. Wie eine kurze Recherche ergeben hat, bezieht sich das Adjektiv «rosa» oft auf die sogenannte «Monte Rosa Hütte», eine Berghütte im Monte-Rosa-Massiv der Walliser Alpen. Es handelt sich bei diesen Ausdrücken also nicht um echte Beschreibungswörter. Somit stellt sich im Rahmen dieser Arbeit eine andere Eigenschaft von Ortsnamen als Schwierigkeit heraus als ursprünglich angenommen. Toponyme bestehen oft aus zwei Teilen. Beim ersten Teil handelt es sich um einen Ausdruck, welcher den eigentlichen Namen beinhaltet oder auf den Standort schliessen lässt. Der zweite Teil beinhaltet hingegen häufig den generischen Begriff respektive das Landschaftselement. Diese Wortkombinationen werden von SentiTours fälschlicherweise als Landschaftsbeschreibungen erfasst.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**	**Cluster 8**
hoch	gross	steil	steil	markiert	markiert	gut	klein
erst	umgestürzt	offen	klein	gleich	berner	schön	rosa
klein	klein	flach	gross	steil	steil	brüchig	schön
umliegend	geschlossen	abschüssig	schön	gut	gut	los	erst
schön	los	kurz	flach	breit	schmal	deutlich	weit
letzte	tief	hoch	weit	direkt	schön	fest	gross
nah	schön	felsig	erst	schmal	licht	vorhanden	nah
weit	riesig	heikel	aper	schön	offiziell	schwach	vierwaldstätter
gross	sichtbar	einfach	saftig	weit	los	wunderschön	tief
markant	umgefallt	weit	breit	einfach	dicht	griffig	rauschend

Tab. 8: Markierung einiger clustertypischer Adjektive.

Im Cluster 8 wird das Adjektiv «klein» als wichtigstes Beschreibungswort aufgelistet. In Kapitel 7.2.1.2 wurde Skepsis darüber geäussert, dass der Begriff «See» so negativ bewertet wird. Basierend auf dieser Beobachtung lässt sich die negative Beurteilung der Gewässer jedoch vermutlich besser nachvollziehen. Diese Vermutung wird sogleich überprüft, indem die durchschnittlichen Sentiment-Werte der Landschaftsbegriffe erneut berechnet werden (siehe Abbildung 40), jedoch diesmal unter Ausschluss des Adjektivs «klein».

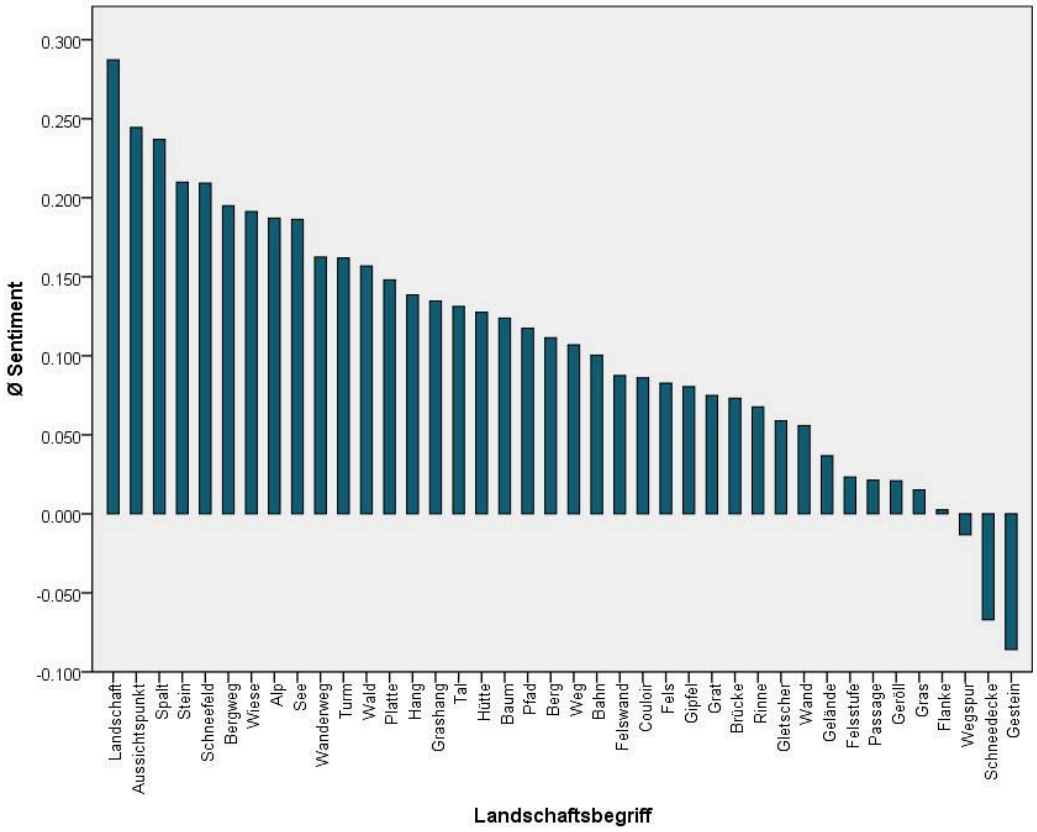


Abb. 40: Sentiment-Werte der 40 häufigsten Landschaftsbegriffe ohne das Adjektiv «klein».

Der Begriff «See» wird bei 245 von insgesamt 911 Beschreibungen mit dem Adjektiv «klein» beschrieben. Durch die Entfernung aller Beschreibungen mit dem Wort «klein» hat sich der Sentiment-Wert markant ins Positive verändert. Dies macht deutlich, wie wichtig ein «sinnvoller» Polaritätswert besonders bei hochfrequenten Beschreibungswörtern ist. Bezüglich der zweiten Forschungsfrage ist dies ein Hinweis darauf, dass die Arbeit mit einer nicht an die Domäne angepassten Sentiment-Bibliothek kaum aussagekräftige Resultate liefert, was Aussagen in der Literatur stützt (*Blitzer et al.*, 2007). Das Ergebnis zeigt aber, dass sich dennoch die qualitativen Unterschiede zwischen den Beschreibungen von Landschaftsbegriffen herauskristallisieren lassen, trotz der Dominanz einiger allgemeiner Wörter.

Um die Aussagekraft der Sentiment-Werte noch etwas genauer zu untersuchen, wird im nächsten Kapitel die Interpolationsdarstellung basierend auf den durchschnittlichen Gefühlspolaritäten der Landschaftsbegriffe diskutiert.

7.3.3 Interpolation der Sentiment-Werte

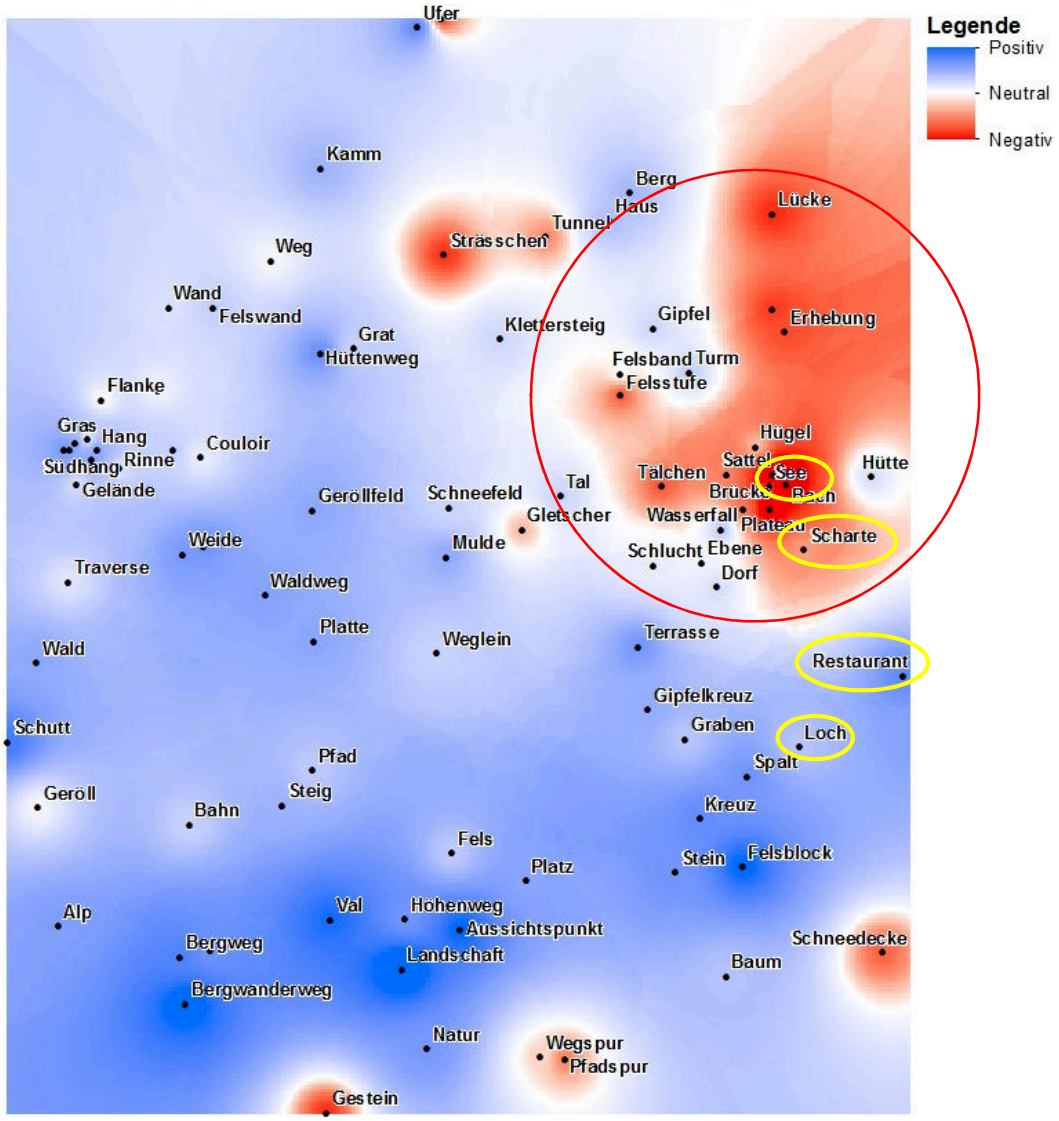


Abb. 41: Hervorhebung von Auffälligkeiten in der erzeugten Interpolationsdarstellung.

In der Interpolationsdarstellung ist ein grobes räumliches Muster zu erkennen. Wie in Abbildung 41 verdeutlicht (roter Kreis), werden die Landschaftsbegriffe im rechten oberen Bereich der Darstellung vorwiegend mit negativen Adjektiven beschrieben, währenddem die restlichen Begriffe überwiegend positiv bewertet werden. Stützt man sich auf die Clustereinteilung in Abbildung 36, entsprechen die vorwiegend negativen Gebiete den Clustern 1 und 8 und die überwiegend positiven Gebiete den restlichen Clustern. Landschaftsbegriffe desselben Clusters haben also ähnliche Sentiment-Werte, was nicht weiter verwunderlich ist, da die Begriffe ja basierend auf den verwendeten Adjektiven im Raum angeordnet wurden.

In den Kapiteln 7.2.1.2 und 7.3.2 wurde bereits die Aussage gemacht, dass die errechneten Sentiment-Werte für die einzelnen Landschaftsbegriffe keine plausiblen Resultate liefern. Die Interpolationsdarstellung bestätigt diese Meinung. So wird noch einmal deutlich, dass die Ähnlichkeiten

nur wenig auf Adjektiven beruhen, welche die Meinung gegenüber den Landschaftselementen beschreiben. Andernfalls wären Begriffe wie «Scharte» und «See» oder «Restaurant» und «Loch» (in Abbildung 41 gelb markiert) kaum in unmittelbarer Nähe angeordnet.
Zur Evaluation der ermittelten Sentiment-Werte wurde folglich auch keine tiefer greifende Auswertung zum Beispiel anhand manuell annotierter Datensätze durchgeführt. Dies weil, wie demonstriert wurde, die Verwendung einer nicht an die Domäne von Tourenberichten angepassten Sentiment-Bibliothek wenig aussagekräftige Resultate liefert.

7.4 Vergleiche im echten geographischen Raum

In diesem Kapitel wird nun die Aussagekraft der extrahierten Landschaftsbeschreibungen bei Vergleichen zwischen verschiedenen Regionen diskutiert. Dabei ist von Interesse, inwiefern sich allfällige Muster im echten geographischen Raum manifestieren. Da die jeweils 10 wichtigsten Beschreibungswörter respektive Landschaftsbegriffe nicht allzu viele Vergleichsmöglichkeiten bieten, wurden die Tabellen 6 und 7 in den Tabellen 9 und 10 um 20 Begriffe erweitert.

7.4.1 Vergleich der Adjektive zwischen den Regionen

Vergleicht man die Rangierung der Beschreibungswörter, welche in den verschiedenen Regionen verwendet werden, fällt vor allem die unterschiedliche Gewichtung des im gesamten Korpus häufigsten Adjektivs auf. Das Wort «steil» (in Tabelle 9 gelb markiert) wird im Safiental als charakteristischstes Wort erachtet. Bei den Berichten aus der Region Valsertal steht dasselbe Adjektiv an zweiter Stelle. Bei den beiden Mittellandregionen hingegen ist es unter den ersten 10 Begriffen nicht vertreten. Es scheint ausserdem, dass sich in den Berichten der alpinen Regionen mehr Beschreibungen befinden, welche auf Unternehmungen in Winterlandschaften schliessen lassen. In den beiden Alpenregionen sind nämlich einige Adjektive wie «unverspurt», «jungfräulich» oder «verschneit» (in Tabelle 9 blau markiert) zu finden, welche typischerweise in diesem Zusammenhang benutzt werden. Jedoch lässt sich nicht immer nur anhand des Beschreibungswortes erkennen, ob beispielsweise ein schneebedecktes Element beschrieben wird. So kann der Ausdruck «tief» sich genauso auf eine Schneedecke beziehen. Ob die Beobachtung dennoch zu bestätigen ist, darüber wird die Auflistung der Landschaftsbegriffe womöglich mehr Aufschluss geben.

	Valsertal
1	markiert
2	steil
3	klein
4	sichtbar
5	gross
6	hoch
7	rotweissrot
8	brüchig
9	gleich
10	flach
11	gut
12	weit
13	markant
14	einfach
15	genannt
16	spitz
17	jungfräulich
18	werdenden
19	unscheinbar
20	los
21	angenehm
22	umliegend
23	westlich
24	oberen
25	leuchtend
26	unübersichtlich
27	schön
28	ganz
29	erst
30	erkennbar

	Safiental
1	steil
2	lang
3	schön
4	klein
5	breit
6	windgeschützt
7	begehbar
8	schmal
9	flach
10	gut
11	einfach
12	link
13	ausgesetzt
14	unverspurt
15	ansteigend
16	geröllig
17	markiert
18	gross
19	hoch
20	letzte
21	eingezeichnet
22	idyllisch
23	oberste
24	deutlich
25	oberen
26	verschneit
27	weit
28	alt
29	sanft
30	interessant

	Thurgau
1	klein
2	markiert
3	gelb
4	rutschig
5	offiziell
6	schön
7	weit
8	nah
9	tannegger
10	erkennbar
11	steil
12	alt
13	erst
14	umgestürzt
15	eigentlich
16	mittelalterlich
17	richtig
18	schmal
19	flach
20	hoch
21	direkt
22	kurz
23	asphaltiert
24	tief
25	licht
26	moosbewachst
27	andern
28	ausgebaut
29	unwegsam
30	märchenhaft

	Schaffhausen
1	vorhanden
2	klein
3	gut
4	markiert
5	hoch
6	gross
7	schön
8	offen
9	unmarkiert
10	bemalt
11	prominent
12	andere
13	grün
14	schmal
15	wunderschön
16	wert
17	weit
18	steil
19	einfach
20	malerisch
21	richtig
22	super
23	26m
24	öffentlich
25	hart
26	unbefestigt
27	kühl
28	zügig
29	entsprechend
30	horizontal

Tab. 9: Markierung charakteristischer Adjektive in den vier Regionen.

7.4.2 Vergleich der Landschaftsbegriffe zwischen den Regionen

Die Begriffe in Tabelle 10 bestätigen, dass in den alpinen Regionen häufiger Winterlandschaften beschrieben werden. So befinden sich die Wörter «Schneedecke» und «Schneefeld» (in Tabelle 10 blau markiert) in den Berichten der beiden alpinen Regionen unter den wichtigsten 16 Begriffen. In den Mittellandregionen sind sie jedoch nicht zu finden. Wenn man die Bedeutung der Landschaftsbegriffe in den einzelnen Regionen betrachtet, scheinen die Resultate noch bei weiteren Elementen ein typisches Muster aufzuweisen. Wie in Kapitel 5.10.2 bereits erwähnt, erwartet man einen «Grat» (grau markiert) vorwiegend in gebirgigen Regionen. So nimmt dieser Begriff in den Berichten aus den Bergregionen auch einen höheren Stellenwert ein. Wie vermutet erhalten in den Texten der alpinen Regionen auch Berge (grün markiert) mehr Gewicht als im Mittelland. Das

Landschaftselement «Berg» ist nur in der Auflistung der beiden alpinen Regionen zu finden. Der Begriff «Hügel» (ebenfalls grün markiert) hingegen kommt unter den 30 wichtigsten Landschaftselementen nur in den beiden Mittellandregionen vor. Ein weiteres Beispiel ist der «Wanderweg» (rot markiert). Dessen tiefe Bewertung in den alpinen Regionen deutet eventuell darauf hin, dass zumindest offizielle Wege in den Gebieten, wo sich die Leute bewegt haben, rar sind. Ansonsten handelt es sich, wie in den Berichten des Valsertals erwähnt, um einen «Bergwanderweg» (ebenfalls rot markiert). Zuletzt sind Hänge (gelb markiert) deutlich gewichtiger in den Alpinregionen. Wie die Auflistung der Adjektive in Tabelle 9 zeigt, werden diese wohl oft als steil beschrieben, weil Hänge in Bergregionen grundsätzlich steiler sind als im Mittelland.

	Valsertal
1	Bergwanderweg
2	Weg
3	Hang
4	Gelände
5	Gipfel
6	Grat
7	Kapelle
8	Schneefeld
9	Pfadspur
10	Gletscher
11	Felsblock
12	Felskopf
13	Passage
14	Berg
15	See
16	Schneedecke
17	Ebene
18	Seelein
19	Wegmarkierung
20	Traverse
21	Bach
22	Nordgrat
23	Felsriegel
24	Gipfelkreuz
25	Hütte
26	Fels
27	Pfad
28	Talseite
29	Stein
30	Gestein

	Safiental
1	Hang
2	Weg
3	Grat
4	Piz
5	Couloir
6	Gelände
7	Mulde
8	Gipfel
9	Passage
10	Tälchen
11	Grashalde
12	Gras
13	Landschaft
14	Tal
15	Schneefeld
16	Osthang
17	Kirche
18	Kamm
19	Gipfelgrat
20	Felsturm
21	Talkessel
22	Winterwanderweg
23	Wald
24	Wanderweg
25	Berg
26	Pfad
27	Weglein
28	Aussichtspunkt
29	See
30	Stall

	Thurgau
1	Weg
2	Wanderweg
3	Baum
4	Tobel
5	Kirche
6	Hof
7	Höhle
8	Klettergarten
9	Wegspur
10	Gelände
11	Wald
12	Kapelle
13	Anhöhe
14	Kuppe
15	Hang
16	Loch
17	Pfad
18	Dorf
19	Felsstufe
20	See
21	Wiese
22	Haus
23	Abzweigung
24	Grat
25	Weglein
26	Hügel
27	Weiler
28	Burg
29	Waldstück
30	Weggabelung

	Schaffhausen
1	Weg
2	Turm
3	Wanderweg
4	Fels
5	Hügel
6	Wald
7	Gipfel
8	Pfad
9	Loch
10	Haus
11	Naturschutzgebiet
12	Graben
13	Treppe
14	Feld
15	Restaurant
16	Abzweigung
17	Dörfchen
18	Stein
19	Gipfelchen
20	Schlucht
21	Abhang
22	Berhang
23	Blume
24	Steig
25	Meer
26	Alp
27	Gelände
28	Wand
29	Terrain
30	Burg

Tab. 10: Markierung charakteristischer Landschaftsbegriffe in den vier Regionen.

Es scheint, dass Unterschiede in den Regionen einfacher anhand der Landschaftsbegriffe abzulesen sind. Ein Grund dafür ist wohl, dass die Diversität der Beschreibungswörter (5‘772) viel grösser ist als diejenige der Landschaftsbegriffe (273). Es soll aber nicht nur untersucht werden, was die Leute in den Regionen sehen, sondern auch, wie sie es beschreiben. Obwohl in zwei verschiedenen Regionen derselbe Begriff beschrieben werden kann, handelt es sich dabei um unterschiedliche Entitäten mit unterschiedlichen Eigenschaften.

Um dies etwas zu verdeutlichen, wurde auf nicht repräsentative Weise anhand einiger Landschaftselemente überprüft, ob ihnen in verschiedenen Regionen markant unterschiedliche Eigenschaften zugesprochen werden. Wie sich herausstellte, trifft dies auch in einigen Fällen zu. Als Beispiele sollen die Vergleiche der Begriffe «Dorf» und «Restaurant» zwischen den Kantonen Graubünden und Tessin angeführt werden. Restaurants werden in Graubünden im Vergleich zum Tessin auffällig oft mit den Worten «urchig», «urig» oder «urtümlich» charakterisiert. Betrachtet man die Beschreibungswörter des Begriffs «Dorf», fällt auf, dass im Kanton Graubünden oft das Wort «höchstgelegen» fällt. Im Tessin ist dies hingegen nie der Fall. Eine kurze Nachforschung hat ergeben, dass es sich beim Dorf «Juf» im Kanton Graubünden um die höchstgelegene Siedlung der Schweiz handelt, welche ganzjährig bewohnt ist *(www.bfs.admin.ch)*.

Am meisten Aussagekraft über die Wahrnehmung der Landschaft geben folglich die kompletten von SentiTours extrahierten Landschaftsbeschreibungen, d.h. die Kombination zwischen Beschreibungswort und Landschaftsbegriff. Aus diesem Grund sollen in einem erneuten Vergleich die kompletten Beschreibungen zwischen den Mittelland- und Alpinregionen einander gegenübergestellt werden. Da die Wahrscheinlichkeit für das mehrfache Vorkommen von Wortkombinationen geringer ist als dasjenige einzelner Wörter, werden die Beschreibungen der beiden Mittellandregionen sowie der beiden alpinen Regionen zusammengefasst. Aufgrund der starken Dominanz der Wegbeschreibungen werden alle Wegelemente von dieser Analyse ausgeschlossen. Die gemäss Tf-idf-Mass wichtigsten Wörter in den beiden Regionen sind der Tabelle 11 zu entnehmen.

Mittellandregionen	
1	tannegger Grat
2	eng Tobel
3	nah Hof
4	flach Kuppe
5	rutschig Hang
6	weit Hof
7	umgestürzt Baum
8	hoch Erhebung
9	klein Loch
10	mittelalterlich Burg

Alpinregionen	
1	steil Gelände
2	steil Hang
3	klein See
4	geschlossen Schneedecke
5	schmal Grat
6	unübersichtlich Gelände
7	flach Passage
8	ganz Hang
9	brüchig Grat
10	markant Felskopf

Tab. 11: Die 10 wichtigsten Landschaftsbeschreibungen in den Mittelland- und Alpinregionen.

Subjektiv betrachtet passen die Beschreibungen gut zu den Vorstellungen, welche man von den beschriebenen Regionen hat. Die Ausdrücke in der rechten Auflistung haben klar einen alpineren Charakter. Auffallend ist, dass im Mittelland unter den ersten zehn Beschreibungen keine Erwähnung eines Sees auftaucht. Dies liegt möglicherweise daran, dass die Seen, allen voran der Bodensee, häufiger beim Namen genannt und deshalb nicht erfasst werden. In den alpinen Regionen hingegen sprechen wohl vor allem Nicht-Einheimische bei den zahlreichen Gletscherseen einfach von einem «kleinen See». Im Gegensatz zum Mittelland («Hof», «Burg») sind Beschreibungen von anthropogenen Elementen in den Alpinregionen weniger wichtig. Wie schon beobachtet, ist die Steilheit von Hängen und Gelände in den Alpinregionen von grosser Bedeutung («steil Gelände», «steil Hang», «flach Passage»). Des Weiteren werden öfters Aktivitäten in verschneiten Landschaften unternommen, worauf die Beschreibung «geschlossene Schneedecke» hindeutet. Dies ist ausserdem ein Hinweis darauf, dass bei einer zukünftigen Analyse eine Stratifizierung der Aktivitäten sinnvoll wäre. Denn viele Winteraktivitäten wie zum Beispiel Skifahren können im Mittelland im Gegensatz zu den Alpen kaum unternommen werden. Charakteristisch ist zudem die Beschreibung des Landschaftselements «Tobel», welches interessanterweise nur im Thurgau beschrieben wird. Als Tobel bezeichnet man in der Schweiz ein trichterförmiges Tal, welches sich durch einen engen, schluchtartigen Ausgang kennzeichnet (*Schweizerisches Idiotikon*, 1961). So sprechen auch die Autoren auf Hikr oft von einem «engen Tobel». Das einzige in beiden Tabellen auftauchende Landschaftselement ist der «Grat». Dieser wird in den beiden Regionen aber unterschiedlich beschrieben. In den alpinen Regionen wird er oft als «schmal» und auch als «brüchig» bezeichnet. Im Mittelland handelt es sich jedoch um ein Toponym, den «Tannegger Grat». Der Grat wird dort also keineswegs am häufigsten beschrieben, da es sich gar nicht um eine echte Landschaftsbeschreibung handelt.

Ein Punkt, den es stets zu beachten gilt, ist die Verortung der Beschreibungen. Die beschriebenen Landschaftselemente können sich unter Umständen weit entfernt vom Betrachter oder der Betrachterin befinden. So kann bei einer Wanderung im Thurgau ohne Weiteres der Säntis als «schöner Berg» in den Appenzeller Alpen beschrieben werden.

Zusammenfassend lässt sich sagen, dass die extrahierten Landschaftsbeschreibungen den Vergleich von Landschaften im echten geographischen Raum ermöglichen. Es lassen sich typische Eigenschaften von Regionen oder auch Landschaftselementen in Abhängigkeit der Region herausarbeiten. Im Zusammenhang mit der Landschaftsbewertung wäre zwar eine akkuratere Erfassung der Meinungen nützlich. Dennoch können durch die in diesem Kapitel gemachten Vergleiche Eigenschaften und Elemente der Landschaft ausfindig gemacht werden, welche für die Leute von Bedeutung sind.

7.5 Räumliche Skala und Datenverfügbarkeit

Zum Abschluss dieser Diskussion soll noch das Thema der Datenverfügbarkeit angesprochen werden. Die Datenverfügbarkeit kann nämlich ein limitierender Faktor sein für die Analyse der Landschaftsbeschreibungen. Dies zeigt sich vor allem dann, wenn verschiedene Regionen miteinander verglichen werden sollen. In den Alpenregionen sind sehr viele Berichte vorhanden (z.B. Graubünden: 6‘554 Berichte, Stand: 12.08.2015). Deshalb wird der Kanton auf Hikr auch noch in weitere Subregionen unterteilt (z.B. Oberengadin: 1‘067 Berichte (Stand: 12.08.2015). Der Kanton Thurgau andererseits verfügt nur über 104 Berichte (Stand: 12.08.2015). Problematisch wird dies vor allem bei den potentiell aussagekräftigsten Vergleichen mit den kompletten Beschreibungen, bestehend aus Beschreibungswort und Landschaftsbegriff. Es muss deshalb stets beurteilt werden, ob es sich bei den erhaltenen Resultaten um echte Unterschiede handelt oder ob die Differenzen nur zufällig sind und von der geringen Datenverfügbarkeit herrühren.

Werden Mittellandregionen mit alpinen Regionen verglichen, lassen sich gewisse charakteristische Unterschiede herauskristallisieren. Beim Vergleich zweier alpiner Regionen ist dies jedoch schwieriger, weil die Unterschiede in den Landschaften natürlich geringfügiger sind. Dies könnte sich jedoch ändern, wenn viel kleinere räumliche Ausschnitte miteinander verglichen würden. Als kleinste räumliche Einheiten wurden die von Hikr definierten Subregionen der einzelnen Kantone verwendet. Interessanter wäre es womöglich, verschiedene Routen miteinander zu vergleichen. Dies würde aber wohl wieder durch die Datenverfügbarkeit eingeschränkt.

Zusammenfassend lässt sich sagen, dass ein Trade-Off besteht zwischen der Homogenität der Landschaft und der Datenverfügbarkeit. Dies hat zur Folge, dass beispielsweise für Vergleiche mit Landschaften aus dem Mittelland die Berichte ganzer Kantone verwendet werden müssen, da sonst zu wenige Beschreibungen extrahiert werden können. Damit werden jedoch grössere Gebiete mit verschiedenen heterogenen «Landschaften» zusammengefasst. Hikr verzeichnet in der jüngsten Zeit einen sehr starken Zuwachs an Berichten. Während der Erstellung dieser Arbeit stieg die Anzahl Beiträge innerhalb von 10 Monaten von ca. 65‘000 auf ca. 75‘000. So wird der Textkorpus zunehmend attraktiver für Anwendungen im Bereich des GIR. Bestimmt lassen sich zudem auch englischsprachige Datensätze finden, wie derjenige in *Nowak* (2013), welche mit der vorliegenden Methodik untersucht werden können.

8. Schlussfolgerung

In diesem Kapitel werden mithilfe der Erkenntnisse aus der Diskussion die Forschungsfragen beantwortet. Zum Schluss wird dann basierend darauf rekapituliert, was mit dieser Arbeit erreicht wurde.

8.1 Beantwortung der Forschungsfragen

8.1.1 Extraktion von Landschaftsbeschreibungen

Forschungsfrage 1

Teilfrage 1: Inwiefern lassen sich Beschreibungen von Landschaften aus Texten rechentechnisch extrahieren?

Es hat sich gezeigt, dass die mit SentiTours extrahierten Wortkombinationen aus Beschreibungswort und Landschaftsbegriff in den meisten Fällen plausibel sind. Dies ist eine Bestätigung dafür, dass die vorliegende Methodik verlässliche Resultate liefert. Basierend auf dem Konzept der Dependenzgrammatik und mithilfe eines Dependency Parsers lassen sich relevante Themenbegriffe mit den dazugehörigen Beschreibungswörtern aus einem Text herausfiltern. Im Gegensatz zur Analyse von klassischen Produktebeschreibungen, wie in *Cataldi et al.*, (2013) gezeigt, lässt sich dieses Konzept auch auf Landschaftsbeschreibungen übertragen. Dabei können die Landschaft als Produkt und die einzelnen Landschaftselemente als deren Produkteigenschaften angesehen werden. Die Möglichkeiten wurden dabei in dieser Arbeit noch nicht vollständig ausgeschöpft. So wurden beispielsweise noch keine Verben und Nomen als Beschreibungswörter erfasst. Mit dieser Arbeit wurde aber der Grundstein für eine Weiterentwicklung dieser neuen Methodik zur Extrahierung von Landschaftsbeschreibungen gelegt. Durch eine vertieftere Auseinandersetzung mit den syntaktischen Eigenschaften von solchen Beschreibungen kann das Vorgehen sicherlich verbessert werden.

8.1.2 Muster in den extrahierten Beschreibungen

Forschungsfrage 1:

Teilfrage 2: Welche Muster lassen sich in diesen Beschreibungen erkennen?

Die Ergebnisse zeigen, dass sich in den aus Hikr extrahierten Landschaftsbeschreibungen in verschiedener Hinsicht Unterschiede erkennen lassen. Hinsichtlich der Beschreibungswörter äus-

sern sich diese Unterschiede relativ deutlich beim Vergleich der verschiedenen Landschaftselemente. So lassen sich diese, basierend auf ihren Eigenschaften, in Clustern anordnen. Dann lassen sich typische Merkmale dieser Gruppen oder auch von einzelnen Landschaftselementen herausarbeiten.

Beim Vergleich der Beschreibungswörter zwischen verschiedenen Regionen ist das Sichtbarmachen räumlicher Eigenheiten etwas schwieriger. Es lässt sich zwar erkennen, dass unterschiedliche Adjektive verwendet werden und dass gemeinsam verwendete Adjektive unterschiedlich wichtig sind in den verschiedenen Regionen. Die Frage ist jedoch, welche Aussagekraft diese entdeckten Unterschiede anhand der alleinigen Analyse der Beschreibungswörter haben. Die Diversität der Adjektive ist sehr gross. Durch diese grosse Vielfalt an Adjektiven innerhalb einer Region ist der Vergleich zwischen zwei Regionen schwierig. Einfacher gelingt dies scheinbar mit der Analyse der beschriebenen Landschaftselemente. Wie in der Diskussion erwähnt wurde, repräsentiert ein Begriff jedoch stets eine Entität mit individuellen Eigenschaften. Die ausgewählten Beispiele zeigen dann auch, dass Landschaftselemente in verschiedenen Regionen sehr unterschiedlich beschrieben werden können.

Es bedarf also eines Vergleichs der Wortkombinationen bestehend aus Beschreibungswort und Landschaftsbegriff, um einen zusätzlichen Informationsgehalt über den Charakter einer Landschaft und ortstypische Eigenheiten zu liefern. Ausserdem können diese aufzeigen, wie die einzelnen Landschaftselemente von den Leuten wahrgenommen werden, was nur anhand einer Analyse der Adjektive oder der Landschaftsbegriffe alleine schwierig ist. In der Extraktion genau solcher Wortbeziehungen besteht auch der Mehrwert dieser Arbeit gegenüber zahlreichen Arbeiten, welche sich lediglich mit der Suche von alleinstehenden Begriffen beschäftigen. Die extrahierten Beschreibungen ermöglichen Erkenntnisse darüber, wie Leute ihre Umwelt beschreiben und welche Elemente für sie weshalb von Interesse sind. Dies gilt nicht nur bezüglich der Ästhetik, sondern auch hinsichtlich anderer Aspekte.

8.1.3 Landschaftsbeschreibungen und Sentiment Analysis

Forschungsfrage 2:

Inwiefern ist es möglich, mithilfe der Technik der Sentiment Analysis die Meinungen von Menschen gegenüber beschriebenen Landschaften zu ermitteln?

Es hat sich herausgestellt, dass einige der klassifizierten Sentiment-Wörter in der Domäne der Tourenberichte kaum eine plausible Polarität repräsentieren oder zumindest ohne Kontext wenig Aussagekraft haben hinsichtlich der Meinung der Leute gegenüber der Landschaft. Hinzu kommt, dass sich diese Ausdrücke besonders unter den häufig vertretenen Sentiment-Wörtern, im Speziellen unter den negativen, befinden. Dies zeigt, dass vor allem Ausdrücke, welche allgemein als

negativ beurteilt werden, im Touren-Kontext oft nicht als negativ gelten. Wenn ein Hang als «steil» beschrieben wird, ist es ohne zusätzlichen Kontext schwierig zu beurteilen, ob diese Beschreibung für die Person eine negative oder positive Bedeutung hat. Der Einbezug der Adverbien könnte dabei in gewissen Fällen behilflich sein. Wird ein Hang beispielsweise als «zu steil» beschrieben, so ist dies mit grosser Wahrscheinlichkeit negativ gemeint. Wird er jedoch wiederum als «etwas steil» charakterisiert, ist noch weiterer Kontext vonnöten. Unter den tiefer frequentierten Ausdrücken der Sentiment-Wörter befinden sich durchaus solche, welche eine differenzierte Meinung gegenüber einer Landschaft auch ohne grösseren Kontext wiedergeben können. Dazu gehören Ausdrücke wie «eintönig», «langweilig», «hässlich», «spektakulär», «sanft», «idyllisch» etc. Diese und ähnliche Wörter finden sich auch in den Auflistungen der Regionenvergleiche wieder. Doch deren Häufigkeit ist teilweise etwas gering, so dass Vergleiche schwierig sind.
Aufgrund der genannten Punkte liefern die mit der vorliegenden Methodik extrahierten Gefühlspolaritäten wohl keinen Mehrwert für die Landschaftsbewertung. Durch eine Weiterentwicklung der Anwendung könnten aber aussagekräftigere Resultate erzielt werden. Wie gross deren Nutzen sein kann, gilt es zu überprüfen.

8.2 Was wurde erreicht?

In dieser Arbeit wurde in einer explorativen Vorgehensweise untersucht, inwiefern sich Landschaftsbeschreibungen und deren inhärente Meinungen aus unstrukturierten deutschen Texten automatisiert herausfiltern lassen. Dazu wurden die Theorie der Dependenzgrammatik und Ideen aus der Sentiment Analysis genutzt, um die Anwendung SentiTours zu entwickeln. Die extrahierten Landschaftsbeschreibungen, bestehend aus Beschreibungswort und beschriebenem Landschaftselement, wurden auf deren Plausibilität getestet und anschliessend auf räumliche Muster und andere Eigenschaften untersucht.
Der Beitrag dieser Arbeit besteht darin, dass eine Technik zur Extraktion von Landschaftsbeschreibungen präsentiert wurde, welche nach Kenntnis des Autors bis anhin nicht in der Erforschung der Landschaftswahrnehmung zur Anwendung kam. Es wurde zudem ein erstes Mal überprüft, inwiefern sich Meinungen über Landschaften aus unstrukturierten deutschen Texten erkennen lassen, wenn keine quantitative Bewertung der Landschaft als Referenz vorliegt.

9. Ausblick

9.1 Weiterentwicklung von SentiTours

9.1.1 Extrahierung von Landschaftsbeschreibungen

In einer weiterführenden Arbeit wäre es interessant zu untersuchen, ob eine Möglichkeit besteht, zwischen beschreibenden und nicht beschreibenden Nomen oder Verben zu unterscheiden. Dies zum Beispiel, indem für die ermittelten Wörter quantitativ beurteilt wird, wie oft sie eine beschreibende Funktion einnehmen. Somit könnten weitere Beschreibungswörter respektive weitere Arten von Landschaftsbeschreibungen ermittelt und analysiert werden. Des Weiteren wäre es spannend, Adverbien für die Erfassung von zusätzlichem Kontext mit einzubeziehen. Zur Weiterentwicklung der getesteten Methodik würde sich eine vertieftere Auseinandersetzung mit der Sprache und ihren syntaktischen Eigenheiten anbieten.

9.1.2 Sentiment Analysis

Die Verwendung einer nicht an den Kontext angepassten Sentiment-Bibliothek hat sich als wenig geeignet erwiesen, um sinnvolle Gefühlspolaritäten für die Landschaftsbeschreibungen zu erhalten. Es wäre deshalb sinnvoll, eine Sentiment-Bibliothek zu entwickeln, welche auf die Domäne von Tourenberichten angepasst ist. Dazu könnte zum Beispiel anhand einer repräsentativen Auswahl von Beschreibungen für jedes Wort ermittelt werden, wie häufig es im positiven oder negativen Sinne verwendet wird. Es wäre interessant zu sehen, inwiefern eine solche auf den Hikr-Korpus angepasste Sentiment-Bibliothek die Resultate verbessern würde. Es wird erwartet, dass häufig verwendete Ausdrücke wie «steil» und «klein» dann neutral bewertet würden. Bei der Erstellung einer solchen Bibliothek müssten die Bedürfnisse der verschiedenen Aktivitäten berücksichtigt werden. Denn für einen Kletterer hat ein steiler Hang wohl eine andere Bedeutung als für einen Wanderer.

In dieser Arbeit wurde versucht, sowohl die Beschreibungswörter per se als auch die Gefühlspolaritäten zu untersuchen. Deshalb hat sich die Verwendung einer Sentiment-Bibliothek angeboten. Um die Möglichkeiten einer Sentiment-Analysis weiter zu testen, wäre als Vergleich auch die Anwendung einer Methode des maschinellen Lernens wie beispielsweise in *Pang & Lee* (2004) interessant. Dazu müsste im Falle des Hikr-Korpus eine Teilmenge an Texten respektive Sätzen manuell bewertet werden, um einen Trainingsdatensatz zu erhalten. Danach könnten die Gefühlspolaritäten auf Satzebene ermittelt werden, und zwar für jeden Satz, in welchem ein Landschaftsbegriff vorkommt oder beschrieben wird.

9.2 Analyse der Landschaftsbeschreibungen

Im Rahmen dieser Arbeit konnte das Potenzial der ermittelten Beschreibungen noch nicht ausreichend untersucht werden. Aufgrund der vielen Metainformationen wäre eine vielseitige Stratifizierung des Datensatzes möglich. So könnten die Beschreibungen von verschiedenen Nutzergruppen (Skitourengänger, Biker etc.) miteinander verglichen werden und somit könnte eventuell auf verschiedene Bedürfnisse dieser Nutzergruppen geschlossen werden. Auch eine zeitliche Stratifizierung, beispielsweise nach Jahreszeiten, könnte interessant sein. Wie bereits erwähnt, wurden für die räumlichen Vergleiche die von Hikr definierten Regionen verwendet. Eine weitere Unterteilung und ein Vergleich auf der Ebene von Routen wäre bestimmt interessant, weil somit weniger die Eigenschaften verschiedener Regionen zusammengefasst würden und womöglich auch Vergleiche zwischen alpinen Gebieten mehr Aussagekraft hätten.

Eine andere interessante Anwendung könnte der Vergleich mit tatsächlichen Geländeformen sein. So könnten für verschiedene Regionen mit einem Geländemodell die durchschnittlichen Hangneigungen berechnet werden. Anschliessend könnte überprüft werden, inwiefern sich die Beschreibungen der Neigung in den Berichten mit den Erwartungen decken.

Zusätzlich wäre es spannend, die entwickelte Methodik auf andere Sprachen anzuwenden. So könnte man zum Beispiel Vergleiche zwischen den verschiedenen Landessprachen der Schweiz anstellen. Auch wäre eine Ausweitung auf andere Regionen der Welt und andere Textkorpora möglich.

Grundsätzlich kann die Extrahierung von Landschaftsbeschreibungen, so wie in dieser Arbeit durchgeführt, eine Erweiterung der Analysemöglichkeiten in verschiedenen Bereichen bieten. Sie erweitert Untersuchungen, welche sich bisher vor allem auf die Benennung von Entitäten konzentrierten, um die Eigenschaften, welche diesen zugesprochen werden. Für Arbeiten, welche sich auch schon der Extrahierung von Landschaftsbeschreibungen widmeten, kann die entwickelte Methodik aufgrund der hohen Plausibilität eine interessante Alternative bieten.

Literatur

Antrop, M. (2005). From holistic landscape synthesis to transdisciplinary landscape management. *Landscape Research to Landscape Planning: Aspects of integration, education and application, 27-50.*

Appleton, J. (1980). Landscape in the Arts and the Sciences. University of Hull.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *LREC*, 10, 2200-2204.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2011). Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. Vol. 13, Berlin, Springer.

Balasubramaniam, N. (2009). User-generated content. *In: Proceedings of business aspects of the internet of things, seminar of advanced topics, Zürich: ETH, 28-33*.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *ACL*, 440-447.

Borg, I., & Groenen, P. J. (2005). Modern multidimensional scaling: Theory and applications. New York, Springer.

Buchholz, S., & Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. *In: Proceedings of the Tenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, 149-164.*

Burenhult, N., & Levinson, S. C. (2008). Language and landscape: a cross-linguistic perspective. *Language Sciences*, *30*(2-3), 135–150.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, (2), 15-21.

Cardie, C. (1997). Empirical Methods in Information Extraction. *AI Magazine*, 18(4), 65.

Cataldi, M., Ballatore, A., Tiddi, I., & Aufaure, M. A. (2013). Good location, terrible food: detecting feature sentiment in user-generated reviews. *Social Network Analysis and Mining*, *3*(4), 1149-1163.

CoNLL-X Shared Task: Multi-lingual Dependency Parsing. Tenth Conference on Computational Natural Language Learning. <http://ilk.uvt.nl/conll/> (Stand: Juni, 2006; Zugriff: März, 2015).

Dakin, S. (2003). There's more to landscape than meets the eye: towards inclusive landscape assessment in resource and environmental management. *The Canadian Geographer/Le Géographe Canadien*, *47*(2), 185–200.

Daniel, T. C. (2001). Whither scenic beauty? Visual landscape quality assessment in the 21st century. *Landscape and Urban Planning*, *54*(1-4), 267–281.

Derungs, C., & Purves, R. S. (2013). From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, *28*(6), 1272–1293.

Derungs, C., Wartmann, F., Purves, R. S., & Mark, D. M. (2013). The meanings of the generic parts of toponyms: use and limitations of gazetteers in studies of landscape terms. *In: Spatial Information Theory*, 261-278.

Derungs, C. (2014). From Text to Landscape: Extraction of Landscape Concepts through the Resolution of Ambiguity and Vagueness present in Descriptions of Natural Landscapes. Dissertation. *Geographisches Institut der Universität Zürich, 1-183.*

Duden (2014). Artikel «Landschaft», Bedeutungsübersicht. <http://www.duden.de/rechtschreibung/Landschaft> (Zugriff: November, 2014).

Emerson, R. W. (1987). Ausgewählte Texte, 1. Aufl., München, Goldmann, S. 59.

Feld, S. (1996). Waterfalls of song: an acoustemology of place resounding in Bosavi, Papua New Guinea. *In: Senses of Place*, Feld, S., & Basso, K. H. (Hrsg.), 91-135. Santa Fe NM: School of American Research Press.

Fisher, P. (2000). Sorites paradox and vague geographies. *Fuzzy Sets and Systems*, *113*(1), 7–18.

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*(3-4), 137–148.

Foth, K. A. (2006). Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Fachbereich Informatik. Universität Hamburg.

Granö, J. G. (1997). Pure Geography. Baltimore MD: Johns Hopkins University Press.

Habib, M. B., & van Keulen, M. (2012). Improving Toponym Disambiguation by Iteratively Enhancing Certainty of Extraction. University of Twente. SciTePress.

Hauthal, E., & Burghardt, D. (2014). Mapping Space-Related Emotions out of User-Generated Photo Metadata Considering Grammatical Issues. *The Cartographic Journal*.

Hikr Tourenberichte Website. <http://www.hikr.org/> (mehrere Zugriffe 2014 und 2015).

Hill, S., & Ready-Campbell, N. (2014). Expert Stock Picker: The Wisdom of (Experts in) Crowds. *International Journal of Electronic Commerce*, 15(3), 73-102.

Hollenstein, L., & Purves, R. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21-48.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. *AAAI*, 4(4), 755-760.

Huang, A. (2008). Similarity measures for text document clustering. *In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, 49-56.*

Hurst, M. F., & Nigam, K. (2003). Retrieving Topical Sentiments from Online Document Collections. *Electronic Imaging, International Society for Optics and Photonics*, 27-34.

Jung, W-Y. (1995). Syntaktische Relationen im Rahmen der Dependenzgrammatik. Beiträge zur germanistischen Sprachwissenschaft, Vol. 9, Hamburg, Buske Verlag.

Manning, C., & Schütze, H. (1999). Foundations of statistical natural language processing. Massachusetts, MIT press.

Mark, D. M., Turk, A. G., Burenhult, N., & Stea, D. (2011). Landscape in language: An introduction. *In: Landscape in language: transdisciplinary perspectives*, Mark, D., Turk, A. G., Burenhult, N., & Stea, D. (Hrsg.). Culture and Language Use, Vol. 4, Philadelphia, 1-24.

Mark, D.M., Turk, A.G., & Stea, D., (2010). Ethnophysiography of Arid Lands. *Landscape Ethnoecology: Concepts of Biotic and Physical Space*, 346, 27-45.

Naveh, Z., & Lieberman, A. S., (1984). Landscape ecology theory and applications. New York, Springer.

Nouah's Ark. Graphik der geographischen Lage der Schweiz in Europa. <http://www.nouahsark.com/en/infocenter/worldwide/europe/switzerland/switzerland_location.php> (Zugriff: Mai, 2015).

Nowak, M. (2013). Modellierung der Landschaftsqualität mit Opinion Mining. Diplomarbeit. *Geographisches Institut der Universität Zürich*, 1-132.

Ochoa, X., & Duval, E. (2008). Quantitative analysis of user-generated content on the web. *In: Proceedings of webevolve 2008: web science workshop at WWW2008*, Peking, China, 1–8.

Palka, E. I. (1995). Coming to grips with the concept of landscape. *Landscape Journal*, 14(1), 63-73.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, *Association for Computational Linguistics*, 271-279.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, *Association for Computational Linguistics*, 10, 79–86.

ParZu - The Zurich Dependency Parser for German. <https://github.com/rsennrich/parzu> (mehrere Zugriffe 2014 und 2015).

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *In: Proceedings of the First Instructional Conference on Machine Learning.*

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. *LREC*, 1168-1171.

Safiental. Inserat über die Gemeinde Safiental; Gesamtfläche. <http://www.safiental.ch/fileadmin/user_upload/customers/safiental/safien_gemeinde/Dokumente/Diverse_Dokumente/Inserat_Verwaltung_2014.pdf> (Zugriff: Juni, 2014).

Scheider, S., & Purves, R. (2013). Semantic place localization from narratives. *In: Proceedings of the First ACM SIGSPATIAL International Workshop on Computational Models of Place, COMP'13, 16-19.*

Schweizerisches Bundesamt für Statistik. Statistisches Lexikon. <http://www.bfs.admin.ch/bfs/portal/de/index/infothek/lexikon/lex/0.html> (Zugriff: Juni, 2015).

Schweizerisches Bundesgesetz über die Raumplanung (Raumplanungsgesetz, RPG) vom 22. Juni 1979 (Stand am 1. Mai 2014).

Schweizerisches Idiotikon. Band XII, Spalten 116–122, Artikel «Tobel». <https://digital.idiotikon.ch/idtkn/id12.htm#!page/120115/mode/1up> (Zugriff: September, 2015).

Sennrich, R., & Schneider, G. (2009). A new hybrid dependency parser for German. *In: Proceedings of the German Society for Computational Linguistics and Language Technology, 115-124.*

SentiWS. Wortschatz Universität Leipzig. <http://asv.informatik.uni-leipzig.de/download/sentiws.html> (Zugriff: April, 2015).

Tarvainen, K. (2000). Einführung in die Dependenzgrammatik. Germanistische Linguistik, Vol. 35, Tübingen, Niemeyer.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558.

Tulloch, D. (2007). Many, many maps: Empowerment and online participatory mapping. *First Monday*, 12(2), 5.

Turney, P. D. (2001). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 417-424.

Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, *15*(1), 121–149.

Võ, M. L. H., Jacobs, A. M., & Conrad, M. (2006). Cross-validating the Berlin Affective Word List. *Behavior Research Methods*, *38*(4), 606–609.

Vals. Gemeinde-Steckbrief; Gesamtfläche. <http://www.vals.ch/gemeinde-region/gemeinde/steckbrief/> (Zugriff: Juni, 2014).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05, Association for Computational Linguistics*, 347–354.

Yamashita, S. (2002). Perception and evaluation of water in landscape: use of Photo-Projective Method to compare child and adult residents' perceptions of a Japanese river environment. *Landscape and Urban Planning*, *62*(1), 3–17.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *In: Proceedings of the 2003 conference on Empirical methods in natural language processing, Association for Computational Linguistics,* 10, 129–136.

Zipf, G. K. (1935). The psycho-biology of language: an introduction to dynamic philology. Boston, Houghton Mifflin company.

Anhang

Anhang A

Landschaftsbegriffe aus *Derungs & Purves* (2013).

rank	nat. features	count $T+B$	rank	nat. features	count $T+B$	rank	nat. features	count $T+B$
1	gipfel	29635	36	baum	1955	71	felsgrat	882
2	berg	27037	37	flanke	1784	72	gipfelgrat	861
3	alp	24840	38	südwand	1768	73	schutthalde	833
4	gletscher	17849	39	weide	1710	74	westwand	810
5	fels	17522	40	schneefeld	1687	75	steilhang	792
6	grat	14337	41	fluss	1653	76	paß	787
7	wand	14202	42	geröll	1645	77	vorgipfel	754
8	tal	10273	43	ostgrat	1608	78	kuppe	753
9	spitze	6544	44	horn	1590	79	gletscherzunge	747
10	thal	5705	45	wiese	1567	80	südostgrat	727
11	stein	5626	46	westgrat	1514	81	talboden	722
12	hang	5551	47	nordgrat	1511	82	nordostgrat	691
13	wald	5199	48	abgrund	1429	83	nordflanke	670
14	see	4967	49	felsblock	1405	84	südwestgrat	666
15	gebirge	4822	50	abhang	1386	85	küste	630
16	platte	4078	51	südgrat	1386	86	alpweide	593
17	gestein	3717	52	überhang	1385	87	wüste	558
18	landschaft	3614	53	bergschrund	1364	88	einzugsgebiet	551
19	pass	3580	54	loch	1364	89	nordwestgrat	527
20	schlucht	3418	55	schrund	1335	90	westflanke	526
21	spalte	3345	56	plateau	1319	91	waldgrenze	515
22	felswand	3169	57	massiv	1308	92	südflanke	511
23	bach	3103	58	insel	1269	93	talseite	487
24	scharte	2800	59	wasserfall	1187	94	wasserscheide	486
25	gelände	2662	60	passhöhe	1167	95	ostflanke	474
26	meer	2637	61	hauptgipfel	1118			
27	pfad	2610	62	feld	1097			
28	kamm	2585	63	schutt	1069			
29	hochgebirge	2479	64	ostwand	1060			
30	rinne	2477	65	matten	1060			
31	moräne	2312	66	eiswand	1044			
32	nordwand	2217	67	blume	950			
33	ebene	2074	68	gebirgswelt	911			
34	sattel	2049	69	hügel	909			
35	quelle	2009	70	terrain	894			

Anhang B

Aufbau der Studie zur Ermittlung von Landschaftsbegriffen im Hikr-Korpus.

Das Ziel dieser Aufgabe ist das Identifizieren von Landschaftsbegriffen aus einer Liste mit Nomen, welche aus einem Textkorpus mit Tourenberichten extrahiert wurden. Wird ein Wort als Landschaftsbegriff erachtet, ist im Feld neben dem Wort ein Kreuz zu machen. Die Felder der übrigen Wörter sind leerzulassen. Jedes Vorkommen eines Landschaftsbegriffs muss annotiert werden, unabhängig davon ob an früherer Stelle ein ähnlicher/deckungsgleicher Begriff bereits vorgekommen ist (z.B. Wald, Waldstück). Die Liste besteht aus 1500 Wörtern. Die benötigte Zeit zur Bearbeitung aller Wörter spielt keine Rolle.

Arbeitsdefinition Landschaftsbegriff:

Als Hilfestellung zur Annotation der Landschaftsbegriffe in strittigen Fällen werden nachfolgend einige Regeln definiert. Die Regeln leiten sich aus der eigenen Arbeitsdefinition des Landschaftsbegriffs sowie aus den Annotationsregeln für natürliche Objekte aus *Derungs & Purves* (2013) ab:

Ein Landschaftsbegriff…

…*bezeichnet ein Landschaftselement.* Ein Landschaftsbegriff wird verwendet, um einen Ausschnitt oder ein Element der Landschaft zu beschreiben. Es kann sich dabei sowohl um ein anthropogenes oder anthropogen geprägtes (z.B. Dorf, Trampelpfad), als auch um ein natürliches Element (z.B. Grat, Fels) handeln.

…*bezeichnet keine Aktivität.* Manchmal können Nomen sowohl Aktivitäten als auch Landschaftselemente sein. In solchen Fällen wird das Wort als Aktivität angesehen und somit nicht angekreuzt. Folgendes Vorgehen kann dabei als Entscheidungshilfe dienen: Wenn ein Nomen in direkter Beziehung zu einem Verb mit der gleichen Bedeutung steht (Aufstieg → aufsteigen) wird es nicht als Landschaftsbegriff annotiert. Ändert sich bei der Umformung in ein Verb der Wortsinn, wird das Wort als Landschaftsbegriff gekennzeichnet (Berg → bergen).

…*ist generisch.* Er bezeichnet demnach Vertreter einer Objekt-Klasse und nicht individuelle Objekte. Infolgedessen ist zum Beispiel Wiese ein Landschaftselement. Beim Rothorn hingegen handelt es sich um ein Individuum.

…*bezeichnet kein Phänomen oder Qualität.* Bei Landschaftselementen handelt es sich um weitgehend unabhängige Existenzen. Phänomene oder Qualitäten sind meist Spezifikationen von Landschaftselementen. Die Begriffe Schnee oder Eis werden zum Beispiel verwendet, um den Zustand eines Hangs näher zu beschreiben. Ein Eisfeld hingegen ist ein unabhängiges Landschaftselement.

Anhang C

Auflistung aller als Landschaftselemente definierten Begriffe. Diese setzt sich zusammen aus der Identifikation von Landschaftsbegriffen aus den 1500 häufigsten Nomen des Hikr-Korpus und der Liste aus *Derungs & Purves* (2013). Basierend auf dieser Liste hat das Programm SentiTours nach Beschreibungen der entsprechenden Begriffe gesucht. Zusätzlich sind deshalb jeweils die ermittelten Häufigkeiten und die daraus resultierende Rangnummer angegeben.

Rang-Nr.	Landschaftselement	Häufigkeit
1	Weg	14941
2	Gipfel	5328
3	Gelände	4422
4	Grat	3689
5	Wanderweg	3499
6	Hang	3473
7	Fels	3467
8	Pfad	3265
9	Berg	2545
10	Passage	1914
11	Wald	1656
12	Alp	1626
13	Hütte	1604
14	Couloir	1410
15	Schneefeld	1300
16	Rinne	1288
17	Wegspur	1256
18	Gras	1246
19	Bergweg	1237
20	Landschaft	1092
21	Flanke	1061
22	Wiese	972
23	Stein	952
24	Tal	925
25	See	914
26	Gletscher	872
27	Platte	852
28	Wand	802
29	Felswand	763
30	Turm	673
31	Felsstufe	672
32	Baum	671
33	Gestein	656
34	Spalt	637
35	Grashang	624
36	Geröll	613
37	Brücke	583
38	Schneedecke	560
39	Aussichtspunkt	557
40	Bahn	553

Rang-Nr.	Landschaftselement	Häufigkeit
41	Haus	519
42	Platz	494
43	Bach	460
44	Klettersteig	437
45	Weglein	435
46	Steig	430
47	Gipfelkreuz	421
48	Dorf	420
49	Talseite	412
50	Restaurant	411
51	Wasserfall	404
52	Sattel	395
53	Felsband	390
54	Schlucht	390
55	Ebene	372
56	Traverse	364
57	Val	359
58	Loch	352
59	Hügel	351
60	Schutt	351
61	Scharte	340
62	Erhebung	325
63	Ufer	322
64	Piz	311
65	Grasflanke	307
66	Mulde	307
67	Weide	301
68	Pfadspur	296
69	Südflanke	292
70	Gipfelhang	287
71	Waldweg	279
72	Terrasse	273
73	Nordflanke	266
74	Bergwanderweg	264
75	Plateau	260
76	Strässchen	237
77	Kreuz	236
78	Vorgipfel	230
79	Lücke	230
80	Tunnel	229

Rang-Nr.	Landschaftselement	Häufigkeit
81	Geröllfeld	228
82	Hüttenweg	228
83	Felsblock	225
84	Südhang	223
85	Natur	219
86	Höhenweg	213
87	Seelein	207
88	Graben	203
89	Kamm	202
90	Tälchen	201
91	Alpweg	195
92	Treppe	195
93	Pass	194
94	Felsriegel	192
95	Kirche	189
96	Kapelle	189
97	Seilbahn	187
98	Gipfelgrat	186
99	Felsbrocken	184
100	Abzweigung	183
101	Geröllhalde	183
102	Piste	180
103	Höhle	180
104	Firngrat	178
105	Steilhang	176
106	Blume	172
107	Kante	172
108	Firnfeld	172
109	Alphütte	171
110	Grashalde	171
111	Tobel	171
112	Kuppe	169
113	Waldstück	165
114	Holzbrücke	165
115	Gebäude	163
116	Steg	162
117	Felsplatte	162
118	Viertausender	162
119	Hindernis	158
120	Hotel	156
121	Westflanke	155
122	Nordhang	154
123	Hochebene	152
124	Klettergarten	151
125	Ostgrat	150
126	Felsturm	149
127	Südgrat	147
128	Ostflanke	145
129	Horn	145
130	Westgrat	145

Rang-Nr.	Landschaftselement	Häufigkeit
131	Felsaufschwung	144
132	Forstweg	142
133	Talkessel	141
134	Gratweg	139
135	Weiler	138
136	Gipfelplateau	136
137	Beiz	133
138	Südwand	132
139	Moräne	131
140	Terrain	131
141	Winterwanderweg	129
142	Hängebrücke	129
143	Nordgrat	129
144	Feld	125
145	Alpweide	125
146	Gratrücken	125
147	Dörfchen	124
148	Steinmännchen	122
149	Bahnhof	117
150	Hof	116
151	Felskopf	114
152	Stollen	114
153	Nordwand	113
154	Lawine	111
155	Felsgrat	110
156	Lawinenkegel	110
157	Feuerstelle	109
158	Normalweg	109
159	Bachbett	107
160	Wegmarkierung	105
161	Monte	105
162	Lago	103
163	Bächlein	101
164	Schutthalde	101
165	Siedlung	101
166	Ruine	99
167	Gipfelchen	99
168	Trampelpfad	95
169	Mauer	94
170	Stall	94
171	Gasthaus	88
172	Skigebiet	88
173	Senke	86
174	Anhöhe	86
175	Brunnen	85
176	Fluss	84
177	Karrenfeld	84
178	Station	81
179	Stadt	81
180	Skipiste	80

Rang-Nr.	Landschaftselement	Häufigkeit
181	Fahrsträsschen	79
182	Überhang	79
183	Schneegrat	76
184	Spaltenzone	75
185	Gebirge	74
186	Verbindungsgrat	74
187	Osthang	72
188	Spitzkehren	71
189	Stausee	69
190	Lawinenverbauung	69
191	Ostwand	69
192	Blockgrat	68
193	Bähnli	66
194	Abhang	66
195	Abgrund	66
196	Hauptgipfel	66
197	Südseite	65
198	Berggasthaus	65
199	Autobahn	64
200	Massiv	64
201	Bachlauf	62
202	Talboden	62
203	Schotter	62
204	Bergschrund	61
205	Verzweigung	60
206	Luftseilbahn	59
207	Einsattelung	58
208	Naturschutzgebiet	58
209	Zaun	55
210	Standseilbahn	55
211	Bergrestaurant	54
212	Bänkli	54
213	Sommerweg	53
214	Sessellift	53
215	Steinbruch	53
216	Lac	52
217	Krete	52
218	Burg	51
219	Staumauer	50
220	Gipfelkopf	50
221	Bergstation	49
222	Bergbahn	48
223	Waldrand	47
224	Gondelbahn	46
225	Nordostgrat	45
226	Gratkante	45
227	Wegweiser	44
228	Südwestflanke	44
229	Westwand	42
230	Meer	42

Rang-Nr.	Landschaftselement	Häufigkeit
231	Nordwestgrat	41
232	Südwestgrat	40
233	Nordwestflanke	40
234	Sesselbahn	39
235	Gipfelfels	39
236	Weggabelung	37
237	Gipfelbereich	35
238	Ostgipfel	35
239	NW-Grat	33
240	Kloster	32
241	Südostgrat	32
242	Insel	32
243	Berghaus	32
244	Südgipfel	31
245	NE-Grat	31
246	SE-Grat	29
247	Nordgipfel	24
248	Westgipfel	24
249	Gletscherzunge	21
250	Joch	20
251	Schrund	20
252	Talgrund	20
253	Eiswand	19
254	Grätli	18
255	Wasserscheide	16
256	Passhöhe	14
257	Seilbahnstation	14
258	Delta	13
259	Kurhaus	13
260	Talstation	13
261	Bahnstation	12
262	Passo	12
263	Wandfuss	11
264	Waldgrenze	8
265	Hochgebirge	8
266	Gebirgswelt	7
267	Mittelgipfel	5
268	Wasseraue	4
269	Wüste	4
270	Einzugsgebiet	4
271	Küste	1
272	Mittelstation	0
273	Thal	0

Anhang D

Dokumentation der Anwendung SentiTours erstellt mithilfe von «pydoc».

SentiTours Dokumentation

#coding: UTF-8
#Anwendung: SentiTours
#Autor: Ramón Huldi
#Datum: 30.05.2015
#Version: 1.0
#Entwicklungsumgebung: Linux Ubuntu
#Eine Anleitung zur Installation des Dependency Parsers ParZu befindet sich auf
https://github.com/rsennrich/parzu

Modules

os **re** **BeautifulSoup** **Detect**
Decimal

Classes

__builtin__.object
HtmlGroup
HtmlParser
Report
ReportGroup
RelevantSentence
SentenceFinder
DependencyRelation
SentimentFinder

class **HtmlGroup**(__builtin__.object)
#Klasse HtmlGroup: Vorbereitung aller Dateien/Texte für den Parser

Methods defined here:

Prepare_Files(self)
#Speichert den gesamten Quellcode jeder html Datei als Element in einer Liste

class **HtmlParser**(__builtin__.object)

#Klasse HtmlParser: Die relevanten Inhalte aus den vorbereitenen html Dateien werden mit separaten Methoden herausgelesen

Methods defined here:

Find_Activity(self)
#Herauslesen der Aktivität

Find_Author(self)
Herauslesen des Autors

Find_Difficulty(self)
#Herauslesen der Schwierigkeit

Find_Region(self)
#Herauslesen der Region

Find_Text(self)
#Herauslesen des Texts

Find_Time(self)
#Herauslesen des Zeitbedarfs

Find_TourDatum(self)
#Herauslesen des Tour-Datums

Find_Waypoints(self)
#Herauslesen der Wegpunkte

---

class **Report**(__builtin__.object)

#Klasse Report: Bericht-Objekt mit ID, Region, Text, Metainformationen, Liste mit den relevanten Sätzen, Liste mit den gefundenen Wortbeziehungen, Liste mit den gespeicherten Sentiment-Werten

Methods defined here:

__init__(self, reportID, region, text, date, activity, difficulty, waypoints, time, author, sentenceList, relationList, sentimentList)

---

class **ReportGroup**(__builtin__.object)

#Klasse ReportGroup: Erstellen und Speichern der Berichte mit ihren Attributen in einer Liste

Methods defined here:

Report_Creator(self)

---

class **RelevantSentence**(__builtin__.object)

#Klasse RelevantSentence: Jeder Satz, welcher ein Keyword (Landschaftsbegriff) enthält: Objekt mit reportID, einer eigenen SatzID, dem Landschaftsbegriff, und dem Satz im CoNLL Format

Methods defined here:

__init__(self, reportID, sentenceID, keyword, CoNLLsentence)

---

class **SentenceFinder**(__builtin__.object)

#Klasse SentenceFinder: Suchen und speichern der Sätze, in welchen Landschaftsbegriffe vorkommen

Methods defined here:

Find_Sentence(self)

#Findet die Sätze mit Landschaftsbegriffen und erstellt aus diesen ein Objekt der Klasse RelevantSentence
#Diese Objekte werden bei den Bericht-Objekten in das Attribut sentenceList gespeichert
#Dazu wird jeder Text der Bericht-Objekte in Sätze aufgesplittet und dann wird geprüft, ob ein gesuchter Landschaftsbegriff in diesem Satz vorkommt

---

class **DependencyRelation**(__builtin__.object)

#Klasse DependencyRelation: Implementation der verschiedenen Methoden, welche es zum Auffinden der verschiedenen Beziehungstypen in den CoNLL-Sätzen braucht. Die gefundenen Wortbeziehungen werden an die Elemente der sentenceList der Bericht-Objekte gehängt und neu in die relationList gespeichert

Methods defined here:

Attr_Relation(self)

#Methode zum Auffinden von Attributbeziehungen (attr)
#Beispiel: Dies ist ein schöner Berg

Pred_Relation(self)

#Methode zum Auffinden von Prädikatbeziehungen (pred)
#Beispiel: Dieser Berg ist schön

PredPerfect_Relation(self)

#Methode zum Auffinden von Prädikat Beziehungen in der Vergangenheitsform Perfekt
#Beispiel: Dieser Berg ist schön gewesen

---

class **SentimentFinder**(__builtin__.object)

#Klasse SentimentFinder: Sucht nach den Sentimentwerten der beschreibenden Wörter. Die gefundenen Sentiment-Werte (können auch None sein) werden an die Elemente der relationList der Bericht-Objekte gehängt und neu in die sentimentList gespeichert. Die sentimentList enthält schlussendlich die fertigen Outputs von SentiTours.

Methods defined here:

Find_PositiveSentiment(self)

#Prüfen, ob das Beschreibungswort unter den positiven Ausdrücken in SentiWS zu finden ist

Find_NegativeSentiment(self)

#Prüfen, ob das Beschreibungswort unter den negativen Ausdrücken in SentiWS zu finden ist

Find_NoneSentiment(self)

#Hier werden auch diejenigen Beschreibungswörter gespeichert, welche nicht in SentiWS gefunden wurden

Persönliche Erklärung:

«Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und die den verwendeten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.»

Datum, Ort: Der Autor: