Preprint (unkorrigiert, vorläufiges Layout) von: Bubenhofer, Noah/Scharloth, Joachim (eingereicht): "Korpuspragmatische Methoden für kulturanalytische Fragestellungen", in: Linguistik als Kulturwissenschaft: Korpus - Kommunikation - Kultur, Giessen Contributions to the Study of Culture.

Korpuspragmatische Methoden für kulturanalytische Fragestellungen

Noah Bubenhofer & Joachim Scharloth

Während in vielen wissenschaftlichen Disziplinen die Analyse von großen Datenmengen inzwischen zum Standardrepertoire gehört, ist in den Geisteswissenschaften erst ein zögerlicher Methodenwandel erkennbar. Zwar werden gerade in den Sprachwissenschaften korpuslinguistische Methoden inzwischen für viele Fragestellungen eingesetzt, doch bei kulturanalytischen Forschungsprojekten liegt das Potenzial quantitativer Analysen noch weitgehend brach. Zudem verbergen sich hinter dem Label ‚Korpuslinguistik' eine Vielzahl von unterschiedlichen Paradigmen und Zugängen zu empirischen, sprachlichen Daten. So werden natürlich auch bei kulturanalytisch ausgerichteten Arbeiten mehrheitlich Korpora verwendet, um Belege für bestimmte Phänomene zu finden und zu dokumentieren. Weitgehend ungenutzt sind jedoch datengeleitete Verfahren unter Ausnutzung der zur Verfügung stehenden computerlinguistischen, informatischen und statistischen Technologie, die die Voraussetzung für einen ‚Data-driven Turn' in den Sozial- und Geisteswissenschaften sind.

Wir wollen im Folgenden aufzeigen, wie ein korpuspragmatischer, datengeleiteter Zugriff auf Textdaten hilft, kulturanalytische Fragestellungen zu bearbeiten. Dies erfolgt am Beispiel eines Korpus alpinistischer Literatur aus der Schweiz, dem Text+Berg-Korpus, das im Folgenden noch genauer beschrieben wird.

1. Korpuspragmatik: Der Data-driven Turn

Wir haben bereits dargelegt (Scharloth und Bubenhofer 2011), dass wir Forschung als korpuspragmatisch verstehen, die wie folgt charakterisiert ist: Die Korpuspragmatik deutet signifikant häufig auftretende sprachliche Muster in Korpora als Ergebnis rekurrenter Sprachhandlungen der Autorinnen und Autoren der im Korpus enthaltenen Texte bzw. der sie autorisierenden Institutionen und Gruppen. Sie geht davon aus, dass sich pragmatische Informationen "im pragmatischen Mehrwert oder Gebrauchswert von Einheiten aller sprachlicher Strukturbereiche" (Feilke 2000: 78) zeichenhaft manifestieren. Damit werden pragmatische Spuren an der sprachlichen Oberfläche, in die sich ein Gebrauchswert eingeschrieben hat, als Muster sichtbar (vgl. zu dieser Rehabilitierung der Textoberfläche auch den Sammelband von Feilke und Linke 2009). Diese "Sprachgebrauchsmuster" (Bubenhofer 2009) werden damit als Ergebnis von sprachlich-sozialem Handeln gelesen und gedeutet.

Attraktiv an dieser Orientierung an sprachlichen Mustern auf der Textoberfläche sind die methodischen Möglichkeiten, die sich dadurch ergeben. Dank der Verfügbarkeit von großen Korpora und der maschinellen Verarbeitung von Textdaten wird es möglich, quantitativ zu

arbeiten und Algorithmen zu entwickeln, um die Musterhaftigkeit in den Daten induktiv zu entdecken. Digitale Korpora sind somit nicht nur "Belegsammlungen oder Zettelkästen in elektronischer Form", sondern ermöglichen eine eigene "korpuslinguistische Perspektive" (Perkuhn und Belica 2006: 2).

Was ist mit dieser 'korpuslinguistischen Perspektive' gemeint? Die oben gemachten Ausführungen deuten es bereits an. Offensichtlich ist das quantitative Vorgehen: Nicht der Einzelbeleg interessiert, sondern die Musterhaftigkeit von Belegen, die als Ergebnis rekurrenter Sprachhandlungen gelesen werden können. Doch es kommt ein weiterer Aspekt hinzu: In Ergänzung zu den klassischen datenbasierten korpuslinguistischen Analysen, die der Überprüfung von Forschungshypothesen dienen ('corpus-based'-Paradigma), ist es fruchtbar, ein induktives Vorgehen der datengeleiteten Analyse ('corpus-driven'-Paradigma) hinzuzuziehen. Dieses Paradigma wird von Tognini-Bonelli (2001: 84ff.) vor dem Hintergrund der Arbeiten von Sinclair (1991) expliziert und im deutschen Sprachraum von mehreren Forscherinnen und Forschern verbreitet (Belica und Steyer 2006; Bubenhofer 2009; Perkuhn u. a. 2005; Steyer 2004; Scharloth u. a. im Druck; Bubenhofer und Scharloth 2013). Statt eine Hypothese mit vorher festgelegten Analysekategorien zu überprüfen, werden in einem Korpus sämtliche Zeichenkonfigurationen berechnet, die sich bei der Anwendung vorher festgelegter Algorithmen ergeben. Beispielsweise werden die Frequenzen von Wörtern, Mehrworteinheiten, aber auch abstrakteren Größen wie Frames, berechnet. Im Vergleich mit einem Referenzkorpus können nun die linguistischen Einheiten berechnet werden, die im Untersuchungskorpus besonders auffällig – also statistisch gesehen überzufällig häufig – sind. Damit geraten häufig Evidenzen in den Fokus, die entweder quer zu den vorher existierenden Erwartungen stehen und die Grundlage für neue Hypothesen sind, oder im besten Fall sogar solche Evidenzen, die die Bildung neuer interpretativer linguistischer Analysekategorien nahelegen.

2. Sprechen über Berge

Die Wahrnehmung der Berge und deren Funktion im kulturellen Haushalt spielen in der Schweiz eine bedeutende Rolle, so die generelle Hypothese der vorliegenden Arbeit. Texte, die vom Umgang des Menschen mit den Bergen handeln, widerspiegeln auch auf sprachlicher Ebene eine Mentalitätsgeschichte (Hermanns 1995; Linke 1996; Scharloth 2005) – und prägen diese mit. Eine Reihe von Arbeiten hat alpinistische Literatur bereits hinsichtlich dieser Hypothese gedeutet (Günther 1998; Stremlow 1998; Bätzing 2003), wobei bis anhin kaum linguistisch orientierte Zugangsweisen vorhanden sind. In alpinistischer Literatur müssten sich aber, so die Hypothese, auf sprachlicher Ebene Umbrüche beobachten lassen, die mentalitätsgeschichtlich gedeutet werden können. Diese Umbrüche zeigen sich auf verschiedenen Ebenen, wobei hier zwei in den Blick genommen werden sollen:

1. Wortebene: Zeitspezifisches Vokabular.
2. Phrasenebene: Zeitspezifische Mehrworteinheiten und syntagmatische Muster.

Das Korpus ‘Text + Berg‘ (Bubenhofer u. a. 2011) umfasst alle von 1864 bis 2009 erschienenen Ausgaben des Jahrbuchs bzw. der Zeitschrift des Schweizer Alpenclubs SAC und ist deshalb eine nützliche Grundlage, um Mentalitätsveränderungen über einen etwa 150 Jahre dauernden Zeitraum linguistisch zu erfassen. Von den Verfassern (in unterschiedlichen Konstellationen) liegen bereits einige Analysen und Beschreibungen des Korpus vor (Bubenhofer und Scharloth 2011; Bubenhofer und Schröter 2012; Bubenhofer 2013); in der vorliegenden Arbeit präsentieren wir Aspekte dieser Analysen, allerdings ergänzt um weiterführende methodische Vorschläge und Analysen zu Wortschatzveränderungen.

2.1 Das Text+Berg-Korpus

Der Schweizer Alpenclub (SAC) wurde 1863, kurz nach dem British Alpine Club gegründet und entwickelte sich, ähnlich wie die anderen Alpenvereine, von einem kleinen Zirkel von Bergsteigern zu einem großen Dienstleistungsverein für breite Schichten von Bergbegeisterten. Von 1864 bis heute gibt es die kontinuierliche Publikationsreihe ‘Jahrbuch des Schweizer Alpenclub‘, später ‘Die Alpen‘, in der über die Tätigkeiten der Clubmitglieder und andere Themen im Bereich Berge, Bergsport, Natur und Kultur im alpinen Raum berichtet wird. [1] Die Publikationsreihe ist für diachrone Untersuchungen besonders interessant, da sie lückenlos über knapp 150 Jahre besteht und bezüglich der vorkommenden Themen und Textsorten – etwa Tourenberichten, biographischen Artikeln, philosophischen Essays, Protokollen und Finanzberichten des Vereins – relativ homogen ist.

Das Gesamtkorpus umfasst 196 Bände, was einer Menge von 35,8 Mio. laufenden Wortformen entspricht. Die Publikation ist teilweise parallel in französischer Sprache erschienen. Das ganze Korpus steht in einer XML-Version zur Verfügung, in der Artikelgrenzen und Metainformationen erfasst und zudem Wortarten- und Lemma-Informationen annotiert sind.[2] Weiter wurden Namen von Personen, Bergen, Hütten und Gletschern mit maschinellen Methoden ausgezeichnet (Ebling u. a. 2011; Jitca u. a. 2011; Volk u. a. 2010b). Die Texte wurden mittels OCR-Software maschinell erkannt, was jedoch kein fehlerfreier Prozess ist. Die Texte wurden vollständig mit halbautomatischen Methoden (Volk u. a. 2010a, 2011) und teilweise zusätzlich manuell korrigiert, trotzdem ist mit Erkennungsfehlern zu rechnen. Sie sollten bei den großen Textmengen jedoch nicht ins Gewicht fallen.

Für die folgenden Analysen wurde nicht mit dem Gesamtkorpus gearbeitet, denn alle nicht-deutschen Texte mussten ignoriert werden. Zudem zeigte sich vor allem bei den Analysen auf der Wortebene, dass die Veränderung der Textzusammensetzung in den Bänden über die gesamte Periode unerwünschte Auswirkungen auf die Berechnung der Frequenzverläufe hat (dazu weiter unten). Deshalb erschien es sinnvoll, nur Texte mit einer Mindestlänge von 2000 Wörtern ins Untersuchungskorpus aufzunehmen. Daraus ergibt sich das in Tabelle 1 aufgeführte Korpus.

---

[1] Die Beiträge zum „Jahrbuch des Schweizer Alpenclub“ und den „Alpen“ betreffen nicht nur die Alpen, sondern auch andere Gebirge.

[2] Das Korpus wurde mit Hilfe des „TreeTaggers“ (Schmid 1994) und den standardmäßig verfügbaren deutschen Trainingsdaten, die auf dem STTS-Tagset (Schiller u. a. 1995) beruhen, annotiert.

Jahrzehnt	Wörter	Texte	Wörter/Text
1860	863136	130	5757.632244
1870	1100998	181	6070.823886
1880	1200639	213	5655.046676
1890	1700271	227	7594.176847
1900	1465082	185	7971.130431
1910	1338310	147	8310.672573
1920	1115393	178	6791.344292
1930	1242801	292	4286.711133
1940	997577	290	3420.388707
1950	1147438	282	4072.678278
1960	1204197	269	4501.058688
1970	1025434	272	3772.508064
1980	926693	194	4820.190173
1990	908918	236	3910.587373
2000	879946	270	3213.30627
Total	**17116833**	**3366**	**5343.217042**

Tabelle 1: Untersuchungskorpus 1864 bis 2009, Texte mit mindestens 2000 Wörtern; Zahlen gegliedert nach Jahrzehnten. In den Jahren 1870, 1915, 1924 erschienen keine Jahrbücher.

In Ergänzung zu den Analysen in diesem Untersuchungskorpus greifen wir auf Resultate vergangener Studien zurück (Bubenhofer und Scharloth 2011; Bubenhofer und Schröter 2012), in denen wir mit anderen Korpuszusammensetzungen arbeiteten, da die dort eingesetzten Methoden (Berechnung komplexer n-Gramme) die Bildung von Teilkorpora bedingen. In Tabelle 2 sind diese Teilkorpora mit den dazugehörigen Informationen aufgeführt.

Name	Periode	Wörter[3]	Texte
Korpus A[4]	1880 bis 1899	3.394.471	1021
Korpus B[4]	1930 bis 1949	2.795.149	1124
Korpus C	1960 bis 1979	2.607.881	870
Korpus D	1990 bis 2009	3.840.933	3891
Korpus komplett	1864 bis 2009	20.138.876	7385

Tabelle 2: Umfang der Korpora der in der Übersicht (der Studien Bubenhofer und Scharloth 2011; Bubenhofer und Schröter 2012)

[3] Die für Korpus A und B genannten Zahlen der laufenden Wortformen differieren leicht von den in Bubenhofer/Schröter (2012) genannten (A dort: 3.463.496; B dort: 2.915.548), da dort die (wenigen) nicht-deutschen Artikel nicht ausgesondert wurden. Um die Korpora aber mit den Korpora C und D vergleichen zu können, werden hier die Zahlen genannt, die sich ergeben, wenn die Korpora auf die gleiche Weise zusammengestellt werden wie die Korpora C und D.

[4] Korpus A und B wurden in Bubenhofer/Schröter (2012) verwendet.

Das Gesamtkorpus ist (nach Registrierung) für wissenschaftliche Zwecke frei verfügbar. Informationen dazu finden sich unter www.textberg.ch.

2.2 Methodik

Auf der Wortebene zielen die eingesetzten Analysemethoden darauf, zeitspezifisches von universell eingesetzten Lexemen zu unterscheiden. Die einfachste Möglichkeit ist, das für ein bestimmtes Teilkorpus im Vergleich zu einem anderen Teilkorpus oder dem Gesamtkorpus typische Vokabular zu berechnen. Dazu wird ein statistischer Signifikanztest verwendet, der die Abweichung der bei gleichmäßiger Verteilung erwarteten Häufigkeiten von den tatsächlich beobachteten Häufigkeiten bemisst. Je größer die Abweichung ist, desto signifikanter ist der Frequenzunterschied und desto typischer das Lexem für das jeweilige Teilkorpus.

Eine weitere Möglichkeit ist die Verwendung eines Maßes, um die Gleichmäßigkeit der Verteilung eines Lexems über das Korpus hinweg zu berechnen. Es existiert eine Vielzahl von solchen Verteilungsmaßen, wobei sich für linguistische Analysen Gries' ‚Deviation of Proportion' (DP) als besonders geeignet erwiesen hat (Gries 2008a, 2009). Das Maß dient vor allem dazu, stabiles von eher variablem Vokabular zu unterscheiden. Um mentalitätsgeschichtliche Umbrüche zu finden, ist das variable Vokabular von besonderem Interesse.

Weiter lassen sich die Frequenzreihen der Lexeme über die Zeitspanne des Korpus clustern, um Gruppen von Vokabular zu finden, das sich bezüglich der Verwendungshäufigkeiten ähnelt (Gries 2008b: 294ff.). Mit diesem Verfahren löst man sich von der Beobachtung einzelner Lexeme und kann Wortfelder finden, die über die Zeitspanne hinweg entweder eher variabel oder aber stabil verwendet werden.

In der vorliegenden Untersuchung kombinierten wir die oben skizzierten Verfahren wie folgt: In einem ersten Schritt wurden für jedes Jahr des Korpus die Frequenzen aller vorkommenden Lexeme der Wortarten Nomen, Adjektive, Verben und Personalpronomen ausgezählt.[5] Die Wahl der genannten Wortarten orientiert sich an Beobachtungen vergangener Analysen, bei denen sich diese Wortarten als besonders interessant erwiesen (Bubenhofer und Scharloth 2011; Bubenhofer und Schröter 2012). Die absoluten Frequenzen wurden in Relation zur jeweiligen Korpusgröße gesetzt und normalisiert, damit alle Werte zwischen 0 und 1 liegen und damit bezüglich ihrer Verlaufsform über die Zeitspanne vergleichbar werden, auch wenn die Frequenzen insgesamt unterschiedlich sind. Zudem wurden die Frequenzen eines Lexems mit der Berechnung eines gleitenden Durchschnittes (Fenstergröße fünf Jahre) geglättet, so dass starke Ausschläge in einem Jahr nicht so stark ins Gewicht fallen. Es wurden aber nur Lexeme berücksichtig, die im Durchschnitt über die ganze Zeitperiode hinweg häufiger als 30 Mal pro Mio. Wörter belegt sind. So ergibt sich eine Liste von 2113 Lexemen.

---

[5] Dafür wurde ein Script in der Programmiersprache Perl (Wall u. a. 2000) geschrieben.

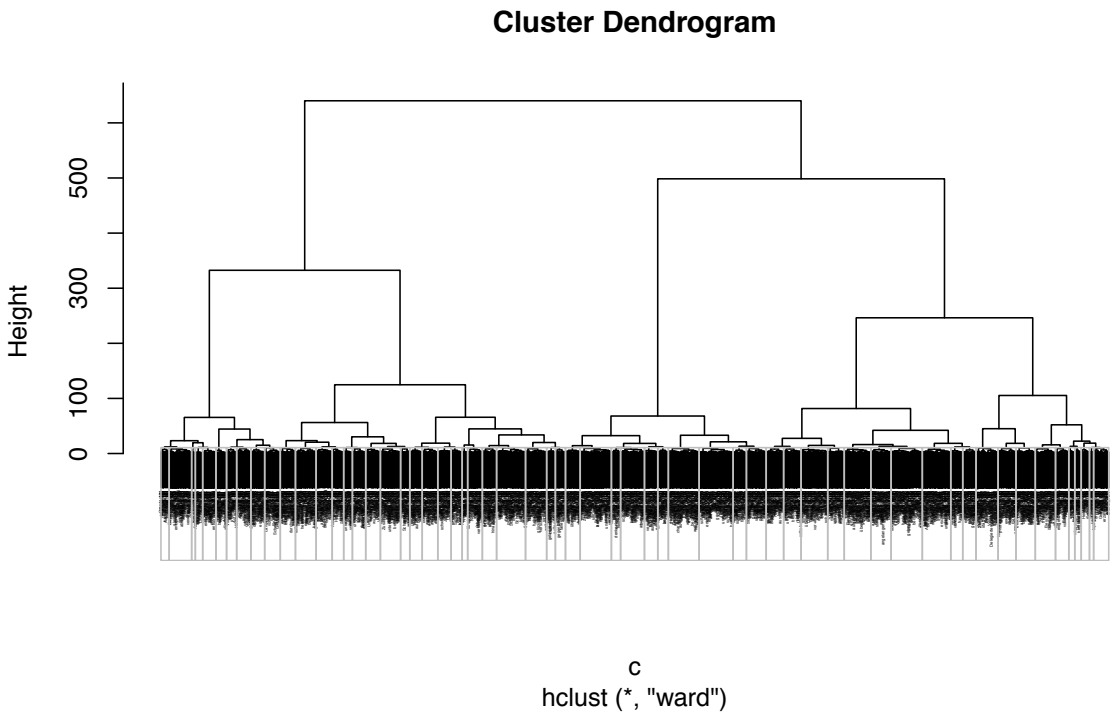

Abbildung 1: Dendrogramm der Clusteranalyse; Bildung von 60 Gruppen (graue Umrandungen auf der untersten Ebene)

Zusätzlich wurde die Anzahl der Lexeme weiter eingeschränkt, indem besonders gleichmäßig und extrem ungleichmäßig verteilte Lexeme ausgeschlossen wurden. Die Schwellenwerte wurden experimentell bei DP-Werten von 0,3 und 0,9 festgelegt; 2064 Lexeme liegen zwischen diesen Werten.

Anschließend wurde eine Clusteranalyse durchgeführt (hierarchisches Clustering, Methode ‚ward'), um die Lexeme zu finden, die sich bezüglich Verlauf über die Zeitspanne stark ähneln (Gries 2008b: 305).[6] Die gefundenen Cluster lassen sich in einem Dendrogramm visualisieren, wobei das Dendrogramm für die weitere Analyse auf der Höhe beschnitten wird, bei der die Differenzierung in die Cluster bereits genügend fein, aber noch nicht zu fein ist, so dass zu wenige Vertreter in einem Cluster vorhanden wären. Experimentell erwies sich ein Schnitt bei 60 Gruppen als sinnvoll (vgl. Abbildung 1).

Die Visualisierung der Verlaufskurven aller Lexeme einer Gruppe hilft dabei, die Verlaufskurven der Gruppen zu kategorisieren. In Abbildung 2 ist beispielsweise eine Gruppe von Lexemen visualisiert, die alle eine über den Zeitraum steigende Verlaufskurve aufweisen; in Abbildung 3 hingegen eine Gruppe mit fallender Verlaufskurve und in Abbildung 4 ist eine Verlaufskurve mit einem Ausschlag nach oben um 1900 herum dargestellt. Ein paar wenige Gruppen enthalten auch zu wenige Lexeme, um für die Analyse von Interesse zu sein oder andere sind zu diffus, um interpretiert werden zu können (vgl. Abbildung 5).

[6] Die folgenden Schritte wurden in R (Adler 2010) programmiert.

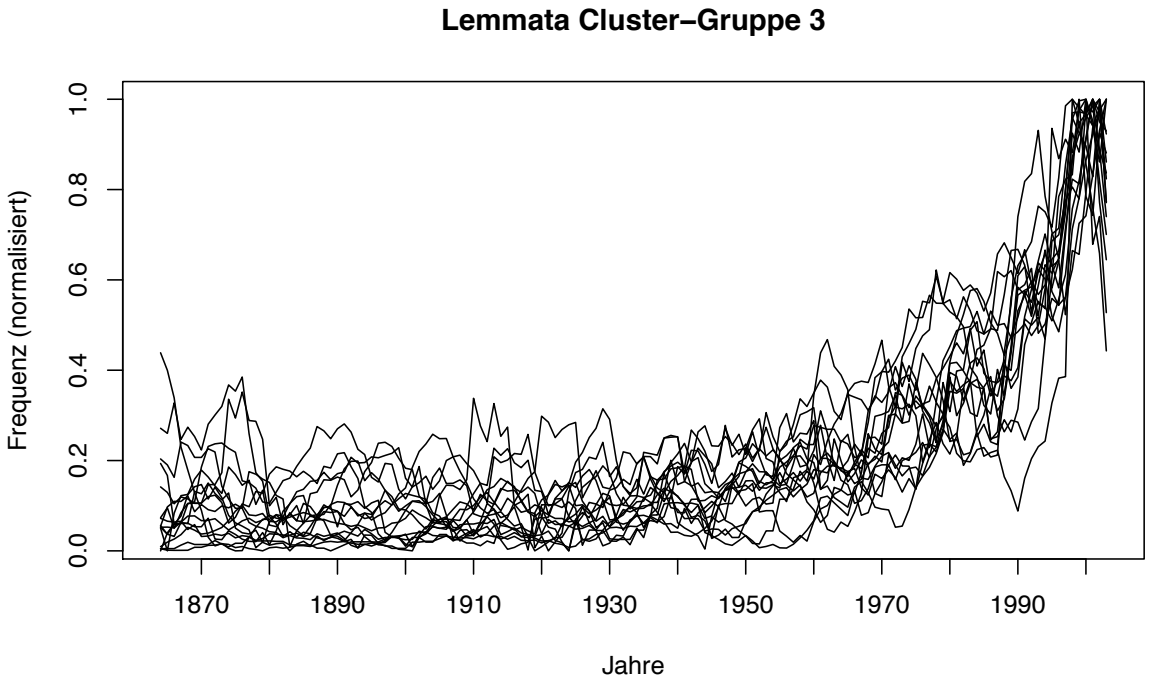

Abbildung 2: Beispiel für steigende Verlaufskurven

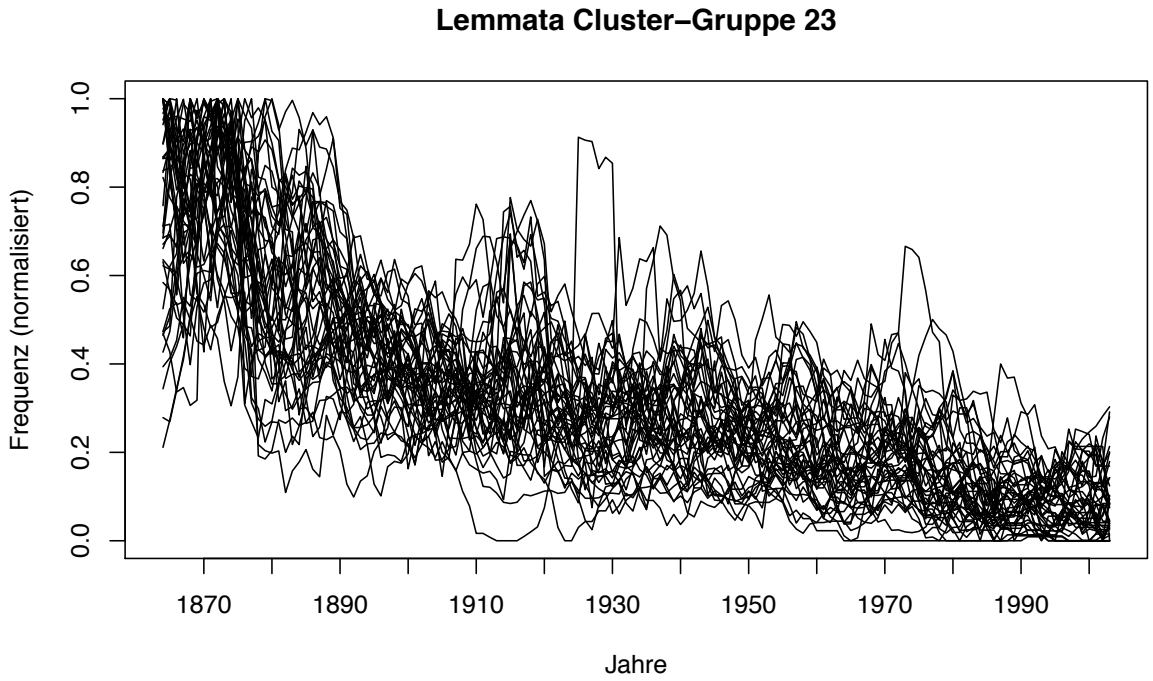

Abbildung 3: Beispiel für fallende Verlaufskurven (vgl. auch Abbildung 6)

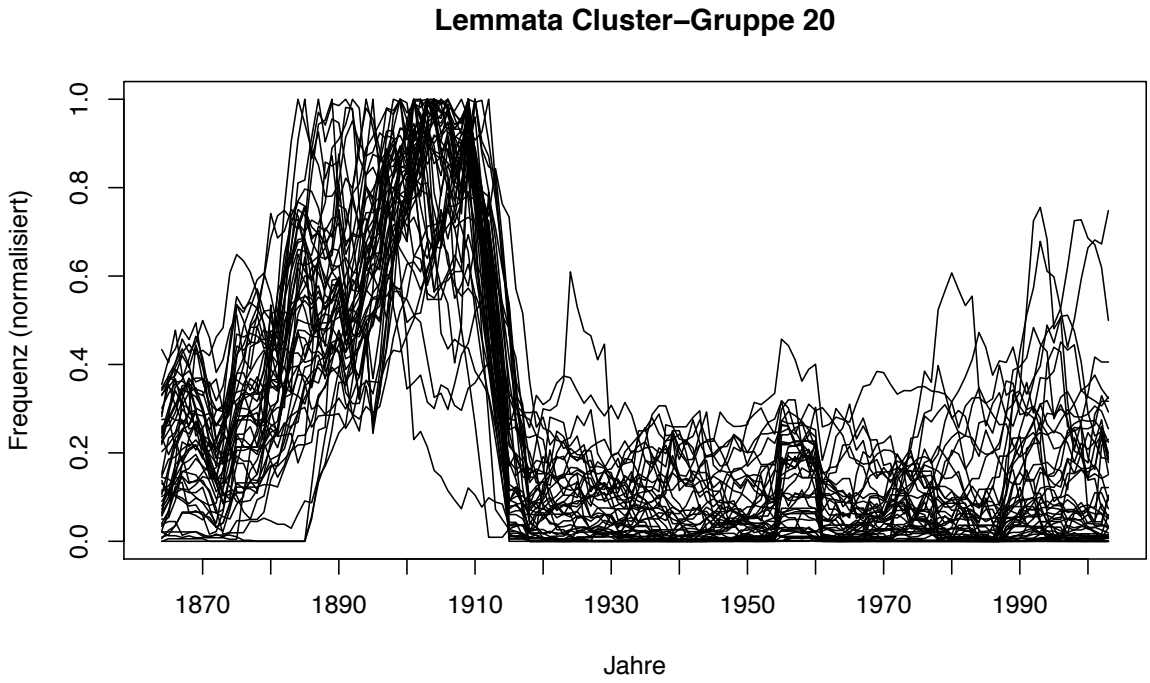

Abbildung 4: Beispiel für Verlaufskurven mit einem Ausschlag um 1900

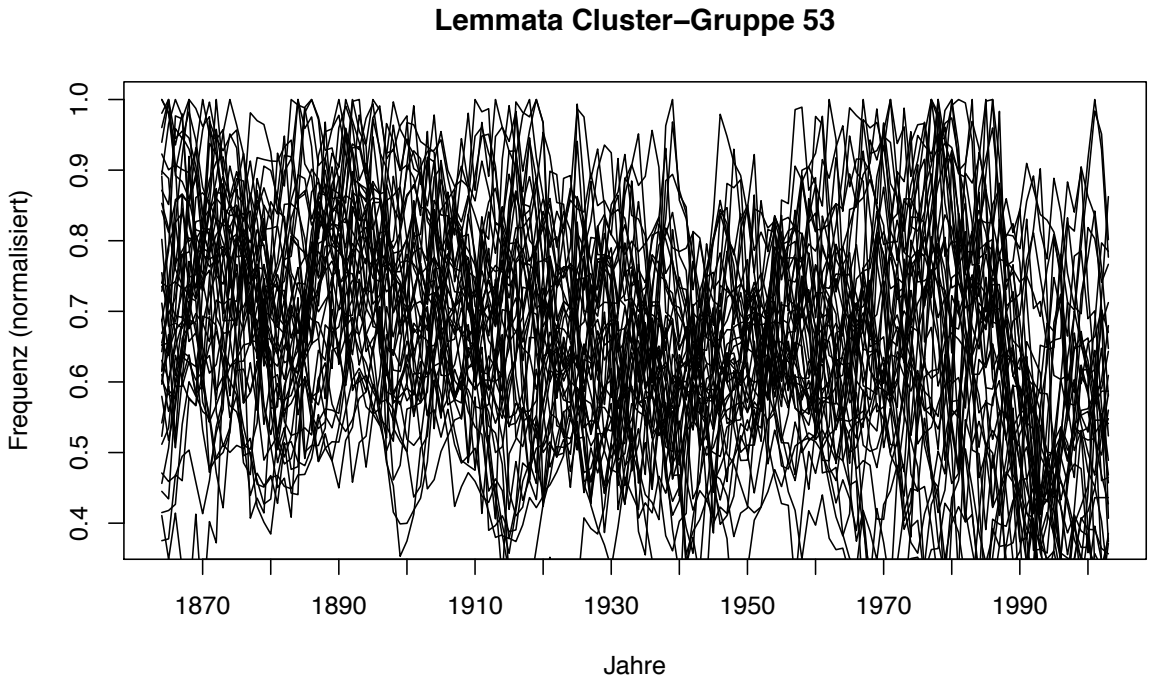

Abbildung 5: Beispiel für diffuse Verlaufskurven

Bei der Interpretation werden nun die Cluster der Verlaufskurven mit dem darin vertretenen Vokabular verglichen und gedeutet. Es stellen sich die folgenden Fragen:

1. Für welchen Zeitraum ist das Vokabular typisch?
2. Aus welchen Gründen ist das Vokabular für den Zeitraum typisch?

Ebenfalls als Interpretationshilfe wird das Vokabular der jeweiligen Gruppe als sog. ‚Wortwolke' visualisiert, wobei die Wortgröße die Gleichmäßigkeit der Verteilung als DP-Wert ausdrückt: Je größer das Wort abgebildet ist, desto ungleichmäßiger ist es im Korpus verteilt. Weiter werden die vier vorkommenden Wortartgruppen farblich markiert. Abbildung 6 zeigt das Vokabular von Cluster 23 (Abbildung 3), das also besonders für die Zeit von 1864 bis etwa 1890 typisch ist. Im Analyseteil gehen wir dann auf mögliche Interpretationen ein.

Wortwolke Gruppe 23

Abbildung 6: Wortwolke mit dem Vokabular von Cluster 23 (vgl. auch Abbildung 3)

Das oben skizzierte kombinierte Verfahren bietet Vorteile gegenüber der klassischen Methode, bei der im Vorab bestimmte Perioden bezüglich ihres typischen Vokabulars untersucht werden (angewandt in Bubenhofer und Scharloth 2011; Bubenhofer und Schröter 2012). Auch dort werden mit einem statistischen Verfahren die Lexeme berechnet, deren Frequenzen sich signifikant von den erwarteten Werten unterscheiden und deshalb für die jeweilige Periode im Vergleich zur anderen Periode typisch sind. Dies bedingt jedoch, dass bereits vorher bestimmt werden muss, welche Perioden miteinander verglichen werden sollen. Zudem prüft dieser Vergleich nur die Hypothese, ob die Lexeme in der einen Periode signifikant häufiger vorkommen, nicht jedoch, ob es über den gesamten Zeitverlauf Lexeme mit ähnlichen Frequenzverläufen gibt. Das kombinierte Verfahren entfernt sich damit auch vom Fokus auf Einzelwörter, da kovariierende Lexeme entdeckt werden können.

Trotzdem wäre eine solche Analyse, die die Wortebene fokussiert, einseitig. Ergänzend dazu sind deshalb Untersuchungen zu typischen Sprachgebrauchsmustern hilfreich (Bubenhofer 2009). Eine mögliche Operationalisierung von Sprachgebrauchsmustern ist die Berechnung von komplexen n-Grammen (Scharloth und Bubenhofer 2011). Komplexe n-Gramme sind eine erweiterte Form von n-Grammen (Bubenhofer 2009: 118; Manning und Schütze 2002: 192ff.), also Mehrworteinheiten, die nicht nur auf einer Folge von Wortformen bestehen, sondern auch aus einer Kombination von Wortformen und Wortart-Informationen bestehen können.[7] Während ein n-Gramm beispielsweise als Wortformenkette *so verbringen wir* definiert ist, werden bei der Berechnung von komplexen n-Grammen die Wortarten mit einbezogen, so dass eine Reihe von ähnlichen Wortformen-n-Grammen abstrakter als „*so* – finites Verb – Personalpronomen“ gefasst werden.[8]

Die Berechnung von für ein Untersuchungskorpus typischen n-Grammen folgt dem Paradigma datengeleiteter Analysen: Zunächst werden für das Untersuchungs- und das Referenzkorpus je alle möglichen Wortkombinationen einer bestimmten Länge *n* berechnet. Wie bereits oben für die Berechnung von typischem Vokabular kann nun mit einem statistischen Signifikanztest die Menge an n-Grammen extrahiert werden, die im Untersuchungskorpus signifikant häufiger als erwartet vorkommen. Im Fall der vorliegenden Untersuchung wurden Teilkorpora bestimmter Zeitperioden gebildet und die jeweils typischen komplexen n-Gramme im Vergleich zu jeweils einem Teilkorpus einer anderen Zeitperiode berechnet. Die große Menge an signifikanten n-Grammen wurde anschließend manuell gesichtet und kategorisiert.

---

7 Vgl. das ähnliche Konzept der ‘Collostructions’ von Stefanowitsch und Gries (2004).

8 Zur Berechnung der komplexen n-Gramme wurde ein von der Forschergruppe semtracks (www.semtracks.org) entwickeltes Programm „cwb-n-grams“ verwendet, das eine Erweiterung zur Open Corpus Workbench ist (Evert und The OCWB Development Team 2010).

3. Analysebeispiel: Veränderung des narrativen Stils

3.1 Wortebene

Von den 60 berechneten Clustern jeweils ähnlicher Frequenzverläufe von Lexemen erwiesen sich bei der Interpretation 30 unmittelbar als erhellend und plausibel deutbar. Die restlichen Gruppen enthielten teilweise zu wenige Lexeme, so dass sie weniger im Hinblick auf Mentalitäten interpretierbar sind, sondern im Zusammenhang von zu spezifischen Texteigenschaften (Veränderung der Terminologie, der Textsorte etc.) stehen. Teilweise enthielten diese Gruppen jedoch auch Lexeme, die im Korpus generell sehr gleichmäßig verteilt sind und deshalb für unsere Fragestellung weniger interessant sind.

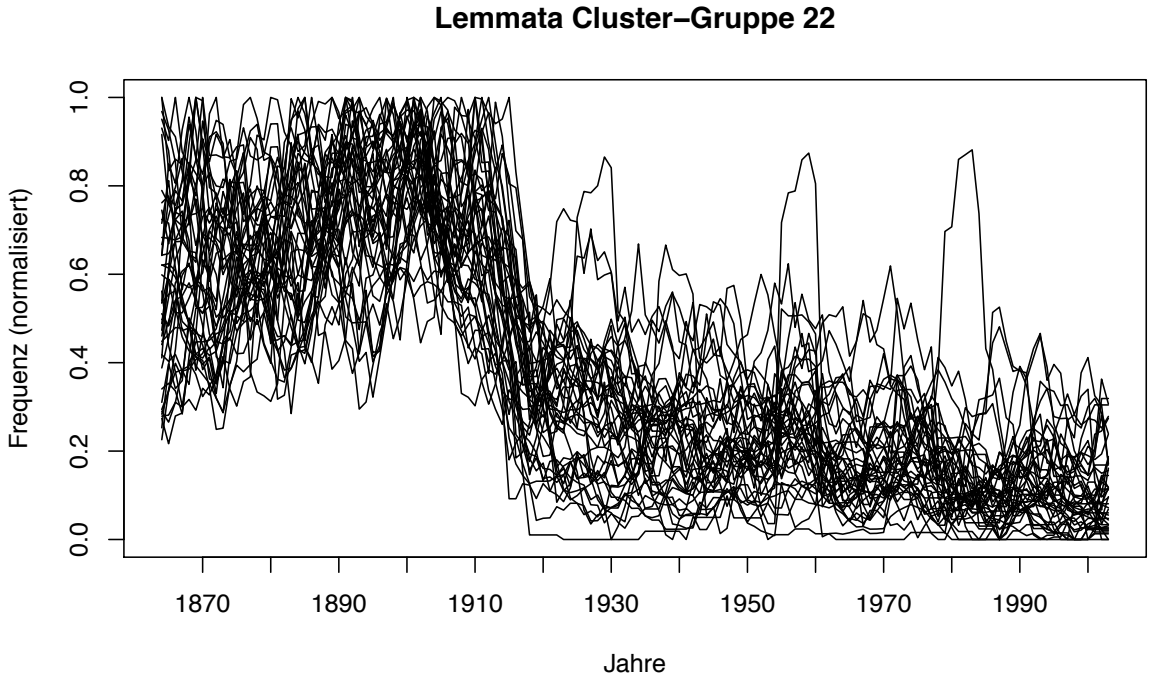

Abbildung 7: Cluster 22 mit Lexemen, die bis etwa 1910 häufig sind

Wortwolke Gruppe 22

Abbildung 8: Wortwolke der Lexeme von Cluster 22 (vgl. Abbildung 7), die bis etwa 1910 häufig sind

Unsere vergangenen Untersuchungen zeigten bereits, dass sich die Art der Narration in den alpinistischen Beschreibungen stark verändert (Bubenhofer und Scharloth 2011: 256ff.; Bubenhofer und Schröter 2012: 270ff.); in den 1930/40er- und den 1960/70er-Jahren identifizierten wir Merkmale eines deutlich expressiveren und persönlicheren Erzählstils als in den Texten davor und danach. Diese Beobachtungen lassen sich nun mit den weiteren Analysen noch differenzieren. Cluster 22 (Abbildung 7) vereinigt Lexeme, deren Frequenzen mehrheitlich bis in die 1910er-Jahre relativ hoch sind und dann deutlich sinken. Ähnlich verhalten sich vier weitere Cluster; insgesamt decken die fünf Cluster 256 Lexeme ab. In Abbildung 8 sind die Lexeme des Cluster 22 abgebildet. Auffällig sind Ortsangaben und Wegmarken (*Westseite, Punkt, Ostseite, östlich, Paßhöhe* etc.), evaluative Ausdrücke (*stattlich, vorzüglich, malerisch, hübsch*) und *Exkursion* und *Partie*, die Bezeichnungen für die Bewegungsformen in den Bergen ausdrücken. Abbildung 9 zeigt Cluster 24 mit Lexemen, deren Verlaufskurven über die gesamte Zeitperiode des Korpus kontinuierlich sinken. In dieser Gruppe befinden

sich textstrukturierende Lexeme, die typisch für Narration sind (*sodann, brechen … auf, gerade, indessen, hinan*), weitere Wegmarken (*Felswand, Alphütte, Dörfchen*) und landschaftsbeschreibendes Vokabular (*Kamm, Fluh, Aussicht, Schneide* etc.).

Wortwolke Gruppe 24

Abbildung 9: Wortwolke der Lexeme von Cluster 24, deren Frequenzen über die Gesamtperiode kontinuierlich sinken

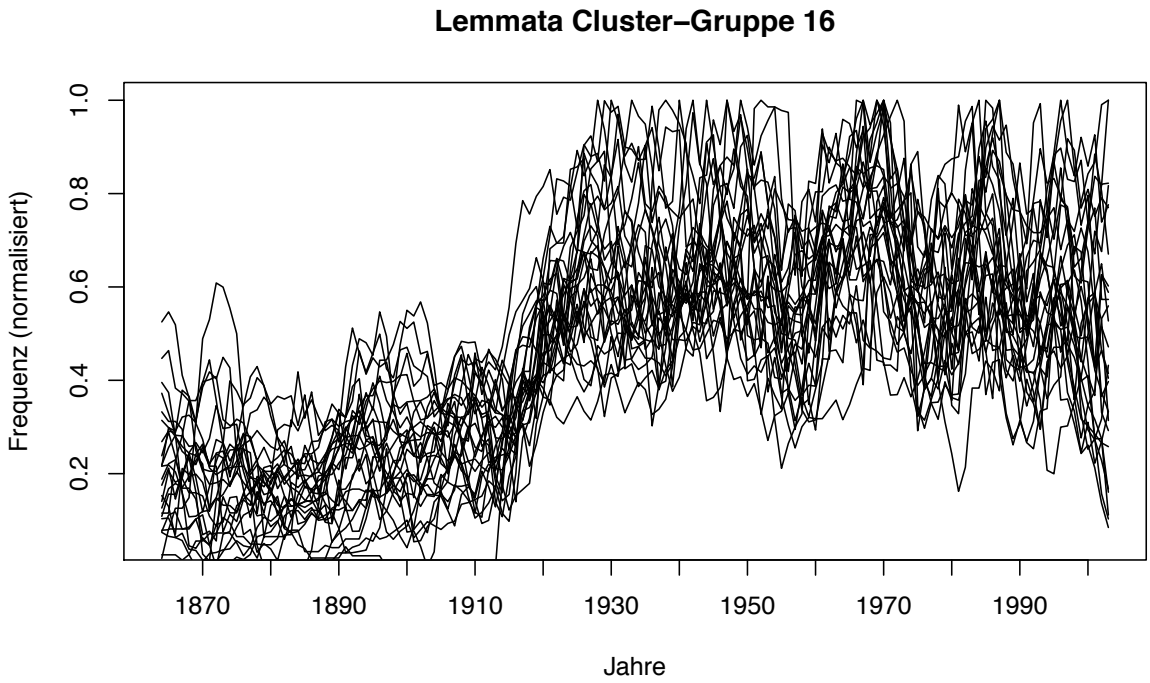

Abbildung 10: Cluster 16 mit Lexemen, die ab den 1920er-Jahren häufiger sind

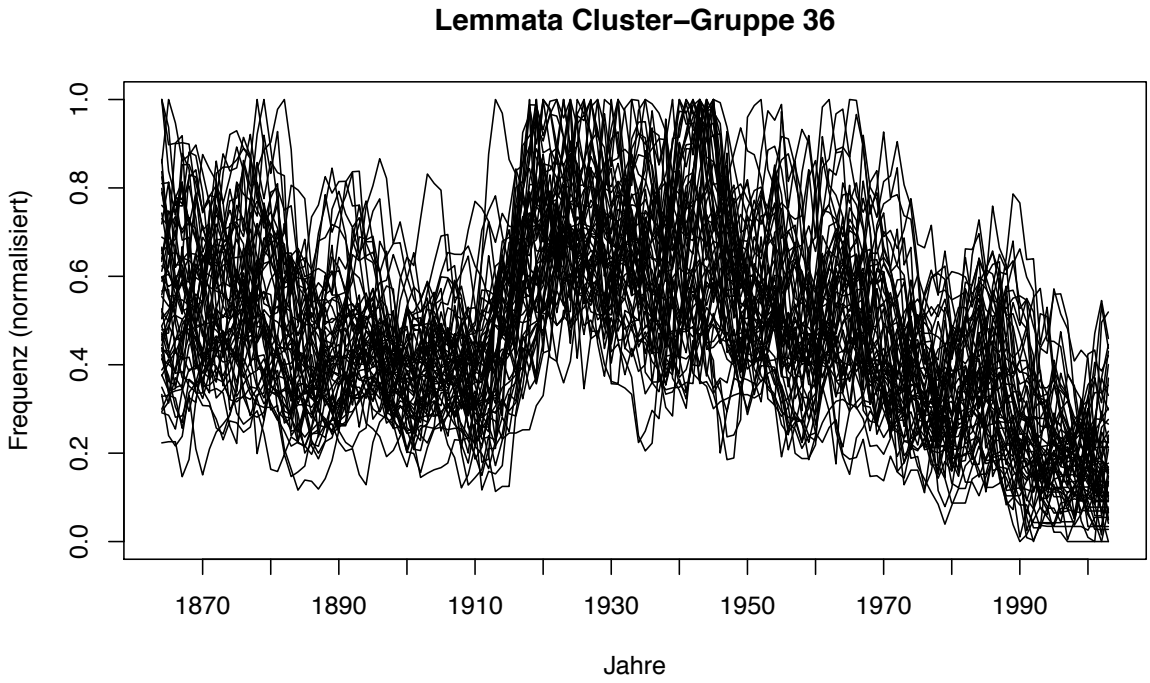

Abbildung 11: Cluster 36 mit Lexemen, die zwischen 1920 und 1950 hohe Frequenzen aufweisen (vgl. auch Abbildung 12)

In Kontrast zu diesen Clustern gibt es einige, bei denen die Frequenzen ab den 1920er-Jahren höher sind und es bis 2009 bleiben (z.B. Cluster 16, Abbildung 10) oder nur zwischen 1920 und etwa 1950er- bis teilweise 1970er-Jahre hohe Frequenzen aufweisen (z.B. Cluster 36, Abbildung 11). Ein wichtiger Grund für den Anstieg ab den 1920er-Jahren ist das immer populärer werdende Winter- und Skibergsteigen, das allerdings bereits ab der Jahrhundertwende praktiziert wird (Grupp 2008: 107f.). Interessanter sind jedoch häufiger auftretende Lexeme wie *Einsamkeit, lieben, Spannung, Einsamkeit* (Cluster 16, ohne Abbildung), *Schönheit, Lust, Körper* (Cluster 36, Abbildung 12), *Gedanke, Freude, Wunsch, Sinn, Gefährte, Freund, Ruhe* (Cluster 33, ohne Abbildung), *Liebe, Gott, Herz, Wesen, Geist* (Cluster 42, Abbildung 13) und *menschlich, Erlebnis, Gefühl, Sorge, empfinden, fühlen, Begeisterung* (Cluster 49, ohne Abbildung). Die meisten dieser Cluster (33, 36, 42, 49) ähneln sich in ihrem Verlauf und verzeichnen für die Zeit von etwa 1920 bis maximal 1970 höhere Frequenzen als davor und danach (vgl. Abbildung 14 für den Verlauf von Cluster 42), wobei Spitzen um 1925 und 1948 zu verzeichnen sind. Die erwähnten Lexeme stammen aus dem Bereich des Gefühls- und Seelenlebens und thematisieren Körperlichkeit (*Körper, Lust, Sinn, menschlich*).

Abbildung 12: Wortwolke der Lexeme von Cluster 36 (vgl. Abbildung 11), die zwischen 1920 und 1950 hohe Frequenzen aufweisen

Abbildung 13: Wortwolke der Lexeme von Cluster 42 (vgl. Abbildung 14), die zwischen 1920 und 1950/70 hohe Frequenzen aufweisen

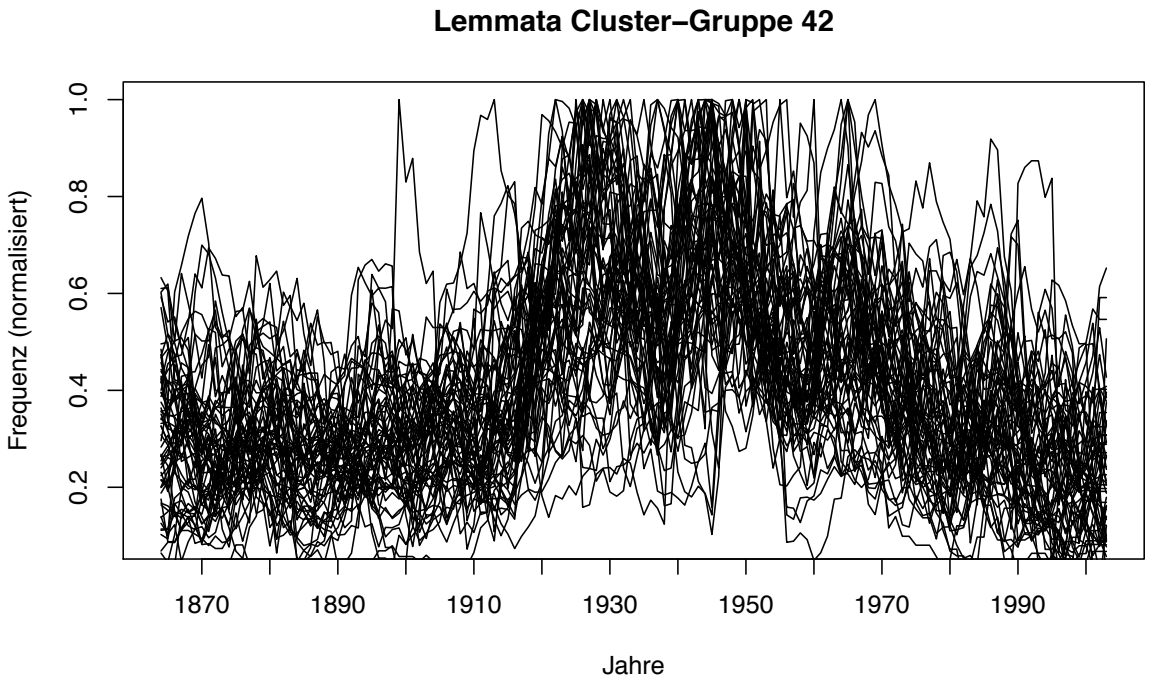

Abbildung 14: Cluster 42 mit Lexemen, die zwischen 1920 und 1950/70 hohe Frequenzen aufweisen (vgl. auch Abbildung 13)

Die Verteilung auf die Wortarten in den Zeitperioden 1870-1910 und 1920-1970 ist ebenfalls unterschiedlich (gemessen an den Clustern 15, 18, 22, 23, 24 bzw. 16, 33, 36, 42, 49): Der Anteil der Adjektive an den erfassten Wortarten (neben den Adjektiven: Nomen, Verben, Personalpronomen) ist in der älteren Zeitperiode höher (Mittelwert 41%, SD 7) als in der jüngeren Zeitperiode (Mittelwert 26%, SD 8).

3.2 Phrasenebene

Die Berechnung von komplexen n-Grammen ergänzt die oben gemachten Beobachtungen: In Bubenhofer/Schröter (2012: 273ff.) beschrieben wir eine Reihe von Mustern, die auf veränderte narrative Formen hinweisen, ebenso in Bubenhofer/Scharloth (2011: 251ff.), wobei wir in den Studien unterschiedliche Zeiträume abdeckten, wie Tabelle 2 in Kapitel 2.1 zeigt. Die vier Korpora decken also zu einem Teil die oben postulierten unterschiedlichen Perioden 1870-1910 und 1920-1970 ab und stellen zudem den Kontrast zu den neusten Daten dar.

Typische komplexe n-Gramme für Korpus A (1880–1899)	**Beispiele**[9]
ADJA[10] Stunde ADJA (NN)	halben Stunde angenehmer Steigung
	halbe Stunde steilen Anstieges
	halbe Stunde langem Zeitaufwand
	halben Stunde weiteren Weges
	halbe Stunde langer
ADV APPR CARD NN CARD	schon um 9 Uhr 30
ADV APPR CARD Uhr	circa um 1 Uhr 30
	selbst um 12 Uhr 10
	bereits um vier Uhr
	Endlich gegen 9 Uhr
ADV erreichten PPER ART NN	endlich erreichten wir den Aaresattel
	(Bald) nachher erreichten wir den Guggistafel
	Nun erreichten wir das Gebiet (des Kalkfelsens)
VVFIN APPR CARD Uhr	stiegen um 12 Uhr
	erreichten um 2 Uhr
	verließen um 3 Uhr
an der ADJA NN des	an der linken Seite des
	an der rechten Seite des
	an der anderen Seite des
	an der breiten Wand des
APPR ART Nähe ART NN	in der Nähe des Gipfels
	in der Nähe der Grenze
	in der Nähe des Muttensees
ADJA Weg APPR ART NN	alten Weg über den Feegletscher
	anderen Weg auf das Gabelhorn
	ausgetretenen Weg durch den Moränenschutt
APPR ART Führer NE NE	mit den Führern Alois Pinggera
	neben dem Führer Alphons Supersaxo
Dr. NE NE	Dr. Emil Burckhardt

9 Vgl. für Belege er Beispielrealisierungen das online zugängliche Korpus unter www.textberg.ch.

10 Vgl. für die Abkürzungen der Wortarten das STTS-Tagset (Schiller u. a. 1995).

APPR Prof. Dr. NE NE	von Prof. Dr. K. Schulz
	von Prof. Dr. G. Meyer

Tabelle 3: Auswahl komplexer n-Gramme, die für das Korpus A (1880–1899) typisch sind

Auffallend im älteren Korpus A sind Zeitangaben (vgl. Tabelle 3). Natürlich spielen Zeitangaben auch im jüngeren Korpus B eine Rolle, sie sind jedoch im Korpus A signifikant häufiger. Ebenfalls typisch sind Phrasen, die der Wegbeschreibung dienen und Wege charakterisieren können. Die für das ältere Korpus typische Bezeichnung *Führer* taucht in den n-Grammen auf und wird als Funktionsbezeichnung vor Namen verwendet. Weiter fällt eine Häufung von akademischen Titeln auf, die Personennamen (meist den Autorennamen) vorangestellt sind. Alleine der Titel *Dr.* wird in Korpus A 923 Mal verwendet, in Korpus B nur 185 Mal pro Mio. Wortformen.

Typische komplexe n-Gramme für Korpus B (1930–1949)	**Beispiele**
CARD NN ADJA NN	7700 m hohem Gipfel
	5400 m hohen Sasirpass
	3000 m hohe Nordwestflanke
ADV KOKOM CARD Meter	mehr als 1000 Meter
	mehr als hundert Meter
dann VVFIN ART NN	dann kündete die Gipfelglocke
dann VVFIN ART ADJA	dann geschieht das Wunder
dann VVFIN PPER auf	dann folgt ein heikler
	dann standen wir auf (dem kühnen Gipfel)
KOUS PPER APPR ART NN	Wie ich in den Riss einstieg
VVFIN	als wir in der Gabel anlangten
KOUS PPER ART NN VVFIN	Bevor wir in das Couloir hinübersteigen
	während wir der Hütte zustrebten
	während wir die Steigeisen ablegten
	Als wir die Passhöhe erreichten
ADV VVFIN ART ADJA NN .	Draussen erwachte ein neuer Tag .
	Dann kam ein trüber Tag .
	Nun naht das schwierigste Stück .
ADV VVFIN PPER VVINF	So lasst uns eilen
	jetzt heisst es handeln
Schönheit ART NN	Schönheit der/unserer Ber-
Schönheit PPOSAT NN	ge/Heimat/Alpen/Natur/Landschaft/
ADJA Schönheit	Landschaftliche/besondere/grenzenlose/
	verborgene/unendliche/wilde/natürliche/
	winterliche/einzigartige/alpine/
	überwältigende/eigenartige/unsagbare/grosse
	Schönheit
ADJA Bewunderung	ästhetischen/hohen/ehrfürchtigen Bewunderung
ADJA Empfinden	feinem/wahrsten/schweizerischen Empfinden
ADJA Enttäuschung	grosse/bittere/gewisser/schwere Enttäuschung

Tabelle 4: Auswahl komplexer n-Gramme, die für das Korpus B (1930–1949) typisch sind

Während für Korpus A Zahlen in Form von Uhrzeiten typisch waren, sind es in Korpus B Zahlen, die Höhen oder Höhenunterschiede angeben (vgl. Tabelle 4). Weiter fallen viele Muster auf, die ein subjektiveres und expressiveres Erzählen widerspiegeln. Eine weitere Klasse von Mustern formuliert subjektive Urteile und Affekte, wobei Lexeme erscheinen, die als

dem Wortschatz des Seelenlebens und der Gefühle zugehörig identifiziert werden können, wie *Schönheit, Bewunderung, Enttäuschung* etc. (Bubenhofer und Schröter 2012: 275).

Auch die 1960er und 1970er-Jahre sind durch einen expressiven, subjektiven Erzählstil geprägt: Auffällig im Vergleich zu den 1990-2009 ist z.B. das Muster ‚Adverb – finites Verb – Personalpronomen', das im Korpus hochfrequent ist (5383 Mal im älteren Teilkorpus) und im älteren Korpus hoch signifikant häufiger vorkommt als im neueren (LLR [Log Likelihood Ratio] 1667, $p < 0.00001$). Es steht für Realisierungen wie *so verbringen wir*, *so sehnten wir*, *dann stiegen wir* etc. (Bubenhofer und Scharloth 2011: 251ff.). Für eine ganze Reihe ähnlicher Muster ist zudem gemeinsam, dass alle Personalpronomen enthalten. Tabelle 5 stellt diese Muster mit den jeweiligen Belegen dar.

Typische komplexe n-Gramme für Korpus C (1960-1979)	**Beispiele**
ADV VVFIN PPER	So verbringen wir
	Dann stiegen wir
	Damals umschwirrten uns
,/$, ADV dass/KOUS PPER	so dass wir
	so dass ich
ADV VMFIN PPER	Zuerst mussten wir
	[Doch] inzwischen dürften sie
	Dort musste ich
	Jetzt müssen wir
./$. Als/KOUS PPER	. Als ich
	. Als wir

Tabelle 5: Auswahl komplexer n-Gramme die für das Korpus C (1960-1979) typisch sind

Viele weitere Muster, z.B. bestehend als Partizipien in Nebensätzen (*angelangt, geniessen wir; zugedeckt, versuchen wir*) oder Konnektoren und Präpositionen (*und wehmütig nahm ich; und verwaschen wölbte sich; Nach vierstündigem Marsch; Nach kurzer Beratung*), verweisen auf den typischen narrativen Stil dieser Zeitperiode (vgl. Bubenhofer und Scharloth 2011: 253f.).

Die narrativen Strukturen sind teilweise durch emotionale Ausrufe geprägt, wie Muster von Adjektiv-Nomen-Ellipsen zeigen oder am Satzanfang stehende Konnektoren (Tabelle 6).

Typische komplexe n-Gramme für Korpus C (1960-1979)	**Beispiele**
ADJA NN !/$.	[War das ein] wertvoller Fund!
	[Ein] wunderbares Wandergebiet!
	[Aber auch hier ein] schiefgestelltes Kreuz!
	[Eine] wunderbare Welt!
	[Ein] faszinierender Anblick!
	[Welch] andere Welt!
	[Was für] verlorene Zeit!
./$. ADV ,/$,	. Da, [ein Knall]
	. Nun, [für den Normalweg reicht es allemal]
	. Jetzt, [da sich]

Tabelle 6: Auswahl komplexer n-Gramme die für das Korpus C (1960-1979) typisch sind

Die präsentierten komplexen n-Gramme sind nur ein Teil der deutenswerten Muster. Zudem ist es lohnenswert, noch weitere Phänomene in den Daten in den Blick zu nehmen. Dazu gehören beispielsweise grammatische Merkmale: In Bubenhofer/Scharloth (2011: 247) konnten wir zeigen, dass zwischen etwa 1930 und 1980 weniger Passivformen verwendet werden als davor und danach. Der Häufigkeitsunterschied des durchschnittlichen Passivkoeffizienten zwischen 1960-1979 und 1990-2009 ist hochsignifikant.[11] Auch die Verwendung von sog. Intensivierern des „absoluten" und „extrem hohen Intensivierungsbereichs" (Os 1989: 134ff.) wie *absolut, gänzlich, komplett* bzw. *höchst, äußerst, zutiefst* etc. streut zwar über das Gesamtkorpus relativ stark, nimmt in der Tendenz über die gesamte Zeitperiode allerdings ab (1960-1979 gibt es signifikant mehr Intensivierer als 1990-2009)[12] (Bubenhofer und Scharloth 2011: 247).

4. Fazit

Wenn die verschiedenen Beobachtungen oben zusammengefasst werden, wird deutlich, dass sich das 'Sprechen über Berge' in den SAC-Jahrbüchern von 1864 bis 2009 stark verändert hat (vgl. ausführlich dazu Bubenhofer und Scharloth 2011: 256ff.). Um die bisher abstrakten Analysen etwas zu veranschaulichen, seien an dieser Stelle Ausschnitte aus Beispieltexten unterschiedlicher Epochen zitiert:

> 1) Wir hielten uns möglichst hoch auf der nördlichen, nach Süden fallenden Seite von Vaplona, gewannen das kleine Tobel, das gegen den Punkt 2547 m ansteigt, wateten durch dasselbe hinaus und erreichten den genannten Punkt um 8 Uhr 45 Min. Hier sahen wir etwa 200 m unter uns den Schottensee mit Schnee und Eis bedeckt. *Schottenseefurke* wäre vielleicht, der *Wildseefurke* (2515 m) entsprechend, der passendste Name für diese Scharte zwischen den Punkten 2647 m und 2650 m. Wir blieben nicht lange hier, sondern stiegen bald wieder links auf, theils über Schnee, theils über Verrucanotrümmer, und betraten den Gipfelpunkt 2650 m um 9 Uhr 15 Min. Es ging ein schwacher Windzug, und das Thermometer zeigte auf -2 C.
>
> *Jahrbuch 1888-1889: Eine Sectionsfahrt auf den Piz Sol (J. J. Schiesser)*
>
> 2) [N]ochmal wird angegriffen. Und endlich gelingt es mir, für die linke Hand ganz oben einen guten Griff zu schaffen. Die Rechte bohrt sich mit dem Eisbeil in der ersehnten Rampe ein Loch, der rechte Fuss steigt nochmal nach, und ich sehe über den Rand, indessen der Körper schwer nach aussen hängt. Sich ganz auf die rechte Hand verlassend, fährt die linke blitzschnell weit über den Rand in den Firn, ein Ruck, und ich liege verschnaufend auf dem Bauch, während die Füsse in der Luft baumeln.
>
> *Die Alpen 1932: Piz Bernina-Nordostflanke (Karl Schneider)*
>
> 3) Und da der Tag gut ausgefüllt war, beschliessen wir zu biwakieren. Schon summt der Kocher, und alles wäre in Ordnung, wenn sich nicht über dem Brenva-Becken schwere, schwarze Wolken entleeren würden. Wir sind auf der Höhe der Poire. Der ganze Mont-

---

[11] t=-13,2002, df=1997,302, p<2,2e-16

[12] t=9,8323, df=2258,657, p < 2,2e-16

Blanc-Gipfel steckt in den Wolken, während die Sicht nach unten gut ist. Da erleuchtet ein Blitz den ganzen Berg. Es ist Nacht geworden, aber eine Stunde lang zuckt ein Blitz nach dem andern. Mit einer Zeltplane zugedeckt, versuchen wir, uns an den Berg zu kleben. Bei jedem Blitzstrahl halten wir einen Augenblick den Atem an. Leise beginnt es zu schneien. Hie und da sehen wir Lampen in der Wand. Eine Seilschaft klettert in den Ausläufern der Major-Route trotz schlechtem Wetter und Lawinengefahr. Wir wissen, was wir an ihrer Stelle tun würden! Für uns gibt es keinen anderen Ausstieg als den Gipfel des Pilier. Dann werden wir weiter sehen.

SAC-Jahrbuch 1975: Pilier d'Angle: Bonatti/Gobbi-Führe (Michel Vaucher)

4) Bei P. 2508 den Weg direkt nordwärts einschlagen, der Hirschplatten quert und den Tällistock umrundet. Achtung, die Tunnels von Hirschplatten können Anfang Saison noch mit Schnee gefüllt sein. In diesem Fall muss man absteigen, denn sie können nicht umgangen werden. Rechts ist der Muttgletscher zu bewundern, der die Rhone speist. Weiter auf dem Höhenweg bis in den Furkapass. Bis Anfang des 20. Jahrhunderts war die Furka der längste befahrbare Pass der Schweiz. Vom Pass können Sie, wenn die Füsse nicht mehr tragen, ins Postauto steigen und gemütlich nach Andermatt hinunterfahren.

Die Alpen 2006: Zu Fuss um den Pizzo Rotondo. Gebirgige Wasserlandschaft (Vincent Theler)

Sowohl in den 1930er/40er als auch in den 1960er/70er werden Geschichten aus sehr persönlichen Erzählperspektiven heraus erzählt. Vokabular des Gefühls- und Seelenlebens wird in dieser Zeit häufiger verwendet als davor und danach und Körperlichkeit sowie Sinneswahrnehmungen werden stärker thematisiert. Die häufige Verwendung von Intensivierern in den 1930er/40er-Jahren könnte als Marker für Emotionalität gedeutet werden.

Dies steht im Kontrast sowohl zu den Daten aus jüngster Zeit (1990 bis 2009) als auch zu den ältesten analysierten Daten (1870er bis 1890er). Erstere widerspiegeln den Professionalisierungsschub der SAC-Publikation, wo heute, wie in vielen journalistischen Erzeugnissen, distanziertere Berichte ohne Involvierung des Erzähler-Ichs und mit systematischen Info-Zusammenstellungen als Serviceangebot für die Leserschaft Verbreitung finden. Damit einher geht eine häufigere Verwendung von Passivformen.

Die ältesten untersuchten Texte lesen sich dagegen wie Laborberichte, in denen von wissenschaftlichen Experimenten, nämlich dem Besteigen von Bergen, die Rede ist (Bubenhofer und Schröter 2012).

In einem weiteren Kontext könnte man diese Befunde als Zeichen dafür interpretieren, dass sich die Motivation, sich in den Bergen zu bewegen, verändert hat. Während im ausgehenden 19. Jahrhundert der wissenschaftliche Entdeckergeist vorherrschte, mit dem in sachlicher und systematischer Weise die Bergwelt entdeckt wurde, steht ab den 1930er-Jahren der Mensch im Vordergrund: Es ist seine innere Welt, die in Reaktion zum äußeren Erleben im Vordergrund steht.[13] Kameradschaft und Gemeinschaft scheinen besonders wichtige Werte zu sein, die als wichtige Tugenden des Bergsteigens konstruiert werden. Dieses Wir-

[13] Günther (1998: 13, 161), die sich in ihrer Studie des „bürgerlichen Alpinismus (1870-1930)“ stark auf die Schriften des Deutschen und Österreichischen Alpenvereins (DÖAV) stützt, spricht dementsprechend davon, dass „der Entdeckeralpinismus [um die Jahrhundertwende] vom Schwierigkeitsalpinismus abgelöst wird“.

Gefühl in Kombination mit der Verherrlichung der Natur in religiös-romantischer Weise könnte besonders während der nationalsozialistischen Zeit als Form des Eskapismus – aber natürlich auch der Geistigen Landesverteidigung (Jorio 2008) – gedeutet werden.

In neuster Zeit ab den 1990er-Jahren zeigt sich ein utilitaristischer Zugang zu den Bergen, bei dem diese als Objekte der Freizeitgestaltung und des sportlichen Wettkampfs betrachtet werden und daher Tipps und Vorschläge gefragt sind, wie die Freizeit am besten gestaltet werden kann. Vergleichbar damit beobachtet Stremlow (1998: 278f., 265), in "Magazinen einzelner Aktivsportarten" des späten 20. Jahrhunderts, dass die "Landschaft der Alpen" dort "zur Steigerung des Ich-Erlebnisses funktionalisiert" werde. Dass die "postmoderne Freizeitgesellschaft" die Alpen als "Sportgerät zur Auslösung von körperlichen Erlebnissen" auffasst, bemerkt auch Bätzing (2003: 19).

Diese Hypothesen decken sich teilweise mit anderen Forschungsergebnissen und Einschätzungen, wie oben verwiesen. Aus sprachwissenschaftlicher Sicht ist es jedoch interessant, die Hypothesen linguistisch fassen zu können: Die Veränderungen des sozialen Handelns findet einen Niederschlag in den untersuchten Texten und die Texte konstituieren gleichzeitig Deutungsmuster, mit denen soziales Handeln bewertet wird. Gleichwohl ist klar, dass die gewonnenen Hypothesen wiederum durch weitere Analysen validiert werden müssen. Methodisch gesehen bietet es sich deshalb auch an, das bisher eher datengeleitete in der Folge um ein hypothesengeleitetes Vorgehen zu ergänzen.

Methodisch hoffen wir gezeigt zu haben, dass ein datengeleiteter, korpuslinguistischer Zugang linguistische Phänomene zutage fördert, die pragmatisch gedeutet werden können. Die Clusteranalysen der Frequenzverläufe von ungleichmäßig verteiltem Vokabular im Korpus führte zudem zu ermutigenden ersten Möglichkeiten, datengeleitet Trends und Umbrüche in den Daten zu entdecken. Die eingesetzten statistischen Verfahren sind vergleichsweise trivial – das Potenzial wäre aber enorm: Gries (2010: 269) stellt ebenso fest, dass die Korpuslinguistik zwar meistens quantitativ arbeitet, statistische Möglichkeiten der Datenanalyse jedoch nur zögerlich und unvollständig einsetzt, obwohl dies bei der quantitativen Ausrichtung naheliegen würde.

Wir sind fern davon, den idealen Methodenapparat für kulturanalytische Forschungsfragen entwickelt zu haben. Das Entwickeln von Messmethoden unter Einbezug der ganzen Bandbreite von statistischem, computerlinguistischem und sprachwissenschaftlichem Knowhow dürfte aber ein lohnenswertes Unterfangen sein, um zu neuen Erkenntnissen zu Mentalitäten und kulturellen Umbrüchen zu gelangen.

Bibliographische Angaben

Adler, Joseph. 2010. *R in a Nutshell*. 1. Aufl. O'Reilly.

Bätzing, Werner. 2003. *Die Alpen: Geschichte und Zukunft einer europäischen Kulturlandschaft*. C.H.Beck.

Belica, Cyril/Steyer, Kathrin. 2006. *Korpusanalytische Zugänge zu sprachlichem Usus*. In: *AUC (Acta Universitatis Carolinae), Germanistica Pragensia*. XX .

Bubenhofer, Noah. 2013. *Quantitativ informierte qualitative Diskursanalyse. Korpuslinguistische Zugänge zu Einzeltexten und Serien*. In: Roth, Kersten Sven/Spiegel, Carmen (Hrsg.) *Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven*. Berlin: Akademie-Verlag (Diskursmuster - Discourse Patterns). 109–134.

Bubenhofer, Noah. 2009. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin, New York: de Gruyter (Sprache und Wissen).

Bubenhofer, Noah/Scharloth, Joachim. 2013. *Korpuslinguistische Diskursanalyse: Der Nutzen empirisch-quantitativer Verfahren*. In: Warnke, Ingo/Meinhof, Ulrike/Reisigl, Martin (Hrsg.) *Diskurslinguistik im Spannungsfeld von Deskription und Kritik*. Berlin: Akademie-Verlag (Diskursmuster – Discourse Patterns). 147–168.

Bubenhofer, Noah/Scharloth, Joachim. 2011. *Korpuspragmatische Analysen alpinistischer Literatur*. In: Elmiger, Daniel/Kamber, Alain (Hrsg.) *La linguistique de corpus – de l'analyse quantitative à l'interpretation qualitative / Korpuslinguistik – von der quantitativen Analyse zur qualitativen Interpretatino*. Neuchâtel: Institut des sciences du langage et de la communication (Travaux neuchâtelois de linguistique). 241–259.

Bubenhofer, Noah/Schröter, Juliane. 2012. *Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalye*. In: Maitz, Péter (Hrsg.) *Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate*. Berlin/Boston: de Gruyter (Studia Linguistica Germanica). 263–287.

Bubenhofer, Noah/Volk, Martin/Althaus, Adrian u. a. (Hrsg.). 2011. *Text+Berg-Korpus (Release 145)*. Institut für Computerlinguistik, Universität Zürich.

Ebling, Sarah/Sennrich, Rico/Klaper, David u. a.. 2011. *Digging for names in the mountains: Combined person name recognition and reference resolution for German alpine texts*. In: *5th Language & Technology Conference*.

Evert, Stefan/The OCWB Development Team. 2010. *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*.

Feilke, Helmuth. 2000. *Die pragmatische Wende in der Textlinguistik*. In: Brinker, Klaus (Hrsg.) *Text- und Gesprächslinguistik/Linguistics of Text and Conversation*. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science). 64–82.

Feilke, Helmuth/Linke, Angelika (Hrsg.). 2009. *Oberfläche und Performanz. Untersuchungen zur Sprache als dynamische Gestalt*. Berlin, New York: de Gruyter.

Gries, Stefan Thomas. 2008a. *Dispersions and adjusted frequencies in corpora*. In: *International Journal of Corpus Linguistics*. 13 . 403–437(35).

Gries, Stefan Thomas. 2009. *Dispersions and adjusted frequencies in corpora: further explorations*. In: *Language and Computers*. 71 (1). 197–212.

Gries, Stefan Thomas. 2008b. *Statistik für Sprachwissenschaftler*. Göttingen: Vandenhoeck & Ruprecht (Studienbücher zur Linguistik).

Gries, Stefan Thomas. 2010. *Useful statistics for corpus linguistics*. In: *A mosaic of corpus linguistics Selected approaches*. S. 269–291.

Gries, Stefan Thomas/Stefanowitsch, Anatol. 2004. *Extending collostructional analysis*. In: *International Journal of Corpus Linguistics*. 9 (1). 97–129.

Grupp, Peter. 2008. *Faszination Berg die Geschichte des Alpinismus*. Köln: Böhlau.

Günther, Dagmar. 1998. *Alpine Quergänge: Kulturgeschichte des bürgerlichen Alpinismus (1870 - 1930)*. Campus Verlag.

Hermanns, Fritz. 1995. *Sprachgeschichte als Mentalitätsgeschichte. Überlegungen zu Sinn und Form und Gegenstand historischer Semantik*. In: Gardt, Andreas/Mattheier, Klaus/Reichmann, Oskar (Hrsg.) *Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien*. Tübingen: Niemeyer S. 69–101.

Jitca, Magdalena/Sennrich, Rico/Volk, Martin. 2011. *From historic books to annotated XML: Building a large multilingual diachronic corpus*. In: *Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*. Universität Hamburg (Arbeiten zur Mehrsprachigkeit, Folge B. Working Papers in Multilingualism, Series B). 75–80.

Jorio, Marco. 2008. *Geistiges Landesverteidigung*. In: *Historisches Lexikon der Schweiz (HLS)*.

Linke, Angelika. 1996. *Sprachkultur und Bürgertum: zur Mentalitätsgeschichte des 19. Jahrhunderts*. Stuttgart: Metzler.

Manning, Christopher D/Schütze, Hinrich. 2002. *Foundations of Statistical Natural Language Processing*. 5. Aufl. Cambridge, Massachusetts: The MIT Press.

Os, Charles van. 1989. *Aspekte der Intensivierung im Deutschen*. Tübingen: Narr (Studien zur deutschen Grammatik).

Perkuhn, Rainer/Belica, Cyril. 2006. *Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik*. In: *Sprachreport*. 22 (1). 2–8.

Perkuhn, Rainer/Belica, Cyril/al-Wadi, Doris u. a.. 2005. *Korpustechnologie am Institut für Deutsche Sprache*. In: Schwitalla, Johannes/Wegstein, Werner (Hrsg.) *Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003*. Tübingen: Niemeyer S. 57–70.

Scharloth, Joachim. 2005. *Sprachnormen und Mentalitäten. Sprachbewusstseinsgeschichte in Deutschland im Zeitraum von 1766 bis 1785*. Tübingen: Niemeyer (Reihe Germanistische Linguistik).

Scharloth, Joachim/Bubenhofer, Noah. 2011. *Datengeleitete Korpuspragmatik: Korpusvergleich als Methode der Stilanalyse*. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hrsg.) *Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen von Texten und Gesprächen*. Berlin, New York: de Gruyter S. 195–230.

Scharloth, Joachim/Eugster, David/Bubenhofer, Noah (im Druck): *Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn*. In: Busse, Dietrich/Teubert, Wolfgang (Hrsg.) *Linguistische Diskursanalyse. Neue Perspektiven*. Wiesbaden: VS Verlag.

Schiller, Anne/Teufel, Simone/Thielen, Christine. 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Stuttgart.

Schmid, Helmut. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Steyer, Kathrin. 2004. *Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven*. In: Steyer, Kathrin (Hrsg.) *Wortverbindungen – mehr oder weniger fest*. Berlin, New York: de Gruyter (Institut für Deutsche Sprache. Jahrbuch 2003). 87–116.

Stremlow, Matthias. 1998. *Die Alpen aus der Untersicht: von der Verheissung der nahen Fremde zur Sportarena : Kontinuität und Wandel von Alpenbildern seit 1700*. Haupt.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: Benjamins (Studies in Corpus linguistics).

Volk, Martin/Marek, Torsten/Sennrich, Rico (2010a): *Reducing OCR errors by combining two OCR systems*. In: *ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*. S. 61–65.

Volk, Martin/Furrer, Lenz/Sennrich, Rico. 2011. *Strategies for reducing and correcting OCR error*. In: Sporleder, Caroline/Bosch, Antal van den/Zervanou, Kalliopi (Hrsg.) *Language Technology for Cultural Heritage*. Berlin: Springer (Theory and Applications of Natural Language Processing). 3–22.

Volk, Martin/Bubenhofer/Althaus, Adrian/Bangerter, Maya u. a. (2010b): *Challenges in building a multilingual alpine heritage corpus*. In: *Seventh International Conference on Language Resources and Evaluation (LREC), Malta, 19 May 2010 – 21 May 2010*.

Wall, Larry/Christiansen, Tom/Orwant, Jon. 2000. *Programming Perl*. O'Reilly.