# Der Europabegriff auf außereuropäischen Webseiten.

Ein Vergleich des semantischen Kontextes ausgewählter Domains mithilfe rechnergestützter Textanalysemethoden.

**Diplomarbeit**

**im Diplomstudiengang Geographie**

**in der Fakultät Geistes- und Kulturwissenschaften**

**der Otto-Friedrich-Universität Bamberg**


| | |
|---|---|
| Verfasser: | Dominik Kremer |
| Erstgutachter: | Prof. Dr. Boris Braun |
| Zweitgutachter: | Prof. Dr. Christoph Schlieder |

„...wenn man in einem Denksystem an einem Begriff zieht,
dann kommt alles nach, weil zwischen allen Teilen Fäden gespannt sind.“

Georges Dumézil

# Inhaltsverzeichnis

# 1 Europa liegt (auch) im Web

Wo liegt Europa? Wie selbstverständlich wandert der Finger zur großen Wandkarte und umreißt vage ein bestimmtes Gebiet. Doch wo liegt Europa genau? Vielerlei Bemühungen wurden über die Jahrhunderte unternommen, um die Grenzen Europas zum Rest der eurasischen Landmasse hin exakt zu bestimmen (vgl. Schmale 2000, 45ff.; Schultz 1999, 23ff.) – alle mussten letztlich unvollständig bleiben. Weder gibt es eine klare Küstenlinie, noch einen durchlaufenden Gebirgszug, der Europa von Asien trennt, und schon gar nicht einen homogenen Kulturraum, dessen Charakteristika sich ab einem bestimmten Grenzstein urplötzlich auf allen Wahrnehmungsebenen extrem verändern (vgl. Schulze 1999, 11). Gibt es den „Kontinent" Europa überhaupt (noch)? Wenn ja, besteht dieser Zusammenhang im wörtlichen Sinne sicher nicht in der Abgeschlossenheit seiner Landmasse.

Die Frage nach einer Abgrenzung Europas muss also genauer lauten: Was ist Europa? Es wird im weiteren Verlauf noch zu zeigen sein, dass sich die Antwort auf diese Frage im Lauf der Geschichte mehrfach verändert hat (Kap. 2.1). Mehr als auf manch anderen „Raum" trifft also auf Europa zu, dass es sich dabei vor allem um eine Idee handelt, mit der von unterschiedlicher Seite unterschiedliche Hoffnungen, Erwartungen oder gar Befürchtungen verbunden sind. Aktuelle geographische Publikationen zum Europabegriff sprechen daher bereits im Titel nur noch vorsichtig von der *Annäherung an ein umstrittenes Konstrukt* (Reuber et al. 2005)[1]. Eine klare Definition von Europa kann also ebenfalls nicht erwartet werden (vgl. Schmierer 1996, 176). Vielmehr gilt es, die verschiedenen Sichtweisen offenzulegen und transparent zu machen, wie sich diese über die Zeit verändern.

Eines der machtvollsten Konstrukte der jüngeren Geschichte, dem im oben genannten Themenband unter dem Begriff „Europa" fast die ausschließliche Aufmerksamkeit gilt, ist dabei zweifelsohne die Europäische Union. Als Raum gleichen Rechts setzte diese Vertragsgemeinschaft bereits bei ihrer Gründung klare Grenzen. Ursprünglich als Instrument zur Stabilisierung zwischen verfeindeten Nationalstaaten konzipiert, waren bald Kriterien nötig, um die Idee Europa in ihrem raschen Wachstum einzudämmen. Heute, da die Nachfrage nach einem EU-Beitritt ungebrochen scheint, stellt sich die Frage nach den Grenzen der Erweiterbarkeit. Wieder einmal sind es die willkürlichen Landmarken zu Asien

[1] Kursive Zeichensetzung soll im Folgenden Zitate kennzeichnen. Anführungszeichen dienen nur zur Hervorhebung von Einzelbegriffen.

hin. Auch wenn die Handlungsfähigkeit der Union durch den Vertrag von Lissabon selbst bei einer stetig wachsenden Mitgliederzahl gesichert scheint, muss es somit eher als politische denn als topographische Frage gelten, ob Europa in absehbarer Zeit an den Irak grenzt.

Die Tatsache, dass nicht nur der symbolische Gehalt, sondern auch Lage und Ausdehnung des Begriffs „Europa“ frei verhandelbar erscheinen, hat zu unzähligen unterschiedlichen Sichtweisen auf ihn geführt. Ist es sonst allzu einfach, sich anhand griffiger Stereotype vom Nachbarn abzugrenzen, um z.B. den Zusammenhalt einer nationalen Gemeinschaft zu festigen (vgl. Berting/Villain-Gandossi 1999, 26), ist die Lage beim transnationalen Europabegriff ungleich schwieriger: Es ist immer auch ein Europa der „anderen“, dessen Einfluss auf das „Eigene“ stetig wächst. So wohnen Europaeuphorie und Europaskeptizismus mitunter im wahrsten Sinne des Wortes Tür an Tür, wobei die eine wie die andere Seite Anhänger in den unterschiedlichsten Gebieten Europas finden. Nicht zuletzt deshalb bleibt es somit auf absehbare Zeit lohnender Gegenstand verschiedenster wissenschaftlicher und insbesondere geographischer Untersuchungen, den Diskurs über den Europabegriff zu verfolgen und zu begleiten.

Selbst wenn es also einen klar abgegrenzten Europabegriff schon seinem Wesen nach gar nicht geben kann, müssen bei einer vollständigen Betrachtung des Gegenstandes zumindest alle Sichtweisen berücksichtigt werden, die zu seiner ständigen Weiterentwicklung beitragen. Insofern wir als Bürger der Europäischen Union allzu sehr in die Thematik involviert sind, können Beiträge von dieser Seite als befangen gelten. Was liegt also näher, als ganz im Sinne von Adam Smith (vgl. Smith 2004) äußere Beobachter zu Rate zu ziehen, die, wenn schon nicht als idealisierter neutraler Schiedsrichter, so doch zumindest in ihrer unbeteiligten Lesart ihren Beitrag zu einem tieferen Verständnis des Topos Europa liefern können. Wie aber kann es gelingen, auch ohne aufwändige Auslandsaufenthalte und Fallstudien zumindest einen Einblick, wenn nicht einen vergleichenden Überblick über die Sichtweisen zu erhalten, die sich dem Betrachter von außen auf Europa bieten?

In den letzten 20 Jahren hat sich das Internet und hier insbesondere das World Wide Web rasant zu einem wesentlichen Faktor gesellschaftlichen Lebens entwickelt. Ursprünglich zur Verteilung der Rechenlast zwischen mehreren Computern in sensiblen Bereichen von Forschung und Sicherheit konzipiert, wurde schnell das Potential erkannt, Kommunikationsprozesse jeder Art zu unterstützen. Den ersten Diskussionsforen folgten rasch die ersten Mail-Dienste und neben gewerblichen fanden auch private Webseiten

eine zunehmende Verbreitung. Längst haben dabei dynamische Strukturen die ersten statischen Seiten abgelöst. Heute macht das Schlagwort vom „Web 2.0“ die Runde, wenn in Form von Blogs, Podcasting und Online-Communities eine zunehmende Personalisierung und zugleich soziale Vernetzung das Web auf allen Bereichen erfasst. Was läge also näher, als die Fülle an Dokumenten auf außereuropäischen Webseiten dazu heranzuziehen, ein Bild von Europa zu gewinnen, das unseren Beobachtungen normalerweise verschlossen bleibt?

Ziel der vorliegenden Arbeit ist es somit, den Europabegriff auf einigen dieser außereuropäischen Webseiten zu untersuchen und die unterschiedlichen Sichtweisen nach Möglichkeit untereinander zu vergleichen. Da die Geographie für derartige Untersuchungen (noch) nicht über erprobte Standardverfahren verfügt, müssen dafür Anleihen aus verschiedenen Nachbarwissenschaften gemacht werden. Hier sind diverse Herangehensweisen möglich. Analog zu üblichen Datenerhebungen wären z.B. Online-Befragungen oder die teilnehmende Beobachtung von Online-Foren denkbar (vgl. z.B. Chayko 2002, 181ff.). Um jedoch den Aufwand zu minimieren und da die gesamte Kommunikation im World Wide Web – gleich ob bidirektional mit einem bestimmten oder unidirektional mit einem beliebigen Adressaten – ohnehin textbasiert und in digitaler Form stattfindet, bieten sich computergestützte Verfahren und hier insbesondere Methoden der automatischen Textanalyse an. Diese ermöglichen neben der Erzeugung eines Textkorpus und der Extraktion der die Textsemantik tragenden Begriffe auch die kookkurrenzbasierte Berechnung einfacher Verbindungen zwischen diesen. Dadurch kann ein Netzwerk von Begriffen abgeleitet werden, das den semantischen Kontext des Europabegriffs in diesem (Teil-)Korpus repräsentiert und mit dem semantischen Kontext eines anderen (Teil-)Korpus verglichen werden kann.

Sicher kann der gewählte Ansatz eine hermeneutisch-qualitativ arbeitende Studie nicht ersetzen. Allein die Möglichkeit, sehr große Datenmengen in kürzester Zeit zu verarbeiten und konsistent reproduzierbare, wertfreie und somit ohne Weiteres vergleichbare Ergebnisse zu erzielen, lohnt aber den ergänzenden Einsatz dieser freilich sehr reduktiven Methode. Um die Anschlussfähigkeit an bewährte geographische Arbeitsweisen zu demonstrieren, wird zusätzlich gezeigt werden, dass solche computergestützten und somit quantitativen Verfahren problemlos sowohl unterstützend als auch verifizierend in den Arbeitsprozess üblicher qualitativer Studien integriert werden können.

Die Arbeit gliedert sich folgendermaßen: Zunächst wird anhand einiger Beispiele dargelegt, unter welchen Gesichtspunkten Fremdbildforschung zum Europabegriff

betrieben wird und welches die dabei vorherrschenden Methoden sind (Kapitel 2). Hierauf wird in Abgrenzung dazu und unter Berücksichtigung der Einschränkungen und Möglichkeiten einer webbasierten Untersuchung eine eigene Methode zur Beantwortung der gestellten Frage nach dem Europabegriff auf außereuropäischen Webseiten zusammengestellt (Kapitel 3). Dabei wird als methodischer Rahmen zunächst aus üblichen Verfahren qualitativer Textanalyse ein allgemeiner Arbeitsprozess abgeleitet, in den die computergestützten Textanalyse-Methoden problemlos integriert werden können. Da zu diesem speziellen Einsatzziel noch keine geeignete Software existiert, wird in einem nächsten Schritt ein Computerprogramm entworfen, das die Umsetzung der beschriebenen Analyseschritte ermöglicht. Die während der Implementierung getroffenen Entscheidungen werden dabei begründet und dokumentiert (Kapitel 4). Diese Software wird daraufhin dazu verwendet, die gestellte Forschungsfrage anhand ausgewählter Domains zu beantworten (Kapitel 5). Abschließend werden die Ergebnisse der Studie zusammengefasst (Kapitel 6) und Anschlussmöglichkeiten für weitere Forschungsarbeiten beschrieben (Kapitel 7).

# 2 Sichtweisen auf Europa

Es ist bereits deutlich geworden, dass es sich bei „Europa“ um einen konstruierten Begriff handelt, für den weder klare Grenzen noch genaue Abgrenzungskriterien genannt werden können. Gerade weil wir als „Europäer“ aber durch unseren tagtäglichen Gebrauch des Begriffs ein weitreichendes Vorverständnis für ihn entwickelt haben, ist es unerlässlich, sich die Vielschichtigkeit bewusst zu machen, die bereits aus der Innensicht eine immer neue Akzentuierung ermöglicht. Da jedoch das Selbstbild von Europa ein gewachsenes ist, sollen zumindest in aller Kürze die wesentlichen Stationen der geschichtlichen Entwicklung und die damit verbundenen Konnotationsverschiebungen des Europabegriffs nachvollzogen werden, die in ihrer Gesamtheit im heutigen Bild nachwirken. Erst vor diesem Hintergrund macht es Sinn, nach einem Fremdbild des Europabegriffs zu fragen.

## *2.1 Geschichte des Europabegriffs*

Da eine Geschichte des Europabegriffs naturgemäß die Arbeit von Geschichtswissenschaftlern ist, muss sich in diesem Fall *der Geograph beim Historiker [umsehen]* (Schultz 1999, 74). Zwar sind sich auch die Autoren der Standardwerke auf diesem Gebiet durchaus bewusst, dass Europa *weder als geographische noch als politische Wirklichkeit zu erfassen* (Schulze 1999, 11) ist, dennoch berichten sie vorrangig von *eine[r] Geschichte Europas, nicht eine[r] des Europagedankens* (Salewski 2000, 9), also von den historischen Entwicklungen in einem wie auch immer abgegrenzten Europa, nicht denjenigen des Europabegriffs an sich. Eine Ausnahme ist hier die *Geschichte Europas* von Schmale (Schmale 2000), auf die sich die weiteren Ausführungen im Wesentlichen stützen.

Einen ersten Überblick über die verschiedenen Dimensionen des Europabegriffs und ihre zeitliche Entwicklung bietet Abb. 1. Da es sich bei der vorliegenden Arbeit um eine Textanalyse handelt, soll der Schwerpunkt dabei auf Ausdrucksmitteln liegen, die eine textuelle Darstellungsform aufweisen können.

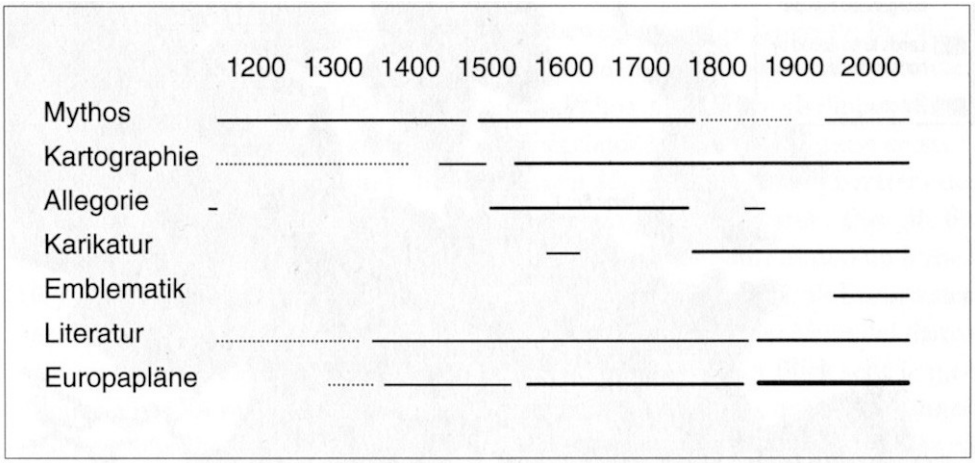



*Abb. 1: Dimensionen des Europabegriffs und ihre zeitliche Entwicklung (Schmale 2000, 83)*

Am Anfang steht der antike Mythos der Europa, nicht als Opfer des Zeus, sondern vielmehr als Inbegriff von *Auserwähltheit, Glück und Fruchtbarkeit* (Schmale 2000, 26). Als Bezeichnung für einen Erdteil findet sich der Begriff erstmals in den Historien des Herodot. Obwohl „Europa“ in seiner Weltkarte auch die Gebiete weiter nördlich kennzeichnet, dient die Bezeichnung vor allem zur Abgrenzung zwischen Europa als der griechischen und Asien als der persischen Einflusssphäre (Schmale 2000, 22; Schulze 1999, 511). Während

der gesamten Dauer des römischen Imperiums blieb es jedoch bei dieser mythischen bzw. unbestimmt geographischen Lesart (Schmale 2000, 28).

Mehrere Bedeutungsverschiebungen erfährt der Begriff zu Beginn des Mittelalters. Zunächst wird die Metapher der Abgrenzung durch die *europenses* fortgeschrieben, die 732 die Schlacht von Poitiers gegen die vorrückenden Muslime gewinnen (Schmale 2000, 29). Karl der Große instrumentalisiert den Begriff dagegen gleich in doppelter Hinsicht neu. Das regnum Europae soll zum einen seinem Flächenstaat unter fränkischer Herrschaft Identität verleihen, zum anderen setzt er es als weltlichen Machtbereich dem nur noch spirituellen imperium entgegen (Schmale 2000, 29f; Le Goff 2004, 48). Obwohl der Europabegriff im Mittelalter stark mit den Begriffen der Christenheit und des Abendlandes verwoben ist (Schmale 2000, 30), verstärkt sich zur selben Zeit durch den Bilderstreit der Status von Byzanz als weiterem abgrenzenden Pol. Die aus dem Reich Karls des Großen hervorgehenden *Europae regna* sind für Byzanz bestenfalls Barbaren, die die byzantinische Oberhoheit über die gesamte Christenheit nicht anerkennen (Le Goff 2004, 44; Schmale 2000, 30ff.). Während des Mittelalters war der Begriff Europa allerdings kaum einer breiten Bevölkerung geläufig (Schmale 2000, 30).

Zu Beginn der Neuzeit verändert sich auch der Europabegriff erneut. Noch bevor Erasmus die Dame Europa als organische Körpermetapher einführt (Schmale 2000, 87), an die sich weitere Darstellungen von Europa als symbiotische Allegorie anschließen, entwickelt der böhmische König Georg von Podiebrad im 15. Jahrhundert bereits eine anachronistisch anmutende Europakonzeption: Die europäischen Staaten sollten untereinander auf Krieg verzichten, es solle eine Ratsversammlung eingerichtet werden, in der jeder Staat genau eine Stimme habe; ferner sollten ein gemeinsames Wappen, ein Siegel und ein eigener Beamtenstab eingeführt werden (Le Goff 2004, 253f.; Schmale 2000, 86f.). Georg von Podiebrad ging es wohl vor allem darum, die Macht von Papst und Kaiser zu schwächen und Böhmen seine dauerhafte Eigenständigkeit zu sichern (Schmale 2000, 87). Dennoch entwickeln sich von nun an in rascher Folge unterschiedlich akzentuierte Variationen dieser Grundidee. Im 17. Jahrhundert kommt Sully am Hof Heinrichs IV. von Frankreich zu dem Ergebnis, dass für ein ausgewogenes Staatensystem 15 europäische Staaten nötig sind, und erarbeitet dazu auch gleich einen entsprechenden kartographischen Entwurf. Russland, das sich als *drittes Rom* (Schultz 1999, 33) in der Tradition von Byzanz sieht, aber auch große Teile des Balkans werden dabei nur als bedingt zu Europa gehörend anerkannt; ihre spätere Einbeziehung wird aber nicht grundsätzlich ausgeschlossen (Schmale 2000, 84 u. 88f.).

Kurzzeitig unterbrochen werden die Pläne durch die Bemühungen Napoleons, ein vereintes Europa unter der Führung Frankreichs zu errichten. Er sieht sich dabei in der Tradition Karls des Großen – und auch seine Hegemonie-Bestrebungen scheitern (Schmale 2000, 94f.; Le Goff 1999, 48). Dennoch wird nicht zuletzt durch ihn eine entscheidende Neuordnung Europas eingeleitet. Die zunehmend geschwächte Position des Begriffs der Christenheit findet ihren Ausdruck in der Säkularisation und auch die alte Trennung zwischen regnum und imperium wird mit dem Ende des Heiligen Römischen Reiches endgültig zu Grabe getragen. Auf dem Wiener Kongress fallen die verschiedenen Gleichgewichtsideen also auf fruchtbaren Boden, als Metternich erstmals neben Russland auch das Osmanische Reich als stabilisierenden Faktor in die Überlegungen zur Neuordnung mit einbezieht (Schmale 2000, 96). Dennoch bleibt die Zugehörigkeit beider Staaten zu Europa umstritten. Wurde das Osmanische Reich einfach als *fremdes Volk* (Schultz 1999, 33) abgetan, versuchte die Geographie des 19. Jahrunderts vornehmlich in der Landesnatur Russlands den Beweis bald für die Lage in Europa, bald in Asien zu finden (Schultz 1999, 32ff.).

Während des Vormärz werden eifrig Pläne für Vereinigte Staaten von Europa geschmiedet. Diese scheitern allerdings am erstarkenden Konzept der Nation, das spätestens nach 1848 eng an die Seite des Europabegriffs tritt (Schmale 2000, 101). Da der Nationenbegriff zu dieser Zeit jedoch vor allem für Freiheit und Souveränität des Volkes steht, können trotz der beginnenden Entstehung der Nationalstaaten nationales und europäisches Bewusstsein nebeneinander existieren (Schmale 2000, 102).

Der erste Weltkrieg stört die bereits vorhandenen Verflechtungen vor allem wirtschaftlicher Natur empfindlich, aber auch wenn die Idee der Vereinigten Staaten von Europa vorrangig in den Völkerbund einfließt, der seinem Anspruch nach nicht nur auf Europa beschränkt ist, bleibt die Idee Europa erhalten. Im 1923 erschienenen Pan-europäischen Manifest und im daran anknüpfenden Briand-Plan finden sich erstmals konkrete Überlegungen zu föderativen Strukturen unter wirtschaftlichen, sozialen und politischen Gesichtspunkten, bevor unter dem Nazi-Regime der Europabegriff wieder einmal zur Durchsetzung eines hegemonialen Herrschaftsanspruchs instrumentalisiert wird (Schmale 2000, 110ff.).

Auch wenn die EU gegenwärtig die wirksamste Europakonzeption repräsentiert, darf nicht vergessen werden, dass diese mitnichten der einzige Ansatz zur Verwirklichung einer europäischen Integration in der Nachkriegszeit war. Wesentlich zur kulturellen Integration trug auch der Europarat bei und selbst im ureigenen Feld der wirtschaftlichen

Zusammenarbeit existierte mit der EFTA lange ein von Großbritannien initiierter Gegenentwurf, der weniger stark auf Integration ausgerichtet war (Schmale 2000, 231 u. 240f.). Zudem kann man die EU erst nach zahlreichen Erweiterungen und Zusatzverträgen vor allem nach dem Fall des Eisernen Vorhangs als wirkliche Europakonzeption betrachten.

Obwohl sich die Nebenbedeutungen Europas vom Mythos über die Christenheit bis hin zum Europa der Nationen über die Zeit erheblich gewandelt haben, können aus geschichtlicher Sicht zusammenfassend doch zwei Konstanten im Europabegriff festgehalten werden: zum einen die negative Definition zur *Abwehr von wirklichen oder empfundenen Bedrohungen* (Schulze 1999, 508) von außen und zum anderen das relative Gleichgewicht der Staaten, das bisher stets *die Übermacht eines einzelnen zu verhindern* (Schulze 1999, 509, vgl. auch Le Goff 2004, 48) wusste. Vereinigte Staaten von Europa wird es wohl nicht geben. Aber auch wenn die EU weder eine Föderation noch ein Bundesstaat ist (Schmale 2000, 245), darf sie doch zumindest als ein *innovatives institutionelles Arrangement* (Heeg/Oßenbrügge 2005, 102) gelten, dessen Potential erst noch weiter ausgelotet werden muss.

## *2.2 Vielschichtigkeit des Europabegriffs und daraus resultierende Schwierigkeiten*

Wenn im folgenden Kapitel einige Beispiele für verschiedene Blickwinkel auf Europa und den Europabegriff vorgestellt werden, muss klar sein, dass es sich dabei im vorliegenden Rahmen nur um eine sehr beschränkte und auch keinesfalls repräsentative Auswahl aus der Fülle an Publikationen handeln kann. So findet die Offenheit der Diskussion über den Europabegriff, der gerade aus der Innensicht hauptsächlich als kultureller Begriff verhandelt wird, ihren Niederschlag in der gesamten Breite der Kulturwissenschaften. Beitrag der Geschichtswissenschaften ist neben der bereits eingeführten Begriffsgeschichte des Europabegriffs immer auch das Fremdbild Europas aus historischer Sicht (z.B. Harper 1996, Nagel 1990, Chen 2001). Literaturwissenschaftliche Forschung liefert daneben wesentliche Einblicke in innovative Europaentwürfe sowie die Rezeption von und Kritik an vorherrschenden Konzeptionen in Geschichte und Gegenwart (z.B. Delvaux/Pepiór 1996, Glaser/Schnackertz 2005, Conter 2004, Segebrecht et al. 2003, Keller/Rakusa 2004, Fendler/Wittlinger 1999). Im Sinne einer *Politischen Geographie Europas* (Reuber et al. 2005) soll der Fokus bei den

anschließenden Blickwinkeln auf Europa vor allem auf politische Europakonzeptionen der Gegenwart gelegt werden, die jedoch durchaus um andere für eine geographische Untersuchung relevante Aspekte ergänzt werden können. Die angeführten Beispiele sollen dabei illustrieren, welche Unterschiede zwischen den verschiedenen Sichtweisen auf Europa bestehen und dass folglich ein Vergleich zwischen unterschiedlichen Sichtweisen ein lohnendes und aussichtsreiches Unterfangen darstellt.

Wie im bisherigen Verlauf bereits deutlich geworden ist, handelt es sich beim Europabegriff um ein höchst dynamisches Konzept. Dadurch ergeben sich für den weiteren Verlauf der Untersuchung drei wesentliche Schwierigkeiten. Der zunächst allgemeinere Europabegriff wird vielerorts immer stärker durch das Konzept der EU verdeckt. Bereits Umfragen zu Anfang bzw. Mitte der 90er Jahre des letzten Jahrhunderts belegen, dass eine deutliche Mehrheit der freien Assoziationen zum Begriff „Europa“ bei Jugendlichen und jungen Erwachsenen zumindest innerhalb der EU auf die EU selbst bzw. Resultate ihrer Politik verweist (Riketta/Wagenhut 1998, 5). In vielen der im Folgenden vorgestellten Publikationen wird „Europa“ daher auch einfach synonym zu „EU“ verwendet. Obwohl der allgemeine Europabegriff im Zentrum dieser Untersuchung steht, muss natürlich auch der EU-Begriff als eine und gegenwärtig wirkungsvollste der verschiedenen möglichen Europakonzeptionen Berücksichtigung finden.

Eine zweite Schwierigkeit, die aus der Dynamik des Europabegriffs erwächst, ist die Abgrenzung zwischen Beiträgen, die eine Innen- und solchen, die eine Außensicht repräsentieren. Da Europa aus den eingangs genannten Gründen keine klare Grenze aufweist, sondern je nach der Wirkkraft der vorherrschenden Konzeptualisierung in benachbarte Regionen ausstrahlt, kann auch nur äußerst behutsam versucht werden, die Beiträge je nach der eigenen Wahrnehmung um ein imaginäres Zentrum zu gruppieren.

Die dritte Schwierigkeit besteht darin, eine geeignete Maßstabsebene auszuwählen, auf der die Konzeption des Europabegriffs untersucht werden kann. Hierbei muss berücksichtigt werden, dass eine niedrige Maßstabsebene zu einer unüberschaubaren Vielfalt von Meinungen und Ansichten führt, während eine sehr abstrakte Maßstabsebene eher selbst ein Produkt einer Regionalisierung denn einen konsistent handelnden Akteur darstellt. Bevor also im Folgenden der Nationalstaat als Ebene der Betrachtung verwendet werden kann, muss man sich zunächst bewusst machen, dass seine Existenz immer darauf beruht, dass ihn seine Bürger ebenso wie staatliche Institutionen im wahrsten Wortsinn alltäglich regionalisieren (vgl. Werlen 1997, 329ff.). So gibt es nicht die „japanische“ Sichtweise, vielmehr werden viele Einzelansichten zu einer Gesamterzählung

verwoben, wobei die Gefahr natürlich groß ist, bei der Charakterisierung eines solchen Fremdbildes wiederum der Zuschreibung von Stereotypen (vgl. Kap. 1) zu erliegen. Sollen also Nationalstaaten als Maßstabsebene verwendet werden, muss zunächst hinterfragt werden, welche Akteure diese Sichtweise geprägt haben, und ferner darauf geachtet werden, dass die Vergleichbarkeit der Äußerungen gewährleistet werden kann. Dies darf bei den Äußerungen der Vertreter staatlicher Institutionen als gegeben betrachtet werden.

Nach einem kurzen Blick auf die unterschiedlichen Sichtweisen innerhalb der gegenwärtigen EU sollen im Folgenden die Blickwinkel einiger Staaten des ehemaligen Jugoslawien und der islamischen Welt beleuchtet werden, bevor US-amerikanische, japanische und afrikanische Betrachtungen das Bild abrunden.

## *2.3 Über Europa sprechen – Einige Beispiele*

Einen guten Einblick in die allgemeine Grundstimmung bezüglich der Wahrnehmung der EU in den einzelnen Mitgliedsländern bietet die halbjährlich durchgeführte Europabarometer-Umfrage, die zuletzt ein wachsendes Vertrauen einer Mehrheit der EU-Bürger in die Organe der EU (Europäische Kommission 2007, 23ff.) und auch eine Zwei-Drittel-Mehrheit für eine Europäische Verfassung belegt (Europäische Kommission 2007, 34). Regionale Unterschiede ergeben sich vor allem bei den Hauptsorgen der EU-Bürger, zu denen in allen Regionen die Angst vor Arbeitslosigkeit und Kriminalität zählt. Während in den 12 neuen Mitgliedsstaaten jedoch vor allem *Wirtschaft, Gesundheitssystem und in geringerem Maße Inflation/steigende Preise* (Europäische Kommission 2007, 11) als drängende Probleme wahrgenommen werden, sorgt sich die Bevölkerung der 15 etablierten Staaten mehr als die Einwohner der 12 neuen Mitgliedsländer um die Themen Terrorismus, Einwanderung und Umwelt (Europäische Kommission 2007, 12).

Bei den von Schott (2005) vorgestellten *[g]eopolitischen Regionalisierungen der Europäischen Union und Europas in der Welt* (Schott 2005, 74) handelt es sich dagegen im Wesentlichen um strategische EU-Konzepte hochrangiger politischer Vertreter verschiedener EU-Mitgliedsstaaten. Da die Außensicht in der vorliegenden Untersuchung im Vordergrund steht, soll dies jedoch genügen, um den kurzen Eindruck von verschiedenen Europabildern aus der unmittelbaren Innensicht abzurunden. Während z.B. zwischen Deutschland und Frankreich Einigkeit darüber besteht, dass die EU als globales Einflusszentrum (Schott 2005, 87) gestärkt und als Gegengewicht zu den USA (Schott 2005, 95) ausgebaut werden muss, gibt es doch Unterschiede bei der Frage nach einer

möglichen weitergehenden Osterweiterung. Frankreich erachtet den Erweiterungsprozess abgesehen von einem möglichen Beitritt der verbleibenden Balkan-Staaten als abgeschlossen und macht dafür mit unterschiedlicher Religion, Staatsform oder fehlender Stabilität neben politischen auch kulturelle Gründe geltend (Schott 2005, 86). In Deutschland wird dagegen immer wieder die strikte Trennung zwischen EU und Europa betont, für das die EU nicht in vollem Umfang Verantwortung übernehmen könne, da der europäische Kulturraum prinzipiell auch ganz Russland in seinen heutigen Grenzen umfasse. Eine Erweiterung der EU auf solche Regionen sei zwar nur auf sehr lange Sicht und nach behutsamer Heranführung, aber doch grundsätzlich für möglich (Schott 2005, 95). Auch eine letztliche Ausdehnung des Schengenraumes auf den nahen Osten und Nordafrika sei nicht ausgeschlossen (Schott 2005, 98). Ganz andere Interessen verfolgen dagegen Vertreter Großbritanniens. Von einer raschen Erweiterung der EU erhoffen sie sich bewusst eine weitere Heterogenisierung der Union, um so die Reichweite der deutsch-französischen Integrationspolitik zu schwächen. Dadurch würde die EU im Wesentlichen auf den gemeinsamen Markt beschränkt bleiben und Großbritannien könnte im Sinne nationaler Interessen die eigene einflussreiche Stellung in NATO und Commonwealth wahren, ohne dass dabei seine Macht als global player zu sehr in den Entscheidungsstrukturen der EU aufginge (Schott 2005, 80ff.). Bereits bei langjährigen Mitgliedsstaaten finden sich also bemerkenswerte Unterschiede in der Sichtweise auf die EU im Besonderen und Europa im Allgemeinen.

Betrachtet man Tschechien und Ungarn beispielhaft für die Staaten der jüngsten EU-Osterweiterungen, fällt auf, dass die Europa-Euphorie der frühen 90er Jahre des letzten Jahrhunderts, in denen die EU als Inbegriff Europas Wohlstand, Freiheit und Sicherheit versprach (Klunkert 1996, 238f.), in der Zwischenzeit merklich abgekühlt ist. Wurde damals in Ungarn unter dem Schlagwort *Zurück nach Europa* die eigene europäische Tradition unter Rückgriff auf die historischen Vorlagen wie die gemeinsame Abwehr der Bedrohung durch das Osmanische Reich legitimiert (Klunkert 1996, 238f.) und in der Verfassung verankert (Bedi 2004, 75ff.), wird die durch das „zurück nach“ implizierte Reisemetapher mittlerweile vor allem bei Schriftstellern aus verschiedenen Gründen kritisch gesehen. Während in ungarischer Literatur Mitteleuropa als Illusion verspottet wird (Friedl 2006, 41), wird von tschechischer Seite bestritten, dass überhaupt eine Reise nötig sei, da man von jeher im Herzen Europas geblieben sei. Gerade aus diesem Selbstverständnis erwächst aber auch die Überzeugung, dass es neben dem bloßen *Körper* EU auch eine *Seele* Europa gebe (Hanshew 2006, 10f.), sowie der daraus

resultierende Wunsch, an der weiteren Gestaltung der EU aktiv mitwirken zu können. Nicht zuletzt um das Gewicht der eigenen Stimme zu stärken, proklamieren politische Vertreter Tschechiens nach innen die strikte Gleichwertigkeit aller Mitglieder und nach außen eine *Festung Europa*, deren Grenzen nach Osten nun definiert seien (Schott 2005, 77f.).

Ein äußerst zerrissenes Bild von Europa findet sich bei den Balkanstaaten Bosnien-Herzegowina, Serbien und Kroatien. Obgleich man sich in allen drei Staaten von Europa, in das auch hier größtenteils die EU projiziert wird, Sicherheit, Freiheit und Modernisierung erhofft und sich darüber im Klaren glaubt, dass der Weg nach Europa ein alternativloser sei (Džihić 2006, 285 u. 288), verbleibt doch eine tiefe Abneigung. Dabei vermengen sich verschiedene Motive. Während Europa auf der einen Seite hinterlistig und unmoralisch erscheint, weil es in den Balkankriegen der 90er Jahre des letzten Jahrhunderts entweder zu spät oder – im Falle Serbiens – auf der „falschen“ Seite eingegriffen hat, ergibt sich wieder einmal unter Bezugnahme auf historische Ereignisse der Vorwurf, Europa wisse die Leistung des Balkans bei der Abwehr des Islam nicht zu würdigen (Džihić 2006, 286). Die Zerrissenheit äußert sich letztlich darin, das eigene Land in Europa, die jeweils beiden anderen Staaten aber auf dem Balkan und somit außerhalb Europas zu verorten, während zur selben Zeit das eigene Land als Teil des Balkan gegenüber dem reichen und überheblichen Europa als moralisch überlegen konstruiert wird (Džihić 2006, 286f.).

Durch das Europabild der islamischen Welt kann die eingangs skizzierte Geschichte des Europabegriffs hervorragend aus einer seit Jahrhunderten eng benachbarten Außensicht kontrastiert werden. Galt Europa zunächst als Land der Ungläubigen und der Inquisition, wurde die beginnende Expansion im 17. und 18. Jahrhundert argwöhnisch beobachtet. Zugleich boten die technologischen Errungenschaften dieser Zeit jedoch erstmals einen Anreiz, Europa zu Studienzwecken zu bereisen. Der im 19. Jahrhundert noch offener zu Tage tretende Kolonialismus verstärkte aber weiter das Bild von Europa als Aggressor (Köse 1999, 180ff.). Nach dem Zweiten Weltkrieg präsentierte sich Europa den ausgewanderten Muslime als ein Land des Überflusses und des Luxus (Köse 1999, 182ff.). Dennoch führten internationale Konflikte wie die Zypernkrise, der Israelkonflikt oder zuletzt der Bosnienkrieg zu einer schrittweisen Verschlechterung des Europabildes, da sich in der islamischen Welt die Überzeugung festigte, der Westen entscheide im Gefühl der Überlegenheit stets zu Ungunsten der jeweiligen islamischen Bevölkerungsgruppen (Köse 1999, 185f.). Dabei wird ähnlich zur Sichtweise auf dem Balkan der säkularen, westlichen Welt die moralische Überlegenheit der islamischen

Kultur gegenübergestellt (Köse 1999, 186ff.). Sowohl im Falle des Balkan als auch anhand des Bildes in der islamischen Welt kann somit am Beispiel unmittelbarer Nachbarn belegt werden, wie sehr das machtvolle Konstrukt Europa auf der einen Seite Neugier und Bewunderung hervorruft, auf der anderen Seite aber zur Wahrung der eigenen Identität auch eine klare Abgrenzung provoziert.

Das am besten dokumentierte außereuropäische, wenngleich in seiner Tradition ebenfalls eng mit ihm verknüpfte Bild von Europa ist zweifelsohne das US-amerikanische. Daher soll es an dieser Stelle etwas ausführlicher beleuchtet werden. Auch hier lohnt zunächst ein Blick auf die Geschichte, umso mehr, als sich die Vereinigten Staaten bereits bei ihrer Gründung auf die Europapläne dieser Zeit berufen. Als Realisierung von Sullys Vision gefeiert (Hendrickson 2006, vgl. dazu Ash 2004, 125 – vgl. auch Kap. 2.1), diente die *novo ordo seclorum* (Schwabe 2000, 62) zunächst dazu, sich von der Kabinettspolitik des alten Kontinents mit Bündnissystemen und Geheimverträgen abzusetzen. Im Bewusstsein der Vorzüge der eigenen Verfassung wurde daher zu Beginn eine ausgesprochene Isolationspolitik betrieben. Diese Einstellung wandelte sich aber bald dahingehend, dass die Vereinigten Staaten der restlichen Welt zumindest aus der Ferne als Beispiel für diese freiheitliche Ordnung dienen wollten (Schwabe 2000, 62). Im späten 19. Jahrhundert kam zusätzlich immer mehr der Gedanke auf, die Freiheitsidee müsse nicht nur als Beispiel, sondern auch im Rahmen aktiver Maßnahmen verbreitet werden. Dieser Missionarismus entfaltete erstmals im Ersten Weltkrieg seine Wirksamkeit, als Wilson vor allem Deutschland politische Zurückgebliebenheit attestierte und in seinen Vierzehn Punkten eine neue Ordnung für Europa forderte, aus der schließlich der Völkerbund hervorging (Schwabe 2000, 63f.). Selbst Roosevelt dachte am Anfang des Zweiten Weltkrieges jedoch noch nicht an ein dauerhaftes Engagement in Europa (Schwabe 2000, 65). Erst der Beginn des Kalten Krieges und die Bedrohung durch sowjetische Atomraketen veranlassten die USA, neben dem Verteidigungsbündnis der NATO nachhaltig in eine weitergehende europäische Integration zu investieren (Schwabe 2000, 67). Nachdem ein starkes vereintes Europa nach der Vorstellung Kennedys am nationalen Widerstand Großbritanniens und Frankreichs gescheitert war, beschränkte sich der europäische Integrationsprozess allerdings vorerst auf den wirtschaftlichen Bereich (Schwabe 2000, 68). Dieses Konzept eines in Nationalstaaten zersplitterten Europas, das US-amerikanischer Obhut bedürfe, erhielt sich so nachhaltig (Rosenberger 2005, 190 u. 201), dass auch noch bei Amtsantritt der Regierung Bush das mittlerweile erreichte Potential, das weit über das eines reinen Wirtschaftsraums hinausreicht, völlig unterschätzt wurde (Dale 2004). Noch 2001 werden

in *Policy Review*, einer Online-Zeitschrift der Hoover Institution an der Stanford University, Aufsätze publiziert, die dies eindrucksvoll belegen. Das durch die *Protestant Revolution* (Casey/Rivkin 2001) längst überwundene Heilige Römische Reich solle restituiert werden, wie die Existenz eines an das Reich Karls des Großen erinnernden Karlspreises sowie einige aus dem Zusammenhang zitierte Äußerungen des damaligen Präsidenten der EU-Kommission Prodis belegen sollen (Casey/Rivkin 2001). Müsse dieses Streben als mittelalterlich bewertet werden, könne die Existenz einer nicht gewählten, zentralisierten und niemandem verantwortlichen Verwaltung bestenfalls als absolutistisch gelten (Casey/Rivkin 2001; Gonzalez 2001). Dies stelle einen dramatischen Verlust der Souveränität des Volkes dar (Casey/Rivkin 2001). Hinzu komme noch ein übertriebener Wohlfahrtsstaat, dessen einziger Erfolg hohe Arbeitslosigkeit sei (Gonzalez 2001; vgl. Rosenberger 2005, 199) sowie die Weigerung, *one of Europe's most important states* (Casey/Rivkin 2001), die Türkei, in die EU aufzunehmen, *while not one square inch of either Britain or Ireland, both EU-members, touches the Continent* (Casey/Rivkin 2001)[2]. Die USA dürften jetzt aber dieses hauptsächlich von Frankreich gesteuerte Modell nicht einfach gewähren lassen, sondern müssten versuchen, über Großbritannien Einfluss auf die Entwicklung zu nehmen (Gonzalez 2001, Dale 2004).

Das Fehlen einer wirklich eingehenden Beschäftigung mit Europa wird auch dadurch belegt, dass die Ende 2002 beschlossene und in Europa als historisch empfundene EU-Osterweiterung in der Washington Post mit der lapidaren Schlagzeile quittiert wurde, die EU hätte sich entschieden, ihre Freihandelszone auszuweiten (Dale 2004). Erst als Rumsfeld aus *historischer Unwissenheit* (Rosenberger 2005, 198) zwischen altem und neuem Europa polarisiert und getreu der Sichtweise der EU als *“pool“ from which U.S. allies can be selected* (Dale 2004, vgl. Rosenberger 2005, 201) ein Bündnis für den gewaltsamen Sturz Saddam Husseins zu schmieden versucht, kommt es zum *clash over Iraq* (Dale 2004). Auch wenn es sicher falsch wäre, die vorherigen Sichtweisen als „amerikanische“ zu pauschalisieren, finden sich zumindest in *Policy Review* erst danach differenziertere Europadarstellungen. Einen wichtigen Wendepunkt stellt dabei zweifelsohne auch das Erscheinen von Kagans *Macht und Ohnmacht* (Kagan 2003) dar, wo erstmals ein Autor, der konservativen Kreisen zugerechnet werden muss (Ash 2004, 76), fordert, die gegenseitige *Verleugnung und Polemik hinter sich zu lassen* (Kagan 2003, 12), und sich auf die Suche nach einem tieferen Verständnis in der wechselseitigen

[2] Hier tritt neben die eingangs erwähnten eine weitere Lesart, bei der Europa offensichtlich als das aus der Philosophiegeschichte bekannte Kontinentaleuropa der anglophonen Welt gegenübergestellt wird.

Sichtweise zwischen Europa und den USA begibt (vgl. dazu Hendrickson 2006). Die wesentliche Triebkraft für das politische Handeln sei in der *Psychologie von Macht und Schwäche* (Kagan 2003, 34; vgl. auch Hendrickson 2006) zu suchen. Hier sei es im Laufe der letzten beiden Jahrhunderte zu einer dramatischen Verschiebung der Machtverhältnisse von den europäischen Großmächten zu den Vereinigten Staaten gekommen (Kagan 2003,14f.). Da der geopolitische Einfluss Europas in keinem Verhältnis mehr zu seiner ökonomischen Stärke stehe (Kagan 2003, 26f.; vgl. auch Ash 2004, 133), würde dessen Handeln vor allem von Verhandlungen geprägt, während die USA es sich durch die ihnen zur Verfügung stehenden Machtmittel leisten könnten, sofort selbst aktiv zu werden (Kagan 2003, 38ff.). Europa könne sich heute überhaupt nur als *kantische Welt des ewigen Friedens* (Kagan 2003, 68) feiern, da die Vereinigten Staaten *Sicherheit von außen gewährleisten* (Kagan 2003, 69). Vor diesem Hintergrund erscheint Europa allerdings zumindest als undankbar (Kagan 2003, 52).

Zwar wird daher auch weiterhin die Ansicht vertreten, der Irakkrieg sei vor allem für Frankreich lediglich ein willkommener Anlass gewesen, sich als *counterweight* (Dale 2004) zu profilieren. Dennoch muss das Projekt der Europäischen Union als ein *astonishing success* (Dale 2004) betrachtet werden, auch wenn die daraus resultierende Eigenständigkeit Europas z.B. in Umweltfragen zu den US-amerikanischen Überzeugungen konträre Interessen hervorgebracht hat (Dale 2004; Ash 2004, 133). Zudem wächst das Bewusstsein, dass *American conservatives who read British conservative newspapers and journals find their own assessments reinforced without having to examine the reality of what ist actually happening* (Dale 2004) und dass sich nicht alle Aspekte der europäischen Einigung von vorn herein als notwendigerweise auf die USA bezogen betrachtet werden dürfen (Dale 2004). Hendrickson (2006) resümiert im Anschluss an die von Kagan (2003) diagnostizierte *Psychologie der Macht und Ohnmacht* (Kagan 2003, 34ff.) gar, es sei zu einem *great reversal* (Hendrickson 2006) gekommen. Neben der Umkehrung des Kräfteverhältnisses dienen ihm dabei auch einige frappierende Ähnlichkeiten zwischen den Problemen der frühen Vereinigten Staaten und dem jetzigen europäischen Einigungsprozess als Argument. Während sich die Vereinigten Staaten jedoch allmählich von den ursprünglich föderalen zu eher zentralen Strukturen entwickelt hätten, vertrete nun Europa das Ideal eines hochangepassten und zeitgemäßen föderalen Systems. Der Sache nach gelte also in jederlei Hinsicht: *Europe has ended up where America began* (Hendrickson 2006).

In der englischsprachigen Forschung sind naturgemäß US-amerikanische und britische Beiträge sehr dominant, mit denen ja auch die selbst artikulierte Innensicht von Europa in stetigem Austausch steht (vgl. Bialasiewicz 2005, 368; vgl. dazu Ash 2004, 76). Da anderssprachige Quellen, so z.B. umfangreiche Publikationen in französischer Sprache, aufgrund mangelnder Sprachkenntnisse leider nicht hinzugezogen werden konnten, sollen zwei weitere Blickwinkel als Beispiele für eine Sichtweise auf Europa außerhalb dessen unmittelbarer Interessenssphären ausreichen.

Je weiter außerhalb der unmittelbaren Einflusssphäre Europas eine Region liegt, desto mehr sinkt naturgemäß die Notwendigkeit, sich überhaupt ein dauerhaftes Bild von Europa zu machen. Das Beispiel Japan belegt, dass die Zeit der Europa-Faszination, in der Japan nicht nur die effizienten Produktionsmethoden, sondern auch *a wide range of European institutional and judical models* (Hernádi 1996, 6) importierte, bereits mit dem Ersten Weltkrieg wieder ein Ende fand. Zu sehr waren japanische Europareisende zu dieser Zeit von der Armut und der Hoffnungslosigkeit in der Bevölkerung erschüttert (Hernádi 1996, 6). Da in Japan traditionell wieder die Einzigartigkeit der eigenen Kultur betont wird (Hernádi 1996, 7; Furth 1993, 300f.), beschränkte sich sein Interesse an Europa nach dem Zweiten Weltkrieg auf wirtschaftliche Aspekte. Dabei fürchtete Japan vor allem die mit Handelsbeschränkungen gesicherte *Festung Europa* (Furth 1993, 299; vgl. Hernádi 1996, 4). Als die europäische Wirtschaft in den 70er Jahren des letzten Jahrhunderts relativ zu Japan zurückfiel, hatte dies ein weitgehendes Abwenden von Europa zur Folge. Fortan wurden die Europäischen Gemeinschaften unter dem Schlagwort Eurosklerose verspottet (Fuhrt 1993, 298) und die eigene technologische Überlegenheit betont, mit der sich Europa nicht messen könne (Hernádi 1996, 11). Auch als Japan nach dem Zusammenbrechen der sozialistischen Staaten und der Verabschiedung der Einheitlichen Europäischen Akte ein rasches Erstarken der Europäischen Union und damit der Festung Europa befürchtete (Hernádi 1996, 14), ging die Betrachtungsweise über die eines Handelspartners kaum hinaus. Das Ende des Kalten Krieges beförderte durch das Wegfallen des gemeinsamen Feindbildes eine weitere Orientierung Japans auf die asiatische Welt (Hernádi 1996, 12).

Ein weiteres Beispiel dafür, dass das Thema Europa außerhalb seines unmittelbaren Umfelds durchaus nur von nachrangigem Interesse ist, kann anhand afrikanischer Medienberichte belegt werden (Mytton 2006, 122). Mytton (2006) zeigt, dass trotz der großen Bedeutung europäischer Rundfunksender für das afrikanische Informationswesen

(Mytton 2006, 113f.) ein hoher Grad an Informiertheit über Europa nur dann gegeben ist, wenn dadurch die eigenen Interessen unmittelbar betroffen sind (Mytton 2006, 110).

## *2.4 Typisierung der bisher vorgestellten Sichtweisen*

Obwohl die bisher vorgestellten Quellen nur eine kleine Auswahl von Untersuchungen zum Europabegriff darstellen[3], soll dies als Einführung in den Gegenstand genügen. Im Folgenden werden die wesentlichen Besonderheiten und ihre Auswirkung auf diese Arbeit kurz zusammengefasst. Grundsätzlich lassen sich aufgrund der bisherigen Beobachtungen vier Typen von Sichtweisen unterscheiden. Zum Ersten ist dies die erklärte Innensicht von Vertretern all derjenigen Staaten, die sich am kontinuierlichen Aufbau Europas aktiv beteiligt sehen. Beim zweiten Typ derjenigen, die sich auf verschiedene Weise, sei sie wirtschaftlicher, kultureller oder räumlicher Art, mit Europa eng verbunden glauben, muss man zunächst solche Ansichten unterscheiden, die sich auf fast allen Bereichen einem übermächtigen Europa gegenübersehen, wie dies auf dem Balkan[4] oder in islamisch geprägten Staaten der Fall ist. Dem gegenüber steht die Sichtweise der USA, an deren Mittel diejenigen Europas bzw. der EU politisch bei Weitem noch nicht heranreichen. Letzeres ist bestenfalls der Blickwinkel eines interessierten, aber besorgten *Onkel[s]* (Ash 2004, 133), der mit den Launen eines *pubertierenden Kinde[s]* (Ash 2004, 133) zu kämpfen hat. Der vierte Typ beschreibt die Gesamtheit all der Staaten, die zwar im Rahmen der Globalisierung mit Europa in Kontakt stehen, deren Interesse an Europa sich aber auf wenige für sie relevante Aspekte erschöpft. In dieser Arbeit soll die besondere Aufmerksamkeit diesem letzten, zweifelsfrei als außereuropäisch zu charakterisierenden Typ gelten.

Auffällig ist bei der Betrachtung der verschiedenen Sichtweisen, dass bestimmte Motive immer wieder auftauchen und so gleichsam in doppeltem Sinne von außen reflektiert werden. Dabei kann es sich um historische Aspekte genauso wie um aktuelle Äußerungen handeln, die aufgegriffen und neu akzentuiert werden. So werden geschichtliche Vorlagen verwendet, um die EU mit dem Reich Karls des Großen zu vergleichen (Casey/Rivkin 2001) oder die Bekämpfung des Osmanischen Reiches wiederholt als einigendes Moment der europäischen Völker zu betonen (Džihić 2006, 286; Klunkert 1996, 238f.). Aus

[3] Zudem wird an vielen außereuropäischen Universitäten, deren Webseiten ebenfalls einen Blick auf Europa vermitteln, ein Studiengang „European Studies“ angeboten. Auf diese wird in Kap 4.5 näher eingegangen.

[4] Der Balkan liegt natürlich nur außerhalb des Europas der EU. Die Typisierung ist jedoch durch das im Kapitel zuvor dargestellte Europabild der dort genannten Balkanstaaten wohlbegründet.

zeitgeschichtlicher Perspektive kann es natürlich in den USA nicht ungehört bleiben (Dale 2004), wenn gegenwärtig in Frankreich davon die Rede ist, Europa zum Gegengewicht der Vereinigten Staaten auszubauen (Schott 2005, 95). Daneben finden sich auch noch weitere Bespiele, bei denen dasselbe Bild ohne unmittelbaren Bezug mehrmals verwendet wird. So beklagen sich die Balkanländer und die USA gleichermaßen über die Undankbarkeit Europas, wenn erstere ihre Leistung bei der Abwehr des Islam in vergangenen Jahrhunderten (Džihić 2006, 286) und letztere ihr Engagement bei der Abwehr von gegenwärtigen Bedrohungen (Kagan 2003, 52) nicht ausreichend gewürdigt sehen. Bei außereuropäischen Sichtweisen, die sich ja gerade durch einen geringen Grad an Involviertheit auszeichnen, ist es jedoch unwahrscheinlich, dass diese sich in das Geflecht sich wechselseitig referenzierender Motive einflechten lassen.

Unabhängig davon, ob und inwieweit sich bereits ein europäisches Bewusstsein oder gar eine gemeinsame europäische Identität innerhalb Europas entwickelt hat, kann weiterhin zweifelsfrei festgehalten werden, dass Europa als Begriff auch außerhalb Verwendung findet, obwohl ihm im Sinne des „Anderen“ eine Vielzahl teilweise widersprüchlicher Eigenschaften zugeschrieben wird. Natürlich wird Europa dabei in seinem unmittelbaren Wirkungskreis ausführlicher diskutiert als in Regionen, die sich nicht ständig mit einem sich selbst entwerfenden Europa konfrontiert sehen. Dennoch steht weder zu befürchten, dass eine Betrachtung Europas von außen nicht stattfindet, noch dass sich im weiteren Verlauf der Untersuchung durch allzu konforme Sichtweisen triviale Befunde ergeben. Vielmehr darf die nach außen hin abnehmende Anzahl an Europabildern als zusätzliche Motivation gelten, diese Lesarten näher zu beleuchten, auch wenn sie eventuell nur auf einige wenige Aspekte beschränkt sind.

Da in den folgenden Kapiteln die Frage zu diskutieren sein wird, welche Methoden sich dazu eignen, den Europabegriff auf außereuropäischen Webseiten zu untersuchen, lohnt sich auch eine Betrachtung der Methoden, die bei den oben vorgestellten Studien verwendet wurden. Neben den erwähnten geschichts- und literaturwissenschaftlichen Untersuchungen, die naturgemäß anhand von Quellen bzw. Primärtexten arbeiten, kommen dabei vorwiegend klassische sozialwissenschaftliche Methoden zum Einsatz. Auf Befragungen stützen sich dabei zum einen qualitative Analysen, die wie im Beispiel von Schott (2005, 74) die Charakterisierung der strategischen Regionalisierung einflussreicher nationaler Interessensvertreter zum Ziel haben. Quantitative Verfahren finden sich, wenn es darum geht, die Wahrnehmung oder Grundstimmung zum Europabegriff in breiten Bevölkerungsschichten zu ermitteln (Riketta/Wagenhut 1998 oder Europäische Kommission

2007). Alle Studien, denen keine eigene Datenerhebung zugrunde liegt, verwenden dabei als Basis für ihre Untersuchungen neben den üblichen Verweisen auf Sekundärquellen Informationen, die in der einen oder anderen Form als Text verfügbar sind. Die Spannweite reicht dabei von Zitaten wichtiger Interessensvertreter, die vorrangig aus deren öffentlichen Äußerungen entnommen werden (z.B. *Policy Review*), über Presseartikel (Fuhrt 1993, 302) bis hin zu Verfassungstexten (Bedi 2004, 77). Dabei ist neben der unmittelbaren Aussage der zitierten Textstellen mitunter auch deren Analyse auf implizite Sinngehalte Gegenstand der Untersuchung (Casey/Rivkin 2001). Prinzipiell steht bei der Betrachtung des Europabegriffs also ein breites Spektrum für die Wahl einer Methode zur Verfügung.

# 3 Grundlagen rechnergestützter Textanalyse

Wenn Texte auf außereuropäischen Webseiten dazu herangezogen werden sollen, um mittels semantischer Analysen zu einer Aussage über den dort verwendeten Europabegriff zu kommen, muss berücksichtigt werden, dass es sich beim World Wide Web und dementsprechend auch bei seiner sozialwissenschaftlichen Erforschung um ein relativ junges Forschungsfeld handelt. Deshalb wird an dieser Stelle der Frage nach der Wahl einer geeigneten Methode für eine solche Untersuchung etwas ausführlicher nachgegangen und die getroffene Wahl entsprechend begründet.

Wie bei allen Medien findet auch im Web Kommunikation nicht im persönlichen Kontakt, sondern nur vermittelt statt. Zum *disembodiement* (Thurlow et al. 2005, 99) kommt allerdings im Vergleich zu anderen Medien auch noch die vollständige Anonymität, die das Schaffen virtueller Identitäten ermöglicht (Thurlow et al. 2005, 100). Insofern Identität aber ohnehin immer ein Prozess und keine Konstante ist, handelt es sich dabei allerdings nur um eine weitere Ausdrucksmöglichkeit ein und derselben Person (Thurlow et al. 2005, 105). Das muss das Ergebnis einer Analyse jedoch nicht notwendigerweise verfälschen[5]. Dennoch soll die Gefahr dadurch so weit wie möglich umgangen werden, dass nur Texte auf Webseiten von Organisationen bzw. Institutionen verwendet werden. Diese sind zwar in der Regel gezielt auf eine Außenwirkung hin angelegt, repräsentieren dafür aber idealerweise die konsistente Sprechweise eines klar definierbaren Akteurs.

[5] Erstaunlicherweise scheint diese Tatsache auch bei Online-Befragungen mehr Vorteile als Nachteile zur Folge zu haben, da die Antworten in Ruhe formuliert werden können und dadurch im Schutz der Anonymität oft ehrlicher ausfallen als unter Zeitdruck im direkten Gespräch, auch wenn sie nicht mehr durch den zusätzlichen Kontext der Gesprächssituation verifiziert werden können (Chayko 2002, 181).

Die Arbeit mit Texten hat in der Geographie unter dem Begriff der Diskursanalyse seit einiger Zeit Fuß gefasst. Mit der Erkenntnis, dass Raum[6] nicht an sich Bedeutung besitzt, sondern erst indem er *sozial, ökonomisch und politisch interpretiert* (Gebhardt et al. 2003, 3) wird, reichten klassische geographische Herangehensweisen zur zutreffenden Charakterisierung von Räumen nicht mehr aus (Gebhardt et al. 2004, 295ff.). Unter dem Schlagwort der *neuen Kulturgeographie* (Gebhardt et al. 2003, VIII) wird seitdem auch versucht nachzuvollziehen, wie, von wem und aus welchen Beweggründen vorherrschende räumliche Konzeptionen konstruiert und am Leben erhalten werden und welche Folgen dies für das gesellschaftliche Zusammenleben hat (Gebhardt et al. 2004, 295). Dabei muss beachtet werden, dass auch der Geograph als Wissenschaftler sich niemals vollständig hinter den Diskurs zurückziehen kann, sondern immer aktiv an der Produktion von Raumbegriffen beteiligt bleibt (Gebhardt et al. 2004, 304). Zwar muss die Diskursanalyse nach Keller et al. (2001), auf den sich Gebhardt et al. (2003, 16) berufen, aus der Sicht der Soziologie vorrangig noch als Neuland verstanden werden, dessen Methodenbildung noch lange nicht abgeschlossen ist (Keller et al. 2001, 15)[7]. Wenn es jedoch in der Forschungspraxis der Nachbarwissenschaften weniger darum geht, die allgemeinen Regeln aufzudecken, denen der gesellschaftliche Gesamtdiskurs als solcher folgt, sondern die Beantwortung einer konkreten Forschungsfrage anhand vorliegender Texte im Mittelpunkt steht, hat es sich als zweckdienlich erwiesen, verwandte Methoden der Inhaltsanalyse hinzuzuziehen (Flick 2006, 279 u. 295; vgl. dazu Chilla 2005, 187f.)[8].

Insofern der Europabegriff, wie bereits aufgezeigt, ein konstruierter ist, müssen Methoden der Inhaltsanalyse bei einer Betrachtung dieses Gegenstandes somit als primärer Bezugspunkt gelten. Auch wenn solche Studien gewöhnlich rein qualitativ durchgeführt werden (z.B. Chilla 2007, 16, Schott 2005, 74 oder Christmann 2005, 314ff.), darf das Vorhaben einer Inhaltsanalyse nicht von vorn herein darauf beschränkt werden, zumal bei einer Untersuchung von Webinhalten einige Besonderheiten zu beachten sind,

---

6 Der Begriff „Raum“ ist mehrdeutig. In diesem Zusammenhang ist damit der zunächst völlig neutrale Raumbegriff als Lagebezeichnung gemeint, der erst durch die (in der Regel implizite) Zuweisung einer Extension und weiterer Eigenschaften symbolisch aufgeladen wird. Eine Übersicht über die verschiedenen Raumkonzepte findet sich z.B. bei Weichhart (1999, 67).

7 Allgemein lassen sich bei der Diskursanalyse hermeneutisch-interpretative, strukturalistische und poststrukturalistische Ansätze unterscheiden (Mattissek/Reuber 2004, 234). Der hier vorliegende Vergleich von Europabegriffen verschiedener Akteure ist dabei zweifellos handlungszentriert und damit dem interpretativen Ansatz zuzuordnen.

8 Natürlich können Diskursanalyse und Qualitative Inhaltsanalyse nicht einfach gleichgesetzt werden. Ihr Verhältnis bleibt immer dasjenige von formuliertem Anspruch und methodischer Realisierung. Im Folgenden soll jedoch die Qualitative Inhaltsanalyse als mögliche Realisierung der Diskursanalyse als Vorlage für ein generalisiertes Ablaufmodell des üblichen Vorgehens dienen, an dem sich das weitere Vorgehen orientieren wird.

die bereits die Erstellung eines geeigneten Textkorpus im Vergleich zu anderen Medien erschweren. Das größte Problem besteht dabei vermutlich darin, dass diskursive Ereignisse nicht mehr wie bei Printmedien mit einer gewissen zeitlichen Verzögerung einen einfach zu archivierenden Niederschlag finden, sondern in Form von *web storms* (Schneider/Foot 2005, 162) innerhalb weniger Stunden zu einer einschneidenden *alteration of intertextual and interlinked web objects* (Schneider/Foot 2005, 162) führen können. Dem Forscher stellt sich also gleich ein mehrfaches Problem: Da das World Wide Web über keine Historie seiner vorherigen Zustände verfügt (*ephemerality* nach Schneider/Foot 2005, 166), müssen Inhalte für eine genauere Untersuchung offline verfügbar gemacht werden, um diachrone Untersuchungen zu ermöglichen. Dazu müssen allerdings geeignete Analyseeinheiten definiert werden, was wiederum durch die Dynamik der Webinhalte erschwert wird. Eine genaue Abgrenzung der zu beobachtenden Webseiten erlaubt dabei zwar eine konsistente Reproduzierbarkeit der Forschungsergebnisse, ignoriert jedoch völlig die Möglichkeit, dass sich Webinhalte und Links einer Webseite kurzfristig und unvorhersehbar ändern können (Schneider/Foot 2005, 162). Davon sind sowohl die Stichwortsuche in einer Suchmaschine als auch das Identifizieren möglicherweise relevanter Akteure als Mittel zur Abgrenzung des Textkorpus betroffen (Schneider/Foot 2005, 163f.). Da mit dem Europabegriff zumindest der Schlüsselbegriff als einschränkender Parameter bereits vorgegeben ist, würde sich, abgesehen von den genannten Unwägbarkeiten, die Anfrage an eine Suchmaschine allerdings hervorragend zur Abgrenzung eines Textkorpus für die weitere Analyse eignen. Mit Rücksicht auf die beschränkten Zeitressourcen dieser Studie und um die Gefahr unvorhersehbarer Änderungen auszuschließen, soll dabei eine einmalige, gleichzeitige Anfrage nach dem Stichwort „Europa“ auf den Webseiten der gewählten Institutionen im Sinne einer synchronen Untersuchung zur Generierung des Textkorpus ausreichen, da vorrangig die Unterschiede zwischen den verschiedenen Sprechweisen und nicht die Entwicklung einer bestimmten Sprechweise über Europa analysiert werden sollen.[9]

Zumindest für das Zusammenstellen eines geeigneten Korpus scheint der Einsatz eines rechnergestützten, automatischen Verfahrens also empfehlenswert, da das Korpus nicht nur mittels einer Suchanfrage definiert, sondern anschließend möglichst zeitnah archiviert werden muss, um dem Kriterium der Synchronität zu genügen. Doch auch eine daran anknüpfende Inhaltsanalyse kann durch Rechnereinsatz sinnvoll unterstützt werden. So

[9] Ein Beispiel für eine diachrone Untersuchung regelmäßig wiederkehrender und damit gut antizipierbarer diskursiver Ereignisse („Weihnachten“ bzw. „Fasching“) mittels automatischer Anfrage an eine Suchmaschine findet sich bei Kiefer et al. (2006).

können quantitative Verfahren heute weitgehend automatisch komplexe statistische Analysen durchführen (Krippendorff 2004, 289ff.; vgl. auch Kelle 2007, 488), aber auch zur Unterstützung interpretativ-hermeneutischer Methoden hat sich ein weites Spektrum an Softwareprodukten ausgebildet (Krippendorff 2004, 303ff., Kelle 2007, 362).

Sowohl das qualitative als auch das quantitative Forschungsparadigma haben unbestreitbare Stärken. Während rein qualitative Ansätze anhand von Texten behutsam versuchen, diese in Abhängigkeit von den jeweiligen individuellen Absichten und gesellschaftlichen Rahmenbedingungen zu erklären (Gebhardt et al. 2004, 297f.) und somit die *Prinzipien einer sozial konstruierten Welt auf[zu]decken* (Chilla 2005, 187), können quantitative Ansätze, und hier insbesondere Methoden der automatischen Textanalyse, für sich beanspruchen, jederzeit reproduzierbare, konsistente und übersichtliche Ergebnisse in relativ kurzer Zeit zu erzeugen (Adolphs 2006, 7f.; vgl. auch Lamnek 1995, 177). Beide Methoden haben jedoch auch Schwächen. Denn wo qualitative Inhaltsanalyse dem Vorwurf der Beliebigkeit begegnen muss (Gebhardt et al. 2004, 297), ist durch eine quantitative Analyse, so komplex sie auch sein mag, noch nichts erklärt. Krippendorff (2004) spricht folgerichtig von einer *mistaken dichotomy*, wenn er feststellt: *For the analysis of texts, both are indispensable* (Krippendorff 2004, 87). Daher erstaunt es nicht, dass auf dem Gebiet der Inhaltsanalyse kombinierte Verfahren schon länger erfolgreich eingesetzt werden (z.B. Xenos/Foot 2005). Dabei scheinen zwei Kombinationsmöglichkeiten[10] die jeweiligen Stärken qualitativer und quantitativer Methoden besonders gut auszunutzen. Zum einen können die quantitativen Analysen einer qualitativen Studie im Sinne eines *Screening[s]* (Chilla 2005, 188) behilflich sein, um sich einen Überblick über das Aufkommen bestimmter Schlüsselbegriffe zu bilden (Chilla 2005, 188; vgl. Adolphs 2006, 8), zum anderen können die Befunde einer qualitativen Studie durch quantitative Methoden nachträglich bestätigt und erhärtet werden (Adolphs 2006, 8; vgl. Chilla 2005, 188).

Auch bei einer solchen synergetischen Arbeitsteilung bleiben quantitative Methoden natürlich weitgehend auf theorietestende, falsifizierende Verfahren beschränkt, während nur qualitative Ansätze neue Theorien generieren können (Glaser/Strauss 1998, 24; vgl. Lamnek 1995, 191 u. 199). Die größte Herausforderung für den qualitativ Forschenden ist dabei der Anspruch, neue Theorien frei von Vorwissen und lediglich basiert auf dem vorurteilsfreien Fremdverstehen des Textmaterials zu entwickeln (Lamnek 1995, 199). Da

[10] Ein allgemeiner Überblick über verschiedene Möglichkeiten zur Integration qualitativer und quantitativer Methoden in ein Forschungsdesign findet sich bei Flick (2006, 383ff.).

trotz aller Vorteile eine kombinierte Studie im zeitlichen Rahmen dieser Arbeit nicht bewältigt werden kann, wird daher im Folgenden der Versuch unternommen, die bei Verfahren der automatischen Textanalyse trivialerweise gegebene Wertfreiheit dahingehend auszunutzen, ein Verfahren zu entwickeln, mit dem zumindest erste Hypothesen über die textimmanenten Sinngehalte automatisch generiert werden können. Ein solches Verfahren soll und kann keinesfalls die Aufgaben einer qualitativen Studie ersetzen[11], sondern diese in einem höheren Maße als bisher möglich unterstützen.

Um die gewünschte Anschlussfähigkeit an die etablierten qualitativen Arbeitsweisen zu gewährleisten, wird daher im Folgenden anhand verschiedener Verfahren qualitativer Inhaltsanalyse (Kap. 3.1) ein Ablaufmodell der typischen Arbeitsphasen erarbeitet. Dabei muss klar sein, dass aus der Fülle möglicher Verfahren nur einige wenige Vertreter vorgestellt werden können, die jedoch ein möglichst weites Spektrum qualitativen Arbeitens am Text repräsentieren sollen[12]. Anhand dieses Ablaufmodells können Schnittstellen zur Automatisierung von Teilaufgaben formuliert werden (Kap. 3.2). Im Anschluss daran gilt es aus dem mittlerweile reichhaltigen Inventar von Methoden zur automatischen Textanalyse Verfahren auszuwählen, die diese Teilaufgaben wirksam unterstützen können (Kap. 3.3). Die Grundlagen der Funktionsweise dieser Methoden werden in Kap 3.4 eingeführt und erläutert.

## *3.1 Verfahren qualitativer Inhaltsanalyse als methodischer Bezugspunkt*

Bevor einzelne Verfahren der qualitativen Inhaltsanalyse[13] vorgestellt werden, soll an dieser Stelle ein kurzer Hinweis auf die verwendete Terminologie erfolgen, die sich nicht ad hoc erschließt. Zentrales Instrument bei allen Methoden ist das sogenannte Kodieren. Dabei werden Texteinheiten, die ein einzelnes Wort, einen Satz, aber auch eine ganze Textpassage beinhalten können, mithilfe von Kodes annotiert (Flick 2006, 257f. u. 263). Diese Kodes bestehen ihrerseits zumeist aus knappen Nominalphrasen, die die durch sie gekennzeichnete Textstelle erläutern, interpretieren, bewerten oder auch einfach

[11] Dass Computer Texte nicht verstehen, sondern komplexe Berechnungen auf der Grundlage von geeigneten Repräsentationen durchführen, deren Ergebnis vom Menschen bestenfalls konsistent zu den Quelldaten interpretiert werden kann, ist seit geraumer Zeit hinlänglich bekannt (Winograd/Flores 1986, Dreyfus 1989; vgl. dazu Kelle 2007, 488f.).

[12] Einen Überblick über übliche Verfahren liefert z.B. Flick (2006, 310f.).

[13] Flick (2006, 310f.) versteht unter qualitativer Inhaltsanalyse explizit nur das Verfahren nach Mayring (2000) gleichen Namens. Im Weiteren soll dieser Begriff jedoch allgemein die Gesamtheit aller mittels qualitativer Methoden am Text arbeitenden Verfahren kennzeichnen, die bei Flick (2006, 310f.) als Verfahren der Textinterpretation bezeichnet werden.

zusammenfassen[14]. Diese können dann unter Kategorien subsumiert und zur Ausbildung einer gegenstandsbezogenen Theorie verwendet werden (Flick 2006, 258). Durchaus relevant für die Beantwortung der Fragestellung (Jäger 2004, 176ff.), nicht aber für die Anwendbarkeit der Methode an sich ist dabei, ob das Textmaterial bereits vorliegt oder ob und in welchem Rahmen es für die Untersuchung eigens erhoben wurde (Lamnek 1995, 176ff.).

Obwohl unter dem Begriff der *Kritischen Diskursanalyse* (Jäger 2004) naturgemäß Erkenntnisse über den gesamtgesellschaftlichen Diskurs im Vordergrund stehen (Jäger 2004, 166), verwendet auch die gleichnamige Methode nach Jäger (2004) im Detail inhaltsanalytische Methoden. Dabei werden vor der eigentlichen Analyse zunächst ergänzende Informationen zum Textkorpus akquiriert. Darunter fallen neben der Positionierung des Autors im Autorennetzwerk auch Beobachtungen zu formalen und gestalterischen Mitteln, die die Textaussage mit unterstützen können (Jäger 2004, 176ff.). In einem darauf folgenden Schritt wird mithilfe akribischer sprachwissenschaftlicher Analysen anhand von Wortschatz und Stilistik, vor allem aber anhand von Komposition und Logik die Argumentationsstrategie des Textes rekonstruiert (Jäger 2004, 179ff.). Die abschließende Analyse berücksichtigt neben der ideologischen Färbung auch den Adressaten und versucht, Rückschlüsse auf die beabsichtigte Wirkung[15] im diskursiven Kontext zu ziehen (Jäger 2004, 184ff.). In diesem Zusammenhang ist vor allem interessant, dass Jäger als Sprachwissenschaftler vor allem Substantive und Pronomen als Träger der Information innerhalb von Texten ansieht (Jäger 2004, 182f.).

Ein sehr strukturierendes Verfahren der qualitativen Inhaltsanalyse, das seine Anbindbarkeit an quantitative Arbeitsschritte ausdrücklich betont (Mayring 2000, 42), stellt Mayring (2000) vor. Dabei werden nach der Abgrenzung des Textkorpus zunächst formale Charakteristika der gewählten Texte sowie die Rahmenbedingungen, innerhalb derer sie produziert wurden, als ergänzende Informationen hinzugezogen (Mayring 2000, 47; Lamnek 1995, 207). Erst danach wird die genaue Fragestellung festgelegt. Diese darf durchaus geleitet von Hypothesen weiter differenziert werden, muss aber offen formuliert bleiben (Mayring 2000, 52; Lamnek 1995, 208). Auf der Grundlage eines vorformulierten initialen Kategoriensystems, mithilfe dessen die einzelnen Texteinheiten annotiert werden können, werden nun in einem iterativen Prozess die Kernaussagen des Textes herausgearbeitet. Dabei wird das Kategoriensystem durch induktive Kategorienbildung gemäß der

[14] Beispiele für Kodes finden sich z.B. bei Flick (2006, 260ff.).

[15] Diese Sichtweise wird von Mattissek/Reuber (2004, 234) als intentionales Sprachspiel bezeichnet und folgerichtig den handlungsorientierten Ansätzen zugeordnet.

Textinhalte ständig modifiziert und erweitert (Mayring 2007, 472, Mayring 2000, 74ff.). Hier finden drei sich untereinander ergänzende Analysetechniken Anwendung. Bei der Zusammenfassung wird versucht, Textpassagen in einem überschaubaren Kurztext zu paraphrasieren. Die Explikation sieht die Textstelle dagegen in ihrem Kontext und versucht, daraus weitergehende Informationen zu ihrem Verständnis abzuleiten. Die Strukturierung schließlich dient dazu, den Text gezielt nach bestimmten Teilaspekten zu untersuchen (Mayring 2000, 59ff.; Lamnek 1995, 209ff.). Der Text wird dabei mithilfe der verschiedenen Techniken immer wieder durchgearbeitet, bis keine neuen (Unter-)kategorien mehr gefunden werden können (Mayring 2000, 84). In einem abschließenden Schritt werden die gefundenen Ergebnisse auf die Forschungsfrage hin interpretiert und präsentiert (Lamnek 1995, 215).

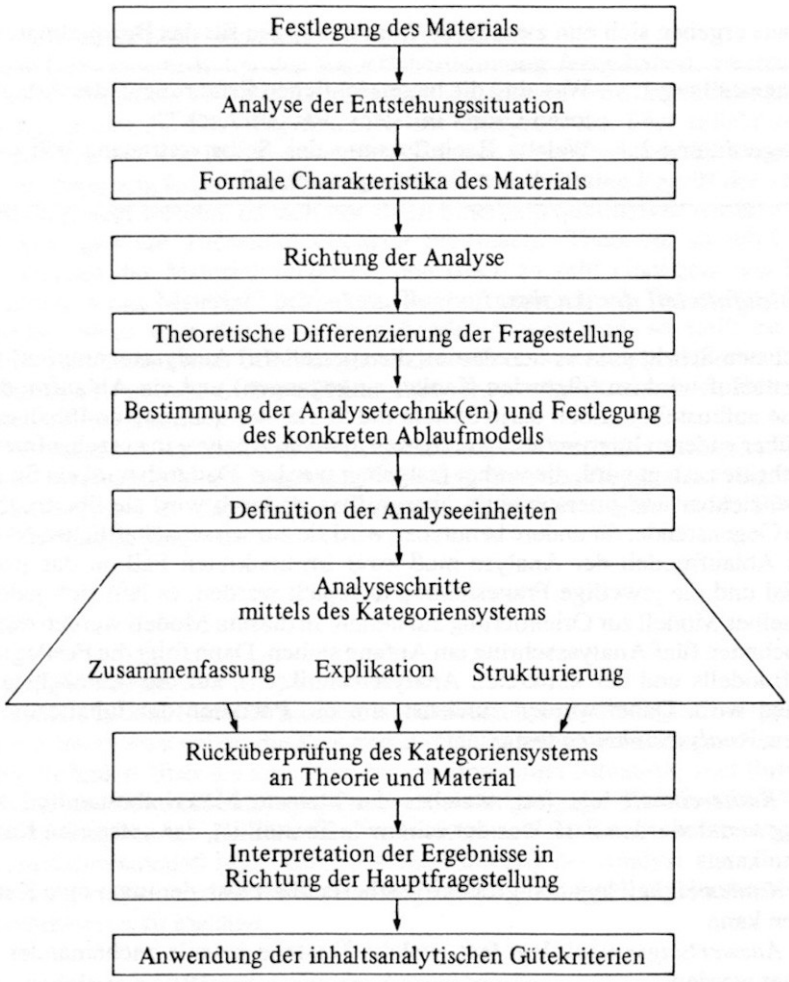



*Abb. 2: Allgemeines inhaltsanalytisches Ablaufmodell nach Mayring (2000, 54)*

Gemäß der Terminologie von Lamnek (1995, 216) handelt es sich bei Mayrings Ansatz zwar um eine qualitative Inhaltsanalyse, die aber aufgrund ihres zusammenfassenden und

abstrahierenden Charakters eher reduktiv als explikativ arbeitet. Auch Mayrings Interesse gilt dabei vorrangig der Herausarbeitung der wesentlichen sinntragenden Elemente, wenngleich in Form von Kategorien, die teilweise bereits vor der Untersuchung definiert sein können und durch induktive Kategorienbildung anhand der Texte schrittweise zu einem Kategoriensystem weiterentwickelt werden.

Eine weitere häufig verwendete Methode[16] zur Analyse von Textinhalten folgt der ursprünglich von Glaser/Strauss (1998) explizit zur Theoriegenerierung entwickelten Grounded Theory (Glaser/Strauss 1998, 12). Böhm (2007) bezeichnet das Verfahren als *Kunstlehre[, die] nicht rezeptartig zu erlernen ist* (Böhm 2007, 476). Im Gegensatz zur Methode nach Mayring (2000) verzichtet die Grounded Theory also bewusst auf die Angabe einer bestimmten Reihenfolge, in der die verschiedenen Schritte ausgeführt werden, da der Text vielmehr wiederholt *aufgebrochen, konzeptualisiert und auf neue Art zusammengesetzt werden* (Flick 2006, 259) soll. Auch wenn es kein schematisches Vorgehen gibt, lässt sich doch ein klares Instrumentarium elementarer Teilaufgaben erkennen.

Nach einigen vorbereitenden Schritten, die unter anderem dazu dienen, das eigene Vorwissen um den Forschungsgegenstand darzulegen und sich einen ersten Eindruck vom Material zu verschaffen, beginnt die eigentliche Auswertung (Krotz 2005, 169ff.). Besonderer Wert wird dabei auf eine lückenlose Dokumentation der eigenen Forschungsarbeit in Form von Memos gelegt. Dadurch soll nicht nur die Nachvollziehbarkeit gewährleistet werden, sondern auch das Potential spontaner Einfälle voll ausgeschöpft werden (Krotz 2005, 173f.).

Als zentrale Verfahren stehen drei Arten des Kodierens zur Verfügung. Beim offenen Kodieren werden zunächst relevante Texteinheiten auf Konzepte verdichtet, die als Kodes an diese annotiert werden (Flick 2006, 259; Krotz 2005, 181). Dabei sollen zunächst nach Möglichkeit In-vivo-Kodes zum Einsatz kommen, also Begriffe, die unmittelbar im Text Verwendung finden (Böhm 2007, 478; Flick 2006, 263). Die große Zahl dadurch erzeugter Kodes wird noch in diesem Teilschritt unter allgemeinere Kategorien subsumiert (Flick 2006, 263), die sich aber ebenfalls in unmittelbarer semantischer Nähe zum Forschungsgegenstand bewegen sollen (Krotz 2005, 181). Das offene Kodieren entspricht daher der Sache nach der induktiven Kategorienbildung bei Mayring (2000, 75).

[16] Im Rahmen einer gemäß der Grounded Theory durchgeführten Studie zeigt Christmann (2005) am Beispiel Dresdens, inwieweit eine Stadt mit ihren Monumenten und den über sie produzierten Bildern, Mythen und Texten eine Art kulturelles Gedächtnis einer Gesellschaft darstellt, das auch die Wahrnehmung ihrer Bewohner nachhaltig beeinflusst.

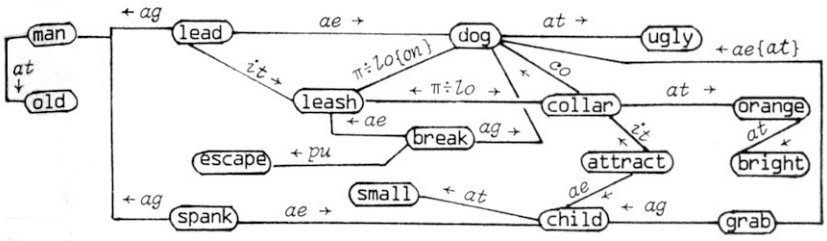
Bei der zweiten Art, dem axialen Kodieren, wird die Beziehung der resultierenden Kodes bzw. Kategorien untereinander untersucht (Flick 2006, 265). Diese Beziehungen können dabei sowohl formaler als auch inhaltlicher Art sein (Böhm 2007, 479). Ziel dieses Schrittes ist eine auf den Forschungsgegenstand bezogene Hierarchie von Kategorien (Krotz 2005, 183), wobei ein besonderes Augenmerk auf den zentralen Kategorien liegt, die für das Handeln und Urteilen der Textproduzenten von Bedeutung sind. Dabei ist es von höchster Wichtigkeit, das Kategoriensystem mit Fundstellen zu belegen und so direkt am Text zu verifizieren (Böhm 2007, 479; Flick 2006, 265).

Beim selektiven Kodieren schließlich werden die zentralen Kategorien oder idealerweise die zentrale Kategorie ermittelt (Krotz 2005, 184). Ein Indiz dafür ist ein hoher Grad an Vernetzung mit anderen Kategorien, der durch Netzwerkdarstellungen visualisiert werden kann (Böhm 2007, 482; vgl. auch Krotz 2005, 186). Ausgehend von den Schlüsselkategorien kann nun eine Theorie über die Zusammenhänge in den untersuchten Texten formuliert werden (Krotz 2005, 184). Da sich diese erste Theorie an weiteren Texten bewähren muss, wird über alle drei Arten des Kodierens in beliebiger Reihung so lange iteriert, bis kein Erkenntnisgewinn mehr eintritt (Flick 2006, 268). In der ursprünglichen Konzeption wird das Vorgehen durch die Formulierung einer soziologischen Theorie abgeschlossen (Glaser/Strauss 1998, 16).

Trotz des eigenen Anspruchs als Kunstlehre weist die Grounded Theory bisher die besten Ansatzpunkte für eine Ausweisung von automatisierbaren Teilschritten auf. Die Suche nach In-vivo-Kodes deckt sich weitgehend mit der Suche nach Substantiven bei Jäger (2004) bzw. initialen Kategorien bei Mayring (2000). Hinzu kommt jedoch der aus der Sicht der semantischen Informationsverarbeitung sehr interessante Schritt, die gefundenen Kodes und Kategorien nicht in Form eines hierarchisierten Kategoriensystems (Mayring 2000), sondern als semantisches Netzwerk darzustellen.

Derartige Visualisierungen sind keine Errungenschaft jüngeren Datums, sondern haben schon früh in erstaunlich formaler Ausprägung Eingang in die Diskursanalyse gefunden. Bereits Beaugrande (1980, 10) moniert, dass sich Textanalysen zuvor primär auf den Satz als zentrale Einheit bezogen haben, der in seinen Eigenschaften nur schwer zu fassen und zu formalisieren ist. Dass semantische Bezüge auch über weite Entfernungen hinweg hergestellt werden können, führt ihn zu folgender Überlegung: *If we define STRUCTURE as a relation between at least two systemic elements in occurrence, it is clear that a theory of language use should be centered upon the notion of CONNECTIVITY* (Beaugrande 1980,

10[17]). Ein wirkliches Aufbrechen der Textstrukturen zur Rekonstruktion semantischer Bezüge erfordert also andere als wieder nur textuelle Herangehensweisen. Folgerichtig entwirft er ein Verfahren, *[to] arrange this much content in a network of concepts and relations* (Beaugrande 1980, 207). Diese Netzwerkdarstellungen werden von Beaugrande (1980, 207ff.) dazu eingesetzt, um die übliche Terminologie eines Begriffsfeldes oder eine Folge von Ereignissen (vgl. Abb. 3) zu erarbeiten, aber auch um ein Sonett mit allen seinen Rückbezüglichkeiten und Andeutungen darzustellen und zu analysieren.




Key: *ae*: affected entity; *ag*: agent of; *at*: attribute of; *co*: containment of;*it*: instrument of; *lo*: location of; *pu*: purpose of; *π*: proximity

*Abb. 3: Netzwerkdarstellung einer Folge von Ereignissen (Beaugrande 1980, 207)*

Derartige Strukturen sind als vollständiges Textabbild gehalten und dadurch natürlich sehr aufwändig. Auch wenn man den Textinhalt gemäß Jäger (2004, 182f.) auf Substantive oder nach einer geeigneten Abstraktion auf Kodes oder Kategorien beschränkt, kann ein solches semantisches Netz noch ein ausdrucksmächtiges Werkzeug zur Darstellung textimmanenter Sachverhalte sein.

Eine Weiterentwicklung und Neuakzentuierung des Vorgehens der Grounded Theory stellt Diaz-Bone (2002, 181ff.) mit seinem auf distinktionstheoretischen Überlegungen (Diaz-Bone 2002, 15ff.) basierenden Verfahren vor. Auch bei dieser Methode wird im Rahmen einer ersten Oberflächenanalyse zunächst nach den im Text enthaltenen Begriffen gesucht, die dann nach und nach in ein Kodesystem eingearbeitet werden. Anschließend wird wie auch bei den vorherigen Ansätzen versucht, das Beziehungssystem zwischen den verwendeten Konzepten und Objekten zu rekonstruieren und dadurch Regelmäßigkeiten der diskursiven Praxis aufzudecken (Diaz-Bone 2002, 203). Wesentlicher Unterschied zu den bereits vorgestellten Ansätzen ist dabei jedoch das Untersuchungsziel. Nicht der Kampf um Macht(mittel) steht im Zentrum des Interesses, sondern vielmehr *die zweite Wirklichkeit der symbolischen [...] Strukturen, die aus den konfligierenden [...] Weltdeutungen sozialer Gruppen besteht* (Diaz-Bone 2002, 16). Dies hat natürlich zur Folge, dass anstelle eines einzigen mehrere verschiedene Korpora verwendet werden, die bewusst so arrangiert werden, dass sie unterschiedliche Kultur-

[17] Großschreibung im Original; Unterstreichung ersetzt die Kursivschreibung des Originals.

bzw. Lebenswelten mit ihren charakteristischen Bedeutungszuschreibungen repräsentieren[18]. Bereits bei der Erarbeitung des Kodesystems gilt die besondere Aufmerksamkeit Begrifflichkeiten, aber auch Metaphern und Symbolen, die in einer grundlegenden Opposition zu verbreiteteren Sprechweisen stehen (Diaz-Bone 2002, 203). Die vorbereitende Textarbeit mündet schließlich in eine vergleichende Gegenüberstellung der Bedeutungszuschreibungen der verschiedenen Kulturwelten auf den Ebenen von Ethik, Ästhetik und Gefühlsstruktur (Diaz-Bone 2002, 204f.).

Insofern es sich auch bei verschiedenen Sichtweisen auf Europa um konkurrierende Bedeutungszuschreibungen handelt, würde sich eine solche Methode im Rahmen einer rein qualitativen Vorgehensweise gut zur Beantwortung einer solchen Fragestellung eignen. Zumindest die Strategie der voneinander getrennten Analyse verschiedener Korpora und die anschließende vergleichende Gegenüberstellung der Ergebnisse wird aber auch in dieser Arbeit Verwendung finden (vgl. Kap 5).

## *3.2 Generalisiertes Ablaufmodell qualitativer Inhaltsanalyse*

Auch wenn sich die zuvor beschriebenen Verfahren der qualitativen Inhaltsanalyse sowohl in ihrem methodologischen Anspruch als auch in den Anwendungsbereichen, für die sie konzipiert wurden, leicht unterscheiden, lässt sich doch, was das methodische Vorgehen bei der Analyse der Texte betrifft, eine gewisse Ähnlichkeit der Arbeitsschritte untereinander nicht bestreiten. Am Anfang steht immer die Zusammenstellung eines Textkorpus, das mit Zusatzinformationen über den Entstehungshintergrund des Textmaterials angereichert sein darf. Danach werden die wesentlichen argumentationstragenden Konzepte aus den Texten herausgearbeitet, wobei teilweise vorgefertigte Kategorien auf den Text angewendet werden (vgl. Verfahren nach Mayring 2000), teilweise aber auch im Text unmittelbar vorkommende Begriffe als Ausgangsbasis dienen (Grounded Theory). Nach wiederholter Abstraktion der Konzepte zu Kategorien und der Verifizierung ihrer Anwendbarkeit auf das Textmaterial liegen nur noch einige wesentliche Schlüsselkategorien vor, deren Zahl nun hinreichend klein ist, um eine Theorie über die textimmanenten Sachverhalte formulieren zu können. Grundlage dafür sind die semantischen Verbindungen zwischen den Schlüsselkategorien, die das Verhältnis beschreiben, in dem diese zueinander stehen. Ein wichtiges Hilfsmittel sind dabei Netzwerkdarstellungen, die diese Abhängigkeiten in graphischer Form visualisieren.

[18] Diaz-Bone (2002, 209ff.) verdeutlicht das Vorgehen am Beispiel der Musik-Genres Heavy Metal und Techno bzw. der durch sie geschaffenen Lebenswelten.

Abb. 4[19] zeigt ein vereinfachtes allgemeines Ablaufmodell qualitativer Inhaltsanalyse. Natürlich stellen Rückkopplungen in Form von mehrfachen Durchläufen derselben Schrittfolge ein wesentliches Element qualitativer Forschung dar, um laufend die Gültigkeit der Zwischenergebnisse für die Ausgangstexte zu überprüfen. Es soll jedoch illustriert werden, welche Phasen üblicherweise mindestens einmal durchlaufen werden, bevor das Textmaterial so weit aufbereitet ist, dass es zur Beantwortung der Forschungsfrage herangezogen werden kann. Eine damit konsistente ausführlichere Darstellung findet sich z.B. bei KRIPPENDORFF (2004, 86).

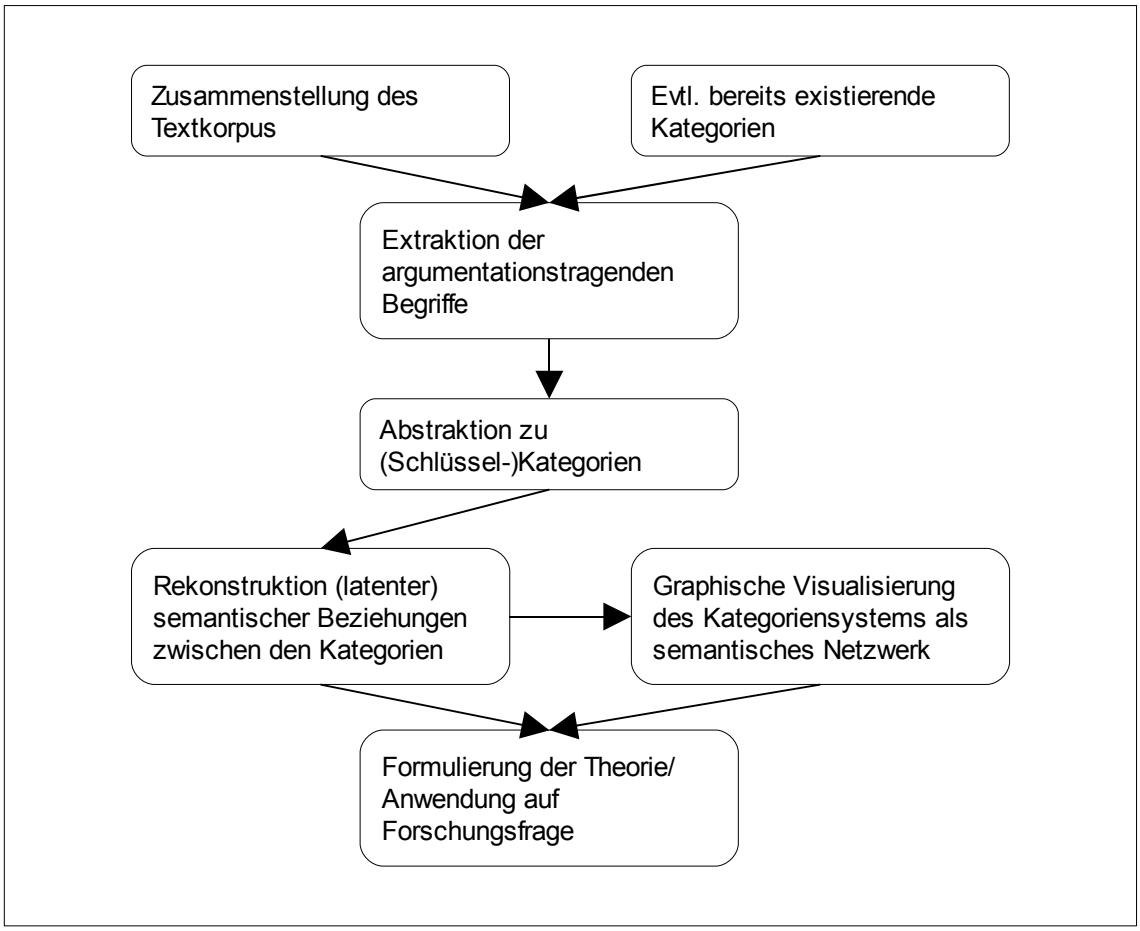



*Abb. 4: Generalisiertes Ablaufmodell qualitativer Inhaltsanalyse*

Die bereits erwähnten handelsüblichen Programme zur interpretativ-hermeneutischen Datenanalyse[20] unterstützen den Forscher bereits heute bei seiner Tätigkeit mit einer großen Spannbreite von Funktionen, die vom Einfügen von Querverweisen über die Annotation von Kodes und die Verwaltung von Memos bis hin zu statistischen Auswertungen und der graphischen Darstellung der Verbindungen zwischen den Kategorien in Form eines semantischen Netzwerks reichen (KELLE 2007, 366; KRIPPENDORFF

[19] Diese und alle folgenden Abbildungen der Arbeit sind eigene Entwürfe.

[20] Bekannte Namen sind z.B. ATLAS/ti, N-Vivo oder NUDIST (FLICK 2006, 373; KRIPPENDORFF 2004, 305 bzw. KELLE 2007, 486).

2004, 303ff.). All diese Funktionen stellen dabei jedoch lediglich effiziente Werkzeuge zur Verwaltung des Textkorpus und der bei seiner Analyse produzierten Daten dar. Die äußerst zeitintensiven eigentlichen Auswertungsschritte müssen dabei komplett vom Forschenden durchgeführt werden.

Da alle Verfahren qualitativer Inhaltsanalyse ohnehin iterativ arbeiten, könnte ein automatisiertes, hypothesengenerierendes Verfahren im Sinne eines konsequent weiterentwickelten Screenings einen wertvollen Beitrag leisten, ohne die Qualität der Untersuchung im Geringsten zu beeinträchtigen. Zunächst könnte es dazu eingesetzt werden, große digitale Textbestände auf ihre Relevanz für ein anstehendes Forschungsprojekt zu überprüfen. Liegt bereits ein definiertes Textkorpus vor, könnte ein solches automatisiertes Verfahren helfen, neben einem ersten Überblick über die zentralen Konzepte eines Textes und deren Fundstellen auch einen Einblick in deren semantische Vernetzung zu erhalten und diese geeignet zu visualisieren. Selbst wenn schon Zwischenergebnisse vorliegen, können diese mit den Ergebnissen der automatischen Analyse verglichen und so verifiziert werden.

## *3.3 Geeignete Verfahren automatisierter Textanalyse*

Will man Aufgaben, die üblicherweise nur durch qualitative Forschung bewältigt werden (können), mittels rechnergestützter und insofern quantitativer Methoden lösen, muss man sich zunächst die weitreichenden Folgen eines solchen Vorhabens bewusst machen. Von zentraler Bedeutung ist dabei zweifelsohne, dass dies einen *shift in methodological emphasis, from solving the human problem of achieving reliable coding for large volumes of text [...] to solving the computational problem of preserving relevant readings of the texts* (Krippendorff 2004, 260) bedeutet. Die bereits thematisierte Wertfreiheit und Nachvollziehbarkeit rechnergestützter Verfahren geht also zu Lasten des semantischen Gehaltes der Texte. Auch wenn solche Verfahren im Sinne eines Screenings in den qualitativen Forschungsprozess eingebettet sind, wäre es also kontraproduktiv *to bypass human coders altogether* (Krippendorff 2004, 259). Als Lösung kommen nur semi-automatische Systeme in Frage (Lobin/Mehler 2004, 1), die entweder für die Probleme eng begrenzter Fachdomänen optimiert sind oder eine Vielzahl allgemeiner, aber relativ einfacher Hilfsschritte zur Verfügung stellen (Krippendorff 2004, 261). Der Vorteil bei Ansätzen letzterer Art ist fraglos ihre Verständlichkeit und intersubjektive Nachvollziehbarkeit: *Content analysts have no difficulty comprehending what they do and*

*employing them where needed – without surrendering their judgement to their results* (Krippendorff 2004, 261). Unter dieser Rücksicht sollen im Folgenden nur solche Verfahren automatischer Textanalyse verwendet werden, die sich in ihrer Funktionsweise leicht nachvollziehen lassen und nicht von vorn herein auf ein bestimmtes fachliches Einsatzgebiet beschränkt sind.

Bevor geeignete Techniken ausgewählt werden können, gilt es zunächst die zum Teil heterogene Terminologie zu erläutern, mit der in diesem interdisziplinären Feld gearbeitet wird. Ansätze zur automatischen Textanalyse werden zumeist unter den allgemeinen Begriffen der *computer-aided text analysis* (Krippendorff 2004, 261) oder der *electronic text analysis* (Adolphs 2006, 1) behandelt. Konkrete Technologien zu deren Umsetzung entstammen zumeist dem Grenzgebiet zwischen *Information Retrieval* (z.B. Beaza-Yates/Ribiero-Neto 1999), *Text Mining* (z.B. Feldman/Sanger 2007) und *Natural Language Processing* (Adolphs 2006, 2) und somit der Schnittstelle zwischen linguistischen (Adolphs 2006, 2) und statistischen, frequenzbasierten (Krippendorff 2004, 289ff.) Methoden.

In allen drei Bereichen werden ähnliche Verfahren eingesetzt, die sich jedoch vorrangig in ihrem Einsatzgebiet unterscheiden. Im Information Retrieval werden effiziente Lösungen gesucht, um verfügbare Informationsbestände durch Schlüsselbegriffe in der Form geeignet zu indizieren, dass für den Nutzer auf eine Anfrage hin relevante und nützliche Dokumente bereitgestellt werden können (Beaza-Yates/Ribiero-Neto 1999, 1). Früher hauptsächlich auf Bibliotheken beschränkt, hat das Information Retrieval nach der Entstehung des World Wide Web durch den Einsatz in Suchmaschinen erheblich an Bedeutung gewonnen (Beaza-Yates/Ribiero-Neto 1999, 2). Natural Language Processing stellt zunächst Verfahren zur Verfügung, mittels derer anhand gegenwartssprachlicher Textkorpora *a better understanding of language in use* (Adolphs 2006, 2) erlangt werden kann. Die daraus gewonnenen Erkenntnisse über die Struktur von Sprache können dann im Anschluss daran dafür verwendet werden, Werkzeuge bereitzustellen, die auch in Nachbarwissenschaften dazu verwendet werden können, anhand der strukturellen Merkmale von Texten Rückschlüsse auf deren semantische Bezüge zu ziehen.

Nicht so klar war dagegen lange, wie die Aufgaben des Text Minings gegen diese beiden Richtungen abgegrenzt werden sollen, da sie häufig auf Ergebnisse des Information Retrieval bzw. Natural Language Processing aufbauen. Als erster betont Hearst (1999, 3ff.) den zusätzlichen explorativen Charakter eines *Text Data Mining[s]* (Hearst 1999, 5; vgl. Mehler 2004, 330), bei dem vorher unbekannte Zusammenhänge aufgedeckt werden können. Fälle, in denen diese Explorationsaufgabe vollständig

unabhängig vom Nutzer bearbeitet wird, sind freilich selten (Mehler 2004, 332). Streng genommen muss von dieser an sich generischen Aufgabe auch die Methode der Informationsextraktion unterschieden werden, da hier mittels *Templates* (Rehm 2004, 435) lediglich die Werte zu im Text bereits vermuteten Variablen[21] gesucht werden. Wendet man Text-Mining-Methoden auf Webinhalte an, ist der Begriff Web Mining allerdings wiederum zu unspezifisch, da dazu neben dem hier gemeinten Web Content Mining auch das Web Structure Mining, bei dem Linkstrukturen von Hypertextdokumenten analysiert werden, und das Web Usage Mining, mittels dessen das Verhalten von Web-Nutzern erforscht wird, gerechnet werden müssen (Mehler 2004, 348; Chang et al. 2001, 94).

Naturgemäß schwierig ist es, bei einer automatisierten Explorationsaufgabe ein Kriterium dafür zu finden, dass die Exploration erfolgreich war. Allgemeines Ziel von Text Mining-Methoden ist es, aus mitunter sehr umfangreichen Dokumentensammlungen die für den Leser nützlichen und interessanten Informationen zu extrahieren (Feldman/Sanger 2007, 1). Versuche, Interessantheit als Maß zu definieren, rekurrieren häufig auf Kriterien wie generelle Gültigkeit, Neuheit, Nützlichkeit und Einfachheit der entdeckten Informationen (Fayyad et al. 1996, 2). Da es bei all diesen Kriterien um den individuellen, kontextabhängigen Wert der Informationen für den konkreten Leser geht, kann man Interessantheit allgemein auch als Abweichung von Vorwissen bzw. der Erwartungshaltung eines Lesers definieren (vgl. Feldman/Sanger 2007, 29).

Anhand des zuvor beschriebenen Ablaufmodells (Kap 3.2) können nun für die einzelnen Teilschritte Möglichkeiten zu ihrer teilweisen Automatisierung diskutiert werden. Wie zu Beginn des Kapitels bereits erwähnt, sollte ein geeignetes Textkorpus für die Analyse der Webinhalte im Rahmen dieser Studie sinnvollerweise mittels einer automatisierten Anfrage an eine Suchmaschine erzeugt werden. Für das weitere Vorgehen ist es aber ohne Belang, ob das Korpus automatisch oder aus gezielt ausgewählten Einzeldokumenten zusammengestellt wurde. Wenn zur Unterstützung qualitativer Forschung sowohl ein initiales Screening als auch ein verifizierender Durchlauf möglich sein sollen, ist es sinnvoll, beide Möglichkeiten anzubieten.

Die Extraktion der argumentationstragenden Begriffe in Textdokumenten ist gegenwärtig bereits eine etablierte Technik im Text Mining bzw. im Information Retrieval (Feldman/Sanger 2007, 6 u. 58ff. bzw. Beaza-Yates/Ribiero-Neto 1999, 169) und somit ohne Probleme möglich. Auch eine weitergehende, streng induktive automatische

[21] Rehm (2004, 435 bzw. 437) nennt die klassischen W-Fragen zu einem Tathergang. Ein anderes Beispiel wären Adressdaten oder Datumsangaben, nach denen gezielt gesucht werden soll.

Kategorisierung wäre auf Grundlage einer vollständigen syntaktischen Analyse unter dem Begriff des *text knowledge mining[s]* (Mehler 2004, 342f.) bereits möglich. Um dem Nutzer jedoch in Form eines Kontrollschrittes die Möglichkeit zu geben, die bis hierhin extrahierten Begriffe zu überprüfen und auf ein evtl. im Rahmen der Gesamtanalyse bereits erstelltes Kategoriensystem hin anzupassen (vgl. Inhaltsanalyse nach Mayring 2000, Kap. 3.1), wird an dieser Stelle darauf verzichtet.

Wurden die sinntragenden Begriffe aus dem Textmaterial extrahiert und vom Forscher geeignet kategorisiert, können sogenannte *association rule[s]* (z.B. Feldman/Sanger 2007, 25ff.) errechnet werden. Bei dieser Methode des Text Minings bildet ein häufiges gemeinsames Auftreten das Kriterium, um auf einen semantischen Bezug zwischen Konzepten bzw. Kategorien schließen zu können (vgl. z.B. Hipp et al. 2000, 59f.). Das Verfahren des Text Minings ist zwar durchaus insofern kritikwürdig, als es nicht die gesamte Textsemantik repräsentiert (Mehler 2004, 342). Da jedoch auch bei (reduktiven) qualitativen Arbeitsweisen vorwiegend Zusammenhänge zwischen Kodes bzw. Kategorien und nicht die gesamte Textstruktur erschlossen werden sollen, scheint die Methode gut geeignet. Wesentlicher Vorteil des Verfahrens im Vergleich zu anderen Text-Mining-Methoden[22] ist dabei, dass die Texte eben nicht gezielt nach erwarteten Mustern oder Kategorien durchsucht werden, sondern dass anhand einiger weniger textstatistischer Maße Hypothesen über teximmanente Zusammenhänge generiert werden können. Zudem erfüllt die Methode durch ihre einfache Verständlichkeit (vgl. Kap 3.4) auch das oben erwähnte Kriterium der intersubjektiven Nachvollziehbarkeit.

Die gefundenen Zusammenhänge können anschließend in Form der ja auch in der Grounded Theory verwendeten semantischen Netzwerke visualisiert werden. Graphendarstellungen sind in der Informatik essentiell und ihre Umsetzung im Text Mining (z.B. Feldman/Sanger 2007, 11) daher weit verbreitet. Die Interpretation der Ergebnisse verbleibt natürlich wie bei allen quantitativen Verfahren beim Forscher.

Abb. 5 zeigt ein modifiziertes Ablaufmodell qualitativer Inhaltsanalyse, in dem die vorher genannten Möglichkeiten zur Automatisierung von Teilschritten bereits berücksichtigt sind. Dieses entspricht in der Abfolge seiner automatisierten Teilschritte der *text mining system architecture* von Feldman/Sanger (2007, 15). Während die Zusammenstellung des Textkorpus und die abschließende graphische Visualisierung lediglich notwendige Schritte für die Eingabe der Rohdaten durch den Nutzer bzw. für die Anzeige der entdeckten

[22] Einen Überblick über mögliche Methoden bieten Feldman/Sanger (2007, 19ff.), Mehler (2004, 341ff.) oder Chang et al. (2001, 81ff.).

Informationen darstellen, werden bei der Extraktion der Schlüsselbegriffe und der Berechnung der association rules das Textmaterial und damit auch die Inhalte in erheblichem Maße transformiert. Gemäß dem Gebot der intersubjektiven Nachvollziehbarkeit wird die Funktionsweise dieser beiden Schritte nun im Detail erläutert.

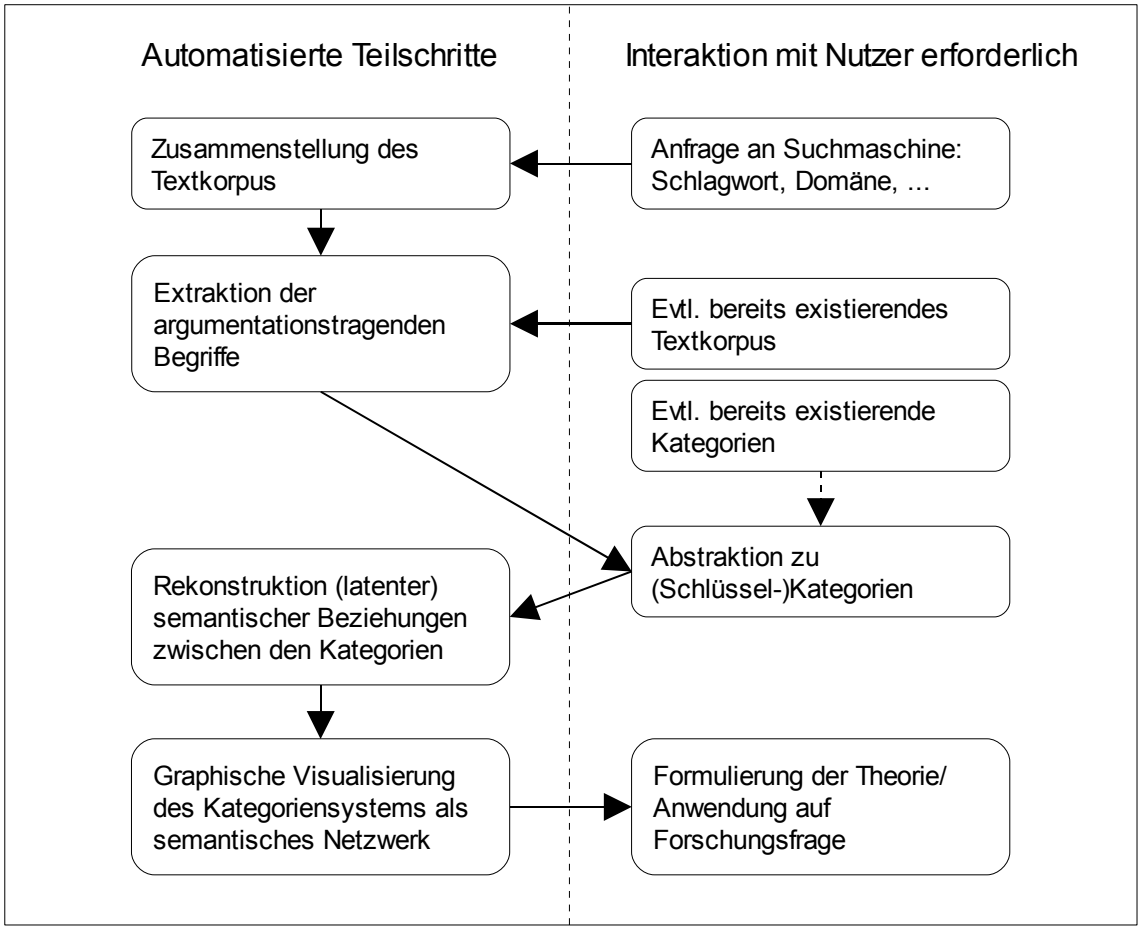



*Abb. 5: Teilautomatisiertes Ablaufmodell qualitativer Inhaltsanalyse*

## *3.4 Methodische Grundlagen automatisierter Textanalyse*

Wie in der Diskursanalyse (JÄGER 2004, 182f.) werden auch im Text Mining vor allem Substantive und deren Attribute als wesentliche Träger der Bedeutung von Texten erkannt (WITSCHEL 2004, 45; DAILLE 1996, 29f.; FELDMAN ET AL. 1998, 9-4; BAEZA-YATES/RIBEIRO-NETO 1999, 169f.). Begriffe können dabei auf unterschiedlichen Abstraktionsebenen extrahiert werden. FELDMAN/SANGER (2007, 4ff.) unterscheiden zwischen einfachen words, den unter Umständen aus mehreren Einzelwörtern gebildeten terms sowie concepts, die terms oder auch ganze Nebensätze repräsentieren können. Dabei müssen concepts nicht aus im Text vorkommenden Wörtern gebildet sein, sondern dürfen auch durch Abstraktion zustande

gekommen sein. Da die extrahierten Begriffe dazu verwendet werden können, den Inhalt eines Dokuments zu repräsentieren, werden sie von Feldman/Sanger (2007, 4ff.) auch unabhängig von der Abstraktionsebene als features bezeichnet. Die Begriffe terms und concepts sind dabei erstaunlich gut auf die Terminologie der qualitativen Inhaltsanalyse abbildbar, wo ebenfalls zwischen unmittelbar im Text vorkommenden Konzepten, den In-vivo-Kodes (s. Kap 3.1) und den aus diesen durch Generalisierung bzw. Abstraktion gebildeten Kategorien unterschieden wird. Weil eine Extraktion von terms aufgrund rein syntaktischer Merkmale möglich ist (Feldman et al. 1998, 9-5; Daille 1994, 30) und ohnehin aus methodischen Gründen auf eine weitergehende automatische Kategorisierung bei diesem Schritt verzichtet wird (vgl. Kap. 3.3), ist eine aufwändige Berechnung der concepts im Rahmen dieser Arbeit nicht nötig[23].

Vor der Extraktion der terms als wesentlicher sinntragender Einheiten der Texte ist zunächst eine umfangreiche Vorverarbeitung des Dokuments unerlässlich. Liegt die Dokumentensammlung bereits in einem einheitlichen Datenformat vor (Feldman/Sanger 2007, 58f.), ist der erste Schritt dabei immer die Zerlegung der Texte in seine einzelnen sprachlichen Einheiten bis hin zur gewünschten Detailebene. Interpunktionszeichen dienen zur Abgrenzung von Sätzen, während Leerzeichen zur Abgrenzung einzelner Wörter verwendet werden können (*lexical analysis* nach Baeza-Yates/Ribeiro-Neto 1999, 165f.; *tokenization* nach Feldman/Sanger 2007, 60). Dies sind natürlich nur Indizien. Abkürzungen und Ordnungszahlen in Datumsangaben[24] sind die häufigsten Beispiele für Zeichenfolgen, die fälschlicherweise ein Satzende anzeigen. Solche Fälle müssen in Form von Ausnahmenlisten explizit ausgeschlossen werden (Heyer et al. 2006, 62ff.).

Unterschiedliche Ansätze können verfolgt werden, wenn es anschließend darum geht, die repräsentativen Schlüsselbegriffe der Dokumente zu extrahieren. Sehr performante Lösungen bieten Information-Retrieval-Methoden an, bei denen die Textmenge rasch auf wenige Schlüsselbegriffe reduziert werden muss. Hierzu wird der Text zunächst mithilfe einer vorab definierten Stoppwortliste gefiltert, die alle Wörter des Textes enthält, die keinen wesentlichen Beitrag zur Beschreibung der Semantik des Textes leisten. Eine solche Liste enthält für das Englische vor allem Präpositionen, Konjunktionen und Artikel (Baeza-Yates/Ribeiro-Neto 1999, 167). Erst daran schließt sich die eigentliche Extraktion der Schlüsselbegriffe (*index terms* nach Baeza-Yates/Ribeiro-Neto 1999) an. Dazu werden

[23] In der englischen Alltagssprache können „term“ und „concept“ synonym verwendet werden. Gemäß der von Feldman/Sanger (2007, 4ff.) getroffenen Differenzierung und unter Berücksichtigung der Terminologie qualitativer Inhaltsanalyse soll im folgenden „term“ mit „Begriff“ und „concept“ mit „Kategorie“ oder „Schlüsselbegriff“ wiedergegeben werden.

[24] Ein Beispiel für das Deutsche: „Dr. Müller traf am 13. Februar ein.“

Substantive, die in einer bestimmten Nachbarschaft im Text auftreten, zu Substantivgruppen (*noun groups* nach Baeza-Yates/Ribeiro-Neto 1999, 169) zusammengefasst. Zuvor können alle Wörter durch Stammformreduktion auf ihren Wortstamm reduziert werden, um zu verhindern, dass sich der semantische Gehalt desselben Schlüsselbegriffs nicht auf unterschiedliche Wortarten oder gar unterschiedliche Flexionsendungen verteilt (Baeza-Yates/Ribeiro-Neto 1999, 168f.). Dies ist vor allem für das Englische anhand relativ weniger Regeln möglich (Baeza-Yates/Ribeiro-Neto 1999, 434).

Die hohe Performanz des Information-Retrieval-Ansatzes geht dabei allerdings zulasten der Semantik, die abgebildet werden kann. Bereits beim Herausfiltern der Stoppwörter kommt es unter Umständen zum Verlust von Präpositionen und Artikeln, die insbesondere im Englischen als Teil von Nominalphrasen bedeutungstragende features konstituieren[25]. Zwar sind vollständige syntaktische Analysen von Sätzen automatisch durchführbar, der hohe Zeitaufwand steht aber dabei in keinem Verhältnis zu der Qualität der Ergebnisse (Feldman/Sanger 2007, 61). Einen vertretbaren Kompromiss zwischen Aufwand und Ergebnis stellt es dar, vor der Extraktion der Begriffe jedem Wort des in Sätze segmentierten Textes zunächst seine Wortart zuzuordnen (Verfahren des *part of speech tagging* nach Feldman/Sanger 2007, 60; Heyer et al. 2006, 109). Zentrales Problem ist dabei, dass ein Wort je nach Kontext unterschiedlichen Wortarten angehören kann (Heyer et al. 2006, 112). Eine Bestimmung, die dies berücksichtigt, kann aber sowohl durch regelbasierte tagger (Brill 1992, 152ff.) als auch durch probabilistische Modelle (z.B. *Hidden-Markov-Modelle* – vgl. Heyer et al. 2006, 115ff.)[26] mit einer beeindruckenden Genauigkeit von 96% (Feldman et al. 1998, 9-4) bzw. 97% (Heyer et al. 2006, 115) vorgenommen werden. Die dazu nötigen Informationen können aus großen Referenzkorpora[27] abgeleitet werden (Feldman/Sanger 2007, 60). Wurde ein Text auf diese Weise mit grammatikalischen Zusatzinformationen angereichert, können anhand von bestimmten syntaktischen Mustern ohne Weiteres auch komplexere Nominalphrasen extrahiert werden, *[that] tend to represent more meaningful concepts* (Feldman et al. 1998,

[25] Ein Beispiel im Englischen: „Lord of the Rings“.

[26] Beispiel: „Die verteidigte Burg fiel.“ Kontextisoliert könnte „verteidigte“ zunächst sowohl ein Adjektiv (bzw. genauer Partizip) als auch eine finite Verbform sein. Da es jedoch äußerst unwahrscheinlich ist, dass eine finite Verbform direkt auf einen Artikel („die“) folgt, kann diese Möglichkeit ausgeschlossen werden (Heyer et al. 2006, 132f.).

[27] Beispiele sind das Brown-Korpus für das Englische und das Negra-Korpus für das Deutsche (Heyer et al. 2006, 128).

9-2) als einzelne, nur nach ihrer Häufigkeit ausgewählte Begriffe.[28] Abb. 6 zeigt den gesamten Vorgang der Extraktion von Schlüsselbegriffen im Überblick.


*Abb. 6: Term extraction auf der Grundlage des part of speech tagging*

Da für eine qualitative Inhaltsanalyse jede Form der automatischen Verarbeitung einen Zeitgewinn darstellt und der semantische Gehalt der gewonnenen Informationen gegenüber der Performanz eines Verfahrens als vorrangig zu erachten ist, muss für die Begriffsextraktion ein ihr vorausgehendes part of speech tagging als Methode der Wahl gelten. Weil alle Vorverarbeitungsschritte zwar sprachspezifisch, immer aber unabhängig von den in der Dokumentensammlung zum Ausdruck gebrachten Inhalten sind, können auch diejenigen, die auf explizit kodiertes Zusatzwissen zurückgreifen (vgl. Ausnahmen bei der Satzerkennung bzw. Stoppwortliste), im Rahmen einer qualitativen Inhaltsanalyse als unbedenklich gelten.

Bereits mit recht einfachen Analysen zur Häufigkeit und Verteilung der extrahierten Schlüsselbegriffe können erste Aussagen über Kookkurrenzbeziehungen zwischen diesen gemacht werden, die als Indizien für einen semantischen Bezug zwischen features dienen können. Als kookkurrent gelten dabei zunächst einfach features, die in unmittelbarer Nachbarschaft im Text auftreten (Heyer et al. 2006, 23). Es sind verschiedene Kriterien vorstellbar, um eine solche Nachbarschaft zu definieren. Meist reicht es aus, wenn die Begriffe im gleichen Satz vorkommen oder nur durch eine definierte maximale Anzahl anderer Wörter getrennt liegen (Heyer et al. 2006, 136). Tritt eine bestimmte Anzahl von Begriffen gehäuft gemeinsam im Text auf, kann man eine zufällige Kombination

[28] Sinnvolle Muster für das Englische sind Substantivgruppen und zugehörige Adjektive, die durch Artikel und Präpositionen verbunden sein können: „operating system of a computer“ (Feldman et al. 1998, 9-5).

ausschließen und von einer signifikanten Kookkurrenz[29] sprechen (Heyer et al. 2006, 24). Die Tatsache, dass Wörter in festen Wendungen trivialerweise kookkurrieren, kann vor allem im Englischen als ergänzendes Kriterium dafür hinzugezogen werden, ob bei der Extraktion von features einzelne Begriffe Teil einer noun group sind oder nicht (Feldman et al. 1998, 9-5).

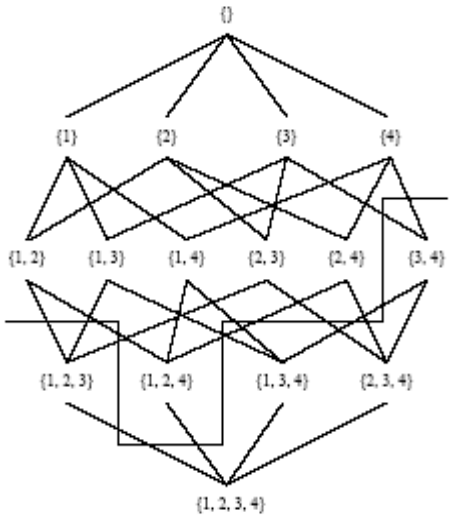



*Abb. 7: Bilden von frequent sets im association rule mining: Der Verband wird dabei von oben nach unten aufgebaut. Die Trennlinie visualisiert den Schwellwert, ab dem aufgrund zu seltenem gemeinsamen Auftreten keine weiteren Kombinationsmöglichkeiten mehr untersucht werden müssen (nach Hipp et al. 2000, 59).*

Ausgehend von solchen Mengen häufig gemeinsam auftretender Begriffe (*frequent sets* nach Feldman/Sanger 2007, 23) lassen sich Assoziationsregeln berechnen, mit deren Hilfe auch die Richtung des semantischen Zusammenhangs zwischen den Begriffen angegeben werden kann. Obwohl Assoziationsregeln ursprünglich nur für die Entdeckung von Zusammenhängen in Datenbanken verwendet wurden (Agrawal et al. 1993, 207ff.), lässt sich die Technik ohne Probleme auf Texte übertragen (Feldman et al. 1996, 950; Rajman/Besançon 1997, 3ff.; Chang et al. 2001, 82; Feldman/Sanger 2007, 25ff.).

Für die Generierung von frequent sets ist es unerheblich, ob die extrahierten Begriffe direkt als Ausgangsmenge dienen oder ob diese zuvor vom Nutzer zu Kategorien abstrahiert wurden. Für alle Begriffe müssen lediglich Anzahl und Stellen ihres Auftretens im Text bekannt sein. Darauf aufbauend kann schrittweise überprüft werden, welche Kombinationen von Begriffen Mengen liefern, bei denen die Anzahl der Kookkurrenz aller in ihr enthaltenen Begriffe (*support* nach Agrawal/Srikant 1994, 5) in der

[29] Heyer et al. (2006, 139f.) zeigt anhand eines geeigneten Maßes, dass auf diese Weise aus Texten extrahierte Begriffspaare sogar in der Stärke ihres Zusammenhangs erstaunlich gut mit den Assoziationen menschlicher Probanden zu einem Stichwort übereinstimmen.

Dokumentensammlung immer noch über einem bestimmten Schwellwert liegt. Dazu werden zunächst alle zweielementigen Teilmengen gebildet und überprüft, dann alle dreielementigen, usw. Prinzipiell werden dabei alle Kombinationsmöglichkeiten untersucht, also die Potenzmenge der Menge der Begriffe (Hipp et al. 2000, 59). Fällt der support einer Teilmenge allerdings unter den Schwellwert, müssen für diese Menge keine weiteren Kombinationsmöglichkeiten mehr untersucht werden (*pruning* nach Hipp et al. 2000, 59).[30] Abb. 7 illustriert das Vorgehen an einer vierelementigen Ausgangsmenge.

Aus den so gefundenen zwei- oder mehrelementigen frequent sets lassen sich nun Assoziationsregeln berechnen. Dazu wird aus jeder Menge wiederum jede mögliche Teilmenge einzeln entfernt und getestet, ob sich bei einem hinreichend großen Anteil (*confidence* nach Agrawal/Srikant 1994, 13) der gemeinsamen Fundstellen der verbleibenden Elemente der Menge auch die Begriffe der zuvor entfernten Teilmenge finden. Ist dies der Fall, wird den Ergebnissen eine Assoziationsregel hinzugefügt, bei der die zuvor entfernte Teilmenge von Begriffen durch die in der Menge verbleibenden Elemente impliziert wird[31] (Algorithmus nach Agrawal/Srikant 1994, 12ff.; vgl. auch Hipp et al. 2000, 59 und Feldman/SAnger 2007, 26f.).

> Formal: Für alle $X \in F$ (X Menge von Schlüsselbegriffen, F frequent set) berechne für alle Regeln der Form $X \setminus Y \rightarrow Y$ ($Y \subseteq X$, $\emptyset \neq Y \neq X$) ihren Konfidenzwert gemäß der Formel support $(X \cup Y)$ / support (X), wobei support (A) allgemein die Anzahl aller gemeinsamen Fundstellen der Begriffe in einer Menge A liefert. Liegt der Konfidenzwert über dem definierten Schwellwert, akzeptiere die Regel. (nach Hipp et al. 2000, 59).

Assoziationsregeln stellen also ein probates und vergleichsweise einfaches[32] Mittel zur Verfügung, wie semantische Zusammenhänge aus einer Dokumentensammlung extrahiert werden können, ohne an den Text zusätzliches Vorwissen über seine Inhalte herantragen zu müssen. Da sowohl bei der Generierung der frequent sets als auch bei der anschließenden Berechnung der Assoziationsregeln abgesehen von der erwähnten Optimierung des prunings alle Kombinationsmöglichkeiten untersucht werden müssen, ist

[30] Beispiel: Der Begriff ‚Prince of Wales' soll in der Menge der Schlüsselbegriffe enthalten sein. Auch die Kombination mit ‚Queen' liefert immer noch einen hinreichenden support. Durch die Hinzunahme von ‚Europa' sinkt dieser allerdings unter den Schwellwert, sodass weitere Kombinationen nicht untersucht werden müssen.

[31] Ein einfaches Beispiel: Bei der frequent set {‚Queen', ‚Prince of Wales'} wird die noun group ‚Prince of Wales' entnommen. Da sie sich bei 50% der Fundstellen von ‚Queen' und damit hinreichend häufig findet, wird der Ergebnismenge eine Regel der Form ‚Queen' → ‚Prince of Wales' hinzugefügt.

[32] Amir et al. (2005) schlägt mit den maximal association rules eine weitere Optimierung vor. Dafür ist jedoch Vorwissen in Form von vor der Analyse festgelegten Kategorien nötig.

das Verfahren durch seinen in Abhängigkeit von der Anzahl der verwendeten Schlüsselbegriffe exponentiell wachsenden Aufwand sehr rechenintensiv.

# 4 Entwurf und Implementierung eines Programms zur semiautomatischen Extraktion semantischer Begriffsnetze

Nachdem alle Grundlagen hinreichend behandelt sind, kann darauf aufbauend ein Programm entworfen und implementiert werden, mit dessen Hilfe die vier in Kapitel 3.3 beschriebenen Aufgaben durchgeführt werden können. Da die Teilschritte qualitativer Inhaltsanalyse zwar aufeinander aufbauen, ihrer konkreten Realisierung jedoch größtmögliche Variabilität eingeräumt werden soll, wurden die Teilaufgaben so weit wie möglich unabhängig voneinander entworfen. Dadurch kann ein Korpus, das automatisch mittels einer Anfrage an eine Suchmaschine erzeugt wurde, genauso unterstützt werden wie bereits vorliegendes digitales Textmaterial. Auch die Suche nach syntaktischen Mustern auf der Grundlage des part of speech tagging zur Extraktion der Schlüsselbegriffe kann durch diese Abstraktion später ebenso durch ein anderes, gegebenenfalls semantisch wertvolleres Verfahren ersetzt werden wie die Methode des association rule minings zur Errechnung der semantischen Zusammenhänge zwischen den einzelnen Begriffen bzw. Kategorien. Um die einzelnen Schritte solchermaßen gegeneinander isolieren zu können, ist die Definition geeigneter Datenaustauschformate erforderlich, mithilfe derer die jeweilige Ausgabe einer Verarbeitungsphase persistent abgelegt werden kann und dem nächsten Teilschritt unabhängig von der Art ihres Zustandekommens als Eingabe zur Verfügung steht. Daneben ermöglicht eine derartige Kapselung durch dauerhaft ablegte Zwischenergebnisse im Zuge des iterativen Arbeitsprozesses qualitativer Inhaltsanalyse wiederholte Programmdurchläufe mit veränderten Parametern, ohne jedes Mal wieder alle elementaren Berechnungen durchlaufen zu müssen.

Unter diesen Vorgaben scheint neben einer Einzelplatzlösung auch eine spätere Server-Client-Architektur machbar, da jeweils nur die Resultate der Arbeitsschritte ausgetauscht werden müssen und die Berechnungen komplett auf einem leistungsfähigen Server durchgeführt werden könnten. Alle zur Konfiguration nötigen Einstellungen könnten dabei in Form eines Webservices angeboten werden, der sich über jeden verbreiteten Browser steuern lässt. Die Umsetzung erfolgte aufgrund ihres experimentellen Charakters zunächst jedoch nur als Einzelplatzlösung. Als Programmiersprache wurde Java[33] gewählt,

[33] JDK 1.6.0, vgl. http://java.sun.com/javase/downloads/index.jsp (20.11.07).

da diese nicht an eine bestimmte Plattform[34] gebunden ist. Alle Entwurfs- und Implementierungsarbeiten wurden dabei mit der Programmierumgebung NetBeans[35] durchgeführt.

Bevor nun der Funktionsumfang, das jeweiliges Resultat und die konkrete Realisierung für die automatisierten Teilschritte beschrieben werden, muss zunächst noch geklärt werden, wie ein Programm des geschilderten Funktionsumfangs in geeigneter Weise unterstützt werden kann, da die softwaretechnische Umsetzung aller vier Teilschritte den zeitlichen Rahmen dieser Arbeit bei Weitem übersteigen würde. Wie die in Kap. 3.3 vorgestellte Aufgabenstellung des natural language processing bereits vermuten lässt, findet sich tatsächlich ein reichhaltiger Fundus von Programmbibliotheken, die bereits alle grundlegenden Funktionen für eine automatische Textanalyse zur Verfügung stellen. Dass solche Programme vorrangig für die englische Sprache verfügbar sind, ist im Rahmen dieser Studie eher nützlich als hinderlich, da sich ausreichend viele Beispiele für außereuropäische Websites in dieser Sprache finden lassen dürften.

Die Wahl geeigneter Hilfsmittel wurde hierbei durch mehrere Kriterien eingeschränkt. Neben der Bereitstellung einer Java-Schnittstelle und ihrer freien Verfügbarkeit spielte bei der Auswahl auch die Qualität der Dokumentation und somit die Nachvollziehbarkeit der Ergebnisse im Sinne der in Kap. 3.3 geforderten intersubjektiven Nachvollziehbarkeit eine entscheidende Rolle. Eine entsprechende Software sollte darüber hinaus nach Möglichkeit alle Vorverarbeitungsschritte bis hin zum part of speech tagging übernehmen, da dieser Schritt auf linguistische Zusatzinformation in Form eines Referenzkorpus (vgl. Kap. 3.4) angewiesen ist und erst auf dieser Grundlage die Extraktion der Schlüsselbegriffe und die Berechnung der association rules ohne weitere Voraussetzungen durchführbar sind.

Steht ein Textkorpus bereits in digitaler Form zur Verfügung, gibt es mehrere Softwareprojekte, die den vorgenannten Kriterien auf einem vergleichbaren Niveau genügen können. Näher betrachtet wurden hierbei LingPipe, OpenNLP und JTextPro[36]. Alle drei bieten mit Satzerkennung, tokenization und part of speech tagging (vgl. Kap 3.4) die benötigten grundlegenden Analysefunktionen. Basis für die Zuweisung der Wortklassen ist dabei jeweils ein probabilistisches Modell (vgl. Kap. 3.4), das mithilfe eines hinreichend großen, allgemeinsprachlichen Referenzkorpus trainiert wurde. Für das Englische dienten das Brown-Korpus (LingPipe, OpenNLP) und das Wall-Street-Journal-

[34] Kombination aus Hardware und Betriebssystem

[35] NetBeans IDE 5.5, vgl. http://www.netbeans.info/downloads/index.php?rs=11 (20.11.07).

[36] Informationen über Funktionsumfang und wesentliche Eigenschaften finden sich unter http://www.alias-i.com/lingpipe (21.11.07), http://opennlp.sourceforge.net (21.11.07) bzw. http://jtextpro.sourceforge.net (21.11.07).

Korpus (JTextPro, OpenNLP) als Vorlage[37]. OpenNLP bietet darüber hinaus auch Funktionen für das Spanische an. Während das probabilistische Modell im Falle von JTextPro eine höhere Genauigkeit erzielt[38], bietet LingPipe bereits von sich aus weitergehende Analysemöglichkeiten an. Alle drei Projekte besitzen eine gut dokumentierte Programmierschnittstelle für Java-Applikationen und sind frei verfügbar[39].

Soll das Textkorpus mittels einer Anfrage an eine Suchmaschine automatisch generiert werden, bietet sich die Verwendung der entsprechenden Programmierschnittstelle von Google[40] an, wie sie auch Kiefer et al. (2006, 45ff.) verwenden. Prinzipiell ist eine ungeprüfte Verwendung eines solchen Korpus schwierig, da nicht alle Seiten, die für die Anfrage relevant sind, gefunden werden (*recall* nach Beaza-Yates/Ribiero-Neto 1999, 75) und nicht alle Seiten, die gefunden werden, für die Anfrage wirklich relevant sind (*precision* nach Beaza-Yates/Ribiero-Neto 1999, 75). Hinzu kommt, dass die konkrete Funktionsweise der Google Soap Search API nicht vollständig dokumentiert ist (vgl. Kiefer et al. 2006, 45ff.), was somit gegen das Gebot der intersubjektiven Nachvollziehbarkeit verstößt. Dennoch stellt die Verwendung im Rahmen des knappen Zeitbudgets dieser Arbeit die effizienteste Möglichkeit dar, durch eine einzige, gezielte Anfrage ein Textkorpus zusammenzustellen.

Damit ein solches mittels der vorgestellten linguistischen Software-Werkzeuge verarbeitet werden kann, reicht es aber nicht aus, von allen oder zumindest einer bestimmten Anzahl der ersten Google-Suchtreffer die entsprechenden Dokumente herunterzuladen und diese miteinander zu verbinden. Da Dokumente im Web gewöhnlich für die Anzeige in einem Browser konzipiert sind, muss zunächst noch die Menge der (Formatierungs-)Auszeichnungen[41] herausgefiltert werden, um an den eigentlichen Text zu gelangen. Bei diesem Vorgang gehen natürlich auch Bilder und andere eingebundene Objekte verloren, die durchaus für die fachliche Fragestellung von Relevanz sein, im Rahmen dieser rein textbasierten Arbeit aber nicht berücksichtigt werden können.

Ideal für die Lösung dieser Aufgaben geeignet ist dabei das Software-Werkzeug JWebPro[42], das auf JTextPro aufbaut, jedoch um zusätzliche Komponenten ergänzt

---

37 Ein Überblick über die Korpora und ihre Bezugsmöglichkeit findet sich auf den Webseiten des Brown Laboratory for Linguistic Information Processing (http://bllip.cs.brown.edu/resources.shtml, 21.11.07).

38 LingPipe verwendet ein Hidden Markov Modell (vgl. Kap 3.4), OpenNLP ein Maximum Entropy Modell; JTextPro basiert auf Conditional Random Fields. Ein Vergleich der drei Technologien mit dem Nachweis der Überlegenheit des letzteren Ansatzes findet sich bei Lafferty et al. (2001, 7).

39 JTextPro unter GNU GPL, OpenNLP unter LGPL, LingPipe unter einer royalty free license (http://www.alias-i.com/lingpipe/licenses/lingpipe-license-1.txt, 21.11.07).

40 Google Soap Search API (beta), vgl. http://code.google.com/apis/soapsearch/ (20.11.07).

41 Im Wesentlichen Html-Tags sowie Anweisungen eingebundener Programmiersprachen.

42 Weiterführende Informationen unter http://jwebpro.sourceforge.net (21.11.07).

wurde, um genau den beschriebenen Anforderungen einer webbasierten Korpusanalyse zu entsprechen. Dabei werden zunächst mittels einer Suchanfrage an die Google Soap Search API potentiell relevante Dokumente ermittelt, automatisch heruntergeladen und unter Verwendung der Funktionalität von HtmlParser[43] auf das reine Textmaterial reduziert. Dieses kann dann wie beschrieben unter Verwendung von JTextPro linguistisch angereichert werden. Wie sein Pendant ist auch JWebPro unter GNU GPL frei verfügbar und bietet eine gut dokumentierte Schnittstelle für Java-Applikationen an. Alternativ dazu reicht die Konfiguration über eine einfache Textdatei bereits aus, um das Werkzeug direkt einsetzen zu können. Trotz der genannten Risiken und obwohl Google die Entwicklungsarbeit an der Google Soap Search API zwischenzeitlich eingestellt hat[44], bietet JWebPro also einen einfachen und eleganten Weg, um mittels einer Google-Suchanfrage ein Textkorpus zu erzeugen, bei dem alle für eine weitere semantische Informationsverarbeitung nötigen Vorarbeiten bereits geleistet sind.

Anhand solcher Teillösungen, die zwar nur unter Einschränkungen verwendet werden können, deren optimale Umsetzung aber im zeitlichen Rahmen dieser Arbeit nicht geleistet werden kann, wird nochmals deutlich, wie wichtig es ist, die tatsächliche technische Realisierung der einzelnen Teilschritte qualitativer Inhaltsanalyse austauschbar zu halten. Unter den bisher beschriebenen Voraussetzungen können nun die Datenformate beschrieben werden, mit deren Hilfe die technischen Realisierungen der aufeinander aufbauenden Schritte gegeneinander isoliert werden können. Ein erstes Zwischenresultat stellen die mittels JWebPro bzw. JTextPro erzeugten Textdateien dar, die bereits mit part of speech tags annotiert sind. Zwar wurde die Zuordnung der einzelnen Wörter zu Wortklassen eigentlich dem Teilschritt der Extraktion von Schlüsselbegriffen zugeordnet (vgl. Kap. 3.4), da dieses Resultat jedoch von den verschiedenen beschriebenen Programmbibliotheken relativ einheitlich zur Verfügung gestellt werden kann, bietet es sich als erste Abstraktionsebene und Ausgangspunkt für die eigene Implementierung an.

[43] Weiterführende Informationen unter http://htmlparser.sourceforge.net (21.11.07).
[44] Vgl. http://code.google.com/apis/soapsearch (21.11.07).

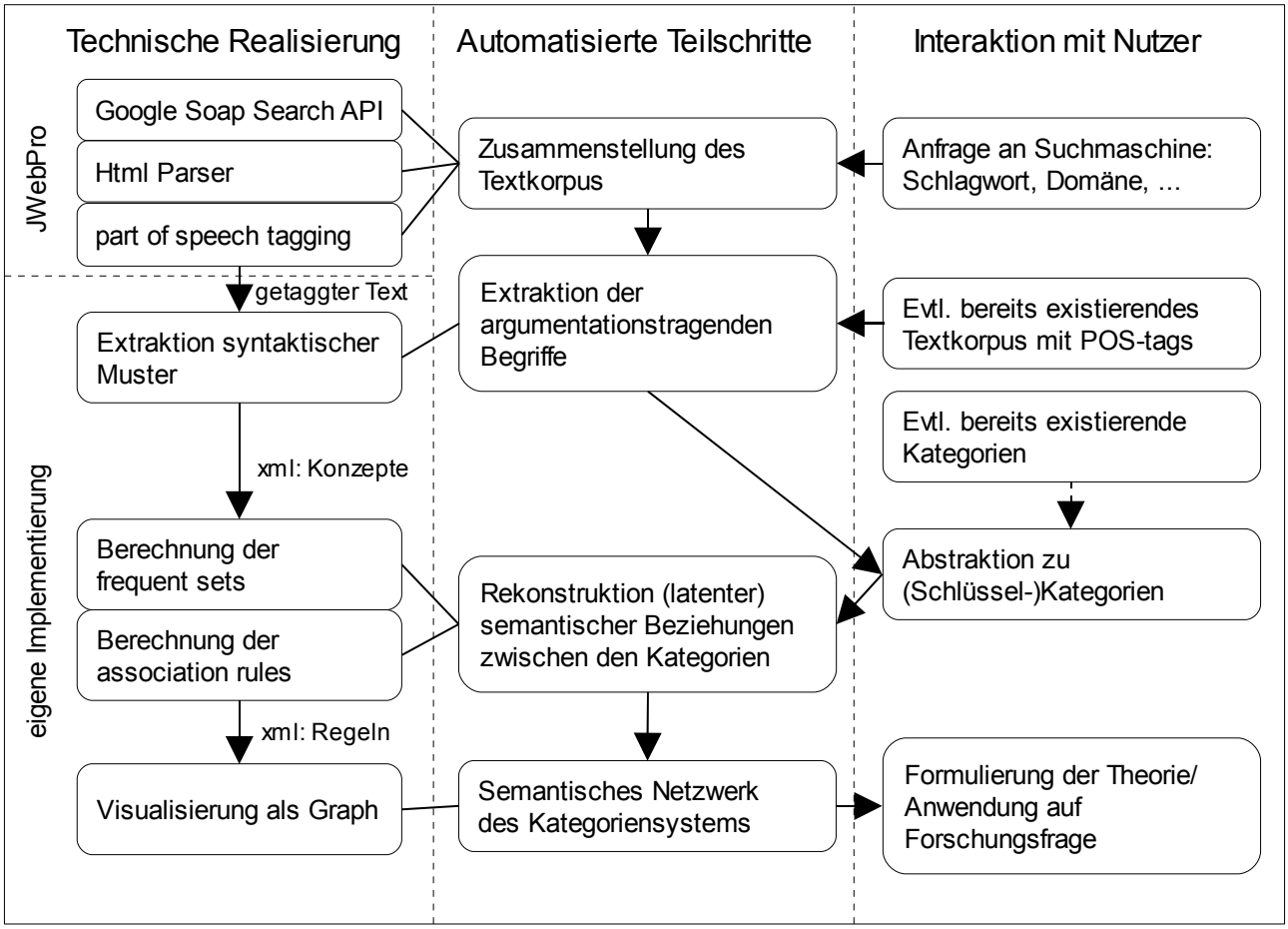



*Abb. 8: Technische Realisierung des teilautomatisierten Ablaufmodells qualitativer Inhaltsanalyse*

Da die extrahierte Terminologie jederzeit wieder aus dem getaggten Textmaterial konsistent reproduziert werden kann, liegt ein zweites sinnvolles Zwischenergebnis erst dann vor, wenn die argumentationstragenden Begriffe vom Nutzer bereits zu allgemeineren Kategorien zusammengefasst wurden. Diese beiden Arbeitsschritte bilden auch deshalb eine logische Einheit, weil es von der konkreten Umsetzung der term extraction abhängt, auf welcher Basis eine weitere Kategorisierung der Schlüsselbegriffe aufbaut. Ein drittes und letztes Resultat liegt schließlich vor, wenn mittels des association rule mining Zusammenhänge zwischen den Kategorien errechnet wurden. Die logische Kapselung dieses Teilschrittes ermöglicht es auch, die Suche nach Assoziationen auf der Grundlage ein und derselben Kategorisierung beliebig oft mit veränderten Parametern zu wiederholen. Als Datenformat für die persistente Ablage der letzten zwei Teilergebnisse wurde xml gewählt[45]. Abb. 8 zeigt das teilautomatisierte Ablaufmodell qualitativer Inhaltsanalyse, das um die konkreten technischen Realisierungen ergänzt wurde.

Das in dieser Arbeit entwickelte Programm trägt nach seinem primären Einsatzziel den Namen Sascet (semi-automatic semantic context extraction tool). Es wurde mit

[45] Xml-Schemata für die jeweiligen Zwischenergebnisse finden sich im CD-Teil dieser Arbeit.

elementaren Funktionen zum Anlegen und Verwalten von Analyseprojekten ausgestattet. Die wesentlichen Teilschritte qualitativer Inhaltsanalyse selbst wurden gemäß ihrer sequentiellen Anordnung als Assistent konzipiert, der den Nutzer Schritt für Schritt durch die Auswertung führt[46]. Als Sprache der Benutzeroberfläche wurde Deutsch gewählt. Nachdem die Rahmenbedingungen hinreichend geklärt wurden, werden nun die im Zuge der Implementierung getroffenen Entscheidungen für die jeweiligen Teilschritte dokumentiert.

## *4.1 Vorverarbeitungsschritt*

Insofern die Vorverarbeitungsschritte durch eingebundene Programmbibliotheken übernommen werden, ist der Funktionsumfang dieses Teilschrittes durch die Entscheidung für JWebPro bzw. JTextPro bereits festgelegt. Im Wesentlichen bieten sich dem Nutzer an dieser Stelle verschiedene Möglichkeiten. Falls bereits ein Textkorpus per Hand zusammengestellt wurde, kann es an dieser Stelle mithilfe von JTextPro mit part of speech tags ausgezeichnet werden. Wenn ein Textkorpus durch eine Google-Suchanfrage erst generiert werden soll, kann neben der genauen Suchanfrage auch angegeben werden, nach wie vielen der ersten Suchtreffer das Herunterladen abgebrochen werden soll und das Korpus als vollständig angesehen werden kann. Dadurch kann verhindert werden, dass bei Suchanfragen mit einer nur geringen Anzahl an Treffern Webseiten von nachrangiger Relevanz Eingang in das Korpus finden. Ein Problem bei der automatischen Generierung eines Korpus besteht allerdings darin, dass mit JWebPro nicht alle Seiten, die Google als Treffer aufführt, auch automatisch heruntergeladen werden können. Betrifft dies einige der ersten Suchtreffer, gehen dadurch unter Umständen hoch relevante Informationen für die weitere Analyse verloren. Da die Funktionalität von JWebPro im Rahmen dieser Arbeit nicht näher betrachtet werden kann, muss diese Einschränkung jedoch hingenommen werden. Die dritte Wahlmöglichkeit besteht darin, ein bereits existierendes, mit part of speech tags versehenes Korpus als Grundlage für die weiteren Schritte zu verwenden. Wurde ein vorhandenes Projekt geöffnet, ist das zugehörige Korpus bereist ausgewählt. Abb. 9 zeigt die entsprechende Sichtweise in Sascet.

[46] Eine kurzes Handbuch zur Bedienung von Sascet sowie eine ausführliche Programmdokumentation finden sich im CD-Teil der Arbeit.

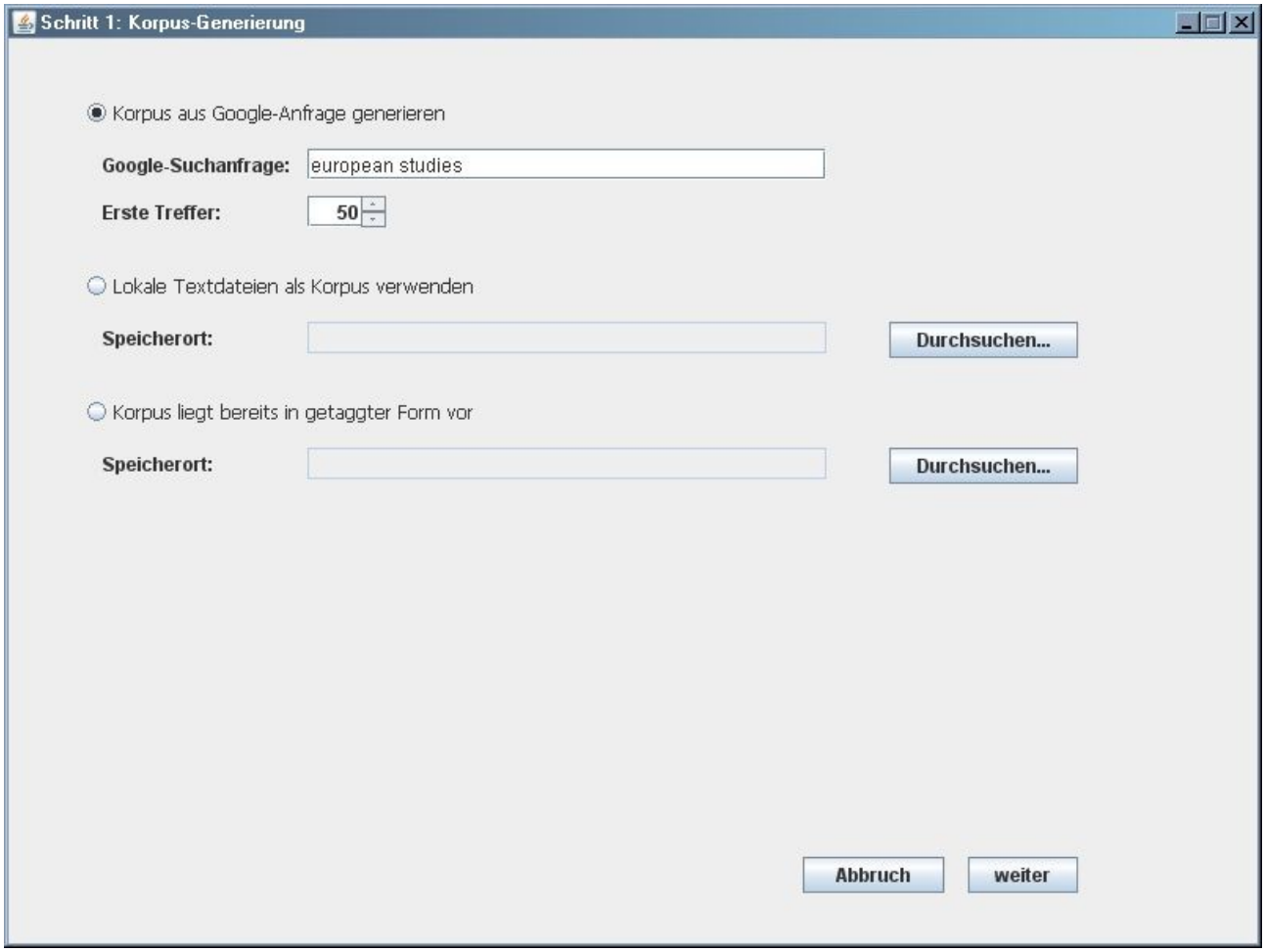


*Abb. 9: Erster Teilschritt: Generieren bzw. Einbinden des Textkorpus*

Als Endprodukt liegt nach der Korpusgenerierung in jedem Fall das mit part of speech tags versehene Textmaterial vor. JWebPro bzw. JTextPro verwenden hierbei das Penn Tagset[47], mit dem nicht nur Wortarten unterschieden, sondern auch weitergehende Differenzierungen, z.B. zwischen verschiedenen Flexions- und Steigerungsformen, getroffen werden können[48].

## *4.2 Extraktion der argumentationstragenden Begriffe*

Wie in Kapitel 3.4 bereits erläutert (vgl. Abb. 6), kann auf der Grundlage der annotierten Wortklassen anhand syntaktischer Muster nach Nominalphrasen gesucht werden, die die sinntragenden Einheiten des Textes repräsentieren. Dabei folgt die Implementierung im Wesentlichen den Vorgaben von Feldman et al. (1998), denen zufolge für eine Extraktion

[47] Eine vollständige Übersicht über die verschiedenen Wortklassen und die ihnen zugeordneten Tags findet sich z.B. unter http://www.computing.dcu.ie/~acahill/tagset.html (23.11.07).

[48] Im Englischen ist es z.B. sehr leicht möglich, Substantive mit Plural- bzw. Genitivendungen zu erkennen und als eigene Klasse auszuweisen. Um der Tatsache Rechnung zu tragen, dass es sich bei den Tags nicht um Wortarten im sprachwissenschaftlichen Sinne handelt, soll im Folgenden der Begriff Wortklasse verwendet werden.

der zentralen Begriffe die Suche nach Substantivgruppen ausreicht, die durch vorausgehende attributiv gebrauchte Adjektive näher bestimmt und durch Präposition und Artikel miteinander verbunden sein können (Feldman et al. 1998, 9-5)[49]. Für alle gemäß dieser Regel extrahierten Begriffe werden neben ihrer Häufigkeit auch alle ihre Fundstellen gespeichert. Als Nachbarschaft für ein mögliches gemeinsames Auftreten der Begriffe wurde dabei die Satzebene gewählt (vgl. Kap. 3.4). Die Vergabe von Satz-IDs erleichtert dabei den späteren Test auf Kookkurrenz im Rahmen der Berechnung der Assoziationsregeln (vgl. Kap. 4.4). Da es vor allem bei der Erstellung eines Korpus mittels einer Google-Suchanfrage wichtig ist, zu wissen, welche Webseiten in das Textmaterial eingeflossen sind, wird für alle Sätze auch die URL ihres Quelldokuments gespeichert.

Die Verkettung mehrerer Substantivgruppen über Präpositionen und Artikel bringt allerdings ein doppeltes Problem mit sich. So weisen zunächst Nominalphrasen, die aus vielen Einzelwörtern gebildet wurden, meist nur eine sehr geringe absolute Häufigkeit auf. In Abhängigkeit davon sinkt aber die Anzahl des Auftretens einfacherer Teilphrasen, da diese teilweise als Konstituenten komplexerer Phrasen auftreten und dadurch nicht mehr als eigene Begriffe berücksichtigt werden[50]. Auch wenn die frequenzbasierten Assoziationsregeln nicht unmittelbar aus diesen Rohdaten errechnet werden, erschwert dies bei der zu erwartenden großen Anzahl von extrahierten Begriffen zumindest eine zutreffende Kategorisierung durch den Nutzer. Zudem werden auf diese Weise auch unvollständige Wortfolgen als Nominalphrasen erkannt[51]. Obwohl komplexe Phrasen natürlich semantisch wesentlich gehaltvoller sind, wurde daher die Standardsuche auf einfache Substantivgruppen mit vorausgehenden Adjektiven beschränkt. Die vollständige Suche nach Feldman et al. (1998) wird jedoch weiterhin als eigene Option angeboten. Um zu verhindern, dass Groß- und Kleinschreibung desselben Begriffs zu einem vergleichbaren Effekt führen, werden alle Begriffe bei ihrer Extraktion konsequent in Kleinschreibung überführt.

---

[49] Ein Beispiel: „relevant terms of the text collection“
Bei solchen Mustern handelt es sich formal um reguläre Ausdrücke:
Adjektiv* Substantiv$^{+}$ ( (Präposition|Artikel)$^{+}$ Adjektiv* Substantiv$^{+}$ )*
Dementsprechend wurde die Suche als Endlicher Automat realisiert.

[50] Ein Zahlenbeispiel: Dürfen Substantivgruppen über Präpositionen bzw. Artikel verbunden sein, sollen die Begriffe „relevant terms“ und „text collection“ 30 bzw. 25 Fundstellen aufweisen. „relevant terms of the text collection“ hat 5 Vorkommen im Text. Ist diese Verbindung nicht erlaubt, finden sich im Text 35 Fundstellen für „relevant terms“ und derer 30 für „text collection“.

[51] Bei der Nominalphrase „conflict between USA and North Korea“ würden so „conflict between USA“ und „North Korea“ als eigene Begriffe erkannt, da „and“ weder eine Präposition, noch ein Artikel ist. Geht das „and“ verloren, werden allerdings auch die engen semantischen Bezüge innerhalb der Phrase zerstört.

Ein ähnliches Problem ergibt sich auch dadurch, dass Substantive häufig in einer flektierten Form im Text auftreten. Falls diese als eigener Begriff gezählt werden, senkt dies natürlich die absolute Frequenz der jeweiligen Grundform. Da jedoch beim Penn Tagset Plural- und Genitivendungen als eigene Wortklassen ausgewiesen werden, können die gefundenen Begriffe bereits vor der Kategorisierung durch den Nutzer durch Rückführung auf ihre Grundform weiter vereinfacht werden. Dies ist für das Englische relativ leicht zu bewerkstelligen, da der Genitiv im Singular wie im Plural regelmäßig gebildet wird und sich die Pluralbildung üblicherweise auf einige wenige Möglichkeiten beschränkt[52]. Abgesehen davon, dass unregelmäßige Bildungen durch Regeln nicht erfasst werden können[53], ergeben sich allerdings zusätzliche Probleme, wenn solchermaßen vereinfachte Formen nicht nur wie sonst im Information Retrieval üblich auf einzelne Schlüsselbegriffe angewendet werden, sondern in komplexeren Nominalphrasen auftreten. Vereinfacht werden darf somit nur das letzte Substantiv der Wortgruppe, da alle übrigen Flexionsformen üblicherweise zur Semantik der Phrase beitragen[54]. Auch am Phrasenende kann eine Vereinfachung jedoch kritisch sein, falls es sich dabei um einen Eigennamen handelt[55]. Werden diese beiden Ausnahmen berücksichtigt, kann die Verwendung der beschriebenen Funktion die folgende Kategorisierung durch den Nutzer erheblich vereinfachen, da nur noch unregelmäßig gebildete Formen mit ihrer Grundform zusammengeführt werden müssen und das Hauptaugenmerk der Aggregation der gefundenen Nominalphrasen zu Schlüsselbegriffen gelten kann.

Eine weitere sinnvolle Ergänzung stellt vor allem im Kontext einer geographischen Arbeit die Erkennung von Ortsnamen dar. Soll diese nicht grundsätzlich auf eine bestimmte Region beschränkt sein, ist dazu der Einsatz eines spatial gazetter erforderlich, der die Informationen über Ortsnamen weltweit enthält[56]. Einen geeigneten und zumindest für alle Toponyme außerhalb der Vereinigten Staaten einheitlichen Datensatz bietet die National Geospatial-Intelligence Agency an[57]. Anhand dessen kann für alle aus dem Textmaterial extrahierten Eigennamen überprüft werden, ob es sich bei diesen auch um

[52] Das verwendete Verfahren orientiert sich am Porter Stemmer (Baeza-Yates/Ribeiro-Neto 1999, 433ff.). Folgende Regeln wurden verwendet:
Pluralformen: sses → ss (z.B. „caresses“), ies → y (z.B. „beauties“), s → ∅ (z.B. „words“)
Genitivformen: 's → ∅ (z.B. „Peter's“), s' → ∅ (z.B. „actors' “)

[53] Einige Beispiele: men, children, thieves.

[54] Ein Beispiel: „Uncle Tom's Cabin“.

[55] Ein Beispiel: „Los Angeles“. Eigennamen können mithilfe des Penn Tagsets von anderen Nomina unterschieden werden.

[56] Das in dieser Arbeit verwendete Verfahren basiert auf dem Vorgehen von Rinck (2007, 42 bzw. 44f.), bei dem allerdings nur Toponymdaten für Deutschland benötigt werden.

[57] Eine ständig aktualisierte Liste ist unter http://earth-info.nga.mil/gns/html/namefiles.htm (25.11.07) verfügbar.

einen Ortsnamen handelt. Da das Gazetteer mehrere Millionen Einträge enthält, ist die Überprüfung allerdings sehr aufwändig. Erschwerend kommt hinzu, dass nicht für alle Ortsnamen eine Angabe über Größe bzw. Bedeutung der Örtlichkeit verfügbar ist und daher eine Vielzahl von Eigennamen teilweise kleinsten Ortschaften zugeordnet wird[58]. Trotz dieser Einschränkungen kann eine Ortsnamenerkennung für die Beschreibung der impliziten Textzusammenhänge von Nutzen sein, um aufzudecken, welche Ortsnamen in Form von Argumentationsmustern immer wieder gemeinsam auftreten.

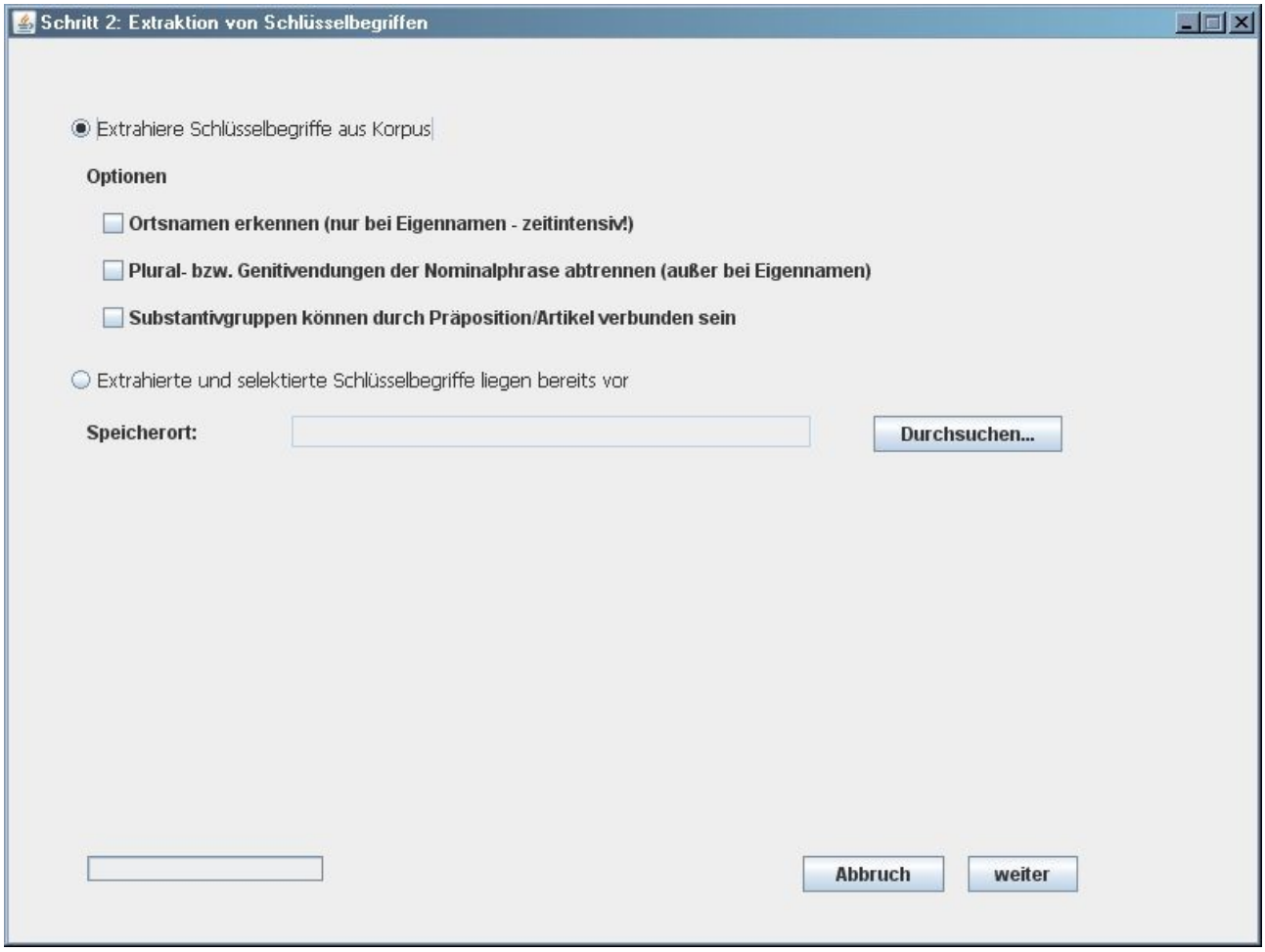



*Abb. 10: Zweiter Teilschritt: Extraktion von Schlüsselbegriffen.*

Allgemein muss natürlich festgehalten werden, dass bei einer Extraktion zentraler Begriffe anhand linguistischer Merkmale immer dann Probleme auftreten, wenn Wörter oder gar ganze Passagen, die in das Korpus mit eingehen, in einer anderen als der durch die Regeln beschriebenen Sprache, in diesem Fall also dem Englischen, verfasst sind. Abb. 10 zeigt die mit dem zweiten Teilschritt assoziierte Sichtweise in Sascet. Gemäß der Abstraktion der Arbeitsphasen untereinander kann hier auch ein extrahiertes bzw. vom Nutzer bereits modifiziertes System von Begriffen bzw. Kategorien wiederhergestellt

[58] Eine Übersicht über mögliche Probleme bei der Arbeit mit spatial gazetters findet sich bei Rinck (2007, 11f.).

werden, sofern die Schrittfolge zuvor bereits bis zu diesem Punkt durchlaufen und die Ergebnisse gesichert wurden.

## *4.3 Kategorisierung der Schlüsselbegriffe*

Da der Schritt der Generalisierung und Kategorisierung der automatisch extrahierten Begriffe ausschließlich in der Verantwortung des Nutzers verbleibt, kann eine programmseitige Unterstützung bei dieser Aufgabe sinnvollerweise nur darin bestehen, die bisherigen Ergebnisse geeignet darzustellen und elementare Funktionen zu deren Bearbeitung zur Verfügung zu stellen. Dementsprechend werden in diesem dritten Teilschritt die extrahierten Nominalphrasen als Liste angezeigt, die wahlweise nach Frequenz oder alphabetisch sortiert werden kann. Während die Sortierung nach Häufigkeit einen ersten Aufschluss über zentrale Begriffe des Textmaterials geben kann, unterstützt die Sortierung nach dem Alphabet den Nutzer beim Auffinden von graphematischen Varianten oder verbleibenden Flexionsendungen ein und desselben Begriffs.

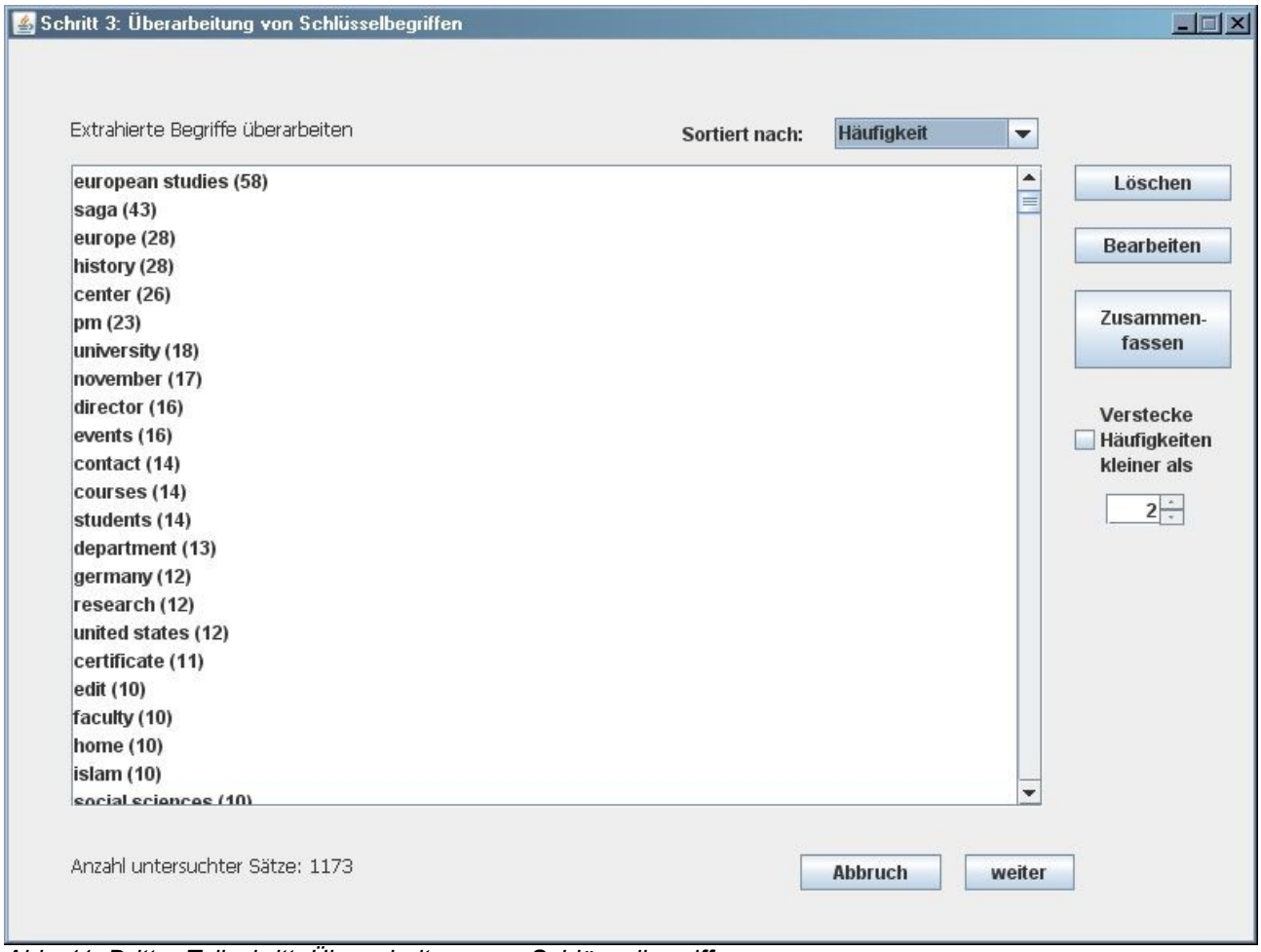


*Abb. 11: Dritter Teilschritt: Überarbeitung von Schlüsselbegriffen*

Diese Liste kann nun in geeigneter Weise editiert werden. Die Möglichkeit, Begriffe zu löschen, ist vor allem bei Korpora relevant, die mittels einer Google-Suchanfrage erzeugt wurden. Auf diese Weise können Html-Auszeichnungselemente, die trotz der Vorverarbeitungsschritte im Textkorpus verblieben sind, ebenso entfernt werden wie Zeichenfolgen, die aufgrund ungewöhnlicher Zeichenkodierungen nicht sinnvoll verarbeitet werden konnten. Letzteres ist auf außereuropäischen Seiten durchaus möglich, betrifft aber nur Webseiten in der jeweiligen Landessprache, die wegen der Beschränkung auf das Englische in dieser Arbeit ohnehin nicht betrachtet werden können. Zentral für die Erstellung eines einfachen Kategoriensystems sind die Funktionen, die das Umbenennen und Zusammenfassen von Begriffen erlauben. Hierunter kann die Zusammenführung von Synonymen oder graphematischen Varianten[59] genauso fallen wie die Aggregation der gefundenen Begriffe zu semantischen Oberbegriffen.

Die Kategorisierung der extrahierten Begriffe hat natürlich eine Auswirkung auf die zugrundeliegenden Daten. So repräsentieren alle Fundstellen der ursprünglichen Begriffe nun gemeinsam die Fundstellen der neu eingeführten Kategorie. Entsprechend ergibt sich die Frequenz des neuen Schlüsselbegriffs aus der Summe der Frequenzen der beitragenden Begriffe. Abb. 11 zeigt die Visualisierung des dritten Teilschritts in Sascet.

## *4.4 Association rule mining*

Bei der Implementierung der Berechnung der Assoziationsregeln stellt vor allem der Aufbau der Potenzmenge (vgl. Kap. 3.4) ein algorithmisches Problem dar. Das verwendete Verfahren orientiert sich am Vorgehen von Hipp et al. (2000). Dort wird als Datenstruktur zum Durchlaufen der Potenzmenge ein Präfixbaum vorgeschlagen (Hipp et al. 2000, 60). Alle Elemente der Ausgangsmenge werden mit einem Index versehen. Anschließend wird der Suchraum von den ein- zu den mehrelementigen Teilmengen hin gemäß der Regel aufgebaut, dass neu hinzukommende Elemente immer einen höheren Index als den höchsten bereits in der Teilmenge vertretenen besitzen müssen. Abb. 12 illustriert, dass auf diese Weise die Potenzmenge vollständig durchsucht wird und gleichzeitig Dopplungen effektiv vermieden werden. Die Trennlinie visualisiert dabei wiederum den Schwellwert, ab dem aufgrund zu geringer Frequenz keine Kombinationen mehr untersucht werden müssen (pruning – vgl. Kap. 3.4).

[59] Dazu müssen auch Tippfehler gerechnet werden, die in gleicher Weise die Frequenz eines Begriffs senken, wie dies Folge unterschiedlicher Flexionsformen sein kann.

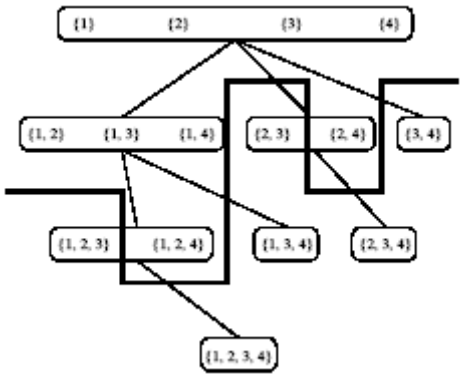



*Abb. 12: Effizienter Aufbau einer Potenzmenge nach Hipp et al. (2000, 60)*

Da das gemeinsame Auftreten der Elemente in einer Teilmenge niemals eine höhere Häufigkeit als die der in ihnen enthaltenen Einzelbegriffe aufweisen kann (*downward closure property* nach Hipp et al. 2000, 59), müssen Begriffe nicht berücksichtigt werden, die bereits für sich genommen zu selten im Text auftreten. Weil aber die Frequenzen der aus dem Textkorpus extrahierten Schlüsselbegriffe zumindest am Anfang für die einelementigen Mengen bereits bekannt sind, macht es Sinn, die Indizes als Ränge der absteigend nach ihrer Häufigkeit sortierten Schlüsselbegriffe zu vergeben. So kann die große Anzahl an Elementen geringer Häufigkeit bereits vor dem Aufbau des Präfixbaumes aus der Ausgangsmenge ausgeschlossen werden. Dies reduziert die Zahl der zu untersuchenden Kombinationen, ohne jedoch das Ergebnis zu verändern (pruning – vgl. Kap. 3.4).

Prinzipiell kann der Aufbau des Präfixbaumes sowohl als Tiefen- als auch als Breitensuche effizient realisiert werden (Hipp et al. 2000, 60f.). Bei Sascet wurde die Breitensuche gewählt, d.h. es werden zunächst alle noch möglichen n-elementigen Teilmengen überprüft, bevor Mengen mit n+1 Elementen gebildet und getestet werden. Wie oft die Einzelbegriffe einer solchen Teilmenge kookkurrieren und ob diese folglich über einen ausreichenden support[60] für eine mögliche Regel verfügt, wird dadurch berechnet, indem die Satz-IDs der Fundstellen der Begriffe miteinander verglichen werden[61].

Vorteil der Breitensuche ist, dass bereits während des Aufbaus des Präfixbaumes alle Teilmengen, deren Elemente hinreichend häufig kookkurrieren, auf Assoziationsregeln überprüft werden können, da der support für alle echten Teilmengen bereits vorliegt. Weil die confidence einer Regel nur vom support der Prämisse und dem support des Konsequenten abhängt (vgl. Kap 3.4), kann dieses Maß für alle möglichen Regeln einer

[60] Zur genauen Definition von support und confidence siehe Kap. 3.4.

[61] Wenn bekannt ist, dass der Begriff „Jerusalem“ in den Sätzen mit der ID 3, 4 und 7 auftritt und „Heilige Stadt“ in den Sätzen mit der ID 2, 7 und 20, kann ohne weitere Betrachtung der Satzinhalte ein gemeinsames Vorkommen in genau einem Fall errechnet werden.

Teilmenge also unmittelbar nach deren Generierung berechnet werden. Überschreitet die confidence einer Regel den nötigen Schwellwert, kann diese zur Ergebnismenge hinzugefügt werden. Eine Regel der Form A → B muss dabei so gelesen werden, dass an einem relativ hohen Anteil der Fundstellen von A auch über B gesprochen wird.

Grundsätzlich fungiert der Schwellwert für confidence also als ein Signifikanzmaß dafür, ab welchem Anteil ein gemeinsames Auftreten mehrerer Begriffe kein Zufall mehr sein kann, während ein Mindestwert für support verhindert, dass eine solche Berechnung auf viel zu kleinen Fallzahlen basiert, um verlässliche Aussagen über die Abhängigkeit machen zu können[62]. Bisher wurden jedoch aus gutem Grund keine exakten Werte für minimalen support bzw. confidence genannt. Diese hängen nämlich in erheblichem Maße von den zugrundeliegenden Daten ab, über denen die Assoziationsregeln berechnet werden sollen. So gibt es zwar Untersuchungen, die für eine Regelgenerierung aus großen Datenbanken für confidence einen Wert von 50% bei einer unteren Grenze von 1% aller Datensätze für support empfehlen (Coenen et al. 2005, 219ff.). Dies ist aber auf Texte nur bedingt übertragbar, da einzelne Begriffe dort ohnehin seltener auftreten als die häufig einem festen Wertebereich entstammenden Einträge eines Datensatzes. Auch bei der Signifikanz müssen bei Texten leichte Abstriche gemacht werden. In ersten Probedurchläufen von Sascet (vgl. Kap. 4.5) hat sich gezeigt, dass zumindest bei der Arbeit mit Webdokumenten Werte von 0,2-0,5% aller Sätze des Korpus für support bzw. Werte von 40-50% für confidence zu sinnvollen Ergebnissen führen. Um dem Nutzer zu erlauben, innerhalb der verbleibenden Spanne brauchbare, zugleich aber auch überschaubare Ergebnisse zu erzeugen, können diese Werte vor der Berechnung der Assoziationsregeln modifiziert werden. Abb. 13 zeigt die entsprechende Sichtweise bei Sascet.

[62] Zwei erläuternde Beispiele: In einem Text treten „Azteken“ und „Kakao“ jeweils genau einmal, jedoch gemeinsam auf. Ohne einen Schwellwert für support würden Regeln generiert, die auf eine hundertprozentige wechselseitige Abhängigkeit schließen lassen.
In einem anderen Text kommen „EU-Kommission“ und „Europaparlament“ in 10 Sätzen gemeinsam vor. Da „Europaparlament“ jedoch in 80, „EU-Kommission“ gar in 100 Sätzen ohne den jeweils anderen Begriff zu finden ist, liegt keine signifikante Kookkurenz der beiden Begriffe vor.

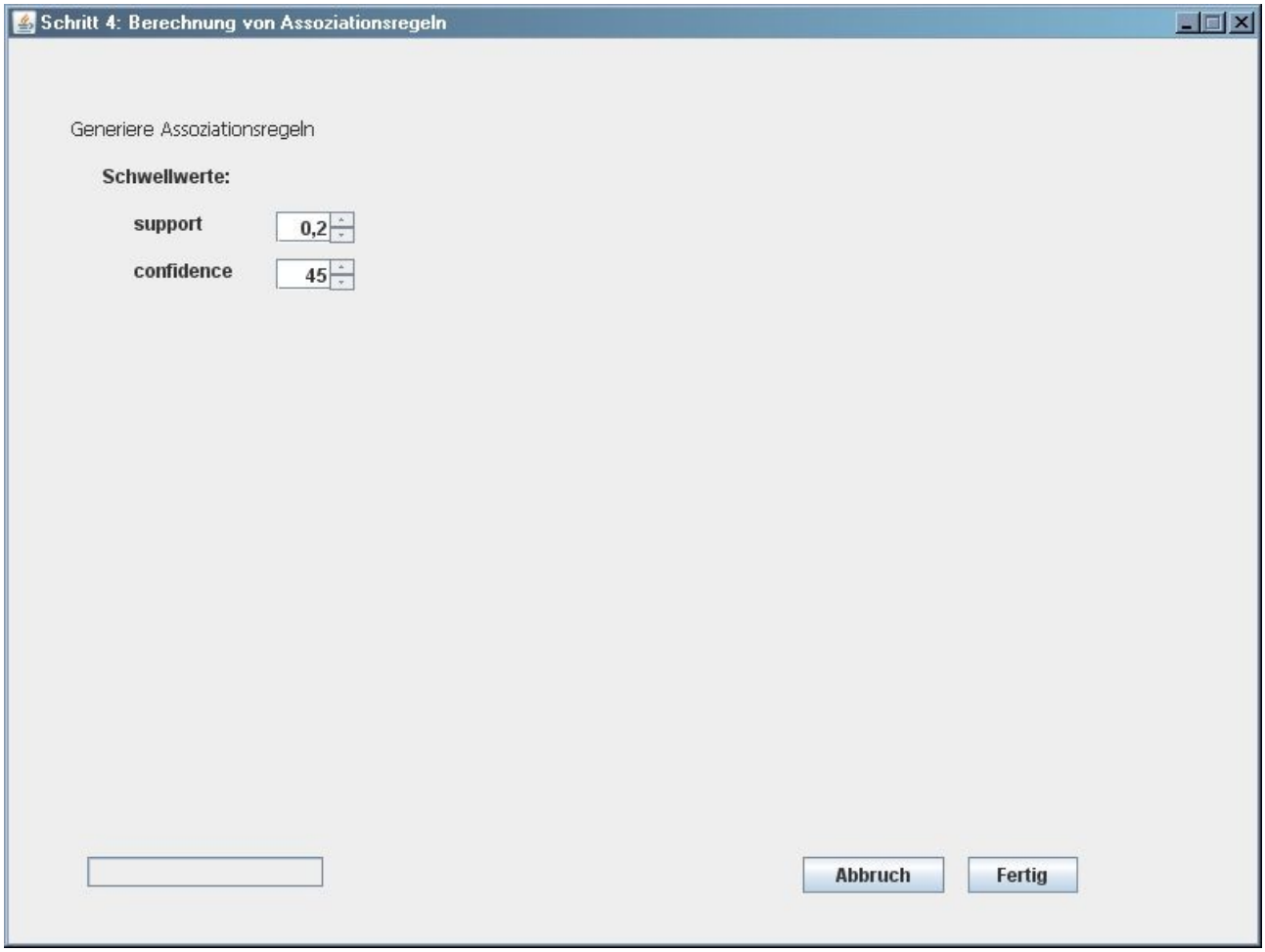


*Abb. 13: Vierter Teilschritt: Berechnung von Assoziationsregeln*

Nach der Berechnung der Assoziationsregeln kann das vollständige Ergebnis eines Programmdurchlaufs in geeigneter Form gespeichert werden. Für jede Regel werden dabei die Begriffe ihrer Prämisse und ihres Konsequenten mit allen zugehörigen Daten für eine geeignete Visualisierung bereitgestellt. Die Qualität jeder Regel wird durch support und confidence näher beschrieben. Als weitere wichtige Information für Optimierungen werden daneben auch die vom Nutzer gewählten Schwellwerte für support und confidence mit im Datensatz vermerkt. Sowohl bei den Schwellwerten als auch im Einzelfall wird hierbei nur für confidence die relative Häufigkeit angegeben. Für den support wird dagegen die absolute Frequenz gespeichert. Dies lässt bei den generell niedrigen Häufigkeiten, wie sie vor allem in heterogenen Textkorpora auftreten können, eine bessere Aussage darüber zu, wie verlässlich der durch die Regel beschriebene Zusammenhang ist, als dies bei einem reinen Prozentwert der Fall wäre. Die gespeicherten Regeln sind auch die Grundlage für die Wiederherstellung der graphischen Visualisierung beim Öffnen eines bereits existierenden Analyseprojektes in Sascet.

## *4.5 Graphische Visualisierung: Möglichkeiten zur Interpretation der Ergebnisse*

Da sich die unter Umständen große Zahl errechneter Assoziationsregeln einer unmittelbaren Interpretation nur schwer erschließt, müssen die gefundenen Zusammenhänge in geeigneter Weise visualisiert werden. Zu diesem Zweck bieten sich Graph- bzw. Netzwerkdarstellungen an, wie sie als originäres Arbeitsmittel ja auch in der qualitativen Inhaltsanalyse verwendet werden (vgl. Kap. 3.1). Schlüsselbegriffe werden dabei als Knoten, die Assoziationen zwischen ihnen als gerichtete Kanten dargestellt, wobei die Richtung der Kante der Implikationsrichtung der Regel entspricht. An die Kante wird dabei der zugehörige confidence-Wert der Regel annotiert. Abb. 14 verdeutlicht das Vorgehen.

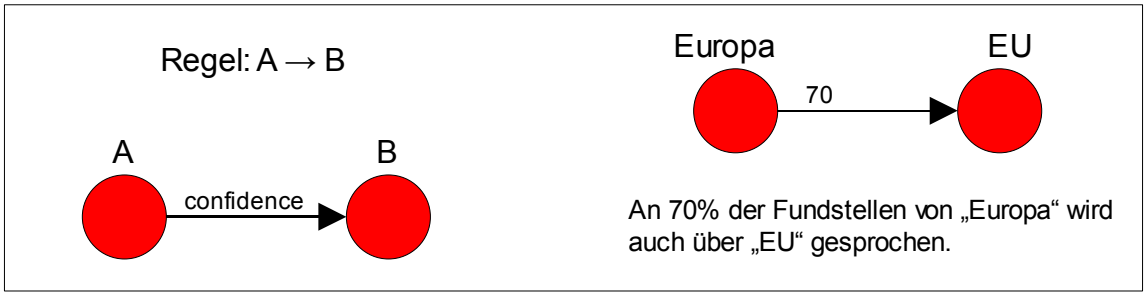



*Abb. 14: Links: Visualisierung der Assoziationsregeln als Graph. Rechts: Ein Beispiel und seine mögliche Interpretation.*

Auf diese Weise können natürlich nur Regeln mit jeweils einem Schlüsselbegriff in der Prämisse bzw. im Konsequenten angezeigt werden. Es wäre zwar durchaus möglich, durch die Einführung besonders hervorgehobener Kanten bzw. Knoten auch komplexere Regeln darzustellen; der resultierende Graph wäre im Gegenzug aber wiederum schwieriger zu interpretieren. Daher werden die übrigen Regeln und ihre zugehörigen Maßzahlen bei Sascet im unteren Teil des Hauptfensters in textueller Form ausgegeben (vgl. Abb. 15).

Für die technische Realisierung des Graphen wurde mit JUNG[63] eine unter freier Lizenz stehende, bereits verfügbare Programmbibliothek verwendet. Auswahlkriterium war hierbei vor allem die hinreichend große Zahl von Softwareprojekten[64], bei denen JUNG bereits erfolgreich und stabil eingesetzt wird, sowie eine ausführliche Dokumentation der äußerst variabel einsetzbaren Funktionen[65]. Der zuvor formal beschriebene Graph wurde bei seiner Implementierung insofern erweitert, dass nach einem Klick auf einen Knoten

[63] Für weitere Informationen siehe http://jung.sourceforge.net (28.11.07).
[64] Vgl. http://jung.sourceforge.net/pmwiki/index.php/Main/ProjectsUsingJUNG (28.11.07).
[65] Vgl. http://jung.sourceforge.net/doc/index.html (28.11.07).

support, Fundstellen und Quellenangaben des durch ihn repräsentierten Begriffs anstelle der komplexen Regeln im unteren Teil des Hauptfensters angezeigt werden. Dies ermöglicht vor allem bei automatisch generierten Korpora eine Abschätzung, welche Aussage das zugrundeliegende Textmaterial über diesen Begriff zulässt[66].

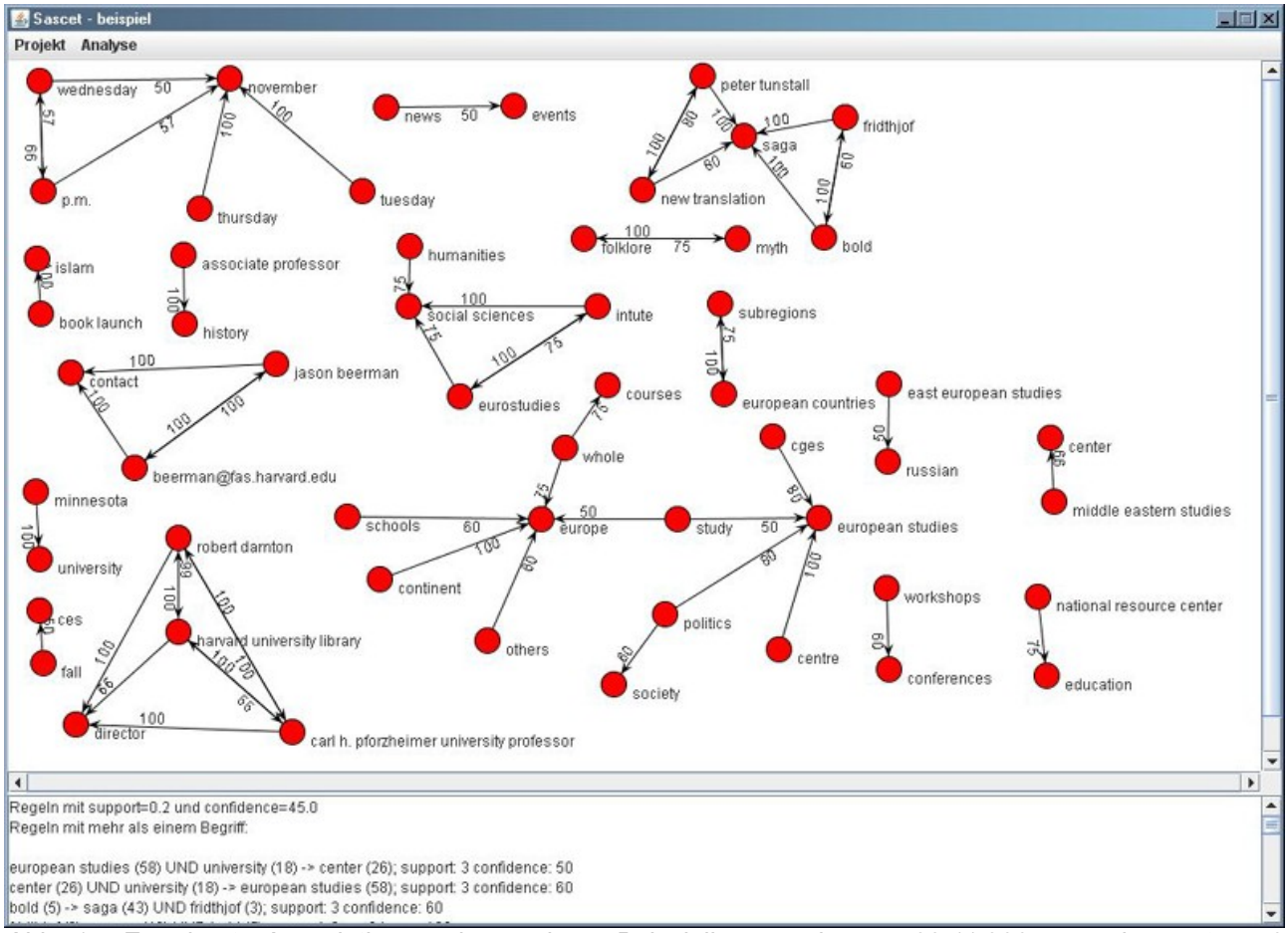



*Abb. 15: Errechnete Assoziationsegeln zu einem Beispielkorpus, das am 22.11.2007 aus den ersten 50 Treffern der Google-Suchanfrage unter dem Stichwort „European Studies" generiert wurde.*

Abb. 15 zeigt für ein Korpus, das mittels einer allgemeinen Google-Suchanfrage unter dem Stichwort „European Studies“ generiert wurde, die Visualisierung der errechneten Regeln in Sascet. Auch wenn bei dieser Beispielanfrage auf eine Kategorisierung der automatisch extrahierten Begriffe verzichtet wurde, lassen sich erste Ansätze zu einer Interpretation eines solchen Graphen ohne Weiteres durchführen und erläutern. Beinahe erwartungsgemäß zeichnet der Graph sowohl die Standorte, an denen unter dem Stichwort „European Studies“ Forschung betrieben wird, als auch die unter diesem Begriff untersuchten Teilregionen nach. Besonders auffällig ist, dass die Kontaktdaten im Falle der Harvard-Universität offensichtlich hinreichend oft in das Korpus Eingang fanden, um als Visitenkarte in Form eines zyklischen Teilgraphen zur Menge der Regeln beizutragen.

[66] So müssen Kookkurrenzen in Überschriften oder Menüpunkten natürlich anders bewertet werden als im Fließtext.

Dass das Web keine Historie, sondern immer nur seinen aktuellen Zustand besitzt (vgl. einleitend Kap. 3), findet auch bei diesem Beispiel seinen Niederschlag in aktuellen Datumsangaben.

Rund um den zentralen Suchbegriff „European Studies“ spannt sich der komplexeste Graph auf. Hier wird unter anderem die organisationstechnische Einordnung des Forschungsbereiches in den Kontext von „humanities“ und „social sciences“ bzw. „politics“ und „society“ sichtbar. Intuitiv verständlich ist der Verweis von einem Center of German & European Studies („CGES“) an der „University“ of „Minnesota“ auf „European Studies“ allgemein. Interessant ist daneben auch, dass unter dem Stichwort „East European Studies“ in der Hälfte der Fälle auch von „Russian“ gesprochen wird.

Diese intuitive Art und Weise, den Ergebnisgraphen zu lesen, ist jedoch auch mit Risiken behaftet. So wird natürlich kein vollständiges Abbild der Textinhalte, sondern nur eine Zusammenschau signifikanter Zusammenhänge dargestellt. Wenn also mit dem Stichwort „East European Studies“ nur „Russian“ verknüpft ist, darf das keinesfalls vorschnell so interpretiert werden, dass in 50% der Fälle „nur“ von Russland die Rede sei. So können an diesen Fundstellen durchaus weitere Aussagen mit den beiden Begriffen verknüpft sein, die im Graphen allerdings nicht mehr sichtbar sind. Aufschluss über den tatsächlichen Sachverhalt kann nur ein Blick auf die vollständige Liste der Fundstellen im unteren Teil des Hauptfensters geben.

Ein weiteres Problem ist die zunächst kontraintuitive Darstellung der Knoten mit mehreren eingehenden Kanten (z.B. „Europe“, „European Studies“ oder „Social Sciences“). So mutet es befremdlich an, dass überall dort, wo allgemeine Begriffe wie „schools“, „continent“ oder „study“ verwendet werden, signifikant häufig auch über Europa gesprochen wird. Vergegenwärtigt man sich jedoch, dass das Korpus unter gezielter Anfrage nach dem Stichwort „European Studies“ erzeugt wurde, wird klar, dass sich der Sachverhalt unter der Annahme einer natürlich unvollständigen Aufzählung auch so lesen lässt, dass sich an den Fundstellen des Begriffs „Europe“ mit hoher Wahrscheinlichkeit zumindest eines der Stichworte „schools“, „continent“ bzw. „study“ findet[67]. Selbst wenn die Summe der confidence der Kanten den Wert 1 weit überschreitet, besteht aufgrund von Überlappungen natürlich noch immer die Möglichkeit, dass sich auch Fundstellen ohne einen der verweisenden Begriffe finden. Dennoch kann man festhalten, dass der hohe Vernetzungsgrad eines Begriffs nicht nur ein Indiz für seine zentrale Relevanz darstellt,

[67] Auch formal gilt in einem Universum, sofern es vollständig und ausschließlich aus den Elementen A, B, ... , Z gebildet wird: $(A \rightarrow B) \cap (C \rightarrow B) \cap ... \cap (Z \rightarrow B) = B$.

sondern dass sich darüber hinaus ein umso differenzierteres Bild vom semantischen Kontext eines Begriffs ergibt, je mehr andere Begriffe in Form einer Assoziationsregel auf ihn verweisen.

# 5 Der Europabegriff auf außereuropäischen Webseiten – die Beispiele Nigeria, Indien und Australien

Nachdem die geeigneten methodischen Mittel für eine Untersuchung des Europabegriffs auf außereuropäischen Webseiten eruiert, in Form eines Programms implementiert und die dabei getroffenen Entscheidungen dokumentiert wurden, können sie nun zur Beantwortung der gestellten Forschungsfrage eingesetzt werden. Davor gilt es noch, die zunächst bewusst allgemein gehaltene Fragestellung genauer zu definieren und dadurch geeignete Korpora für die Untersuchung abzugrenzen.

Wie bereits in Kap. 2.2 diskutiert, wird unter dem Stichwort „Europa“ häufig ausschließlich die Europäische Union referenziert. Nur insofern die beiden Begriffe immer öfter in eins gesetzt werden, kann man davon sprechen, dass der Balkan (noch) nicht in Europa liegt (vgl. Kap 2.3). Um derartigen Regionalisierungen keinen Vorschub zu leisten bzw. um ein mögliches Resultat dieser Untersuchung nicht schon implizit vorwegzunehmen, wird jedoch am allgemeinen Europabegriff als Gegenstand dieser Arbeit festgehalten.

Auch wurde bereits thematisiert, aus welchen Gründen es problematisch ist, den außereuropäischen Europabegriff nur auf der sehr abstrakten Ebene der Nationalstaaten zu betrachten. Eine solche Eingrenzung kann sinnvoll sein, falls repräsentative und gut vergleichbare Akteure jedes Staates ausgewählt werden (vgl. Kap. 2.2). Dies ist bei einer Untersuchung von Webseiten zumindest nicht für die Textinhalte gegeben, die allgemein unter der entsprechenden top level domain[68] eingestellt werden[69]. Da eine unmittelbare Suchanfrage auf Nationalstaatsebene also nicht sinnvoll möglich ist, muss sie durch die Wahl geeigneter Subdomains weiter eingeschränkt werden. Hier bieten sich natürlich in erster Linie Regierungsinstitutionen an, da diese selbst im Falle einer innenpolitisch unsicheren Lage außenpolitisch als Repräsentanten ihres Landes fungieren. Obwohl eine solche Sichtweise unter Umständen nur die Meinung einer schmalen Schicht

[68] Für Deutschland z.B. das Suffix .de. Top level domains sollen im Folgenden auch kürzer mit „Länderdomäne“ bezeichnet werden können.

[69] Gerade bei kleinen Ländern ist der Handel mit Länderdomänen keine Seltenheit. So lässt sich anhand der Domain .tv eher eine Übersicht über verschiedene Fernsehsender weltweit als eine Aussage über spezifische Sichtweisen einzelner Bewohner oder gar repräsentativer Akteure von Tuvalu gewinnen.

widerspiegelt, besitzt sie doch internationale Wirkkraft, da eine Regierung im Regelfall auch über die nötigen Machtmittel verfügt, um gemäß dieser Sichtweisen zu handeln.

Dass mit Sascet nur eine Auswertung von Webseiten möglich ist, die in englischer Sprache verfasst sind, schränkt die Menge der in Frage kommenden Informationsquellen weiter ein. Gerade in Ländern, in denen Englisch zwar offizielle Amtssprache ist, jedoch daneben auch eine Vielzahl indigener Sprachen existiert, besteht die Problematik, dass Webseiten in englischer Sprache explizit an Leser im Ausland adressiert sind und nicht notwendigerweise die interne Überzeugung der jeweiligen Regierung repräsentieren. Dies erfordert zwar eine entsprechend differenziertere Betrachtung möglicher Ergebnisse, unterscheidet sich aber nicht grundlegend von der Gefahr gezielter Fehlinformation, der sich ein „fremder“ Interviewer in einer vergleichbaren Situation gegenübersehen würde.

Die Suche nach geeigneten Domains innerhalb dieser Vorgaben wird dadurch begünstigt, dass die Webseiten von Regierungsbehörden in englischsprachigen Ländern zumeist unter der Subdomäne „gov“ der entsprechenden Landesdomäne zu finden sind. Im Gegensatz zu den Landesdomänen, die dem freien Handel unterliegen, darf in diesem Fall auch davon ausgegangen werden, dass hier im Sinne der konsistenten Sprechweise eines kollektiven Akteurs eine weitgehend einheitliche, nämlich vorwiegend außenpolitische Sichtweise auf Europa abgebildet wird[70].

Bei der Auswahl geeigneter Staaten wird darauf zu achten sein, dass möglichst gute Voraussetzungen für die Herausarbeitung von Unterschieden gegeben sind, um sie später im Sinne einer vergleichenden Gegenüberstellung (vgl. Kap. 3.1 bzw. Kap. 5.4) zueinander in Beziehung setzen zu können. An dieser Stelle fließt zwar Vorwissen in Form einer gewissen Erwartung an die Sichtweise dieser Webseiten mit in die Studie ein. Im Sinne der Definition von Interessantheit als Abweichung der Ergebnisse von eben dieser Erwartungshaltung (vgl. Kap. 3.3) ist jedoch eine Auswahl von Beispieldomains auf der Grundlage solcher Voreinschätzungen methodisch gerechtfertigt. Da rechnergestützte Auswertungen trivialerweise wertfrei erfolgen (vgl. einleitend Kap. 3), beeinflusst dieses Vorwissen auch nicht die eigentliche Textanalyse, kann aber gut als kontrastierender Bezugspunkt für eine Interpretation der Ergebnisse dienen.

Will man unter den gegebenen Voraussetzungen eine möglichst gleichmäßige Verteilung der Untersuchungsgebiete weltweit erreichen und gleichzeitig die Sichtweise der USA als zu sehr involviert ausschließen, bieten sich die Staaten des Commonwealth

[70] Natürlich kann die Sichtweise verschiedener Ministerien im Detail voneinander abweichen, von einer gewissen strategischen Grundhaltung darf jedoch zunächst ausgegangen werden.

als Kandidaten für die engere Auswahl an, da dort Englisch noch immer zumindest gleichberechtigte Amtssprache ist. Zwar unterhalten auch diese Staaten wirtschaftliche Beziehungen mit Europa, besonders mit dem Vereinigten Königreich. Dies ist aber wiederum darauf zurückzuführen, dass sich die englische Sprache abgesehen von wenigen Ausnahmen[71] eben gerade in denjenigen ehemaligen Kolonien Großbritanniens als Amtssprache erhalten hat, die heute noch im Commonwealth vertreten sind. Soll also vermieden werden, dass bei Staaten mit anderer Amtssprache englischsprachige Texte eigens für ausländische Besucher der Webseiten angeboten werden, ist diese Vorauswahl die einzig sinnvolle Alternative.

Gemäß der beschriebenen Vorgaben werden drei geeignete Staaten im Folgenden kurz vorgestellt. Die verwendeten Hintergrundinformationen haben freilich den Charakter von Basisdaten und können die Länder nur unzureichend beschreiben. Als kurze Zusammenschau der aktuellen Entwicklungen spiegelt jedoch genau dies den Rahmen des üblichen Vorwissens über diese Länder wieder, wie er auch aus der Tagespresse erworben werden kann.

Als erste Beispiele dienen Indien und Nigeria. Abgesehen von den genannten Handelsverflechtungen können die Sichtweisen in beiden Staaten den eindeutig außereuropäischen Sichtweisen zugeordnet werden (vgl. Kap. 2.4). Somit steht zu vermuten, dass sich der Bezug zu Europa auf den Webseiten der jeweiligen Regierungen vor allem in wirtschaftlichen Aspekten niederschlägt. Hierbei macht vor allem Indien zuletzt durch ein konstant hohes Wirtschaftswachstum auf sich aufmerksam. Die hohe Staatsverschuldung zwingt die Regierung jedoch dazu, das Interesse der Investoren aus den Vereinigten Staaten, Südostasien und zunehmend auch Europa zur Sanierung der maroden Infrastruktur zu nutzen. Die als Anreiz für Investitionen eingerichteten Sonderwirtschaftszonen verstärken aber auch die ungleiche Verteilung des Wohlstands und erhöhen dadurch das Gewaltpotential in Indien (Fischer Weltalmanach 2007, 218ff.).

Wichtigste Exportgüter Nigerias sind mit einem Anteil von zusammen 98% Erdöl und Erdgas. Zentrum der Ölförderung ist das Nigerdelta, wo verschiedene bewaffnete und gut organisierte Gruppierungen für eine bessere Beteiligung der Bevölkerung an den Gewinnen der zumeist ausländischen Ölkonzerne kämpfen. Dabei kam es in der Vergangenheit immer wieder zu Geiselnahmen von Ausländern. Einen weiteren Konfliktherd in Nigeria, das von insgesamt 434 Ethnien bewohnt wird, stellen die

[71] Z.B. Zimbabwe oder Liberia (Fischer Weltalmanach 2007, 436 bzw. 309)

wiederholten Zusammenstöße zwischen Christen und Moslems im Norden des Landes dar (Fischer Weltalmanach 2007, 355ff.).

Als drittes, eher kontrastierendes Beispiel wurde Australien gewählt, bei dem zweifelsohne eine wesentlich engere, vor allem historische Bindung zu Europa gegeben ist. Bedingt durch die Lage bestehen jedoch starke wirtschaftliche Verflechtungen mit Südostasien, von wo aus sich das rohstoffreiche, aber dünn besiedelte Australien einem wachsenden Einwanderungsdruck ausgesetzt sieht (Fischer Weltalmanach 2007, 66ff.). In keiner Weise ist also eines der beschriebenen Abhängigkeitsverhältnisse zu Europa (vgl. Kap. 2.4) zu verzeichnen. So darf auch aus diesem Blickwinkel von einer Außensicht, wenngleich ganz anderer Natur, gesprochen werden.

In allen drei Fällen folgen die Adressen der untersuchten Beispieldomains der oben beschriebenen Struktur[72]. Bei der Auswertung ist zu beachten, dass die Bevölkerung in Nigeria (0,7%) und Indien (1,6%) zu einem weit geringeren Anteil Zugang zu Computern hat als in Australien (68, 3%) (Fischer Weltalmanach 2007, CD-Teil). Bei den beiden Erstgenannten muss also beachtet werden, dass Inhalte möglicherweise explizit für eine schmale Bevölkerungsschicht bzw. Leser im Ausland angeboten werden.

Nachdem die Untersuchung erfolgreich auf geeignete Domains eingegrenzt wurde, können nun die einzelnen Korpora mithilfe von Sascet zusammengestellt, analysiert und entsprechend interpretiert werden. Dabei findet folgendes Vorgehen Anwendung: Das Korpus wird jeweils über den ersten 200 Treffern[73] einer Google-Suchanfrage nach dem Stichwort „Europe“ erstellt, wobei die Suche auf die jeweilige Domäne eingeschränkt ist[74]. Für jedes Korpus wird daraufhin zunächst dokumentiert, mit welchem Anteil die verschiedenen Subdomains der jeweiligen Regierungs-Webseiten[75] zum Textumfang des Korpus beigetragen haben (Kap. 5.1). Anschließend erfolgt die eigentliche, zweistufige Analyse des Textmaterials. In einem ersten Schritt werden mithilfe von Sascet zunächst Assoziationsregeln über der noch nicht kategorisierten Menge von Schlüsselbegriffen errechnet (Kap. 5.2). Auf diese Weise können ein erster Überblick über die im Textkorpus vorhandenen semantischen Felder gewonnen und Kandidaten für zentrale Schlüsselbegriffe ermittelt werden (vgl. Kap 4.5). Diese Erkenntnisse dienen in einem zweiten Schritt als Anhaltspunkte für ein Kategoriensystem, das durch sukzessive

[72] Für Nigeria: *.gov.ng, für Indien: *.gov.in und für Australien *.gov.au.
[73] Kleinere Treffermengen lieferten teilweise nicht verwertbare Ergebnisse (vgl. Kap. 5.2).
[74] Für das Beispiel Australien bewirkt die Suchanfrage **Europe site:gov.au**, dass nur auf den Webseiten der australischen Regierung nach dem Stichwort **Europe** gesucht wird.
[75] Durch solche Subdomains können wiederum Akteure innerhalb der Regierungs-Domain wie Ministerien, Institute oder andere Einrichtungen abgebildet und folglich differenziert als solche erfasst werden.

Abstraktion aus den automatisch extrahierten Begriffen abgeleitet werden kann (Kap. 5.3). Hierbei wird eine mittlere Abstraktionsebene angestrebt, bei der zwar semantische Varianten desselben Topos ausgeschlossen werden können, aber noch hinreichend viele unkategorisierte Begriffe für eine Berechnung von Assoziationsregeln zur Verfügung stehen. Die Verteilung der verschiedenen Subdomains im Korpus, die Ergebnisse des ersten Durchlaufs und die Assoziationsregeln, die im Zuge des zweiten Durchlaufs zwischen den Kategorien errechnet wurden, bilden anschließend zusammen mit den bereits vorgestellten Hintergrundinformationen die Grundlage für die Interpretation der jeweiligen Ergebnisse. Abschließend werden die drei Beispieldomains im Sinne einer vergleichenden Gegenüberstellung (vgl. Kap. 3.1) zueinander in Bezug gesetzt (Kap 5.4).

## *5.1 Zusammenstellung des Textkorpus*

Um den Kriterien einer synchronen Untersuchung zu genügen (vgl. einleitend Kap. 3), wurden alle drei Korpora innerhalb weniger Stunden im Zeitraum vom 18.06.2007 (gov.in) bis 19.06.2007 (gov.ng und gov.au) zusammengestellt. Von den 200 ersten Treffern der jeweiligen Google-Suchanfrage konnten zwischen 82% und 88% auch tatsächlich heruntergeladen werden[76] (vgl. Kap. 4.1). Der Gesamtumfang aller heruntergeladenen Dateien reicht dabei von 4,0 MB (gov.au) über 4,2 MB (gov.in) bis zu 6,7 MB (gov.ng). Zu keinem der drei Korpora trug ein einzelnes Html-Dokument mehr als 4% des Gesamtvolumens bei. Wesentlich größer waren nur einige wenige Dateien anderen Formats. Da alle Dateien von JWebPro als reines Textformat interpretiert werden, wurden diejenigen aus dem jeweiligen Korpus entfernt, für die dieses Vorgehen zu keinem sinnvollen Ergebnis führte[77].

Für die solchermaßen modifizierten Korpora kann nun ihre genaue Zusammensetzung näher beschrieben werden. Dies ist deshalb von Bedeutung, da das Korpus automatisch erzeugt wurde und die genaue Bewertungsfunktion, auf deren Grundlage Suchtreffer zu einer bestimmten Suchanfrage zurückgegeben werden, von Google leider nicht offengelegt wird (vgl. einleitend Kap. 4). Weil die Art der Textinhalte jedoch entscheidend

[76] Von *.gov.ng konnten 164, von *.gov.au 168 und von *.gov.in 176 Dokumente der 200 Suchtreffer zum Stichwort „Europe“ erfolgreich heruntergeladen werden. Die Ausfälle verteilen sich dabei zufällig auf die Menge der Treffer. Im Falle Australiens wurde das automatische Retrieval in zwei unabhängigen Versuchen jeweils beim 189. Treffer unterbrochen. Die letzten 12 Dokumente fehlen daher vollständig.

[77] Verteilt über die drei Korpora fanden sich fünf *.doc, drei *.pdf sowie ein *.ppt-Dokument. Aufgrund der Einschränkung von JWebPro bzw. JTextPro auf Textdaten (vgl. einleitend Kap. 4) konnten nur die beiden letzten Formate nicht ausgewertet werden.

vom Ursprung der Quelldaten abhängt, sollen als Kennzahlen zur Beschreibung des jeweiligen Korpus die Anzahl der Treffer und der Textumfang dienen, mit denen konkrete Subdomains zum gesamten Textmaterial beitragen.

Bei der Auswahl der Kennzahlen liegen zwei Annahmen zugrunde. Zum Ersten ist festzuhalten, dass die Subdomains der Regierungsbehörden in ähnlicher Weise konsistentere Sprechweisen bzw. Akteure repräsentieren, wie dies bereits für die Einschränkung der Landesdomäne auf die jeweilige Regierungsdomäne postuliert wurde. Aufgrund des hierarchischen Aufbaus von Web-Domains ist diese Annahme unabhängig von der Art der Differenzierung zwischen den einzelnen Subdomains, sofern diese semantischen Kriterien und nicht nur technischen Gesichtspunkten folgt. Die zweite Annahme beruht daran anschließend darauf, dass die für die einzelnen Subdomains unterstellte, relativ konsistente Terminologie sich in umso höherem Ausmaß im Korpus wiederfindet, je größer ihr Anteil an der Textmenge insgesamt ist. Selbst wenn sich das verwendete Vokabular nur unwesentlich von dem anderer Subdomains unterscheidet, folgt aus einem höheren Anteil am Textumfang zumindest auch ein entsprechend größerer Anteil an den Fundstellen der einzelnen Begriffe, auf denen die Berechnung der Assoziationsegeln fußt.

Der Abgleich der Textgröße mit der Anzahl der Suchtreffer je Subdomain erlaubt zudem einen Rückschluss darüber, aus wie vielen Einzeldokumenten sich dieser Anteil zusammensetzt. Besteht er nur aus wenigen oder gar einem einzelnen Dokument, ist eine konsistente Sprechweise natürlich in höherem Maße gegeben als bei vielen kleinen Teildokumenten, die jeweils von verschiedenen Autoren stammen können.

| ***Suchanfrage: Europe site:gov.ng*** | | | | |
|---|---|---|---|---|
| **Subdomains** | **Suchtreffer** | **Anteil (%)** | **Text in kb** | **Anteil (%)** |
| *.deltastate.gov.ng | 56 | 34,15 | 2037 | 30,57 |
| *.ncc.gov.ng | 4 | 2,44 | 1731 | 25,98 |
| *.namasa.gov.ng | 32 | 19,51 | 1664 | 24,97 |
| *.smedan.gov.ng | 16 | 9,76 | 393 | 5,90 |
| *.fmind.gov.ng | 1 | 0,61 | 324 | 4,86 |
| Sonstige | 55 | 33,54 | 515 | 7,73 |
| **Summe** | **164** | **100,00** | **6664** | **100,00** |

*Tab. 1: Zusammensetzung des Textkorpus für gov.ng*

Tab. 1 zeigt die entsprechenden Kennzahlen für das Teilkorpus zur Domäne gov.ng. Ungefähr ein Drittel der Einzeldokumente wie des Texumfangs stammt von der Domäne deltastate.gov.ng (05.12.07). Dabei handelt es sich um den Webauftritt des gleichnamigen

nigerianischen Bundesstaats[78] im eingangs vorgestellten ölreichen Nigerdelta. Ein weiteres Viertel steuert die *National Maritime Administration and Safety Agency* (namasa.gov.ng, 19.06.07) zum Textmaterial bei. Ein Zehntel der Einzeldokumente wurden von www.smedan.gov.ng (05.12.07) heruntergeladen, der *Small and Medium Enterprise Developement Agency of Nigeria*. Ein im Vergleich zur Anzahl der Suchtreffer großer Textanteil fand von den Seiten der *Nigerian Communications Comission* (ncc.gov.ng, 05.12.07), der *independent National Regulatory Authority for the telecommunications industry in Nigeria*[79]*,* Eingang in das Korpus. Bei dem Einzeldokument von den Seiten des Industrieministeriums (fmind.gov.ng, 19.06.07) handelt es sich die Vorstellung der Aufgaben des *Planning, Research and Statistics Department*[80] im selben Ressort. Die übrigen 13 der insgesamt 18 Subdomains stellen den restlichen Umfang des Gesamtkorpus.

| ***Suchanfrage: Europe site:gov.in*** | | | | |
|---|---|---|---|---|
| **Subdomain** | **Suchtreffer** | **Anteil (%)** | **Text in kb** | **Anteil (%)** |
| *.ap.gov.in | 1 | 0,58 | 1046 | 28,32 |
| *.vigyanprasar.gov.in | 12 | 6,98 | 756 | 20,47 |
| *.mea.gov.in | 13 | 7,56 | 336 | 9,10 |
| *.incois.gov.in | 12 | 6,98 | 259 | 7,01 |
| *.iricen.gov.in | 82 | 47,67 | 94 | 2,54 |
| Sonstige | 52 | 30,23 | 1203 | 32,57 |
| **Summe** | **172** | **100,00** | **3694** | **100,00** |

*Tab. 2: Zusammensetzung des Textkorpus für gov.in*

Die Anteile, mit denen Subdomains der Domäne gov.in zum entsprechenden Teilkorpus beitragen, sind in Tab. 2 aufgelistet. Den größten Anteil davon bildet ein Lehrplan für die sechste Klasse[81] im Bundesstaat Andhra Pradesh (ap.gov.in, 06.12.07), in dem Europa vor allem unter klassischen länderkundlichen Kategorien vorgestellt wird. Immer noch ein Fünftel des Textmaterials steuert das Wissenschaftsportal *Vigyan Prasar* (vigyanprasar.gov.in, 06.12.07) bei, kleinere Anteile entfallen auf das Außenministerium (mea.gov.in, 06.12.07) und das *Indian National Centre for Ocean Information Services* (incois.gov.in, 06.12.07), wo meeresbezogene Daten unterschiedlicher Herkunft aufbereitet und in Form einer kostenpflichtigen Dienstleistung online verfügbar gemacht werden[82]. Fast die Hälfte aller Suchtreffer, alle jedoch von nur geringem Umfang, entfallen

78 http://www.deltastate.gov.ng/admin&governmentframe.htm, 03.12.07
79 http://ncc.gov.ng/index1.htm, 05.12.07
80 http://www.fmind.gov.ng/docs/PRSdept.doc, 05.12.07. Das Dokument enthält lediglich zwei Fundstellen für „Europe[an]“.
81 http://swrs.ap.gov.in/Academic/Syllabus%20Division/6/6%20th%20class%20social%20daywise%20plan.doc (18.06.07)
82 http://www.incois.gov.in/Incois/incois1024/index/index.jsp?res=1024# (06.12.07)

auf Seiten des *Indian Railway Institute of Civil Engineering* (iricen.gov.in, 06.12.07). Was die Anzahl der beitragenden Subdomains angeht, ist das Korpus im Falle von gov.in noch heterogener als im Falle von gov.ng. So stammt das verbleibende Drittel von 31 weiteren der insgesamt 36 im Korpus enthaltenen Subdomains.

| ***Suchanfrage: Europe site:gov.au*** | | | | |
|---|---|---|---|---|
| **Subdomain** | **Suchtreffer** | **Anteil (%)** | **Text in kb** | **Anteil (%)** |
| *.dfat.gov.au | 34 | 20,61 | 841 | 25,74 |
| *.dcita.gov.au | 2 | 1,21 | 659 | 20,17 |
| *.awm.gov.au | 12 | 7,27 | 636 | 19,47 |
| *.austrade.gov.au | 19 | 11,52 | 223 | 6,83 |
| *.defence.gov.au | 6 | 3,64 | 165 | 5,05 |
| Sonstige | 92 | 55,76 | 743 | 22,74 |
| **Summe** | **165** | **100,00** | **3267** | **100,00** |

*Tab. 3: Zusammensetzung des Textkorpus für gov.au*

Der größte Teil des Korpus von den Webseiten der australischen Regierung (vgl. Tab. 3) entfällt auf den gemeinsamen Webauftritt des Außen- und Handelsministeriums (dfat.gov.au, 06.12.07). Teil dieses Frameworks ist auch die *Australian Trade Commission* (austrade.gov.au, 06.12.07), die als ihre Aufgabe *helping more Australians succeed in export and international business*[83] angibt. Das *Department of Broadband, Communications and the Digital Economy* (dcita.gov.au, 06.12.07) ist vor allem durch einen Bericht *on Emerging Market Structures in the Communication Sector*[84] vom Juni 2003 vertreten, in dem die Situation in Australien mit der in Großbritannien verglichen wird. Immerhin ein Fünftel des gesamten Textumfangs stammt von den Seiten des *Australian War Memorial* (awm.gov.au, 06.12.07). Dieser Anteil unter dem Stichwort „Europe“ überrascht wenig, da die Erlebnisse australischer Soldaten während der beiden Weltkriege Anlass für die Gründung dieses Dokumentationszentrums waren[85]. Zu einem kleinen Anteil finden sich auch Seiten des Verteidigungsministeriums (defence.gov.au, 06.12.07) im Korpus. Die übrigen 23 der insgesamt 28 Subdomains stellen ein weiteres knappes Viertel.

Anhand der vorgestellten Verteilungen lässt sich eine genauere Erwartungshaltung darüber formulieren, welchen semantischen Feldern die Begrifflichkeiten des jeweiligen Textkorpus wahrscheinlich entstammen. Am besten ist eine solche Abgrenzung für die Domäne gov.au möglich, bei der überwiegend Begriffe aus dem Bereich des Außenhandels und des War Memorial zu erwarten sind. Für das Korpus, das aus den

[83] http://www.austrade.gov.au/About-Austrade1351/default.aspx, 03.12.07
[84] http://www.dcita.gov.au/__data/assets/word_doc/10984/Meridian_Connections.doc (19.06.07)
[85] http://www.awm.gov.au/aboutus/origins.htm (06.12.07)

Webseiten der indischen Regierungsbehörden gewonnen wurde, werden neben länderkundlichen wohl auch außenpolitische sowie wissenschaftliche Termini dominieren. Selbst wenn die Anzahl der in das Korpus eingeflossenen Subdomains im Falle von gov.ng gut überschaubar ist, können für das Gesamtkorpus nur Vermutungen angestellt werden. Im Falle des Webauftritts des Delta States kann ein Zusammenhang des Stichwortes „Europa“ mit der Ölförderung zwar angenommen werden, ist aber durch die Domäne des Bundesstaates nicht per se impliziert. Auch für die Nigerian Communications Comission bzw. die National Maritime Administration and Safety Agency ist unsicher, in welcher Weise hier ein Zusammenhang mit Europa besteht. Problematisch ist bei diesem Korpus zudem die geringe Anzahl an Fundstellen, die selbst größere Einzeldokumente zum Stichwort „Europa“ aufweisen.

## *5.2 Suche nach semantischen Feldern und ersten Zusammenhängen*

Nachdem die Zusammensetzung der Teilkorpora beschrieben wurde, werden diese Vermutungen nun anhand eines ersten Durchlaufs von Sascet (vgl. Kap 4.5) auf der Grundlage der noch unmodifizierten Begriffe einer ersten Prüfung unterzogen. Gleichzeitig kann der semantische Kontext des Europabegriffs in den jeweiligen Korpora in semantische Wortfelder eingegrenzt und auf Schlüsselbegriffe hin überprüft werden. Dazu lohnt es sich, zunächst einen Blick auf die automatisch extrahierten Begriffe[86] zu werfen, auf denen solch ein unmodifizierter Durchlauf unmittelbar aufbaut. Tab. 4 stellt die häufigsten Begriffe der drei Teilkorpora einander gegenüber.

[86] Bei der konkreten Umsetzung wurde keine der in Kap. 4.2 vorgestellten optionalen Einstellungen verwendet. Es wurden also reine Substantivgruppen extrahiert.

***Extrahierte Begriffe (unmodifiziert)***

| **gov.au** | | **gov.in** | | **gov.ng** | |
|---|---|---|---|---|---|
| Begriff | Fundstellen | Begriff | Fundstellen | Begriff | Fundstellen |
| europe | 513 | india | 327 | guest | 158 |
| australia | 424 | europe | 129 | nigeria | 148 |
| trade | 125 | years | 115 | cialis | 118 |
| skip | 115 | home | 103 | government | 93 |
| cent | 112 | english | 92 | phentermine | 87 |
| war | 112 | science | 89 | reply | 87 |
| press display | 105 | time | 89 | export | 86 |
| search help | 105 | courses | 84 | sitesled.com/ | 86 |
| countries | 101 | correction slip nos. | 82 | september | 85 |
| government | 101 | iricen website | 82 | tue jun | 83 |
| home | 98 | iricen's website | 82 | http://www.deforum.org /forums/topic.asp | 81 |
| united states | 98 | irpwm | 82 | nigerian crude oil | 77 |
| germany | 97 | members download area | 82 | pharmacy | 76 |
| eu | 96 | new facility | 82 | state | 69 |
| department | 95 | events | 81 | home | 67 |
| european union | 94 | faculty | 79 | namasa.com namasa | 62 |
| lib | 90 | art | 75 | viagra | 59 |
| years | 78 | spices | 75 | tourism | 57 |
| services | 72 | calendar | 71 | line | 56 |
| articles | 70 | tenders | 71 | post | 54 |

*Tab. 4: Überblick über die 20 häufigsten extrahierten Begriffe der drei Teilkorpora*

Beinahe erwartungsgemäß nehmen „Europe“, „Australia“ und ein möglicherweise verbindendes „trade“ die ersten Plätze im Korpus der Domäne gov.au ein. Auch das War Memorial ist unter dem Stichwort „war“ erkennbar. Noch vor „EU“ finden sich „United States“ und „Germany“. Addiert man allerdings die Fundstellen der Abkürzung „EU“ zu denen der expliziten Schreibweise „European Union“, rückt dieser Begriff auf Rang 3 der Liste vor. Dies veranschaulicht, welchen Effekt bereits eine einfache Zusammenführung graphematischer bzw. semantischer Varianten bei einer späteren Kategorisierung (vgl. Kap. 5.3) solcher Rohbegriffe haben kann.

Auch im Falle des Korpus zur Domäne gov.in nehmen „India“ und „Europe“ die beiden obersten Plätze ein. Wie anhand der Quelldokumente bereits vermutet, finden sich einige Begriffe aus den Bereichen Wissenschaft und Lehre auf den weiteren Plätzen. Auch die vielfach, wenngleich nur in geringem Umfang in das Korpus eingeflossenen Seiten des Indian Railway Institute of Civil Engineering (IRICEN) finden ihren Niederschlag.

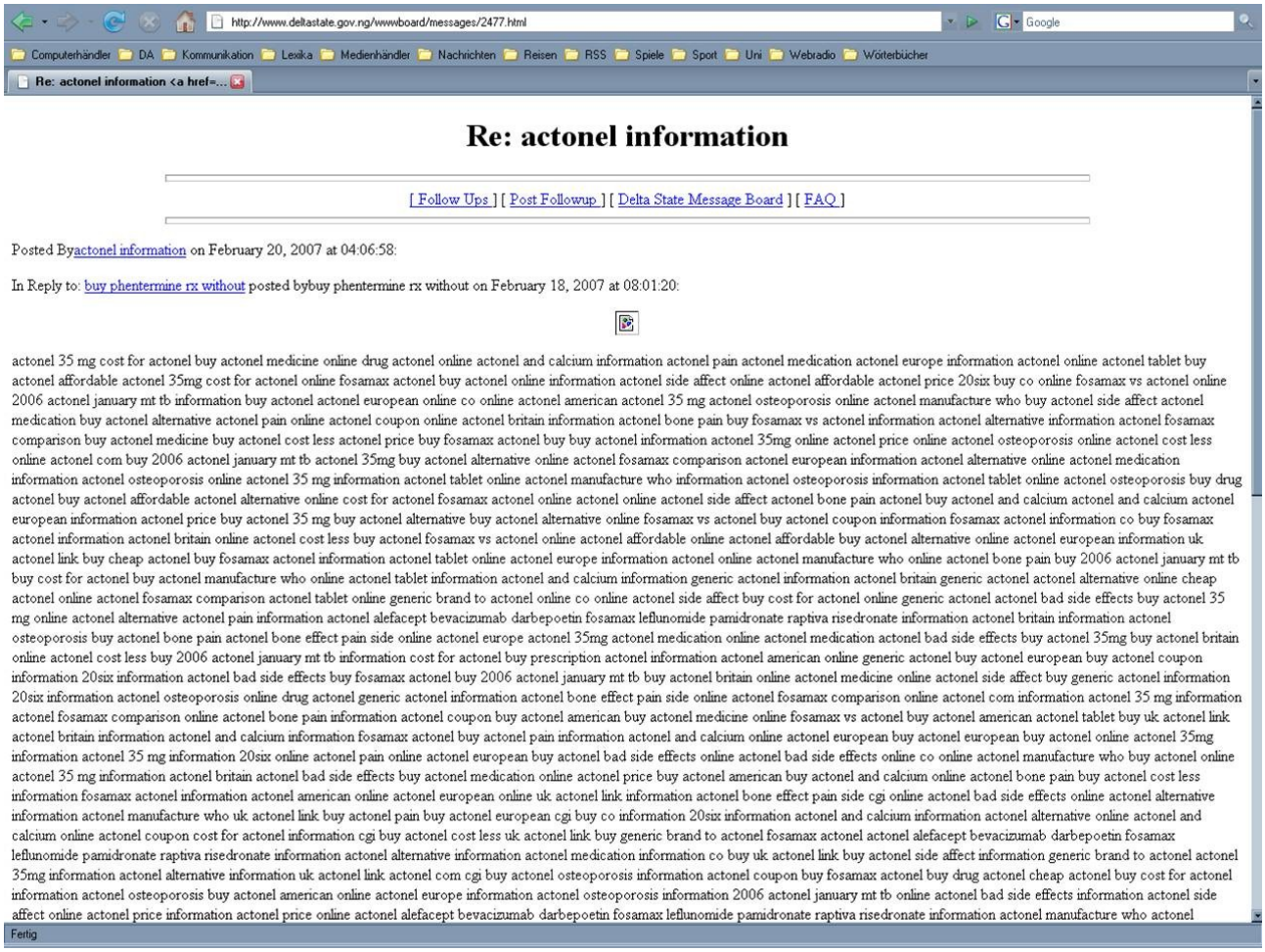



*Abb. 16: Sachfremdes Posting (Spam) auf den Webseiten nigerianischer Regierungsbehörden (www.deltastate.gov.ng/wwwboard/messages/2477.html, 03.07.07)*

Unerwartet ist das Resultat für die Domäne gov.ng. Zwar erscheinen die Stichwörter „Nigerian crude oil“ sowie „export“ unter den häufigsten extrahierten Begriffen, daneben im semantischen Umfeld von „pharmacy“ jedoch auch auffallend viele Medikamentennamen („Cialis“, „Phentermine“, „Viagra“). Dass sich unter den Begriffen auch eine Webadresse zu einem Forum findet, bietet hier einen ersten Anhaltspunkt dafür, dass die tatsächlich von Regierungsbehörden verwendeten Begrifflichkeiten hier teilweise von gezielt sachfremden Postings in Foren überdeckt werden (vgl. Abb. 16). Dies schränkt die zuvor formulierte Annahme einer konsistenten Sprechweise oder gar eines Akteurs bei dieser Domäne natürlich erheblich ein. Das Stichwort „Europe“ erscheint folglich mit einer Häufigkeit von 42 Fundstellen gar nicht mehr unter den ersten 20 Begriffen.

Allgemein kann festgehalten werden, dass die beschriebenen Phänomene größtenteils auf die unterschiedlich große Ausgangsmenge an Treffern zurückzuführen sind, die Google auf den jeweiligen Domänen zu dem Stichwort „Europe“ ermittelt. So liefert eine Suchanfrage auf gov.au 210000 geschätzte Treffer, auf gov.au dagegen 7490 und auf

gov.ng gar nur 1260[87]. Wird das Korpus jeweils über den ersten 200 Treffern gebildet, ist also bezüglich der Relevanz der von Google referenzierten Dokumente ein grundsätzlich verschiedenes Ausgangsniveau gegeben. Ein entsprechend kleineres Korpus konnte jedoch nicht verwendet werden, da bereits bei einem Korpus über den ersten 50 Treffern zur Domäne gov.ng so geringe Häufigkeiten pro Begriff erreicht werden, dass keine sinnvollen Zusammenhänge mehr errechnet werden können. Auch um einen vergleichbaren Textumfang pro Korpus sicherzustellen, wurde an der Korpuserstellung über den ersten 200 Treffern festgehalten.

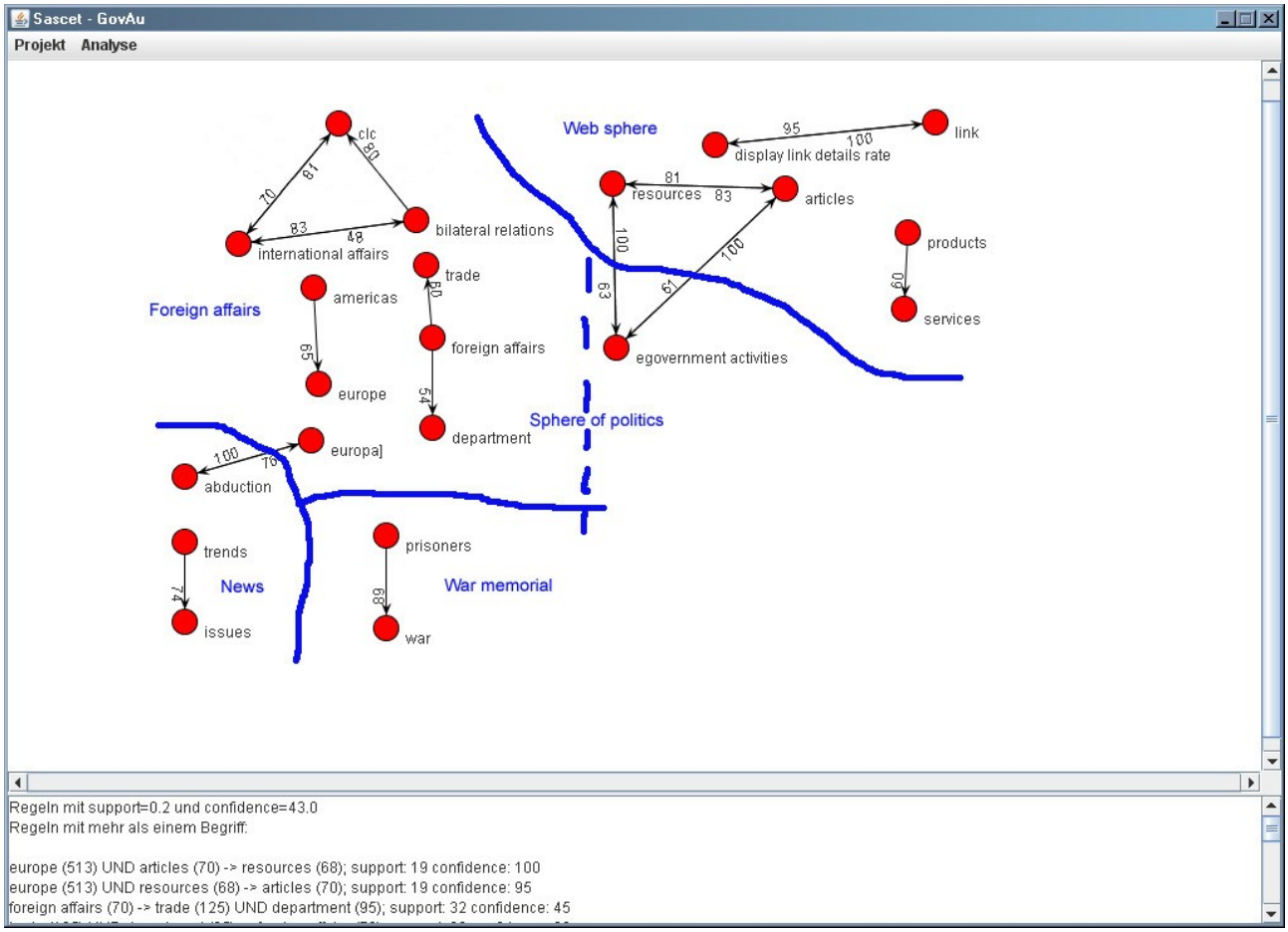



*Abb. 17: Semantische Felder der Anfragedomain gov.au (Schwellwerte: support=0,2%, confidence=43%), nachträgliche Hervorhebung der Wortfelder (blau).*

Die bisherigen Resultate zeigen allerdings klar die Grenzen einer quantitativen Textanalyse auf, bei der lediglich Worthäufigkeiten als Indiz für Textinhalte dienen. Bereits eine erste Berechnung von Assoziationsregeln über der unmodifizierten Menge extrahierter Begriffe hat jedoch eine filternde Wirkung auf die Ergebnisse und kann zudem erste Zusammenhänge aufdecken. In der Tat bestätigen sich bei diesem ersten Durchlauf für das Korpus zur Domäne gov.au die zuvor gesammelten Vermutungen zum Textinhalt

[87] Anfrage jeweils vom 07.12.07.

auf einer wesentlich differenzierteren Ebene (vgl. Abb. 17). Der überwiegende Anteil der Zusammenhänge wird durch die semantischen Wortfelder der Außenpolitik und des Australian War Memorial geprägt. Das Department of Foreign Affairs and Trade findet sich sogar als expliziter Graph wieder. Alle übrigen Zusammenhänge sind entweder der spezifischen Terminologie des Web an sich oder aber allgemeinen aktuellen Ereignissen zuzuordnen. Wirkliche Schlüsselbegriffe können aus dieser Sichtweise aber nicht abgeleitet werden, was auch daran liegt, dass auf dieser Betrachtungsebene die Fülle an Ergebnissen nicht mehr sinnvoll ausgewertet werden kann, wenn die Schwellwerte noch gewählt werden (vgl. Kap. 4.4). Eine Kategorisierung der Einzelbegriffe muss sich jedoch fraglos an diesen semantischen Wortfeldern orientieren.

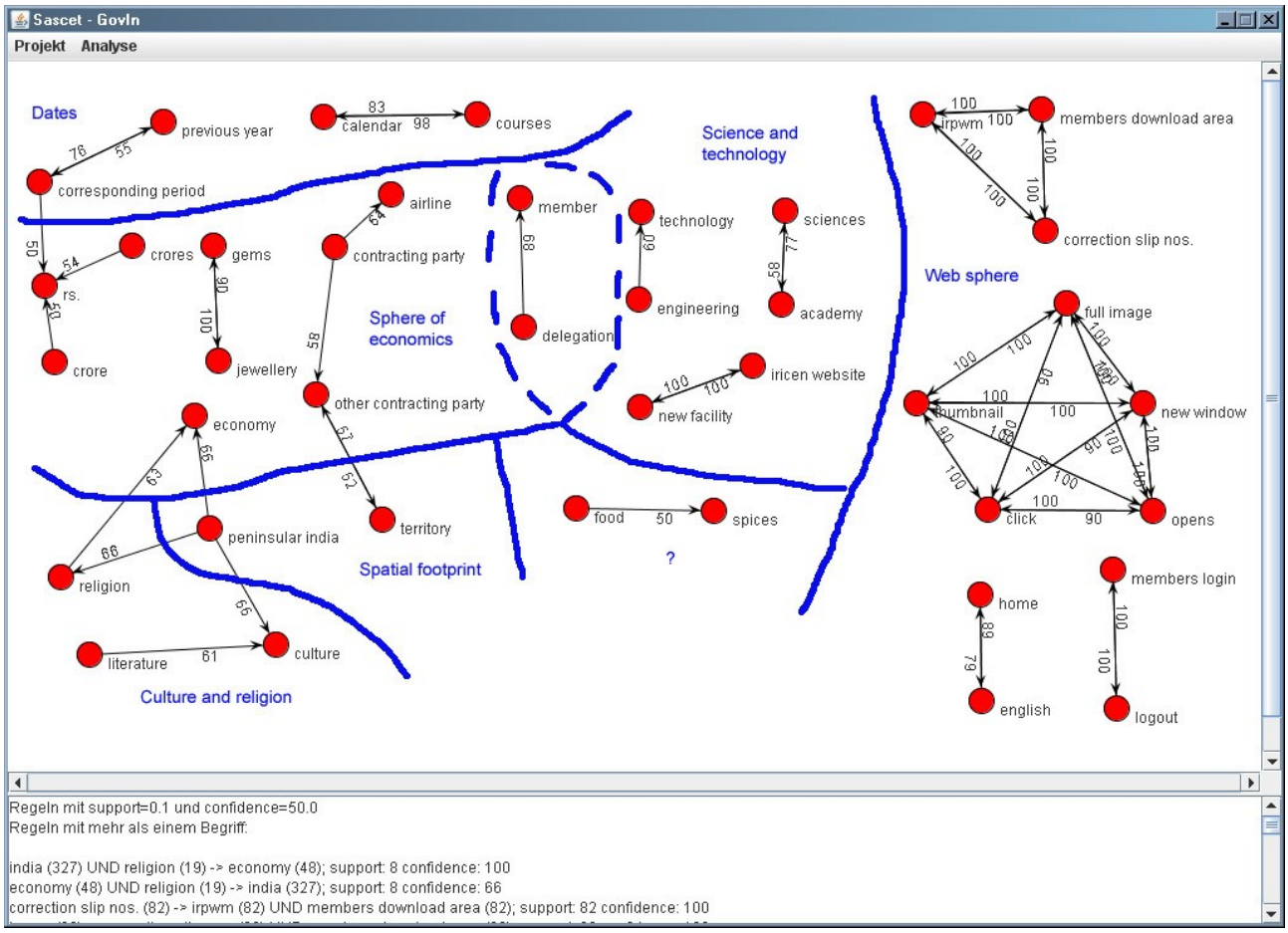



*Abb. 18: Semantische Felder der Anfragedomain gov.in (Schwellwerte: support=0,1%, confidence=50%), nachträgliche Hervorhebung der Wortfelder (blau).*

Ähnlich gut lassen sich die Begriffe der beschriebenen Textquellen für das Korpus zur Domäne gov.in im ersten unkategorisierten Assoziationsgraphen erkennen, selbst wenn hier der Schwellwert für den minimalen support niedriger gewählt werden musste, um sinnvolle Ergebnisse zu ermitteln (vgl. Abb. 18). Klar erkennbar sind die Wortfelder Wirtschaft und Wissenschaft bzw. Technologie. Daneben verweist das „peninsular india“

auch auf kulturelle Themen. Wiederum ergänzen webspezifische Zusammenhänge sowie Zeit- und Teminangaben den Überblick. Klare Schlüsselbegriffe fehlen hier ebenfalls. Als erster Anhaltspunkt für eine Kategorisierung scheinen die Wortfelder jedoch durchaus geeignet.

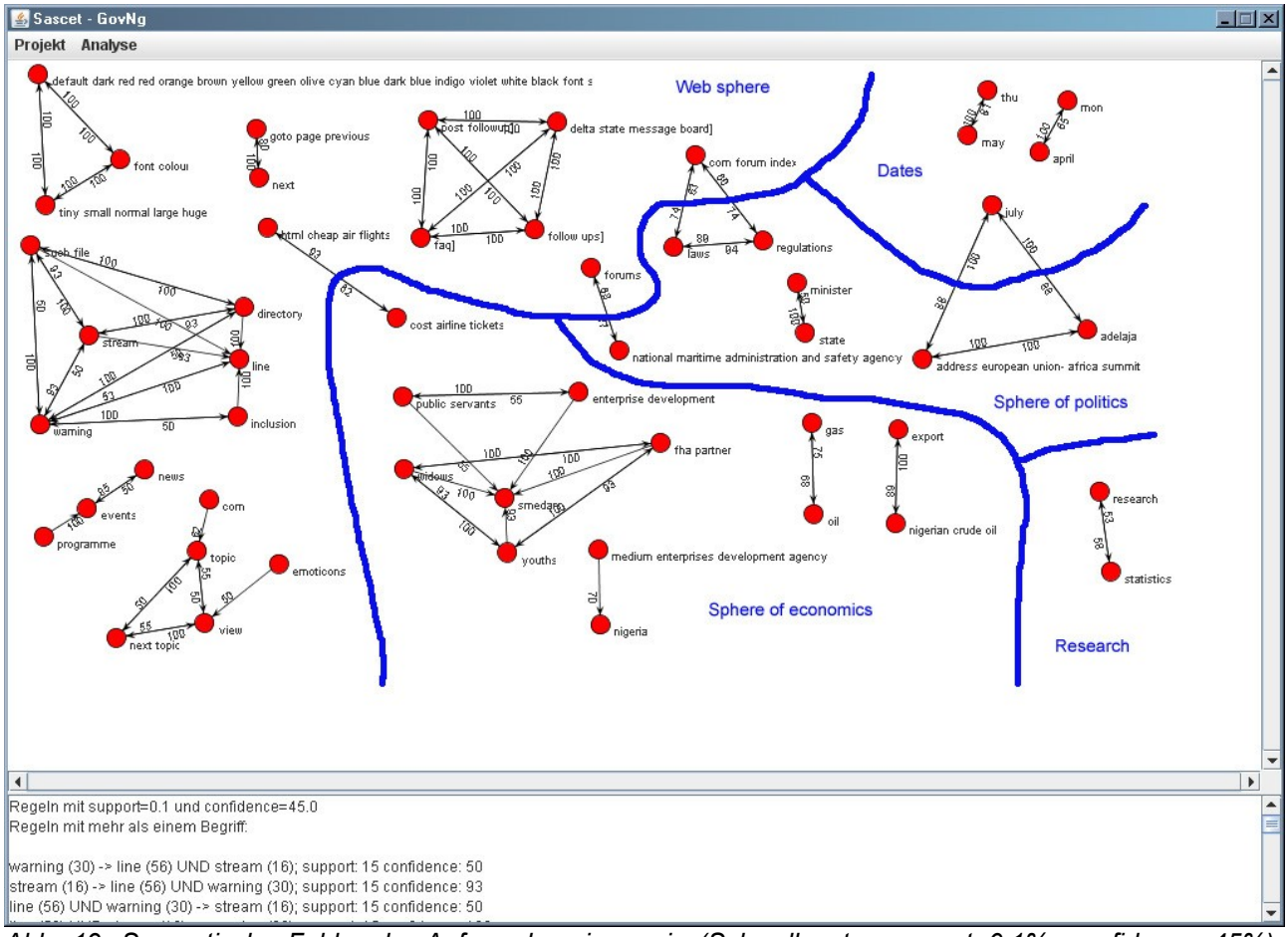



*Abb. 19: Semantische Felder der Anfragedomain gov.in (Schwellwerte: support=0,1%, confidence=45%), nachträgliche Hervorhebung der Wortfelder (blau).*

Für das Korpus zur Domäne gov.ng waren aufgrund der erwähnten Einschränkungen durch die Anzahl der zur Verfügung stehenden Google-Suchtreffer noch geringere Schwellwerteinstellungen erforderlich, um überhaupt Ergebnisse zu erzielen. Dies hat zur Folge, dass die webspezifische Terminologie einen großen Teil der extrahierten Begriffe einnimmt (vgl. Abb. 19). Die werbenden Postings für Medikamente stellen jedoch bereits an dieser Stelle keine signifikanten Zusammenhänge mehr dar. Dafür finden erwartungsgemäß die jeweils überproportional am Korpus beteiligten Subdomains ihren entsprechenden Niederschlag. Unter wirtschaftlichen Gesichtspunkten fällt vor allem die Small and Medium Enterprise Development Agency of Nigeria (SMEDAN – vgl. 5.1) auf. Daneben finden sich auch die vermuteten Hinweise auf den Export von nigerianischem Rohöl. Im Überlappungsbereich zwischen dem Wortfeld Politik und aktuellen

Datumsangaben zeichnet die errechnete Struktur zudem ein afrikanisch-europäisches Gipfeltreffen im Juli dieses Jahres nach.

## *5.3 Semantische Zusammenhänge in der überarbeiteten Begriffsmenge*

All diese ersten Eindrücke wurden auf der Grundlage der unmodifizierten Rohdaten der Begriffsextraktion gewonnen. Im Folgenden wird sich zeigen, dass sich die Ergebnisse durch eine Überarbeitung dieser Begriffe weiter schärfen lassen. Dazu gilt es zunächst die Qualität der extrahierten Begriffe zu verbessern. Um die Begriffsmengen auf ein einheitliches Niveau zu bringen und die nachfolgenden Analysen zu erleichtern, wurden zuerst alle Begriffe, die nicht mindestens dreimal im Korpus vorkommen, aus der Menge der extrahierten Termini entfernt. Dies reduziert die Anzahl der Einzelbegriffe um bis zu 80% auf die besser handhabbare Menge von mehreren hundert bis wenigen tausend. Anschließend wurden Begriffe, die aufgrund fehlerhafter Zeichencodierung unbrauchbar waren, sowie verbleibende html-tags gelöscht. Danach wurden Begriffe, die als Folge von unterschiedlichen Flexionselementen, von Abkürzungen oder auch nur von Tippfehlern als Varianten auftreten, zu einem einzigen Begriff zusammengefasst[88]. Im Falle der Seiten der nigerianischen Regierungsseiten wurden zudem alle Begriffe entfernt, die als offensichtlich sachfremder Spam über Regierungsforen Eingang in das Korpus gefunden haben (vgl. Abb. 16). Zudem fehlen nach der Überarbeitung in allen drei Korpora Begriffe, die offensichtlich und ausschließlich als webspezifische Termini verstanden werden müssen und daher für die weitere Analyse keinen Mehrwert mit sich bringen[89]. Tab. 5 zeigt wiederum die 20 häufigsten Begriffe der drei Korpora nach der Überarbeitung.

[88] So werden z.B. „u.s.a.“, „united states“ und „usa“ zu einem Begriff zusammengefasst.
[89] Z.B. „click“, „font color“ oder „goto previous page“.

***Extrahierte Begriffe (modifiziert)***

| **gov.au** | | **gov.in** | | **gov.ng** | |
|---|---|---|---|---|---|
| Begriff | Fundstellen | Begriff | Fundstellen | Begriff | Fundstellen |
| europe | 528 | india | 354 | guest | 158 |
| australia | 503 | years | 167 | nigeria | 148 |
| eu | 189 | iricen website | 164 | government | 93 |
| united states | 140 | europe | 133 | reply | 87 |
| country | 133 | science | 111 | export | 86 |
| year | 131 | home | 103 | state | 86 |
| trade | 125 | course | 99 | september | 85 |
| skip | 125 | english | 92 | tue jun | 83 |
| cent | 112 | country | 89 | nigerian crude oil | 77 |
| war | 112 | time | 89 | home | 67 |
| press display | 105 | art | 88 | country | 66 |
| germany | 102 | event | 87 | namasa.com namasa | 62 |
| home | 98 | correction slip nos. | 82 | line | 60 |
| department | 95 | irpwm | 82 | tourism | 57 |
| u.k. | 87 | new facility | 82 | delta state message board | 53 |
| minister | 79 | faculty | 79 | ncc | 52 |
| article | 74 | spices | 75 | topic | 51 |
| service | 72 | government | 74 | services | 50 |
| foreign affairs | 70 | calendar | 71 | smedan | 50 |
| resources | 68 | development | 71 | details | 48 |

*Tab. 5: Überblick über die 20 häufigsten modifzierten Begriffe der drei Teilkorpora*

Bereits diese im Vergleich zu Tab. 4 relativ kleinen Verschiebungen führen bei der Berechnung von Assoziationsregeln zu völlig anderen Ergebnissen. Während sich für das ohnehin stark fragmentierte Korpus der nigerianischen Webseiten, bei dem die Aussagekraft aufgrund der geringen Anzahl der Fundstellen für „Europe“ ohnehin zweifelhaft ist (vgl. Kap. 5.2), nun überhaupt keine verwertbaren Ergebnisse mehr erzeugen lassen, gewinnen die Zusammenhänge für die Korpora der Domänen gov.in und gov.au dagegen deutlich an Konturen.

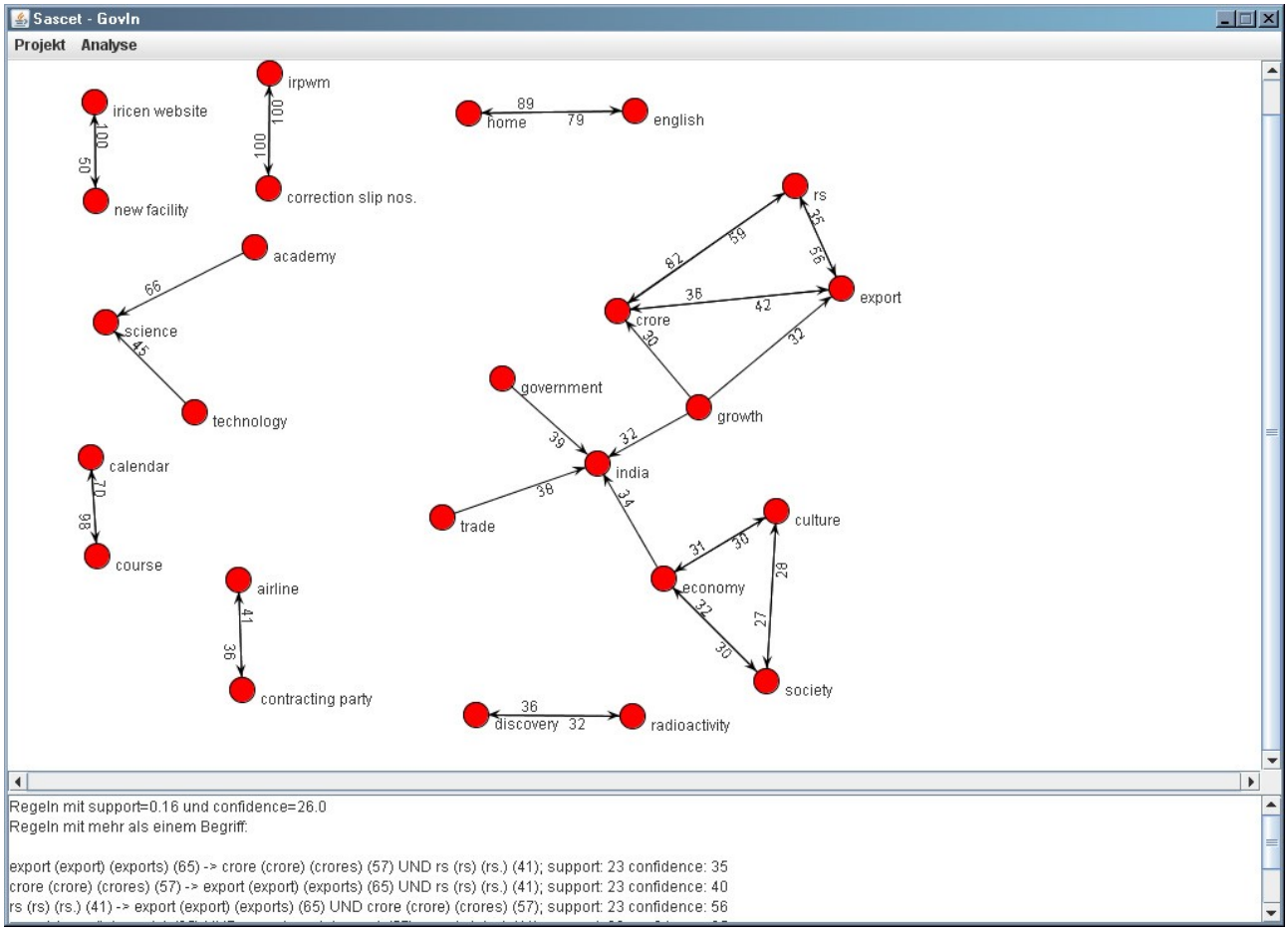



*Abb. 20: Semantische Zusammenhänge der Anfragedomain gov.in (Schwellwerte: support=0,16%, confidence=26%) nach Bereinigung der extrahierten Begriffe.*

So zeigt im Falle von gov.in das Stichwort „India“ nun die charakteristischen Eigenschaften eines Schlüsselbegriffs (vgl. Kap 4.5), dessen eingehende Kanten Aufschluss über seinen semantischen Kontext geben (vgl. Abb. 20). Dass von Indien gesprochen wird, wo auch das Stichwort „government“ zu finden ist, liegt in der Natur der Anfragedomäne. Interessant ist dagegen, dass im Übrigen wirtschaftliche Aspekte dominieren. So wird von „India“ signifikant häufig dort gesprochen, wo von Handel allgemein oder vom Wachstum des Exports um einen bestimmten Betrag die Rede ist. „Economy“ bildet zwar ein schwach signifikantes Cluster mit „culture“ bzw. „society“, nur bei den Fundstellen des Erstgenanntem wird aber hinreichend oft auch von „India“ gesprochen. Zwar erscheint das Stichwort „Europe“ nicht als Teil dieser Zusammenhänge, insofern es jedoch als Suchanfrage zur Korpusgenerierung verwendet wurde, müssen neben diesen wirtschaftlichen auch wissenschaftliche Themen[90] zu dessen Kontext gerechnet werden.

[90] Vgl. die Kookkurrenzen zwischen „science“ und „academy“ bzw. „technology“ oder zwischen „discovery“ und „radioactivity“.

Ein weit unmittelbareres Ergebnis zum Suchbegriff „Europe“ liefert die Berechnung der Assoziationsegeln für das Korpus zu den australischen Regierungsseiten. Hier bildet „Europe“ selbst einen Schlüsselbegriff, dessen semantischer Kontext durch mehrere Begriffe näher bezeichnet wird (vgl. Abb. 21). Mit „Asia“, „America“ bzw. „U.K.“ sind es dabei überwiegend andere Toponyme, an deren Fundstellen sich signifikant häufig auch „Europe“ findet. Hinzu treten noch die Begriffe „Italy“, „Germany“ und „France“, die als EU-Gründungsmitglieder zumindest implizit auf den Europabegriff verweisen. Sicherlich auf den hohen Anteil der Seiten des Australian War Memorial am Korpus zurückzuführen ist auch der Begriff „Second World War“, der aufgrund der Lage der Kriegsschauplätze naturgemäß ebenfalls häufig auf „Europe“ verweist[91]. Außenpolitische Termini wie „international affairs“ oder „bilateral relations“ verweisen zwar nicht direkt auf den Europabegriff, können jedoch ebenfalls als signifikante Aussagen des Korpus festgehalten werden.

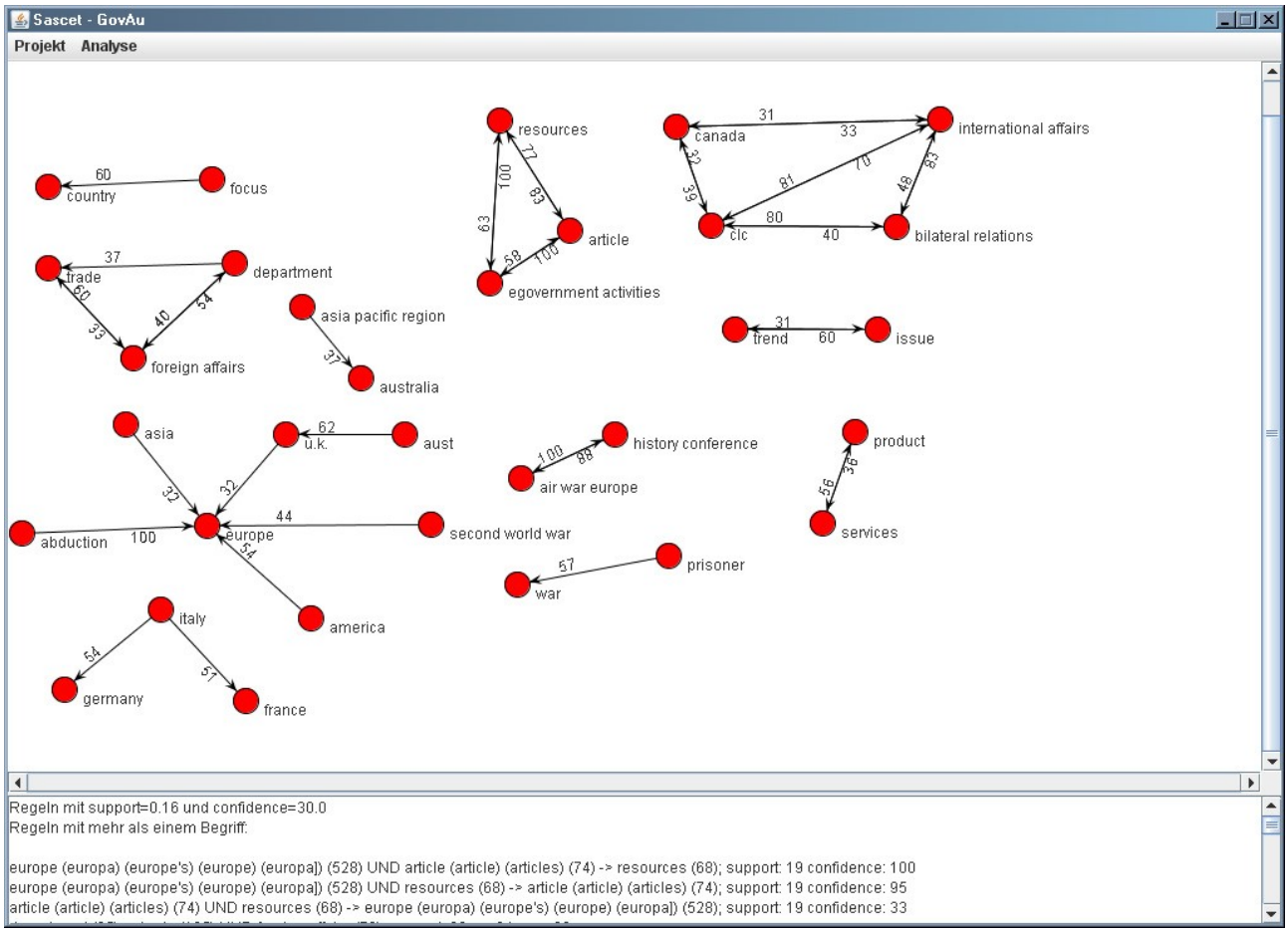



*Abb. 21: Semantische Zusammenhänge der Anfragedomain gov.in (Schwellwerte: support=0,16%, confidence=30%) nach Bereinigung der extrahierten Begriffe.*

[91] Zu diesem semantischen Feld muss neben „war“ und „prisoner“ auch die „history conference“ über den „airwar Europe“ gerechnet werden.

Natürlich stellt die bisherige Überarbeitung der Terminologie noch keine Kategorisierung im Sinne der qualitativen Inhaltsanalyse (vgl. Kap 3.1) dar. So wurden die extrahierten Begriffe bisher in keiner Weise gefiltert und liegen daher auf einem zu geringen Abstraktionsniveau und folglich in zu großer Anzahl vor, um wirklich als In-vivo-Kodes gelten zu können. Dass trotzdem bereits interessante Zusammenhänge aufgedeckt werden können, betont die Stärken des association rule minings gerade bei der Entdeckung interessanter Zusammenhänge in Daten, die in großen Mengen und dadurch relativ unstrukturiert vorliegen.

Gut geeignet für das weitere Vorgehen scheint ein Verfahren, wie es in der qualitativen Inhaltsanalyse nach MAYRING (2000) unter dem Namen Strukturierung Verwendung findet (vgl. Kap. 3.1), bei dem das Textkorpus gezielt auf bestimmte Kategorien hin untersucht wird. Fasst man also nur Begriffe zusammen, die einer dieser Kategorien zuzuordnen sind, kann am resultierenden Assoziationsgraphen leicht abgelesen werden, welche Kookkurrenten aus der Menge der noch unkategorisierten Begriffe mit den neu geschaffenen Kategorien im Zusammenhang stehen. Gemäß der eingangs formulierten Forschungsfrage wird dies nun innerhalb der Begriffsmenge des jeweiligen Korpus für unterschiedliche Akzentuierungen des Europabegriffs untersucht.

Da die Begriffsmenge im Rahmen einer Kategorisierung mehrfach durchgearbeitet werden muss, um sicher zu gehen, dass keine wesentlichen Kategorien übersehen wurden (vgl. Kap. 3.1), wurde vor diesem Schritt der Umfang der Begriffsmengen weiter eingegrenzt. Als Kriterium diente der Wert von 0,1% für den minimalen support[92], da Begriffe unterhalb dieser Schwelle für die Berechnung der Assoziationsregeln ohnehin nicht berücksichtigt werden sollen. Zwar könnten durchaus auch seltenere Begriffe zu den wesentlichen Kategorien der Textinhalte beitragen. Da dies bei einem frequenzbasierten Maß jedoch nur in geringem Umfang der Fall ist, können sie an dieser Stelle vernachlässigt werden. Die Maßnahme reduziert die Begriffsmengen nochmals auf wenige hundert Einzelbegriffe.

Als Vorbereitung für eine geeignete Kategorisierung wurden aus den Rohdaten alle Begriffe, die einen unmittelbaren Bezug zum Europabegriff aufweisen, gesammelt und einander gegenübergestellt. Auch Orts- und Ländernamen wurden hier unter der Annahme mit aufgenommen, dass immer auch von „Europa“ gesprochen wird, wenn von „Britain“ oder „Berlin“ die Rede ist. Es muss klar sein, dass es sich dabei nicht um Vorwissen oder

[92] Dies entspricht einer minimalen Häufigkeit von 14 für das Korpus zur Domäne gov.ng und einer Häufigkeit von jeweils 9 für die beiden anderen Korpora.

gar Grenzen handelt (vgl. Kap. 1), die von außen an das Korpus herangetragen werden. Vielmehr stellt dies nur eine Kategorisierung der Toponyme dar, die sich in diesem Korpus, das ja mithilfe der Suchanfrage „Europe“ erzeugt wurde, ohnehin finden. Die Prominenz bestimmter Einzeltoponyme (vgl. Abb. 22 und 23) kann also im Gegenteil explizit als Instrument zur Abgrenzung Europas in der Sichtweise der jeweiligen Domäne genutzt werden.

Für das Korpus zu den Webseiten der nigerianischen Regierung zeitigt dieses Vorgehen allerdings keinerlei Mehrwert. Außer „Europe“ und dem bereits erwähnten „European Union – Africa summit“ finden sich weder thematische noch weitere Ortsbezeichungen mit einem entsprechenden Bezug. Anders verhält es sich bei den anderen beiden Korpora, die eine Reihe solcher Begriffe enthalten. Die Abb. 22 und 23 stellen die Ergebnisse einander gegenüber. Obwohl der Schriftgrad der Begriffe absolute Häufigkeiten symbolisiert, muss betont werden, dass die Ergebnisse vom Gesamtvolumen her durchaus unmittelbar vergleichbar sind, da die Korpora für gov.in und gov.au in etwa denselben Umfang haben (vgl. Kap. 5.1).

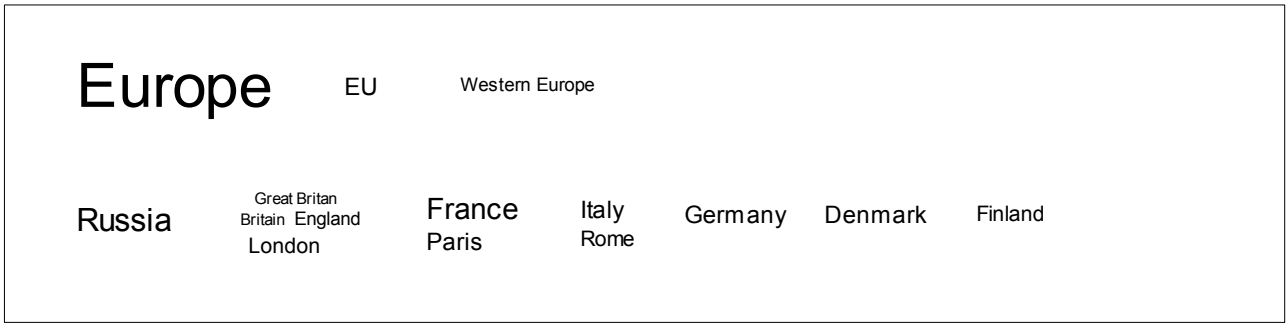



*Abb. 22: Begriffe mit Europabezug für das Korpus zur Domäne gov.in. Der Schriftgrad visualisiert die Häufigkeit des entsprechenden Begriffs im Korpus gemäß der Formel $2f^{0,5}$ (Frequenz f).*

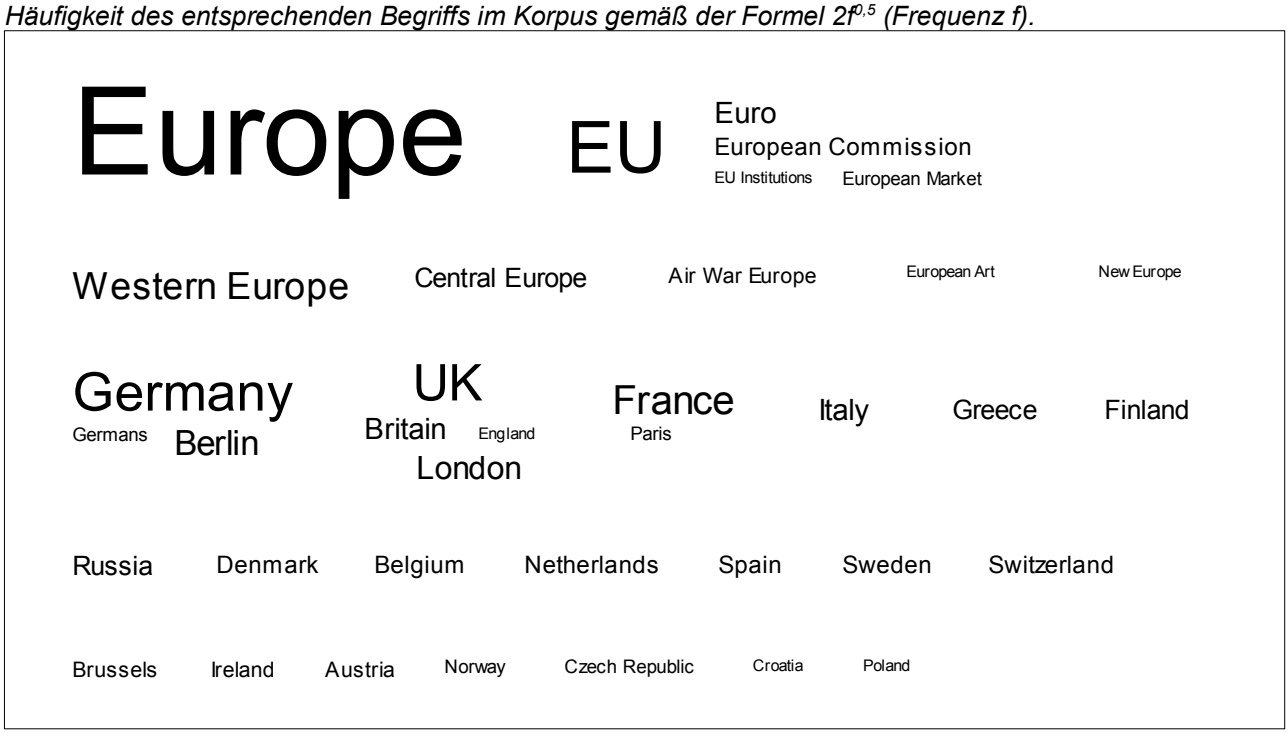



*Abb. 23: Begriffe mit Europabezug für das Korpus zur Domäne gov.au. Der Schriftgrad visualisiert die Häufigkeit des entsprechenden Begriffs im Korpus gemäß der Formel $2f^{0,5}$ (Frequenz f).*

Wenngleich im Falle des Korpus der australischen Regierungsbehörden mehr Begriffe, und diese zudem mit größerer Häufigkeit häufiger auftreten, fällt doch die gute Übereinstimmung der verwendeten Terminologie auf. Neben den zahlenmäßig überwiegenden Ortsnamen ist dabei die „EU“ wie bereits erwartet (vgl. Kap. 1) das dominierende Europakonzept. Näher charakterisiert wird diese im Falle von gov.au durch ihre Institutionen, insbesondere die Europäische Kommission sowie den gemeinsamen Markt und den Euro als gemeinsame Währung. Während „Air War Europe“ auf den Beitrag des Australian War Memorial zum Korpus zurückgeht, finden sich darüber hinaus auch noch Hinweise auf europäische Kunst und den Begriff des „New Europe“. Ein Blick auf die Fundstellen zu diesem Begriff zeigt, dass damit allerdings nicht etwa die polarisierenden Äußerungen jüngeren Datums (vgl. Kap. 2.3) gemeint sind, sondern die Herausforderungen im Zuge der durchaus schon Mitte der 90er Jahre des letzten Jahrhunderts thematisierten Osterweiterung der EU[93].

Auf dieser Grundlage konnten verschiedene Kategorien von Europabegriffen gebildet werden. Dabei werden zunächst alle Begriffe mit einem unmittelbaren Europabezug von den Orts- und Ländernamen unterschieden, die nur insofern einen Europabezug aufweisen, als sie in Europa verortet werden können. Letztere sollen daher bewusst vage zu „European countries“ zusammengefasst werden. Innerhalb der Begriffe mit einem unmittelbaren Europabezug wiederum können bei der Betrachtung des Korpus zu den australischen Regierungsseiten noch diejenigen Begriffe als separate Unterkategorie ausgewiesen werden, die explizit auf die Europäische Union als spezifischeren Europabegriff verweisen.

Ein weiterer Durchlauf mit Sascet erbrachte jedoch noch nicht den gewünschten Verweis der unkategorisierten Begriffe auf die so geschaffenen Schlüsselbegriffe. Deshalb wurden auch die noch verbleibenden Begriffe durch wiederholtes Durcharbeiten der Begriffsmenge in thematische Gruppen eingeteilt. Tab. 6 fasst die neu gebildeten Kategorien übersichtlich zusammen.

[93] Vgl. z.B. http://www.dfat.gov.au/media/speeches/foreign/1996/germany_.html (19.06.07).

| Europabegriffe | Allgemeine Themenblöcke |
|---|---|
| Europe | Topics on trade/export |
| EU | Topics on war |
| European States | Topics on science/technology |
| | Topics on education |
| | Topics on art |
| | Topics on tourism/travel |
| | Topics on culture |
| | Topics on religion |

*Tab. 6: Kategorien mit Europabezug und thematische Kategorien*

Die thematischen Kategorien umfassen dabei die verbleibenden Begriffe natürlich nicht vollständig. So wurden allgemeine und polyseme Begriffe in ihrer lediglich um Varianten bereinigten Form belassen. Sie können im Assoziationsgraphen nicht nur als Brücken fungieren, im Falle polysemer Begriffe kann die Berechnung von Assoziationsegeln sogar explizit zu ihrer Disambiguierung beitragen, da der jeweilige semantische Kontext diese näher bestimmt.

Bereits bei der Bildung der thematischen Kategorien fiel beim Korpus zur Domäne gov.ng die geringe Zahl an geeigneten Begriffen auf. Daher erstaunt es nicht, dass sich für das Korpus auch bei diesem letzten Schritt keine neuen Erkenntnisse ergaben. Nochmals aussagekräftigere Zusammenhänge konnten dagegen für die beiden anderen Korpora errechnet werden. Diese wurden zwar mithilfe noch niedrigerer Schwellwerte als in den vorherigen Durchläufen errechnet, erreichen durch die vorausgehende Kategorisierung bei der absoluten Häufigkeit der Kookkurrenzen jedoch ein stabil hohes Niveau.

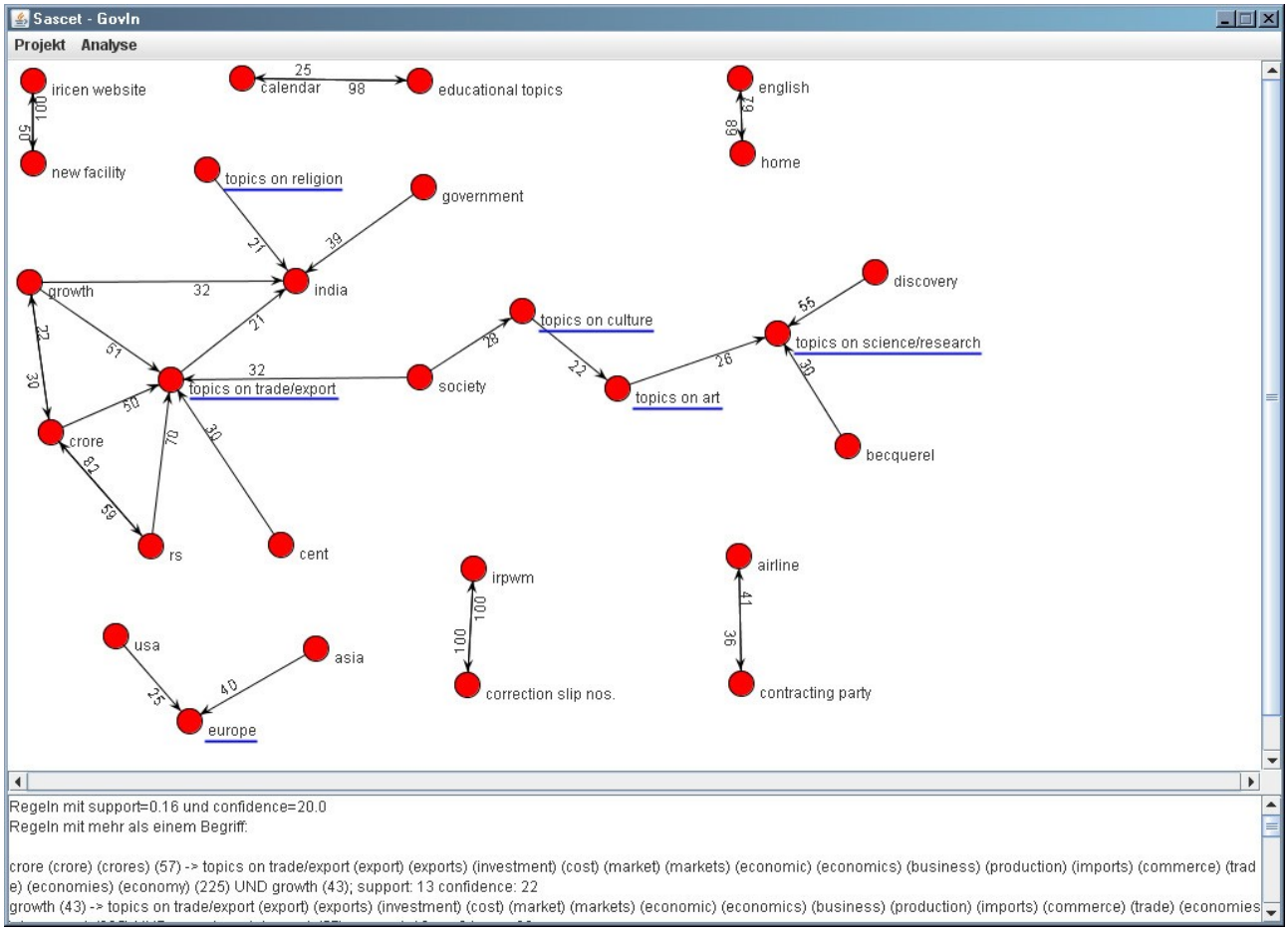



*Abb. 24: Semantische Zusammenhänge der Anfragedomain gov.in (Schwellwerte: support=0,16%, confidence=20%) nach Kategorisierung der extrahierten Begriffe, nachträgliche Hervorhebung der Kategorien (blau).*

Noch stärker als in den vorherigen Abbildungen treten nun „India“ sowie die Themenfelder Handel und Wissenschaft als Schlüsselbegriffe in den Mittelpunkt (vgl. Abb. 24). Erstmals bilden die Zusammenhänge auch größere Strukturen aus. So lassen die Verbindungen zwischen „India“, „growth“, dem Themenblock „trade“ und verschiedenen Maßeinheiten nach wie vor auf Textpassagen schließen, die die wachsenden indischen Exporte zum Inhalt haben.

Neben dem bereits thematisierten „government“ dienen nach der Kategorisierung auch Begriffe mit Religionsbezug als Kontext für das Stichwort „India“. Der Begriff „society“ hingegen verweist über Themen zu Kunst und Kultur schließlich auf wissenschaftliche Themen, zu denen offensichtlich ein Artikel des Wissenschaftsportals Vigyan Prasar zur Entdeckung der radioaktiven Strahlung einen wesentlichen Beitrag leistet[94]. Der eingangs erwähnte Lehrplan (vgl. Kap. 5.1) wird zumindest als Verbindung zwischen dem Themenfeld Erziehung und dem Stichwort „calendar“ abgebildet.

[94] http://www.vigyanprasar.gov.in/dream/apr2001/RADIOACTIVITY.htm (18.06.07)

Durch die Kategorisierung erstmals sichtbar ist mit dem Stichwort „Europe“ endlich auch der zentrale Gegenstand der Untersuchung. Dieser findet sich signifikant häufig allerdings nur im Kontext von „USA“ und „Asia“. Das kann dadurch erklärt werden, dass diese drei Begriffe auf den Seiten der indischen Regierung häufig dazu verwendet werden, um die durch sie bezeichneten Wirtschaftsräume dem indischen vergleichend gegenüberzustellen[95].

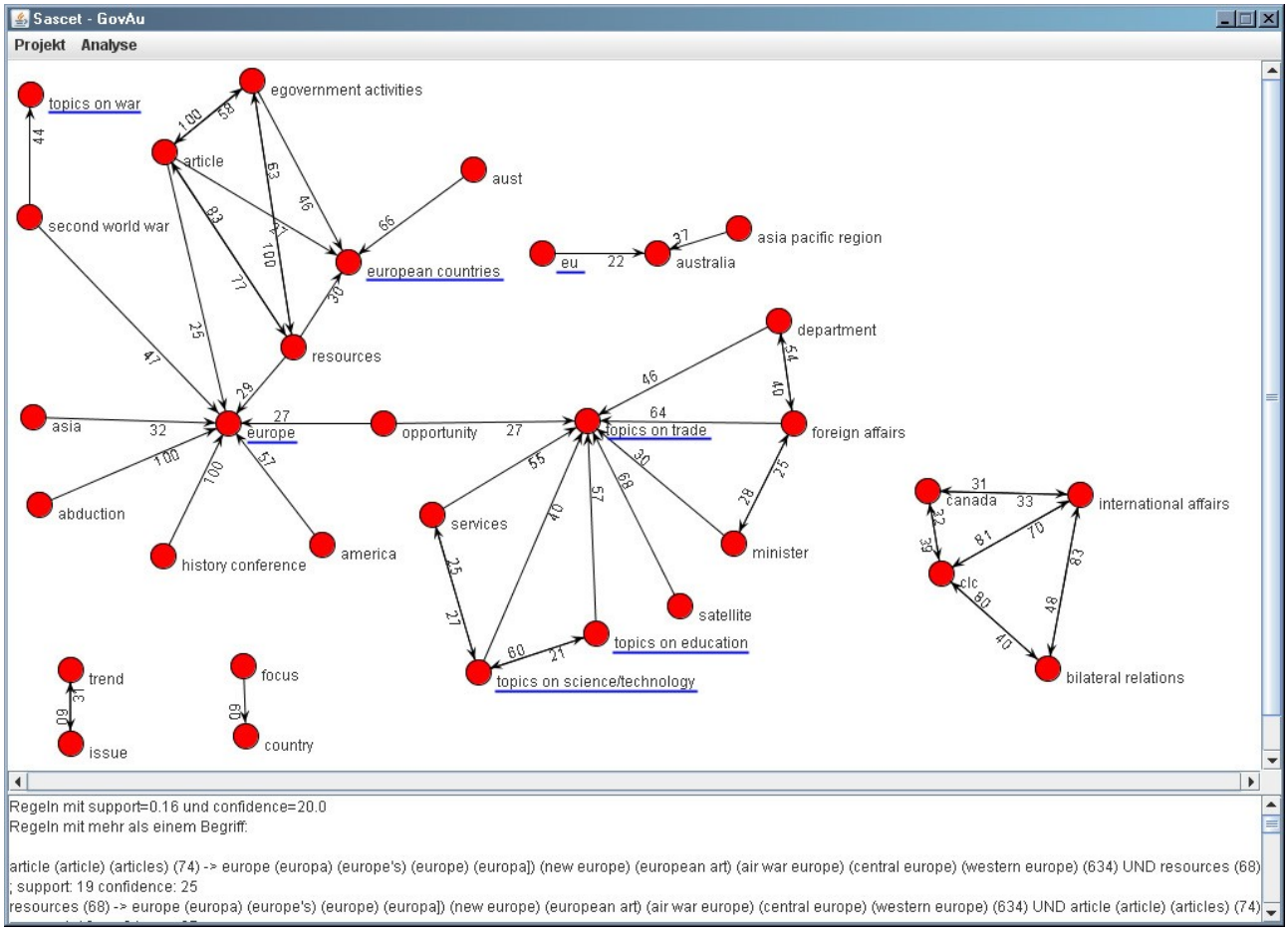



*Abb. 25: Semantische Zusammenhänge der Anfragedomain gov.au (Schwellwerte: support=0,16%, confidence=20%) nach Kategorisierung der extrahierten Begriffe, nachträgliche Hervorhebung der Kategorien (blau).*

Bei gleichen Programmeinstellungen ergeben sich für das Korpus zur Domäne gov.au noch komplexere Zusammenhänge (vgl. Abb. 25). Im Mittelpunkt stehen die Kategorien „trade“, „Europe“ sowie die aus den Vorkommen der einzelnen Orts- und Ländernamen gebildeten „European countries“. Interessant dabei ist, dass sich sowohl „Europe“ als auch der Themenkomplex „trade“ häufig an den Fundstellen des Begriffs „opportunity“ finden. Dies zeigt den auch bei diesem Korpus stark ökonomisch geprägten Kontext des

[95] Ein repräsentatives Beispiel aus dem Korpus: *In dollar terms, Asia and Oceania accounted for 47 per cent of India's total exports, followed by West Europe (22.34 per cent) and America (20.6 per cent) during 2005-06* (http://india.gov.in/sectors/commerce/india_trade.php, 18.06.07).

Europabegriffs[96]. Die hohe Zentralität von Handelsbegriffen ist daran ersichtlich, dass nicht nur das Department of Foreign Affairs and Trade, sondern mit den Themenblöcken Erziehung und Wissenschaft sogar andere Kategorien zu einem hohen Anteil auf ihn verweisen.

Neben dem für dieses Korpus spezifischen Kontext des „Second World War" finden sich auch hier „Europe" als häufiger Begleiter der Begriffe „Asia" und „America", während die Kategorie „EU" ihrerseits auf „Australien" verweist. Ein Blick auf die Fundstellen der betreffenden Begriffe deckt auf, dass die Bezüge mit der Rolle als Mediator begründet werden können, in der sich die australische Regierung zwischen Europa auf der einen und Asien auf der anderen Seite sieht[97].

Die zuvor bereits gezeigte Vielfalt an Orts- und Ländernamen innerhalb Europas (vgl. Abb. 23) erscheint vor allem im Kontext der Begriffe „article", „resources" und „egovernment activities". Hier gibt der ebenfalls im Graph sichtbare Zusammenhang zwischen „country" und „focus" den Hinweis darauf, dass auf einer der beitragenden Webseiten systematisch Länderprofile vorgestellt werden. Tatsächlich findet sich auf den Seiten des *Victoria eGovernment Resource Centre*[98] unter der Rubrik „Focus on Countries" *Articles and resources related to egovernment activities*[99] Informationen über entsprechende Aktivitäten in 45 europäischen Staaten. In diesem Fall hat erst die Kategorisierung der verschiedenen Ländernamen zu „European Countries" zur Entdeckung dieses Sachverhaltes beigetragen.

## *5.4 Vergleich des semantischen Kontextes der drei Domains*

Wie in der abschließenden Interpretation angedeutet, könnte die Untersuchung durch die zu jedem Begriff gespeicherten Fundstellen an dieser Stelle nahtlos in eine qualitative Studie übergehen, mittels der die Ergebnisse weiter fundiert und expliziert werden könnten. Obwohl ein solches Vorgehen wünschenswert gewesen wäre, musste aufgrund der beschränkten Zeitressourcen in der vorliegenden Arbeit darauf verzichtet werden.

---

96 Ein Beispiel aus dem Korpus: *Today, I would like to outline some of the factors behind Australia's strong economic performance, our trade policy agenda, and the challenges and opportunities currently facing the Australia-EU trade and economic relationship* (http://www.trademinister.gov.au/speeches/2004/040513_aus_business_europe.html, 19.06.07).

97 Ein Beleg aus dem Korpus: *This Government has recognised that Australia's relations with Europe - our complex and long-standing network of family, trade, investment, cultural, political and strategic links with Europe - are part of our credentials in Asia. [...] From our position in the Asia-Pacific, we can offer many of the advantages to Asia of our Europe-based technological, educational and cultural experience* (http://www.dfat.gov.au/media/speeches/foreign/1997/france.html, 19.06.07).

98 Vgl. http://www.egov.vic.gov.au (12.12.07).

99 Vgl. http://www.egov.vic.gov.au/index.php?env=-categories:m1699-1-1-8-s-0&reset=1 (12.12.07).

Dennoch wird im Folgenden gemäß der gestellten Forschungsfrage der semantische Kontext, in dem der Europabegriff in den drei Korpora auftritt, zumindest anhand der Ergebnisse der quantitativen Analyse kurz verglichen. Eine solche Interpretation kann zweifelsohne das Textmaterial nur unvollständig durchdringen. Bereits die vorliegenden Resultate bilden jedoch eine solide Basis für schlüssige Erkenntnisse.

Dass es sich dabei eher um eine Gegenüberstellung als um einen Vergleich handelt, ist zum Teil auf die bereits erwähnten ungleichen Voraussetzungen zurückzuführen, die bei der Korpuserzeugung gegeben waren (vgl. Kap. 5.1). So übersteigt die Anzahl der von Google geschätzten Treffer zum Stichwort „Europe“ auf den Seiten der australischen Regierung den Wert der nigerianischen Regierungsseiten um mehr als das hundertfache (vgl. Kap. 5.2). Als Grund kann sicherlich zum Teil angeführt werden, dass die Durchdringung der Gesellschaft mit digitalen Informationsmedien in den drei Ländern eine ganz andere ist (vgl. einleitend Kap. 5). Dass eine sinnvolle Auswertung des Korpus zu den Webseiten der nigerianischen Regierung nicht möglich war, kann jedoch weder mit der insgesamt größeren Webpräsenz der australischen Regierung noch mit der weit größeren Zahl potentieller Rezipienten im eigenen Land vollständig erklärt werden. Schließlich liefert das Korpus zu den indischen Regierungsseiten, das weit weniger Fundstellen zum Stichwort „Europe“ aufweist als das australische, bei einem mit Nigeria vergleichbaren Anteil der indischen Bevölkerung mit Zugang zu Computern trotzdem brauchbare Ergebnisse.

So kann als Erkenntnis festgehalten werden, dass „Europa“ auf den Seiten der nigerianischen Regierung schlicht von nachrangiger Bedeutung ist. Zwar finden sich Hinweise auf den Export von Erdöl im Korpus, jedoch ohne eindeutigen Bezug zu Europa. Einziges nennenswertes diskursives Ereignis war ein Gipfeltreffen zwischen afrikanischen und den EU-Staaten. Dass Europa aus der afrikanischen Sichtweise keine allzu große Rolle spielt, deckt sich dabei aber durchaus mit anderen Forschungsergebnissen (vgl. Kap. 2.3).

Differenzierte Beobachtungen waren hingegen für die beiden anderen Korpora möglich. Sowohl in der Domäne gov.in als auch bei gov.au dominiert dabei die Terminologie des Handels und der Ökonomie. Auch dies überrascht im Falle des indischen Korpus nicht, wurde es doch bereits im Vorfeld der eigentlichen Untersuchung dem Typus der außereuropäischen Sichtweise zugeordnet (vgl. einleitend Kap. 5), bei dem außer internationalen Handelsverflechtungen kaum weitere Berührungspunkte zu finden sind (vgl. Kap. 2.4). Eine ähnliche Akzentuierung kann für die Seiten der australischen

Regierung konstatiert werden, wo in politischen Reden zwar das europäische Erbe beschworen wird, ein Großteil der Terminologie im Umfeld des Europabegriffs jedoch ebenfalls dem ökonomischen Wortschatz zugerechnet werden muss. So verwundert es nicht, dass auf beiden Domains die EU als Wirtschafts- und Währungsunion mit Abstand das bedeutendste Europakonzept innerhalb des allgemeinen Europabegriffs darstellt.

Darüber hinaus findet sich der Begriff „Europa“ auf den Seiten der indischen Regierung explizit als Ort wissenschaftlicher und technologischer Errungenschaften sowie als Gegenstand der Lehre im Schulunterricht. Auch das Korpus zu den Webseiten der australischen Regierung beinhaltet die Themenblöcke Wissenschaft und Bildung. Diese sind dort aber nochmals stärker auf ökonomische Themen ausgerichtet. Eine prägende Rolle für die Sicht auf Europa in dieser Domäne spielt das Australian War Memorial, wo die Schicksale australischer Soldaten während der Weltkriege lebendig gehalten werden.

Zusammenfassend ist festzuhalten, dass wie erwartet die ökonomische Terminologie bei Weitem überwiegt. Zwar konnte das Korpus zu den nigerianischen Webseiten leider nicht sinnvoll ausgewertet werden. Trotz Unterschieden im Detail kann jedoch der hohe Grad an Übereinstimmung zwischen den wichtigsten Themenfeldern im semantischen Kontext des Europabegriffs sowohl auf den indischen als auch auf den australischen Regierungsseiten festgehalten werden.

# 6 Zusammenfassung

Will man Aussagen darüber machen, in welchem semantischen Kontext sich der Europabegriff außerhalb seiner unmittelbaren Einflusssphäre findet, bietet sich der reichhaltige Fundus an außereuropäischen Webseiten als Quellmaterial an. Der schiere Umfang des Textmaterials macht es jedoch erforderlich, rechnergestützte Strategien zur Unterstützung der Auswertung zu entwickeln. Anspruch und primärer Bezugspunkt müssen dabei Verfahren qualitativer Inhaltsanalyse sein, die als methodische Realisierung der Diskursanalyse in der Geographie bereits erfolgreich Einsatz finden. Dabei kann die rechnergestützte, quantitative Verarbeitung der Textdaten die qualitative Analyse nicht ersetzen. Vielmehr gilt es, das noch brach liegende synergetische Potential zwischen qualitativ und quantitativ arbeitenden Methoden der Textanalyse auszuloten und so weit wie möglich zu befördern.

Zu diesem Zweck wurde anhand etablierter Verfahren ein generalisiertes Ablaufmodell der wesentlichen Arbeitsschritte qualitativer Inhaltsanalyse erarbeitet. Für alle Teilschritte

wurde anschließend geprüft, welche rechnergestützten Textanalysemethoden sich zu ihrer Unterstützung eignen. Auf dieser Grundlage konnte mit dem Programm Sascet ein universelles Werkzeug zur Rekonstruktion semantischer Zusammenhänge in Texten entwickelt werden. Als Indiz für diese Zusammenhänge dienen dabei Assoziationsegeln, die mithilfe kookkurrenzbasierter Maße über der Menge automatisch extrahierter sinntragender Begriffe errechnet werden. Die Berechnung von Assoziationsegeln stellt zwar bei Weitem nicht das wirkungsvollste Verfahren zur Extraktion semantischer Zusammenhänge dar, ist aber aufgrund der guten Nachvollziehbarkeit und der daraus resultierenden unmittelbaren Evidenz seiner Ergebnisse als unterstützende Methode für qualitatives Arbeiten gut geeignet. Die besondere Stärke des Verfahrens liegt dabei darin, aus sehr großen, unstrukturierten Textkorpora innerhalb kürzester Zeit sinnvolle Ergebnisse zu errechnen.

Da eine kombinierte qualitative und quantitative Untersuchung im zeitlichen Rahmen dieser Arbeit nicht geleistet werden konnte, wurde lediglich die entwickelte Software eingesetzt, um ein Bild vom semantischen Kontext des Europabegriffs auf außereuropäischen Webseiten zu gewinnen. Hierbei muss eine Reihe von Einschränkungen beachtet werden. Da Sascet auf die Verarbeitung englischsprachiger Texte beschränkt ist, wurden mit Webseiten aus Nigeria, Indien und Australien Textquellen gewählt, in denen diese Sprache Amts- und Verkehrssprache ist. Weil Nationalstaaten keine Akteure, sondern ihrerseits nur Regionalisierungen darstellen, wurde die Untersuchung auf die Webseiten der Regierung des jeweiligen Landes eingegrenzt. Um darüber hinaus dem steten Wandel Rechnung zu tragen, dem Textinhalte im Web unterliegen, wurde die Analyse als synchrone Studie durchgeführt. Die drei Korpora wurden dabei mittels einer Google-Suchanfrage auf den Webseiten der jeweiligen Regierungsbehörde zum Stichwort „Europe“ automatisch erzeugt.

Die anschließende Interpretation der Ergebnisse zeigte, dass im semantischen Kontext des Europabegriffs auf diesen außereuropäischen Webseiten die Themenkomplexe Handel und Export weitaus überwiegen. Das stärkste Europakonzept bildet in diesem Zusammenhang erwartungsgemäß die Europäische Union. Während auf den Seiten der australischen Regierung die Erinnerungen an den Zweiten Weltkrieg ein historisches Europabild formen, schaffen indische Regierungsseiten eine Plattform für technologische und wissenschaftliche Erkenntnisse des Forschungsstandorts Europa. Im Falle der nigerianischen Regierung finden sich nur Dokumente nachrangiger Relevanz im Korpus.

Neben den bereits genannten Einschränkungen hat die verwendete Methode ohne den geforderten Bezug zu qualitativen Arbeitsweisen deutliche Grenzen. So ist eine automatische Zusammenstellung des Textkorpus durch eine Google-Suchanfrage nur bedingt geeignet, da relevante Textteile durch unbrauchbare Passagen verdeckt werden können. Eine gezielte Auswahl geeigneter Texte ist bei einer synchronen Studie von Webinhalten allerdings nur in sehr beschränktem Umfang möglich.

Ein doppeltes Problem ergibt sich durch die Behandlung einer formalen Webdomäne als Akteur. Zum einen muss innerhalb eines Ministeriums, geschweige denn einer Regierung keinesfalls Einigkeit über die Sichtweise auf Europa herrschen. Zum anderen zeigt das Beispiel der nigerianischen Regierungsseiten, dass die zugehörige Domäne an sich nicht uneingeschränkt die Sichtweise der jeweiligen Regierung widerspiegelt, da es durch Foren letztlich jedermann möglich ist, den Textinhalt der betreffenden Domäne zu beeinflussen. Dieses letzte Problem verhindert allerdings auch, dass sich im Rahmen einer automatischen Zusammenstellung von Korpora durch eine sukzessive Eingrenzung der Domäne wirklich konsistentere Sprechweisen finden lassen.

Grundsätzlich kann bei der Bewertung der Methode jedoch ein positives Resümee gezogen werden. Selbst mit recht einfachen Mitteln rechnergestützter Textanalyse ist es möglich, etablierte Verfahren qualitativer Inhaltsanalyse umfassend zu unterstützen. Durch das beschriebene Ablaufmodell ist dabei sichergestellt, dass eine solche Untersuchung jederzeit an übliche qualitative Verfahrensweisen angebunden werden kann. Während die Extraktion der sinntragenden Begriffe eine solide Ausgangsbasis für die weiteren Schritte bereitstellt, liefert die Berechnung der Assoziationsregeln zusammen mit den Fundstellen der Schlüsselbegriffe für sehr große Korpora innerhalb kürzester Zeit eine verlässliche Grundlage für eine erste Interpretation. Vor allem der eigentlich nur zu Kontrollzwecken integrierte Zwischenschritt der Überarbeitung und Kategorisierung der extrahierten Begriffe leistet hierzu einen wesentlichen Beitrag.

# 7 Ansatzpunkte für weitere Forschung

Da die vorliegende Arbeit in ihrem zeitlichen Rahmen nur experimentellen Charakter haben konnte, ergeben sich etliche Ansatzpunkte für weitere Forschung. Technisch ist hier natürlich die Weiterentwicklung und weitere Erprobung von Sascet zu nennen. Durch die Abstraktion der einzelnen Schritte gegeneinander ist dabei sichergestellt, dass die konkrete Realisierung jederzeit dem Stand der Technik angepasst werden kann. Erweitert

werden müssten zudem die bislang nur rudimentären Programmfunktionen zur Überarbeitung und Kategorisierung der automatisch extrahierten Begriffe. So fehlt bislang die Möglichkeit, irrtümlich zusammengeführte Einzelbegriffe wieder zu trennen oder Substantivgruppen mehr als nur einer thematischen Kategorie zuzuordnen. Beides ist auf der Grundlage des zur dauerhaften Speicherung der Ergebnisse spezifizierten Datenformats ohne Weiteres möglich. Ferner wäre es wünschenswert, zur besseren Abschätzung der Verwendungsweise an sich mehrdeutiger Begriffe bereits an diesem Punkt die entsprechenden Fundstellen einsehen zu können. Ebenfalls denkbar ist unabhängig davon eine Erweiterung von Sascet um Funktionen, die bei einem Vergleich verschiedener Korpora als erste Orientierung Maßzahlen zur Beurteilung der Ähnlichkeit oder Unähnlichkeit der jeweils gefundenen Begriffsnetze errechnen.

Weiterhin können die methodischen Einschränkungen, denen diese Studie vor allem bei der Zusammenstellung des Korpus unterlag, schrittweise gelockert werden. So spricht nichts dagegen, nach der Wahl einer geeigneten Archivierungsstrategie auch diachrone Untersuchungen durchzuführen. Einen wesentlichen Fortschritt würde die Erweiterung auf andere Sprachen neben dem Englischen bedeuten. Unabhängig von der weiten Verbreitung des Englischen als Verkehrs- und Wissenschaftssprache ist dies jedoch nur für wenige andere Sprachen in ähnlicher Qualität möglich. Da Sascet bereits jetzt die Auswertung beliebiger digitaler Textdokumente unterstützt, kann das fehlerbehaftete Verfahren der automatischen Zusammenstellung des Textkorpus durch eine Google-Suchanfrage jederzeit ersetzt werden. Die Anwendbarkeit der Analysefunktionen ist dabei grundsätzlich bei jeder Textart gegeben. Der klare Mehrwert ist aber zumindest bei der Verwendung von Assoziationsregeln vor allem bei den sehr großen, heterogenen Textbeständen des Web gegeben.

Auch die fachliche Fragestellung birgt ebenfalls Potential für weitere Studien. Hier könnten qualitative Analysen des Textkorpus dieser Arbeit genauso folgen wie die ergänzende Untersuchung anderer Domains. Das weiteste Feld für eine differenzierte Forschung zum Europabegriff bieten aber nach wie vor konkurrierende Konstruktionen Europas im Inneren. Wird die EU weiter zu Europa? Fühlen sich bestimmte Online-Communities mehr als Europäer als andere? Welche affirmierenden Strukturen finden sich bei der alltäglichen Regionalisierung des Europabegriffs? All diese Fragen können weit besser beantwortet werden, wenn rechnergestützte und qualitative Textanalyse nicht länger als Gegensatzpaar konstruiert werden, sondern sich durch ihre jeweiligen Stärken wechselseitig ergänzen.

# Literaturverzeichnis


Adolphs, Svenja (2006): *Introducing Electronic Text Analysis. A practical guide for language and literary studies.* London, New York: Routledge.

Agrawal, Rakesh und Ramakrishnan Srikant (1994): Fast Algorithms for Mining Association Rules. *In:* Bocca, Jorge B., Jarke, Matthias und Carlo Zaniolo, Hrsg., *Proceedings of the 20th International Conference of Very Large Databases.* VLDB '94. San Fransisco: Morgan Kaufmann, 487-499[100].

Agrawal, Rakesh, Imielinski, Tomasz und Arun Swami (1993): Mining Association Rules between Sets of Items in Large Databases. *In:* Bunemann, Peter und Sushil Jajodia, Hrsg., *Proceedings of the 1993 ACM-SIGMOD International Conference on the Management of Data.* SIGMOD '93. New York: ACM Press, 207-216[101].

Amir, Amihood, Aumann, Jonathan, Feldman, Ronen und Moshe Fresko (2005): Maximal Association Rules: A Tool for Mining Associations in Text. *Journal of Intelligent Information Systems* 25 (3), 333-345.

Ash, Timothy Garton (2004): *Freie Welt. Europa, Amerika und die Chance der Krise.* München, Wien: Carl Hanser Verlag.

Baeza-Yates, Ricardo und Berthier Ribeiro-Neto (1999): *Modern Information Retrieval*. New York: ACM Press.

Beaugrande, Robert de (1980): *Text, Discourse and Process. Toward a Multidisciplinary Science of Texts.* Advances in Discourse Processes IV. Norwood: Ablex Publishing Corporation.

Bedi, Ladislaus (2004): *Europabilder der EU-Beitrittskandidaten: Die Tschechische Republik und Ungarn.* Magisterarbeit. Erlangen-Nürnberg.

Berting, Jan und Christiane Villain-Gandossi (1999): Rolle und Bedeutung von nationalen Stereotypen in internationalen Beziehungen: ein interdisziplinärer Ansatz. *In:* Walas, Teresa, Hrsg., *Stereotypen und Nationen*. Krakau: Internationales Kulturzentrum, 13-31.

Bialasiewicz, Luiza und Claudio Minca (2005): Old Europe, new Europe: for a geopolitics of translation. *Area* 37, 365-372.


[100] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/1451/http:zSzzSzwww.almaden.ibm.comzSzcszSzpeoplezSzragrawalzSzpaperszSzvldb94_rj.pdf/agrawal94fast.pdf (20.06.07)

[101] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/4475/http:zSzzSzwww.cs.uni-bonn.dezSzIIIzSzlehrezSzvorlesungenzSzDataMiningzSzWS97zSz..zSzWS96zSzliteraturzSzagrawal93:mining.pdf/agrawal93mining.pdf (22.04.07)

Böhm, Andreas (2007[5]): Theoretisches Codieren: Textanalyse in der Grounded Theory. *In:* Flick, Uwe, von Kardorff, Ernst und Ines Steinke, Hrsg., *Qualitative Forschung. Ein Handbuch.* Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag, 475-485.

Brill, Eric (1992): A Simple Rule-Based Part of Speech Tagger. *In: Proceedings of the Third Conference on Applied Natural Language Processing.* ANLP '92. San Fransisco: Morgan Kaufmann, 152-155[102].

Casey, Lee A. und David B. Rivkin Jr. (2001): Europe in the Balance. The Alarmingly Undemocratic Drift of the European Union. *Policy Review* 107[103].

Chayko, Mary (2002): *Connecting. How we form social bonds and communities in the Internet Age.* Albany: State University of New York Press.

Chang, George, Healey, Marcus J., McHugh, James A. M. und Jason T. L. Wang (2001): *Mining the World Wide Web. An Information Search Approach.* Boston, Dordrecht u.a.: Kluwer Academic Press.

Chen, Feng (2001): *Die Entdeckung des Westens. Chinas erste Botschafter in Europa 1886-1894.* Frankfurt am Main: Fischer Taschenbuch Verlag.

Chilla, Tobias (2007): Zur politischen Relevanz raumbezogener Diskurse. Das Beispiel der Naturschutzpolitik in der Europäischen Union. *Erdkunde* 61, 13-26.

Chilla, Tobias (2005): 'Stadt-Naturen' in der Diskursanalyse. Konzeptionelle Hintergründe und empirische Möglichkeiten. *Geographische Zeitschrift* 93 (3), 183-196.

Christmann, Gabriela B. (2005): Dresdner Stadtdiskurse und die Wahrnehmung der Stadt. *In:* Keller, Rainer, Hirseland, Andreas, Schneider, Werner und Willy Viehöver, Hrsg., *Die diskursive Konstruktion von Wirklichkeit. Zum Verhältnis von Wissenssoziologie und Diskursforschung.* Schriften zur Wissenssoziologie 10. Konstanz: UVK.

Coenen, Frans, Leng, Paul und Lu Zhang (2005): Threshold Tuning for Improved Classification. Association Rule Mining. *In:* Ho, Tu Bao, Cheung, David Wai-Lok und Huan Liu, Hrsg., *Advances in Knowledge Discovery and Data Mining. Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining.* PAKDD-05. Berlin, Heidelberg, New York: Springer Verlag, 216-225[104].

[102] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/27718/http:zSzzSzacl.ldc.upenn.eduzSzAzSzA92zSzA92-1021.pdf/brill92simple.pdf (27.03.07)

[103] Heruntergeladen von: http://www.hoover.org/publications/policyreview/3476826.html (21.10.07)

[104] Heruntergeladen von: http://www.csc.liv.ac.uk/~frans/PostScriptFiles/pakdd2005coeneLeng.pdf (18.06.07)

Conter, Claude D. (2004): *Jenseits der Nation – Das vergessene Europa des 19. Jahrhunderts. Die Geschichte der Inszenierungen und Visionen Europas in Literatur, Geschichte und Politik.* Bielefeld: Aisthesis Verlag.

Daille, Béatrice (1996): Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. *In:* Klavan, Judith S. und Philip Resnik, Hrsg., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* Cambridge: MIT Press, 49-66[105].

Dale, Reginald (2004): European Union, Properly Constructed. Americans need to take Europe more seriously. *Policy Review* 122[106].

Delvaux, Peter und Jan Pepiór (1996): *Eurovisionen. Vorstellungen von Europa in Literatur und Philosophie.* Amsterdam, Atlanta: Rodopi.

Diaz-Bone, Rainer (2002): *Kulturwelt, Diskurs und Lebensstil. Eine distinktionstheoretische Erweiterung der bourdieuschen Distinktionstheorie.* Opladen: Leske und Budrich.

Dreyfus, Hubert L. (1989): *Was Computer nicht können. Die Grenzen künstlicher Intelligenz.* Frankfurt am Main: Athenäum Verlag.

Džihić, Vedran, Nadjivan, Silvia, Paić, Hrvoje und Saskia Stachowitsch (2006): *Europa – verflucht begehrt. Europavorstellungen in Bosnien-Herzegowina, Kroatien und Serbien.* Wien: Braumüller.

Europäische Kommission (2007): *Europabarometer 67. Die öffentliche Meinung in der Europäischen Union. Erste Ergebnisse*[107]*.*

Fayyad, Usama, Piatetsky-Shapiro, Gregory und Padhraic Smyth (1996): Knowledge Discovery and Data Mining: Towards a Unifying Framework. *In:* Simoudis, Evangelos, Han, Jiawei und Usama Fayyad, Hrsg., *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining.* KDD '96. Menlo Park: AAAI Press, 82-88[108].

Feldman, Ronen und James Sanger (2007): *The Text Mining Handbook. Advanced Approaches in Analysing Unstructured Data.* Cambridge: University Press.

Feldman, Ronen, Fresko, Moshe, Hirsh, Hayam et al. (1998): Knowledge Management: A Text Mining Apporach. *In:* Reimer, Ulrich, Hrsg., *Proceedings of the Second International*

[105] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/27850/http:zSzzSzacl.ldc.upenn.eduzSzWzSzW94zSzW94-0104.pdf/daille94study.pdf (31.03.07)

[106] Heruntergeladen von: http://www.hoover.org/publications/policyreview/3446306.html (21.10.07)

[107] Heruntergeladen von: http://ec.europa.eu/public_opinion/archives/eb/eb67/eb_67_first_de.pdf (26.10.07)

[108] Heruntergeladen von: http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/4848/ftp:zSzzSzftp.research.microsoft.comzSzpubzSzdtgzSzfayyadzSzkdd96zSzfayyad-intro.pdf/fayyad96knowledge.pdf (20.01.07)

*Conference on Practical Aspects of Knowledge Management.* PAKM '98. Zürich: Swiss Group for Artificial Intelligence and Cognitive Science, 9-1 - 9-10[109].

Feldman Ronen, Dagan, Ido und Willi Kloesgen (1996): Efficient Algorithms for Mining and Manipulating Associations in Texts. *In:* Trappl, Robert, Hrsg., *Proceedings of the Thirteenth European Meeting on Cybernetics and Systems Research.* EMCSR '96. Wien: Austrian Society of Cybernetic Studies, 949-954.

Fendler, Susanne und Ruth Wittlinger, Hrsg., (1999): *The Idea of Europe in Literature.* Chippenham, Wiltshire: Antony Rowe Ltd.

Fischer Weltalmanach (2007): *Fischer Weltalmanach 2008. Zahlen Daten Fakten.* Frankfurt am Main: Fischer Taschenbuch Verlag.

Flick, Uwe (2006[4]): *Qualitative Sozialforschung. Eine Einführung.* Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Friedl, Marian (2006): Vom Traum zur Wirklichkeit. *In:* Koschmal, Walter, Hrsg., *Europabilder und Europametaphern.* forost Arbeitspapier 37. München: forost.

Fuhrt, Volker (1993): Perzeptionen und Perzeptions-Defizite: Die gegenseitigen Wahrnehmungen Europas und Japans. *In:* Maull, Hanns W., Hrsg., *Japan und Europa: Getrennte Welten?* Frankfurt am Main, New York: Campus Verlag.

Gebhardt, Hans, Reuber, Paul und Günter Wolkersdorfer (2004): Konzepte und Konstruktionsweisen regionaler Geographien im Wandel der Zeit. *Berichte zur deutschen Landeskunde* 78 (3), 293-312.

Gebhardt, Hans, Reuber, Paul und Günter Wolkersdorfer (2003): Kulturgeographie – Leitlinien und Perspektiven. *In:* Gebhardt, Hans, Reuber, Paul und Günter Wolkersdorfer, Hrsg., *Kulturgeographie. Aktuelle Ansätze und Entwicklungen.* Heidelberg, Berlin: Spektrum Akademischer Verlag, 1-27.

Glaser, Barney G. und Anselm L. Strauss (1998): *Grounded Theory. Strategien qualitativer Forschung.* Bern, Göttingen u.a.: Hans Huber.

Glaser, Brigitte und Hermann J. Schnackertz, Hrsg., (2005): *Europa interdisziplinär. Probleme und Perspektiven heutiger Europastudien.* Würzburg: Königshausen & Neumann.

Gonzalez, Michael (2001): Why Europe Needs Britain. Keeping America's best friend in the European Union. *Policy Review* 108[110].

[109] Heruntergeladen von: http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/10319/http:zSzzSzliawww.epfl.chzSz~lnmainzSzpublicationszSzFeldmanetal98b.pdf/feldman98knowledge.pdf (13.04.07)

[110] Heruntergeladen von: http://www.hoover.org/publications/policyreview/3476456.html (21.10.07)

Hanshew, Kenneth (2006): Europa: banal, fatal oder einfach egal. *In:* Koschmal, Walter, Hrsg., *Europabilder und Europametaphern.* forost Arbeitspapier 37. München: forost.

Harper, John Lamberton (1996): *American Visions of Europe.* Cambridge: University Press.

Hearst, Marti A. (1999): Untangling Text Data Mining. *In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* ACL '99. San Fransisco: Morgan Kaufmann, 3-10[111].

Heeg, Susanne und Jürgen Oßenbrügge (2005): Geopolitische Gesichter des Europäischen Projekts. *In:* Reuber, Paul, Strüver, Anke und Günter Wolkersdorfer, Hrsg., *Politische Geographien Europas – Annäherungen an ein umstrittenes Konstrukt.* Forum Politische Geographie 1. Münster: Lit verlag, 101-116.

Hendrickson, David (2006): Of Power and Providence. The old U.S. and the new EU. *Policy Review* 135[112].

Hernádi, András (1996): *Japan's Image of Europe and Strategy Towards It.* Working Papers of the Institute for World Economics 70. Budapest: Hungarian Academy of Science.

Heyer, Gerhard, Quasthoff, Uwe und Thomas Wittig (2006): *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.* Herdecke, Bochum: W3L-Verlag.

Hipp, Jochen, Güntzer, Ulrich und Gholamreza Nakhaeizadeh (2000): Algorithms for Association Rule Mining – A General Survey and Comparison. *SIGKDD Explorations* 2 (1), 58-64[113].

Jäger, Siegfried (2004[4]): *Kritische Diskursanalyse. Eine Einführung.* Münster: Unrast-Verlag.

Kagan, Robert (2003): *Macht und Ohnmacht. Amerika und Europa in der neuen Weltordnung.* Berlin: Siedler Verlag.

Kelle, Udo (2007[5]): Computergestützte Analyse qualitativer Daten. *In:* Flick, Uwe, von Kardorff, Ernst und Ines Steinke, Hrsg., *Qualitative Forschung. Ein Handbuch.* Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag, 485-502.

Keller, Rainer, Hirseland, Andreas, Schneider, Werner und Willy Viehöver (2001): Zur Aktualität sozialwissenschaftlicher Diskursanalyse – Eine Einführung. *In:* Keller, Rainer, Hirseland, Andreas, Schneider, Werner und Willy Viehöver, Hrsg., *Handbuch*

[111] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/27226/http:zSzzSzacl.ldc.upenn.eduzSzPzSzP99zSzP99-1001.pdf/hearst99untangling.pdf (19.01.2007)

[112] Heruntergeladen von: http://www.hoover.org/publications/policyreview/2913746.html (21.10.07)

[113] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/17185/http:zSzzSzwww.informatik.uni-tuebingen.dezSz~hippjzSzpublicationszSzsigkdd00.pdf/hipp00algorithms.pdf (12.03.07)

*sozialwissenschaftliche Diskursanalyse. Band I: Theorien und Methoden.* Opladen: Leske + Budrich, 7-27.

Keller, Ursula und Ilma Rakusa, Hrsg., (2004): *Writing Europe. What is European about the Literatures of Europe? Essays from 33 European Countries.* Budapest, New York: Central European Press.

Kiefer, Peter, Stein, Klaus und Christoph Schlieder (2006): Visibility Analysis on the Web Using Co-visibilities and Semantic Networks. *In:* Ackermann, Markus, Berendt, Bettina, Grobelnik, Marko u.a., Hrsg., *Semantics, Web and Mining.* Joint International Workshops, EWMF 2005 and KDO 2005 Porto, Portugal, October 3 and 7, 2005. Berlin, Heidelberg: Springer Verlag, 34-50.

Klunkert, Susanne (1996): Europabilder in Mittel- und Osteuropa: Vom Wunschbild zum Abbild der europapolitischen Realität. *In: Europabilder in Mittel- und Osteuropa. Neue Herausforderungen für die politische Bildung.* Bonn: Bundeszentrale für politische Bildung, 231-247.

Köse, Ali (1999): East is East, and West is West. Remarks on Muslim perspectives on Europe and Christianity. *In:* Seufert, Günter und Jaques Waardenburg, Hrsg., *Turkish Islam and Europe. Türkischer Islam und Europa. Europe and Christianity as reflected in Turkish Muslim discourse & Turkish Muslim life in the diaspora.* Papers of the Istanbul Workshop October 1996. Türkische Welten 6. Stuttgart: Franz-Steiner-Verlag, 179-189.

Krippendorff, Klaus (2004[2]): *Content Analysis. An Introduction to Its Methodology.* Thousand Oaks, London u.a.: Sage Publications.

Krotz, Friedrich (2005): *Neue Theorien entwickeln. Eine Einführung in die Grounded Theory, die Heuristische Sozialforschung und die Ethnographie anhand von Beispielen aus der Kommunikationsforschung.* Köln: Herbert von Halem Verlag.

Lafferty, John, McCallum, Andrew und Fernando Pereira (2001): Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In: Proceedings of the 18th International Conference on Machine Learning. ICML'01.* San Francisco: Morgan Kaufmann, 282-289[114].

Lamnek, Siegfried (1995[3]): *Qualitative Sozialforschung. Band 2: Methoden und Techniken.* Weinheim: Psychologie Verlags Union.

Le Goff, Jaques (2004): *Die Geburt Europas im Mittelalter.* München: C. H. Beck.

[114] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/26323/http:zSzzSzwww.aladdin.cs.cmu.eduzSzpaperszSzpdfszSzy2001zSzcrf.pdf/lafferty01conditional.pdf (20.11.07)

Lobin, Henning und Alexander Mehler (2004): Aspekte der texttechnologischen Modellierung. *In:* Mehler, Alexander und Henning Lobin, Hrsg., *Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte.* Wiesbaden: Verlag für Sozialwissenschaften, 1-21.

Mattisek, Annika und Paul Reuber (2004): Die Diskursanalyse als Methode in der Geographie – Ansätze und Potentiale. *Geographische Zeitschrift* 92 (4), 227-242.

Mayring, Philipp (2007[5]): Qualitative Inhaltsanalyse. *In:* Flick, Uwe, von Kardorff, Ernst und Ines Steinke, Hrsg., *Qualitative Forschung. Ein Handbuch.* Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag, 468-475.

Mayring, Philipp (2000[7]): *Qualitative Inhaltsanalyse. Grundlagen und Techniken.* Weinheim: Deutscher Studien Verlag.

Mehler, Alexander (2004): Textmining. *In:* Lobin, Henning und Lothar Lemnitzer, Hrsg., *Texttechnologie. Perspektiven und Anwendungen.* Tübingen: Stauffenberg Verlag, 329-352.

Mytton, Graham (2006): How Africa Sees Europe. *In:* Zöllner, Oliver, Hrsg., *Der Blick der Anderen. Europa in der Wahrnehmung von Medien Afrikas, Asiens und Lateinamerikas.* Bochum: Paragon-Verlag, 109-123.

Nagel, Tilman, Hrsg., (1990): *Asien blickt auf Europa. Begegnungen und Irritationen.* Beiruter Texte und Studien 39. Stuttgart: Franz Steiner Verlag.

Rajman, Martin und Romaric Besançon (1997): Text Mining: Natural Language techniques and Text Mining applications. *In:* Spaccapietra, Stefano, Hrsg., *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics.* DS-7. London u.a.: Chapman & Hall, 50-64[115].

Rehm, Georg (2004): Texttechnologie und das World Wide Web. Anwendungen und Perspektiven. *In:* Lobin, Henning und Lothar Lemnitzer, Hrsg., *Texttechnologie. Perspektiven und Anwendungen.* Tübingen: Stauffenberg Verlag, 434-464.

Reuber, Paul, Strüver, Anke und Günter Wolkersdorfer, Hrsg., (2005) *Politische Geographien Europas – Annäherungen an ein umstrittenes Konstrukt.* Forum Politische Geographie 1. Münster: Lit verlag.

Riketta, Michael und Roland Wakenhut (1998): *Europabild und europäisches Bewusstsein. Bestandsaufnahme der empirischen Forschung und sozialpsychologische Forschungsperspektiven.* ssip-texte 6. Bonn: SSIP.

[115] Heruntergeladen von: http://citeseer.ist.psu.edu/cache/papers/cs/738/http:zSzzSzliawww.epfl.chzSz~lnmainzSzpublicationszSzRajmanBesancon97.pdf/rajman97text.pdf (31.03.07)

Rinck, Jochen (2007): *Entwicklung eines Software-Tools zur Visualisierung und Evaluation von Grounding-Algorithmen für deutsche Webseiten.* Unveröffentlichte Diplomarbeit. Bamberg.

Rosenberger, Sieglinde Katharina (2005): Wie die USA über Europa politisch sprechen. *In:* Öhner, Vrääth, Pribersky, Andreas, Schmale, Wolfgang und Heidemarie Uhl, Hrsg., *Europa-Bilder.* Querschnitte 18. Innsbruck, Wien u.a.: StudienVerlag, 189-204.

Salewski, Michael (2000): *Geschichte Europas. Staaten und Nationen von der Antike bis zur Gegenwart.* München: C. H. Beck.

Schmale, Wolfgang (2000): *Geschichte Europas.* Wien, Köln, Weimar: Böhlau Verlag.

Schmierer, Joscha (1996): *Mein Name sei Europa. Einigung ohne Mythos und Utopie.* Frankfurt am Main: Fischer Taschenbuch Verlag.

Schneider, Steven M. und Kirsten A. Foot (2005): Web Sphere Analysis: An Approach to Studying Online Action. *In:* Hine, Christine, Hrsg., *Virtual Methods. Issues in Social Research on the Internet.* Oxford, New York: Berg, 157-170.

Schott, Michael (2005): Geopolitische Leitbilder und Diskurse als strategische Regionalisierungen in der europapolitischen Diskussion. *In:* Reuber, Paul, Strüver, Anke und Günter Wolkersdorfer, Hrsg., *Politische Geographien Europas – Annäherungen an ein umstrittenes Konstrukt.* Forum Politische Geographie 1. Münster: Lit verlag, 73-100.

Schultz, Hans-Dietrich (1999): *Europa als geographisches Konstrukt.* Jenaer Geographische Manuskripte 20. Jena: Selbstverlag.

Schulze, Hagen (1999): *Phoenix Europa. Die Moderne. Von 1740 bis heute.* Siedler Geschichte Europas Bd. 4. München: Siedler Verlag.

Schwabe, Klaus (2000): Europabilder der Weltmacht Amerika. *In: Europa und Europabilder.* Sammelband der Vorträge des Studium Generale der Ruprecht-Karls-Universität Heidelberg im Sommersemester 1999. Heidelberg: Universitätsverlag C. Winter, 59-73.

Segebrecht, Wulf, Conter, Claude D., Jahraus, Oliver und Ulrich Simon, Hrsg., (2003): *Europa in den europäischen Literaturen der Gegenwart.* Frankfurt am Main, Berlin u.a.: Peter Lang.

Smith, Adam (2004): *Theorie der ethischen Gefühle.* Hrsg. v. Walther Eckstein. Hamburg: Felix Meiner Verlag.

Thurlow, Crispin, Lengel, Laura und Alice Tomic (2005): *Computer Mediated Communication. Social Interaction and the Internet.* London, Thousand Oaks u.a.: Sage Publications.

Weichhart, Peter (1999): Die Räume zwischen den Welten und die Welt der Räume. *In:* Meusburger, Peter, Hrsg., *Handlungszentrierte Sozialgeographie.* Erdkundliches Wissen 130. Stuttgart: Franz Steiner Verlag, 67-94.

Werlen, Benno (1997): *Sozialgeographie alltäglicher Regionalisierungen. Band 2: Globalisierung, Region und Regionalisierung.* Erdkundliches Wissen 119. Stuttgart: Franz Steiner Verlag.

Winograd, Terry und Fernando Flores (1986): *Understanding Computers and Cognition: A New Foundation for Design.* Norwood, New Jersey: Ablex Publishing Corporation.

Witschel, Hans Friedrich (2004): *Terminologie-Extraktion. Möglichkeiten der Kombination statistischer und musterbasierter Verfahren.* Würzburg: ERGON Verlag.

Xenos, Michael A. und Kirsten A. Foot (2005): Politics As Usual, or Politics Unusual? Posotion Taking and Dialogue on Campaign Webites in the 2002 U.S. Elections. *Journal of Communication* 55 (1), 169-186.

# Verzeichnis der verwendeten Webseiten

## *Bei der Softwareentwicklung*

http://code.google.com/apis/soapsearch/ (20.11.07)

http://java.sun.com/javase/downloads/index.jsp (20.11.07)

http://www.netbeans.info/downloads/index.php?rs=11 (20.11.07)

http://www.alias-i.com/lingpipe (21.11.07)

http://www.alias-i.com/lingpipe/licenses/lingpipe-license-1.txt (21.11.07)

http://bllip.cs.brown.edu/resources.shtml (21.11.07)

http://htmlparser.sourceforge.net (21.11.07)

http://opennlp.sourceforge.net (21.11.07)

http://jtextpro.sourceforge.net (21.11.07)

http://jwebpro.sourceforge.net (21.11.07)

http://www.computing.dcu.ie/~acahill/tagset.html (23.11.07)

http://earth-info.nga.mil/gns/html/namefiles.htm (25.11.07)

http://jung.sourceforge.net (28.11.07)

http://jung.sourceforge.net/doc/index.html (28.11.07)

http://jung.sourceforge.net/pmwiki/index.php/Main/ProjectsUsingJUNG (28.11.07)

### *Im Rahmen der Korpuszusammenstellung und -analyse*

http://india.gov.in/sectors/commerce/india_trade.php (18.06.07)
http://swrs.ap.gov.in/Academic/Syllabus%20Division/6/6%20th%20class%20social%20daywise%20plan.doc (18.06.07)
http://www.vigyanprasar.gov.in/dream/apr2001/RADIOACTIVITY.htm (18.06.07)
http://www.dcita.gov.au/__data/assets/word_doc/10984/Meridian_Connections.doc (19.06.07)
http://www.dfat.gov.au/media/speeches/foreign/1996/germany_.html (19.06.07)
http://www.dfat.gov.au/media/speeches/foreign/1997/france.html (19.06.07)
fmind.gov.ng (19.06.07)
namasa.gov.ng (19.06.07)
http://www.trademinister.gov.au/speeches/2004/040513_aus_business_europe.html (19.06.07)
www.deltastate.gov.ng/wwwboard/messages/2477.html (03.07.07)
http://www.deltastate.gov.ng/admin&governmentframe.htm (03.12.07)
http://www.fmind.gov.ng/docs/PRSdept.doc (05.12.07)
ncc.gov.ng (05.12.07)
http://ncc.gov.ng/index1.htm (05.12.07)
www.smedan.gov.ng (05.12.07)
ap.gov.in (06.12.07)
austrade.gov.au (06.12.07)
http://www.austrade.gov.au/About-Austrade1351/default.aspx (03.12.07)
awm.gov.au (06.12.07)
http://www.awm.gov.au/aboutus/origins.htm (06.12.07)
dcita.gov.au (06.12.07)
defence.gov.au (06.12.07)
dfat.gov.au (06.12.07)
incois.gov.in (06.12.07)
http://www.incois.gov.in/Incois/incois1024/index/index.jsp?res=1024# (06.12.07)
iricen.gov.in (06.12.07)
mea.gov.in (06.12.07)
vigyanprasar.gov.in (06.12.07)
http://www.egov.vic.gov.au (12.12.07)

http://www.egov.vic.gov.au/index.php?env=-categories:m1699-1-1-8-s-0&reset=1 (12.12.07)

# Anhang

Sascet 1.0 – Dokumentationsteil (CD-Teil)

**Systemvoraussetzungen**

- Java Runtime Environment 6.0 oder höher
- mindestens 512 MB RAM (1024 MB empfohlen)

**Inhalt: Materialien zur Dokumentation der Software-Entwicklung**

- Das Programm
    - lauffähige Programmversion
    - Kurzes Manual zur Programmbenutzung
    - UML-Klassendiagramm zur Architektur von Sascet
    - Komplettes JavaDoc zu Sascet
    - XML-Schemata zu den Datenformaten der extrahierten Terminologie bzw. der errechneten Regeln
    - Der Quellcode zu Sascet
- Das Textkorpus zu den drei Domains

**Anmerkungen zum Gazetteer:**

Das GNSGazetteer liegt dem Programm aufgrund seiner schieren Größe nicht bei – er muss separat heruntergeladen und unter „Analyse“ > „Einstellungen“ ins Programm importiert werden.

**Hinweis:**

Zum Abrufen der auf die Inhalte verlinkten Übersicht bitte *index.html* auf der beiliegenden CD aufrufen.

„Ich erkläre hiermit gemäß § 28 Abs. 7 DPO, dass ich die vorstehende Diplomarbeit selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe."

Bamberg, den 20.12.2007