CognIEffect: Identifikation kognitiver Effekte in Online-Bewertungen von Arztpraxen

Valentina Stuß

München 2017

CognIEffect: Identifikation kognitiver Effekte in Online-Bewertungen von Arztpraxen

Valentina Stuß

Inauguraldissertation
zur Erlangung des Doktorgrades der Philosophie
an der Fakultät für Sprach- und Literaturwissenschaften
der Ludwig–Maximilians–Universität
München

vorgelegt von
Valentina Stuß
aus Konotop (Ukraine)

München, den 30.11.2017

Erstgutachter: Prof. Dr. Franz Guenthner
Zweitgutachterin: Prof. Dr. Michaela Geierhos
Tag der mündlichen Prüfung: 23.02.2018

Für Elly Nora

Danksagung

Ich bedanke mich bei meinem Erstgutachter, Herrn Prof. Dr. Guenthner, für die zahlreichen produktiven Treffen und seine vielfältigen Hinweise. Außerdem wurde mir eine wertvolle Ressource für diese Arbeit zur Verfügung gestellt. Vielen Dank dafür.
Der Zweitgutachterin, Frau Prof. Dr. Michaela Geierhos, gilt ein besonderer Dank für die Eröffnung der Möglichkeit dieser Promotion. Ihre Unterstützung begleitete mich außerdem im Laufe der ganzen Jahre meines Studiums, meiner Arbeit und des Schreibens dieser Dissertation. Diese drückte sich in einer Reihe von mehrstündigen Diskussionen aus, ihren präzisen Hinweisen und Anmerkungen, ihrer fachlichen und moralischen Unterstützung, ihrem Verständnis und gleichzeitiger Strenge. Danke für ihre enorme Kompetenz und Menschlichkeit.
Bei Thomas Schäfer, dem Systemadministrator des CIS, bedanke ich mich für die zahlreichen Stunden der Pflege meiner Rechner u. ä., was bereits seit dem Beginn meines Studiums seinerseits freundlich und vor allem geduldig immer wieder durchgeführt wurde.
Für die Unterstützung seitens meines Ehemannes, Torsten Stuß, für seine Korrekturen, zahlreiche Anmerkungen, Lob, Kritik und Hilfe bei den Annotationen bedanke ich mich sehr.
Meiner Tochter Elly, die mich die letzten neun Monate meiner Promotion überall begleitete und glücklich machte, danke ich vom ganzen Herzen.
Schließlich bedanke ich mich bei meinen Annotatoren, Cristina Rothenbücher und Evgeniya Badalova für einige mit den aufwendigen Arbeiten verbrachte Stunden.

Inhaltsverzeichnis

Zusammenfassung

Gegenstand der vorliegenden Arbeit ist die automatische Identifikation kognitiver Effekte in Patientenbewertungen von Arztpraxen. Kognitive Effekte stellen sozialpsychologische Phänomene, Denkfehler dar, die den Bewertenden bei der Meinungsbildung unbewusst unterlaufen. Von diesen Denkfehlern gibt es unterschiedliche Arten, was einer Adoption ihrer allgemeinen Definitionen für die gewählte Domäne bedarf, die so noch nie durchgeführt wurde und in dieser Arbeit stattfindet.
In der vorliegenden Arbeit wird versucht, ausgewählte Effekte zu identifizieren und durch Vergabe entsprechender Scores zu klassifizieren. Die Identifikation erfolgt anhand aufgestellter Kriterien. Eines davon ist die Erkennung linguistischer Muster, die Meinungen beinhalten. Um die Muster automatisch zu erfassen, werden Erkenntnisse zweier computerlinguistischer Disziplinen – Informationsextraktion und Stimmungsanalyse – herangezogen. Als Methode wird ein regel- und musterbasiertes Verfahren gewählt, was mit effizienten lokalen Grammatiken umgesetzt wird. In Kombination mit anderen für den jeweiligen Effekt typischen Kriterien – wobei bei verschiedenen Effekten auch gleiche Kriterien vorkommen können – werden Effekte identifiziert und klassifiziert. Interessant sind die erzielten Ergebnisse, die u. a. die Aussagen zum Grad des Auftretens von Effekten innerhalb der Bewertungen darstellen. Es ist nicht neu, sozialpsychologische Phänomene zu identifizieren, zumal eine Reihe von Phänomenen bereits identifiziert wurde: Ironie, Bias, Spam, Inkonsistenzen etc. Neu ist das gewählte Phänomen selbst und dessen automatisches Erkennen in Bewertungstexten zu Arztpraxen. Neu ist außerdem die Erkenntnis, dass mehrere verschiedene Effekte innerhalb einer Bewertung auftreten können und differenziert werden müssen. Die Auswahl und die Zusammensetzung der Methoden zum Zweck der Bewältigung von aufgestellten Forschungsfragen sind dementsprechend innovativ, wobei die allgemeineren und umfassenderen Aussagen getroffen werden können, als dies im Fall der

Auseinandersetzung lediglich mit einem Effekt wäre. Schließlich kann man Effekte als ‚wertvolle Ausreißer' betrachten, was den Anlass zur weiteren Auseinandersetzung mit diesem Phänomen im Rahmen der interdisziplinären Projekte bietet.

Kapitel 1

Einleitung

1.1 Motivation und Problemstellung

Seit der Entstehung des Web 2.0 werden Internetnutzern[1] zahlreiche Technologien für den multi-direktionalen Erfahrungsaustausch angeboten. So haben Bewertungsportale die Funktion einer sogenannten digitalen Mundpropaganda übernommen (Liu, 2010, S. 1; Shailesh, 2015, S. 113). Zahlreiche Rezensionen der Konsumenten, Kunden, Patienten oder anderen Zielgruppen bilden einen riesigen Datenbestand und stellen eine wertvolle Informationsquelle dar (Kaiser, 2012, S. 2; Bretschneider, 2015, S. 1; Wolfgruber, 2015, S. 2), die eine wissenschaftliche Auseinandersetzung erforderlich macht (Geierhos und Stuß, 2015). Aus diesem Grund ist die Erforschung des Konsumentenverhaltens oder der Meinungsbildung von großem wirtschaftlichen Interesse, was z. B. das Treffen der Entscheidungen bei der Produktentwicklung, rechtzeitiges Erkennen von Chancen und Risiken zum Ergreifen entsprechender Maßnahmen (Kaiser, 2012, S. 2) oder „Prognostizierbarkeit von Absatzzahlen“ (Bretschneider, 2015, S. 1) betrifft. Wie und ob Meinungsbildung das Verhalten der Verbraucher und Dienstleistungsnehmer beeinflusst und wie Wissensverarbeitung bei Menschen abläuft, sind Fragen, mit denen sich Sozial- und Kognitionspsychologen beschäftigen (Stürmer, 2009, S. 11; Wolff, 1993, S. 27ff.).
Computerlinguisten haben ein wissenschaftliches Interesse, Erkenntnisse mit-

[1] Aus Gründen der leichteren Lesbarkeit wird auf eine geschlechtsspezifische Differenzierung verzichtet. Entsprechende Begriffe gelten im Sinne der Gleichbehandlung für beide Geschlechter.

tels Informationsextraktion und Stimmungsanalyse (vgl. Kim und Hovy, 2004; vgl. Hatzivassiloglou und McKeown, 1997) aus diesen digitalen Textsammlungen zu gewinnen. Für die Informationsextraktion ist z. B. die automatische Textanalyse interessant: die relevanten Bewertungsinformationen werden aus unstrukturierten Daten extrahiert und in einer strukturierten Form wie z. B. Einträge für Datenbanken oder Frage-Antwort-Systeme umgesetzt (Neumann, 2004, S. 502). Im Kontext computergestützter Auseinandersetzungen mit Online-Texten, die positive oder negative Meinungen zu Produkten, (Dienst)Leistungen u. ä. beinhalten, wurden bereits Arbeiten zur automatischen Erkennung z. B. der Polarität (vgl. Kim und Hovy, 2004; vgl. Hatzivassiloglou und McKeown, 1997), der Subjektivität (vgl. Wiebe et al., 2004) durchgeführt, Systeme für Stimmungsanalyse zum Aufbau von Suchmaschinen oder Dialogsystemen (vgl. Wolfgruber, 2015) oder zur Erforschung des Konsumentenverhaltens im wirtschaftlichen Sinne (vgl. Kaiser, 2012) entwickelt. Andere Arbeiten beschäftigten sich mit der Ausarbeitung möglicher Ressourcen wie Wörterbücher zum Zweck der Einbindung z. B. in Informationsextraktionssysteme (vgl. Remus et al., 2010; vgl. Guenthner und Maier, 1994) oder einer automatischen Erkennung einzelner Phänomene wie Ironie und Sarkasmus (vgl. Schieber et al., 2012; vgl. Bamman und Smith, 2015; vgl. Hallmann et al., 2016) etc. Als Untersuchungskorpora sind bei Stimmungsanalysen Webportale sozialer Netzwerke (vgl. Kaiser, 2012), die Verbraucherbewertungen konkreter Produkte wie z. B. Mobilgeräte (Tablet-PC) (vgl. Schieber et al., 2012) oder der Dienstleistungen (Hotels) (vgl. Wolfgruber, 2015) von wissenschaftlichem Interesse.
Bewertungsportale gewinnen durch ihr rapides Wachstum immer mehr an Bedeutung (Geierhos et al., 2015b, S. 1). Bewertungen von Ärzten bzw. Arztpraxen, die Patienten oder ihre Angehörigen verfassen, bieten den Anlass, die genannten Bewertungstexte auf Online-Portalen als eine eigene Domäne zu betrachten. In der vorliegenden Arbeit bildet sie das Korpus als Grundlage zur Untersuchung des weiter formulierten Phänomens. Ein weiteres Argument – außer der bekannten Tatsache, dass Arztbesuche, Behandlungen und Therapien ein Teil des Alltags von jedem Menschen sind, was die Bedeutsamkeit und die Aktualität der gewählten Domäne verdeutlicht – ist ein ständig wachsender Gesundheitssektor selbst. Dieser ist – wirtschaftlich gesehen – dem Bereich der Dienstleistungsanbieter zuzuordnen. „Das Bedürfnis nach Gesundheit, Wellness und Wohlbefinden hat in der Gesellschaft einen hohen Stellenwert und stärkt die Nachfrage nach personenbezogenen Dienstleistungen. [...] Die Alterung und Individualisierung der Gesellschaft lassen den Be-

darf an sozialen und personenbezogenen Diensten weiter wachsen. Angesichts der wirtschaftsstrukturellen Veränderungen wie Privatisierung, Outsourcing, Konkurrenz und Kostendruck beginnen die Anbieter, sich durch Leistungsdiversifizierung neue Wachstumsmöglichkeiten zu erschließen. Dabei gewinnen vor allem die Rand- und Nachbarbereiche der Gesundheitswirtschaft zunehmend an Relevanz“ (Meifort, 2002, S. 34).

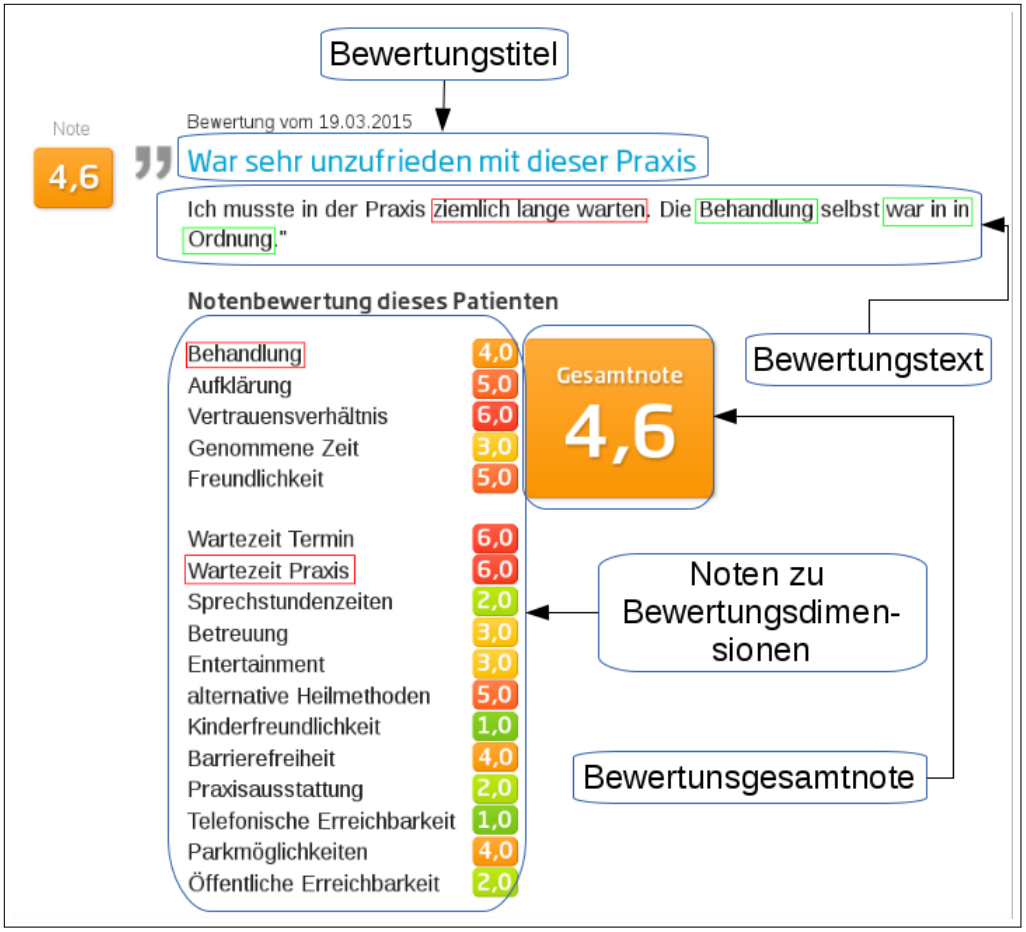

Abbildung 1.1: Struktur und Komponenten einer Arztbewertung (Jameda)

Außer Meinungen zu bestimmten Aspekten oder Gegenständen, die positiv oder negativ sein können (s. Abbildung 1.1), bieten Bewertungstexte ein breites Spektrum an wissenschaftlichen Fragestellungen, die im genannten Kontext zu inter- und multidisziplinären Forschungen veranlassen. So wurde in

Jahren 2013 – 2014 im Rahmen des interdisziplinären Projekts „More than Words“ an der Fakultät für Wirtschaftswissenschaften der Universität Paderborn User Generated Content[2] (UGC) – im Einzelnen von Arztbewertungen – analysiert[3]. Mit Bezug der wirtschaftlich motivierten Fragestellungen wie z. B. der Identifikation von latenten Dienstleistungsqualitätsmerkmalen[4] wurde eine Reihe computergestützter Analysen durchgeführt. Einige davon waren z. B. die semantische Inhaltsanalyse zur Feststellung domänenspezifischer Anforderungen und nutzerspezifischer Polaritätsabweichungen, empirische Ermittlung neuer Bewertungsdimensionen oder ein automatischer Vergleich qualitativer und quantitativer Dienstleistungsbewertungen[5].
Nicht geringere Bedeutung haben sozialpsychologisch motivierte Fragestellungen im Kontext von Meinungsäußerungen, da dadurch z. B. Rückschlüsse auf die Ursachen der (Un)Zufriedenheit der Patienten mit den von Ärzten bzw. Arztpraxen erbrachten Dienstleistungen gezogen werden können. „Die Sozialpsychologie beschäftigt sich – wie die Psychologie ganz generell – mit der Erklärung und Beschreibung von Verhalten und Erleben. Sie unterscheidet sich von den anderen Disziplinen der Psychologie aber dahingehend, daß sie den Ausschnitt von Verhalten und Erleben untersucht, der sich auf zwischenmenschliche Interaktionen bezieht. [...] Einstellungen sind deshalb ein so zentraler Bestandteil der Sozialpsychologie, da sie in starkem Maße unseren Umgang mit anderen bestimmen. So hat unsere Einstellung gegenüber Minderheiten einen Einfluß darauf, ob man mit Mitgliedern dieser Minderheit sozialen Kontakt aufnimmt oder nicht (Stereotype)“[6]. Einige solcher Problematiken in Kundenbewertungen sind z. B. individuelle und kollektive Inkonsistenzen, sprich: Fehler, die aus nicht kongruenten Polaritäten der Bewertungstexte und nummerischer Bewertungen gleicher Aspekte hervorgehen (Geierhos et al., 2015b, S. 2). Diese Inkonsistenzen, Widersprüche, Unregelmäßigkeiten oder kognitive Effekte[7] als sozialpsychologische Phänomene, die von Patienten in ihren Rezensionen hinterlassen werden, lassen sich auf verschiedene wahrnehmungsbedingte Faktoren zurückführen. Eini-

[2] Nutzergenerierte Inhalte.
[3] `http://bit.ly/20HZFDV` (06.02.2016).
[4] ebd.
[5] ebd.
[6] `http://www.spektrum.de/lexikon/psychologie/sozialpsychologie/14576` (28.04.2017).
[7] In der vorliegenden Arbeit wird der Begriff „kognitive Effekte“ oder „Denkfehler“ gebraucht.

ge betreffen individuelle Haltungen, Mentalitäten und Zufriedenheit (Pham and Jung, 2013, zitiert in Geierhos et al. (2015b, S. 2)). Deutlich wird die Tatsache, dass die angedeuteten Effekte Arztbewertungen in eine entsprechende Richtung verzerren und dadurch das Gesamtbild eines Arztes oder einer Arztpraxis verfälschen.

Wie kommen diese Fehler zustande?

Bevor Meinungen gebildet werden können, müssen sämtliche ihnen verfügbare Informationen zu Bewertungsgegenständen verarbeitet werden. Eine Idealvorstellung einer korrekten Informationsverarbeitung wäre „das vernünftige Denken“, das auf den Gesetzen der Logik basiert (Gigerenzer und Gaissmaier, 2006, S. 1), wodurch man richtige Entscheidungen treffen oder objektive Wertungen abgeben könnte. Was geschieht, wenn man keine Zeit hat, sich zu belesen oder zu informieren, bevor man sich zu diesem oder jenem Problem äußert? In diesen Fällen benutzen Menschen sogenannte Heuristiken (Michalkiewicz, 2015). „Heuristiken sind mentale Strategien, Faustregeln oder Abkürzungen, die uns helfen, mit begrenztem Wissen und begrenzter Zeit Entscheidungen zu treffen und Urteile zu fällen“ (ebd.). Die Anwendung dieser Heuristiken haben oft verzerrende Einflüsse auf Urteile bzw. führen zu systematischen Fehlern (Stürmer, 2009, S. 213; Tversky und Kahneman, 1974, S. 1124). Jedoch werden diese in vielen alltäglichen Situationen automatisch, aber auch bewusst verwendet (Michalkiewicz, 2015), da sie rational, schnell, mühelos anwendbar sind und in vielen Situationen zu guten Ergebnissen führen (ebd.).
Auf dem Gebiet kognitiver Effekte und Urteilsheuristiken leisteten Tversky und Kahneman (1974) einen bedeutsamen Forschungsbeitrag. Beide haben die Gründe für stattgefundene kognitive Effekte auf drei Heuristiken zurückgeführt: Repräsentativitätsheuristik, Verfügbarkeitsheuristik und Anker und Anpassung[8]. Außer diesen existieren in der wissenschaftlichen Literatur weitere zahlreiche Heuristiken wie z. B. die Rekognitionsheuristik, die Take-the-Best Heuristik, die Center-of-the-Circle Heuristik, die Hiatus Heuristik (Michalkiewicz, 2015) oder die Positive Teststrategie[9,10].
Solche Verzerrungen, die mit kognitiven Effekten versehenen Bewertungen

[8] `https://www.youtube.com/watch?v=woug36Y4_y8` (04.03.2016).

[9] `http://lexikon.stangl.eu/1546/positive-teststrategie/` (30.03.2016).

[10] In entsprechenden Abschnitten wird lediglich auf die für ausgewählte Effekte relevanten Heuristiken kurz eingegangen.

automatisch festzuhalten, ist eine Aufgabe, deren Bedarf offensichtlich wird. Dass diese Wahrnehmungsfehler in Arztbewertungen vorkommen, wird an einem im nächsten Abschnitt aufgeführten Beispiel deutlich. Zum einen gehören Denkfehler von Patienten zu sogenannten Ausreißern, deren Bedeutung im Rahmen statistischer Erhebungen gemindert wird, um das Analyseergebnis nicht zu beeinflussen[11]. Eine automatische Entfernung solcher Ausreißer aus den Bewertungstexten zur Objektivierung statistischer Analysen wäre eine der möglichen Anwendungen im Rahmen der vorliegenden Arbeit[12]. Zum anderen spielen kognitive Effekte, wie oben angesprochen, bei der Auseinandersetzung mit konkretem Verhalten der Patienten eine entscheidende Rolle. „[...] the information provided can help physicians gain a better understanding of patient concerns“ (Emmert et. al., 2014, zitiert in Geierhos et al. (2015b, S. 2)). Schließlich stellen sie als psychologische Phänomene einen Eigenwert dar, was zu neuen Studien und Untersuchungen wirtschaftlicher und sozialpsychologischer Disziplinen motivieren soll. Computerlinguistisch ist in diesem Sinne die Frage interessant, ob eine automatische Klassifikation kognitiver Effekte denkbar ist.

1.2 Zielsetzung und Einschränkungen

Ziel der vorliegenden Arbeit ist es, die seitens Patienten stattgefundenen kognitiven Effekte in Arztbewertungstexten zu definieren, automatisch zu identifizieren und zu klassifizieren. Im Mittelpunkt des Erkenntnisinteresses stehen dabei folgende *Forschungsfragen*:

- Lassen sich sozialpsychologische Phänomene – im Einzelnen kognitive Effekte – in von Patienten verfassten Bewertungstexten nachweisen? (Existenzfrage)
- (Wie) Können kognitive Effekte in Arztbewertungen automatisch identifiziert werden? (Identifikationsfrage)
- (Wie) Lassen sich kognitive Effekte in der gewählten Domäne auseinander halten bzw. differenzieren? (Klassifikationsfrage)

[11] `http://www.statistik.wiso.uni-erlangen.de/forschung/d0009.pdf` (08.09.2014).

[12] Zu Möglichkeiten des Umgangs mit Ausreißern s. ausführlicher Kapitel 3, Abschnitt 3.3.2.1, Seite 97f.

Um diese Fragen zu beantworten, wird eine Auseinandersetzung mit dem Begriff der kognitiven Effekte im Kontext der schriftlich formulierten Patientenmeinungen notwendig. Damit wird festgestellt, ob kognitive Effekte allein aus Texten „herausgelesen“ werden können, um welche es sich dabei handelt und ob eigene, domänenspezifische Denkfehler von Patienten auffindbar sind. Außerdem wird im Rahmen automatischer Textverarbeitung deutlich, dass die Identifikation kognitiver Effekte allein anhand graphischer, maschinenlesbarer Muster durchführbar sein muss. Zu diesem Zweck werden Methoden und Verfahren der Computerlinguistik nötig, um eine automatische Extraktion dieser Muster zu bewerkstelligen. Gesucht werden dabei Merkmale bzw. Kriterien (s. Kapitel 2, Abschnitt 2.1.2.1; Kapitel 3, Abschnitt 3.2; Kapitel 4, Abschnitt 4.3), durch die eine Bewertung als fehlerhaft interpretiert werden würde. Welche Möglichkeiten bieten Arztbewertungen zur Aufstellung solcher Identifikationskriterien? Um der Antwort näher zu kommen, kann man zunächst ein Online-Bewertungsportal betrachten, um die Struktur und Komponenten einer Arztbewertung nachvollziehen zu können. Wie aus der Abbildung 1.1 ersichtlich, besteht eine Arztbewertung aus einem qualitativen und einem quantitativen Bewertungsbereich (Geierhos et al., 2015b, S. 5). Der qualitative Bereich umfasst zwei Komponenten[13]: Bewertungstitel und Bewertungstext, wobei die Texte in den genannten Feldern von Patienten frei formuliert werden können. Bewertungstitel z. B. impliziert eine Überschrift, die oft eine wertende Aussage zur allgemeinen Zufriedenheit enthält, kann jedoch manchmal aus einer Stellungnahme nur zu einer („netter Arzt“: [„Freundlichkeit“])[14] oder mehreren durch ein Bewertungsportal (hier: Jameda) vordefinierten Bewertungsdimensionen („Kompetenter Arzt ohne lange Wartezeiten“: [„Behandlung“, „Wartezeit (Praxis)“])[15] bestehen sowie die Beschwerden bzw. Diagnosen ohne jegliche Wertung benennen („Herzrythmusstörungen“)[16] etc. Der Begriff der Bewertungsdimension wurde in dem im vorigen Abschnitt angesprochenen Projekt „More than Words“ eingeführt und im Sinne von den zu bewertenden Kategorien (Behandlung, Aufklärung, Vertrauensverhältnis etc.) verwendet. In der vorliegenden Arbeit bleibt der o. g. Begriff erhalten, um Verwechslungen mit den Lexikon-Kategorien oder

[13]Die Komponenten einer Arztbewertung ähneln den Bestandteilen einer Hotelbewertung, die Wolfgruber (2015, S. 15f.) in ihrer Dissertation beschrieben hat.

[14]`http://www.jameda.de/berlin/aerzte/hno-aerzte-hals-nase-ohren/dr-kai-mueller/bewertungen/81160490_1/` (09.03.2016).

[15]ebd.

[16]`http://www.docinsider.de/arzt/bewertungen/simone-henne#/` (09.03.2016).

Kodierungen (s. z. B. Kapitel 4, Abschnitt 4.2.2.3, Kapitel 5, Abschnitt 5.1.2) zu vermeiden. Im quantitativen Bereich befinden sich die nummerischen Bewertungen (hier: Schulnoten 1 bis 6) zu o. g. Dimensionen und eine Bewertungsgesamtnote, die sich aus dem Mittelwert der zu den ersten fünf Bewertungsdimensionen (weiter als Hauptdimensionen bezeichnet) vergebenen nummerischen Werte (hier: Schulnoten) errechnet. Nummerische Werte zu Hauptdimensionen sind für Bewertende verpflichtend, ohne diese kann man keine Bewertung zu einer Praxis abgeben. Aus den getätigten Angaben in den eben beschriebenen Komponentenfeldern sollten erwähnte Indikatoren für mögliche Denkfehler von Patienten auffindbar sein. Wenn man die auf der o. g. Abbildung formulierten Überschrift und Text näher betrachtet, wird es nicht schwer festzustellen, dass sich die Äußerungen, wertende Sätze zumindest teilweise gerade auf die vordefinierten Dimensionen beziehen (hier: „Wartezeit (Praxis)“, „Behandlung“). Ebenso fällt es auf, dass die Bewertung der Dimension „Behandlung“ im Text positiv ausfällt, während die für dieselbe Dimension vergebene Note negativ ist. Dieser eindeutige Widerspruch zeigt ein Beispiel einer in sich inkonsistenten Bewertung (Geierhos et al., 2015a, S. 7). Eine automatische Identifikation und Zuordnung zu einem möglichen kognitiven Effekt (Klassifikation) solcher und möglicher anderer Arten von Unregelmäßigkeiten in Arztbewertungen sind als Ziele der vorliegenden Arbeit zu verstehen.
Sozialpsychologische Phänomene – kognitive Effekte – wurden im Kontext der Arztpraxenbewertungen nicht erforscht, eine theoretische Basis zur beschriebenen Problematik existiert in der wissenschaftlichen Literatur nicht. Das bedeutet, dass es diesbezüglich um eine Reihe von wissenschaftlichen Analysen geht: Die Auseinandersetzung mit kognitiven Effekten und deren Klassifikation, die Überprüfung der Übertragbarkeit des jeweiligen Effekts auf die Domäne der Arztbewertungen, die Festlegung typischer Merkmale übertragbarer Effekte im genannten Kontext und schließlich die Entwicklung eines Informationsextraktionssystems zur Identifikation der Denkfehler von Patienten. Im Rahmen der Dissertation ist eine solche Aufgabe, die alle gerade aufgezählten Aspekte zu 100 Prozent abdecken würde, nicht zu leisten. Daher bedarf es einiger *Einschränkungen*, von denen manche bereits an dieser Stelle genannt werden können:

- Aufgrund der anwendungsorientierten Ausrichtung dieser Arbeit kann eine umfassende Analyse aller existierenden kognitiven Effekte nicht erfolgen. Es wird hier daher eine Auswahl von drei bis vier identifizier-

baren Effekten getroffen.

- Zum oben beschriebenen Vergleich von sprachlichen Äußerungen der Patienten mit Noten, die zu Bewertungsdimensionen vergeben wurden, benötigt man eine umfassende Stimmungsanalyse, die in dieser Arbeit auf Meinungen zu vorgegebenen Dimensionen eines entsprechenden Online-Portals eingeschränkt wird.
- Da bei jedem Bewertungsportal eine eigene Klassifikation der Bewertungsdimensionen existiert (s. Kapitel 2, Abschnitt 2.1.1.1.3, Seite 15ff.), impliziert eine einheitliche Klassifikation für mehrere Online-Portale (s. Kapitel 4, Abschnitt 4.1.5, Seite 106) einen zeitlichen Aufwand, der zum Erreichen des Forschungsziels und zur Beantwortung der aufgestellten Forschungsfragen nicht notwendig ist. Aus diesem Grund wird darauf verzichtet.

Weitere Einschränkungen – sofern nötig – werden an entsprechenden Stellen der vorliegenden Arbeit genannt.

Zusammenfassend, steht im Rahmen der vorliegenden Arbeit die Entwicklung eines automatischen Verfahrens zur Identifikation kognitiver Effekte (CognIEffect) im Vordergrund. Zentrale Rolle spielt dabei die Informationsextraktion (IE)[17], deren Inhalte Stimmungen bzw. Meinungen bezogen auf die vorgegebenen Bewertungsdimensionen bilden. Die automatische Stimmungsanalyse erfolgt mittels lokaler Grammatiken. Durch die Effizienz lokaler Grammatiken in der Beschreibung der Sprache (Nagel, 2008, S. 7), durch die Möglichkeit partieller syntaktischer Analysen der Teilsätze oder Phrasen (ebd.; Geierhos, 2010, S. 30; Wolfgruber, 2015, S. 123), durch ihre einfache Modifizierbarkeit und Wiederverwendbarkeit etc. eignen sich diese zur Modellierung des Zusammenspiels von Syntax und Semantik in Patientenäußerungen, um dann in diesen verschiedene Arten von Stimmungen zu bestimmten Objekten zu lokalisieren, zu klassifizieren und zu annotieren. Kognitive Effekte werden zunächst ausgewählt, ausführlich domänenbezogen analysiert und in Bezug auf ihre automatische Identifizierbarkeit geprüft. Die nicht identifizierbaren Effekte werden verworfen. Die Identifikation der Effekte basiert auf einer automatischen Erkennung der im Vorfeld für jeden

[17] s. eben eingeführte Abkürzung zum Verfahren der automatischen Identifikation kognitiver Effekte CognIEffect, wobei IE symbolisch im Zentrum der Bezeichnung des Verfahrens steht.

identifizierbaren Effekt definierten Kriterien. Die Kriterien setzen sich aus der Kombination von den mit lokalen Grammatiken extrahierten Mustern (s. Kapitel 5, Abschnitt 5.2, Seite 148ff.) und weiteren Korpusanalysen (s. Abschnitt 5.3, Seite 167ff.) zusammen. Die Ergebnisse der Musterextraktion und Denkfehleridentifikation werden im Anschluss an das durchgeführte Verfahren präsentiert und analysiert, Probleme und mögliche Modifizierungen werden diskutiert.

1.3 Aufbau der Arbeit

Um das Ziel zu erreichen und die aufgeführten Forschungsfragen zu beantworten (Abschnitt 1.2), wird eine strukturierte Vorgehensweise notwendig.

1.3.1 Theoretischer Rahmen

Um sozialpsychologische Phänomene, kognitive Effekte, differenziert und in einer speziellen Domäne im Sinne deren automatischen Identifikation erfassbar zu machen, muss es eine Reihe von Überlegungen und Auseinandersetzungen mit dem Phänomen selbst geben. Mit den Möglichkeiten moderner automatischer Informationsverarbeitung sowie mit Problemen und Fragen der Übertragbarkeit von Effekten auf die Domäne der Arztbewertungen. Dies bedeutet, dass zunächst die Entwicklung theoretischer Grundlagen notwendig wird, um eine eigene zielbezogene Konzeption aufzustellen (s. Kapitel 2 bis 4).
Im Kapitel 2 (Multidisziplinäre Grundlagen) werden Aspekte menschlicher Informationsverarbeitung aus der Kognitions- und Sozialpsychologie thematisiert. Es wird ein Erklärungsversuch möglicher ‚falsch' ablaufender Denkprozesse abgehandelt, um der Frage nach der Existenz der aus solchen Prozessen resultierenden Fehler näher zu kommen. Dieselbe Existenzfrage wird deutlicher bei der Definition und Klassifikation kognitiver Effekte in Abschnitten 2.1.2 und 2.1.3, da bei deren aufgestellten Übersicht und dem bisherigen Wissen zur Domäne bereits eine Vorstellung zu Merkmalen möglich ist, durch die die Auffindbarkeit bestimmter Effekte in Arztbewertungen bestätigt bzw. ausgeschlossen wird. Werden graphische maschinenlesbare Merkmale für diesen oder jenen Effekt eindeutig, so wird die Existenzfrage (s. Seite 6) positiv beantwortet.
Außer der Definition zu kognitiven Effekten benötigt man eine Reihe von

zusätzlichen Definitionen z. B. zu Meinungen, Entitäten oder Objekten etc., was ebenfalls im Kapitel 2, Abschnitt 2.2 thematisiert wird. Dort werden vor allem anwendungsorientierte Fragen der Musterextraktion interessant, wodurch die Erkennungsmerkmale des Untersuchungsgegenstandes näher differenziert werden. Im Sinn der im Abschnitt 1.2 (Seite 6) formulierten Frage nach einer Automatisierung der Mustererkennung sowie einer weiteren Verarbeitung dieser Muster zu Identifikations- und Klassifikationszwecken werden theoretische Grundlagen computerlinguistischer Disziplinen Informationsextraktion und Stimmungsanalyse zum Thema.
Im Kapitel 3 (Stand der Wissenschaft) wird sich mit Methoden automatischer aspektbasierter Stimmungsanalyse sowie mit aktuellen Arbeiten im Bereich der Informationsextraktion und automatischer Identifikation sozialpsychologischer Phänomene befasst. Es werden praktische Lösungen und Realisierungen ähnlicher Problematiken vorgestellt und in Bezug auf eigene Fragestellungen diskutiert. Außerdem werden am Beginn des Kapitels einige experimentelle Arbeiten zu kognitiven Effekten betrachtet, die einem Vergleich der sozialpsychologischen und computerlinguistischen Methoden dienen sowie ausgewählte Ansätze zum Erreichen der in der vorliegenden Arbeit gestellten Teilziele liefern.
Im Kapitel 4 (Methodische Vorgehensweise) erfolgt zunächst eine Auseinandersetzung mit ausgewählten Problematiken der Domäne der Arztbewertungen, um den Umgang mit diesen in der vorliegenden Arbeit zu bestimmen. Auf der Basis dieser Auseinandersetzung und durch die Aufstellung der Kriterienkataloge für die begründete Auswahl von kognitiven Effekten (s. Kapitel 2, Abschnitte 2.1.2 und 2.1.3) wird eine methodische Vorgehensweise bestimmt, die die Vorgehensschritte des eigenen Verfahrens darstellt und visualisiert (Abschnitt 4.3). Eine tiefere theoretische Auseinandersetzung mit den im Abschnitt 1.2 angesprochenen lokalen Grammatiken (Seite 9) sowie mit den Möglichkeiten des gewählten Systems *UNITEX* findet statt (Abschnitt 4.2). Dadurch wird die methodische Relevanz von lokalen Grammatiken für die vorliegende Dissertation deutlich gemacht. Zwar werden die Forschungsfragen nach Identifikation und Klassifikation lediglich theoretisch gelöst, jedoch ist dieses bei der anwendungsorientierten Arbeit nur logisch, da die Antworten darauf erst aus den Ergebnissen des durchgeführten Verfahrens sichtbar werden (s. Kapitel 6).

1.3.2 Umsetzung und Qualität

Der anwendungsorientierte Teil dieser Arbeit (Kapitel 5 und 6) stellt die praktische Umsetzung des im Kapitel 4 ausgearbeiteten Verfahrens und dessen Evaluation dar.
Im Kapitel 5 (Umsetzung des Identifikationsverfahrens CognIEffect) wird das Identifikationssystem CognIEffect schrittweise vorgestellt. Zunächst wird auf die Korpora und die Ressourcen eingegangen, die teils aus dem Trainingskorpus selbst gewonnen werden und teils externe Quellen darstellen (Abschnitt 5.1). Es wird ein strukturiertes Graphensystem vorgestellt, das die Extraktion der Meinungen zu den vordefinierten Dimensionen und ggfs. zu Effekten, die durch sprachliche Äußerungen zu erkennen sind, realisiert (Abschnitt 5.2). Nach der Extraktion wird an der Identifikation der ausgewählten kognitiven Effekte gearbeitet (Abschnitt 5.3). Gelingen die automatischen Klassifikation und Identifikation kognitiver Effekte, so werden die beiden zuletzt aufgestellten Forschungsfragen (s. Seite 6) positiv beantwortet.
Eine Evaluation und Diskussion der Ergebnisse sowohl bei der Pattern-Extraktion als auch bei der Identifikation und Klassifikation kognitiver Effekte wird im Kapitel 6 (Evaluation) durchgeführt. Bei der Beschreibung üblicher Qualitätsmaße werden Kriterien für beide Teilverfahren (Musterextraktion und Effekte-Identifikation) festgelegt (Abschnitte 6.1 und 6.2), Ergebnisse zusammengefasst und erläutert (Abschnitte 6.3, 6.4 und 6.5). Hierbei sind die erzielten Qualität sowie Problematiken des vorgestellten und durchgeführten Verfahrens von Interesse. Aufgrund dieser abschließenden Auseinandersetzung mit dem entwickelten Verfahren, durch die Aussagen zur dessen Qualität werden die aufgestellten Forschungsfragen differentzierter bearbeitet, womit die vorliegende Arbeit abgeschlossen werden kann.

1.3.3 Schlussbetrachtung

Im Kapitel 7 (Fazit und Ausblick) erfolgt eine ausführliche Zusammenfassung der gesamten Dissertation. Es wird reflektiert, ob und wie das gesetzte Ziel erreicht und die aufgestellten Fragen beantwortet wurden. Auf die Aspekte der Nutzbarkeit der durchgeführten Arbeit wird eingegangen, die Perspektiven weiterer Forschung werden beleuchtet.

Kapitel 2

Multidisziplinäre Grundlagen

Die Grundlagen der automatischen Erkennung kognitiver Effekte in Arztbewertungen sind multidisziplinär. Einerseits erfolgt in dieser Arbeit die Adaption der Definitionen sozialpsychologischer Phänomene im Kontext der Online-Bewertungen zu Arztpraxen und andererseits müssen computerliguistische Methoden zur Umsetzung des automatischen Identifikationsverfahrens herangezogen werden.
Zunächst wird der Begriff „kognitive Effekte“ domänenbezogen definiert und die Hintergründe des Auftretens näher erläutert. Danach erfolgt die Klassifikation dieser Phänomene in Arztbewertungen. Dabei wird die Auswahl kognitiver Effekte anhand der wissenschaftlichen Literatur getroffen, die in Bewertungstexten empirisch belegt werden können. Die in Bewertungstexten ermittelten Phänomene, die man nach deren Definition ebenfalls zu kognitiven Effekten zählen kann, werden als empirisch ermittelte Effekte klassifiziert. Als Ergebnis des angedeuteten Vorgehens werden relevante Effekte in einer Tabelle (Seite 39) zusammengefasst und definiert. Mit Selektions- bzw. Identifikationskriterien der in den Arztbewertungen erkennbaren Effekte wird sich befasst, die pro Effekt im Kapitel 4 aufgestellt werden. Am Ende werden computerlinguistische Begriffe und Konzepte dargestellt, die im Sinne einer automatischen Identifikation erkennbarer Effekte relevant sind. Erkenntnisse aus zwei Teildisziplinen der Computerlinguistik sind für die vorliegende Arbeit von Bedeutung: die der domänenspezifischen Informationsextraktion und der aspektbasierten Stimmungsanalyse. Durch konzeptuelle Parallelen wird eine Verbindung zwischen sozialpsychologischer und computerliguistischer Erkenntnisse und die Einordnung der hier aufgestellten Problematik in den wissenschaftlichen Kontext deutlich.

2.1 Kognitive Effekte in Arztbewertungen

2.1.1 Theoretischer Hintergrund

In den nachfolgenden Abschnitten wird sich mit der Bewertungskultur von Patienten befasst, die ihre Erfahrungen beim Arztbesuch auf Online-Plattformen schildern, woraufhin eine allgemeine Charakterisierung und Definition dieser Phänomene in der beschriebenen Domäne erfolgt. Auf einige Hintergründe des Zustandekommens von kognitiven Effekten wird eingegangen, um mögliche Hinweise auf automatisch identifizierbare Kriterien für die beschriebenen Phänomene in Arztbewertungen zu analysieren.

2.1.1.1 Überblick zur Domäne der Arztbewertungen

2.1.1.1.1 Bewertungsportale

Wie in dem Artikel[18] der Welt-Zeitung vom 09.09.2012 berichtet wurde, sind seit einigen Jahren Bewertungsportale für Arztpraxen aktiv im Internet präsent. Dabei zeigt sich eine steigende Tendenz in Bezug auf Wachstum und Intensität des Informationsaustauschs unter den Patienten. „Received medical services are incresingly discussed and recommended on physisian rating websites“ (Geierhos et al., 2015b, S. 1). Diese Bewertungsportale bieten Patienten die Möglichkeit, anonym ihre Meinungen zu bestimmten Dimensionen wie „Behandlung“, „alternative Heilmethoden“, „Freundlichkeit“, „Wartezeit“ usw. in Form von selbst formulierten Texten in einem Freitextfeld[19] zu äußern. Ebenfalls können nummerische Bewertungen getätigt werden.
Einige Bewertungsportale im Gesundheitswesen sind: jameda.de, docinsider.de, sanego.de, arzt.weisse-liste.de, topmedic.de, etc.[20]

[18] `http://www.welt.de/finanzen/verbraucher/article109103878/Wenn-der-Patient-dem-Arzt-eine-Sechs-gibt.html` (20.09.2014).

[19] `http://www.welt.de/wirtschaft/article13554994/Patienten-koennen-Aerzte-mit-Schulnoten-bewerten.html` (20.09.2014).

[20] Weitere Bewertungsportale sind bei Geierhos et al. (2015b, S. 1) aufgeführt.

2.1.1.1.2 Bewertungskomponenten

Die Struktur und Komponenten einer Bewertung wurden im Abschnitt 1.2 der Einleitung dargestellt. Darüber hinaus sind bei mehreren Bewertungen zu einer Arztpraxis weitere wertende Elemente auffindbar:

- Gesamtnote aller Bewertungen zur Praxis, die sich aus dem Mittelwert der von Patienten vergebenen nummerischen Werte zu Bewertungsdimensionen errechnet
- Anzahl der vergebenen Bewertungen
- Prozentangaben zur Weiterempfehlung
- Anzahl der Aufrufe
- usw.

Ob und inwiefern diese Angaben zur automatischen Identifikation kognitiver Effekte verwertbar sind, wird sich im Weiteren herausstellen und im Kapitel 4, Abschnitt 4.3 pro Effekt dokumentiert.

2.1.1.1.3 Bewertungssysteme und -dimensionen im Vergleich

Bewertungsportale im Allgemeinen haben eigene Klassifikationen der Bewertungsdimensionen und verschiedene Systeme zur Vergabe nummerischer Werte. In Tabelle 2.1 sind Bewertungssysteme und -dimensionen von drei Bewertungsportalen (Jameda, DocInsider, sanego) (s. o.) im Vergleich dargestellt. Die nummerischen Bewertungen aller aufgeführten Portale weisen Unterschiede sowohl in den Bezeichnungen der Bewertungsskalen als auch in der Anzahl nummerischer Werte auf. Während man bei Jameda Schulnoten von 1 bis 6 vergeben kann, bewertet man Leistungen bei beiden anderen Portalen mit Sternen oder Punkten in aufsteigender Richtung (von einer schlechten zur besten Leistung, s. Abbildung 2.1; z. Vergleich s. Abbildung 4.1).

	Jameda	**DocInsider**	**sanego**
Skalen	6-Noten-System	5-Sterne-System	10-Punkte-System
Dimensionen	Behandlung	Fachliche Kompetenz	Behandlungserfolg, Arztkompetenz
	Aufklärung	Information & Beratung	Arztberatung
	Vertrauensverhältnis	Vertrauensverhältnis	
	Genommene Zeit		
	Freundlichkeit		Team Freundlichkeit
	Wartezeit (Termin)	Wartezeit (Termin) (in Tagen)	Wartezeit (Termin) (in Tagen / Wochen / Monaten)
	Wartezeit (Praxis)	Wartezeit (Wartezimmer) (in Minuten)	Wartezeit (Wartezimmer) (in Minuten)
	Sprechstundenzeiten		
	Entertainment		
	Betreuung		
	Kinderfreundlichkeit		
	Barrierefreiheit		
	Praxisausstattung		Praxisausstattung
	Telefonische Erreichbarkeit		
	Öffentliche Erreichbarkeit		
	Parkmöglichkeiten		
	Alternative Heilmethoden		
	Weiterempfehlung (ja/nein)	Weiterempfehlung (ja/nein)	Empfehlung (nummerisch, nein-ja)
		Einbindung Entscheidungen	Entscheidungen
		Qualität im Allgemeinen	
		Organisation in der Praxis	
			Terminvereinbarung

Tabelle 2.1: Bewertungssysteme und -dimensionen von Jameda, DocInsider und sanego im Vergleich

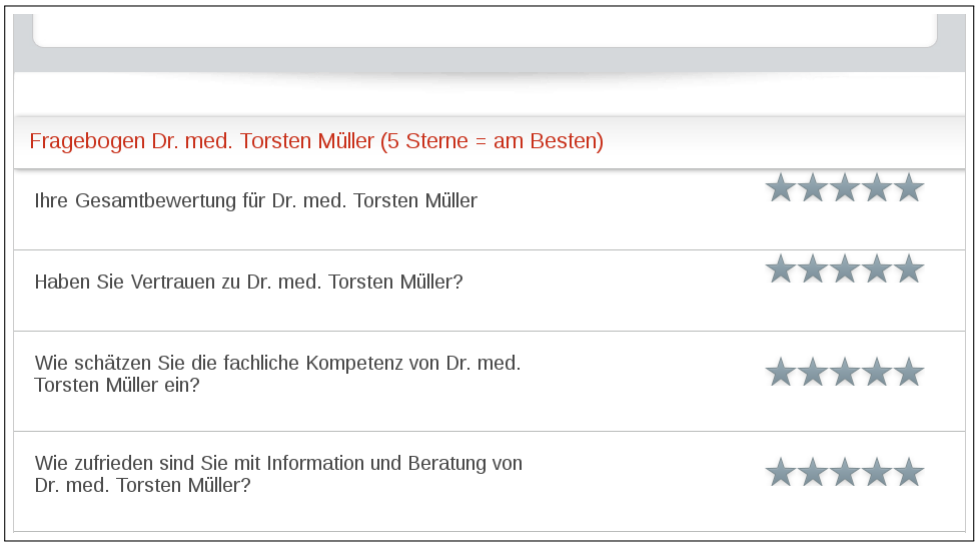

Abbildung 2.1: Bewertungssystem bei DocInsider

Die Einheitlichkeit der nummerischen Bewertungssysteme stellt ein Problem der Skalentransformation für die automatische Textverarbeitung dar (s. Kapitel 4, Abschnitt 4.1.5).
Der Variantenreichtum der Skalen erstreckt sich auch auf die jeweiligen Bewertungsdimensionen. Eine feinkörnige Differenzierung bzw. die Unterteilung ärztlicher Dienstleistungen in verschiedene Dimensionen findet man z. B. bei Jameda. In der Tabelle 2.1 wurde, ausgehend von den Jameda-Dimensionen, eine vergleichende Darstellung mit äquivalenten Dimensionen der beiden anderen Portale realisiert, wobei die nicht definierten Dimensionen entsprechender Portale nicht eingetragen wurden (leere Zellen in der Tabelle). Die Bewertungsdimensionen, die bei allen drei Portalen mit gleichwertigen Bewertungsmöglichkeiten im freien Textfeld[21] vertreten sind, sind „Fachliche Kompetenz“ oder „Behandlung“ sowie „Aufklärung“ oder „Beratung“. Die Wartezeiten auf einen Termin und im Warteraum einer Praxis können jedoch nur bei Jameda mit Noten charakterisiert werden. Bei den anderen genannten Portalen sind nur Angaben in Minuten möglich. Die „Weiterempfehlung“ ist bei allen Portalen vorhanden, kann jedoch nur bei sanego nummerisch nach dem 10-Punkte-System (Skala von ‚nein‘ bis ‚ja‘) charakterisiert werden. Bei den anderen beiden Bewertungsportalen wird sie lediglich in Form

[21] Bei den nummerischen Bewertungen entsprechend den Skalen-Systemen eines jeweiligen Portals.

einer Frage formuliert, auf die mit ‚ja' oder ‚nein' geantwortet werden kann. Bei den Bewertungsportalen DocInsider und sanego ist eine Bewertungsdimension „Entscheidungen" (bzw. „Einbindung Entscheidungen") vorhanden. Diese impliziert die Bewertung der Frage, inwiefern man als Patient in die Entscheidungen des behandelnden Arztes eingebunden wurde[22].

2.1.1.2 Aufstellung der Definition

2.1.1.2.1 Zum Begriff kognitiver Effekte

Kognitive Effekte oder Denkfehler sind in der wissenschaftlichen Literatur unter mehreren Bezeichnungen bekannt. So spricht Schneider (2013, S. 14f.) von Wahrnehmungsdefekten oder Informationspathologien und setzt sich mit den Begriffen „Wahrnehmung", „Pathologien" etc. und deren Synonymen auseinander. Gehrig und Breu (2013, S. 47ff.) sowie Wilkening (2008, S. 1ff.) bezeichnen diese Erscheinungen als Denkfehler, die als „Ausrutscher beim Denken, wie sie von Zeit zu Zeit jedem Menschen unterlaufen" (Wilkening, 2008, S. 1) definiert werden. Treffend formuliert das Phänomen Grams (2006, S. 1) als „Reinfälle", „Denkfallen", „die hinter unseren alltäglichen Irrtümern stecken" (ebd.). Ein Beispiel für solche Irrtümer wäre die Befragung der Schüler zur Selbsteinschätzung ihrer Leistung nach einer Prüfung. Dabei zeigte sich, dass sich die Schüler mit den schlechtesten Ergebnissen deutlich überschätzten, während diejenigen mit besten Ergebnissen ihre Leistungen leicht unterschätzten[23]. Ein anderes Beispiel für alltägliche Denkfehler wäre der Gedanke, dass Senioren schuld daran sein sollen, dass man in öffentlichen Verkehrsmitteln nie einen Sitzplatz findet. „Objektiv gesehen schnappen Ihnen Senioren nicht öfter den Sitzplatz im Bus weg als jüngere Menschen. Aber all den Fällen, welche Ihre Annahme nicht stützen (wenn Ihnen Nicht-Senioren den Platz wegschnappen), geben Sie weniger Gewicht oder ignorieren sie gänzlich bei der Beurteilung Ihrer Annahme"[24].
Ob kognitive Täuschungen (vgl. Schweizer, 2005), Verzerrungen[25], bias[26] oder

[22] Erläuterung des sanego-Portals: „Wurden [S]ie ausreichend in die Entscheidungen einbezogen?"

[23] `http://www.skeptiker.ch/themen/kognitive-verzerrungen/` (04.03.2016).

[24] ebd.

[25] ebd.

[26] `http://psychology.about.com/od/cindex/fl/What-Is-a-Cognitive-Bias.htm` (04.03.2016).

Denkfehler[27] etc., es geht in diesem Zusammenhang darum, eine Definition kognitiver Effekte im Kontext der vorliegenden Arbeit zu finden. Warum im Kontext und nicht eine allgemein gültige Definition? Weil es keine solche gibt! „It does not appear possible today to group all of the phenomena that have been qualified as cognitive biases under one and the same definition“ (Caverni et al., 1990, S. 7f.). Wenn man auf die oben angeführten Beispiele den Fokus richtet, so stellt sich in Bezug auf die Definierbarkeit kognitiver Effekte die Frage, was die Leistungseinschätzung der Schüler und die von Senioren besetzten Plätze im Bus miteinander zu tun haben und was diese Situationen verbindet. Der gemeinsame Nenner ist, dass Menschen in beiden Situationen Meinungen und Einschätzungen zu bestimmten Sachverhalten formulieren und dass diese Meinungen auf irgendeine Art falsch sind. Durch eine fehlerhafte Information in ihren Entscheidungs-, Beurteilungs- oder Bewertungsprozessen treten Wahrnehmungsdefekte auf (Schneider, 2013, S. 14). „A cognitive bias is a systematic error in thinking that affects the decisions and judgments that people make. [...] A cognitive bias is a type of error in thinking that occurs when people are processing and interpreting information in the world around them“[28]. Fehler oder Defekte beim Wahrnehmen, Denken, Interpretieren usw. von Informationen, die jedoch immer noch in völlig verschiedenen Kontexten auftreten (s. oben aufgeführte Beispiele). Da jeder Mensch Denkfehler fast zwangsläufig begeht (Grams, 2006, S. 2), sind auch Patienten, die Meinungen zu ärztlichen Dienstleistungen formulieren, nicht davor geschützt. Das bedeutet, dass von dem Vorhandensein dieser Fehler in Arztbewertungen ausgegangen werden muss, da sie nicht auf einem reinen Zufall basieren. „As such, a bias is detected when derivation from norm is observed. [...] Occasional and accidental errors are obviously not part of the issue of cognitive biases.“ (Caverni et al., 1990, S. 7f.). Kognitive Effekte erfolgen also unbewusst als Ergebnis falsch abgelaufener Informationsverarbeitungsprozesse (Wilkening, 2008, S. 1; Schneider, 2013, S. 14) und manifestieren sich in Meinungen und somit in verfassten Bewertungstexten. „So wie ein Virus Ihren Computer an der korrekten Verarbeitung von Informationen hindert, halten Denkfehler Sie davon ab, Erfahrungen richtig zu bewerten“(Wilkening, 2008, S. 1). Um dem im Abschnitt 1.2 gestellten Ziel gerecht zu werden, sollte man überlegen, auf welche Weise sich solche Effekte

[27] `http://nutzerverhalten-online.de/denkfehler` (01.01.2014).

[28] `http://psychology.about.com/od/cindex/fl/What-Is-a-Cognitive-Bias.htm` (04.03.2016).

in niedergeschriebenen Texten sichtbar machen.

2.1.1.2.2 Kriterien fehlerfreier Bewertungen

Die Auseinandersetzung mit kognitiven Prozessen, Theorien, wie das Wissen organisiert, erworben, abgerufen etc. wird, aber auch was bei diesen Prozessen ‚falsch' ablaufen könnte, finden ihre Bestätigungen und Interpretationen in zahlreichen Studien mit Probanden[29]. Der falsche Ablauf dieser Prozesse impliziert, dass es eine Definition dafür geben sollte, was in diesem oder jenem Kontext als ‚richtig' zu betrachten wäre (Caverni et al., 1990, S. 8). Genauer gesagt, sollte eine erwartbare Norm, wann ein Prozess als ‚korrekt abgelaufen' zu deuten ist, definiert werden. Für die Domäne der Arztbewertungen, konkret für Bewertungstexte muss man eine Norm festlegen, um den Aspekt einer normativen Bewertung bei der domänenspezifischen Definition kognitiver Effekte zu berücksichtigen. Allgemein können zwei Kriterien fehlerfreier Bewertungen festgelegt werden:

- Übereinstimmung wertender Aussagen und nummerischer Werte:
  Bei einer in einem freien Textfeld formulierten ‚korrekten' Bewertung wäre logisch, dass sich Positivität bzw. Negativität (Polarität[30]) wertender Aussagen zu konkreten Bewertungsdimensionen in den entsprechenden nummerischen Werten zu denselben Dimensionen widerspiegeln (Geierhos und Stuß, 2015, S. 238f.). „It is generally assumed that ratings are a numeric representation of text sentimens and their valences are consistent" (Hu et al., 2013, S. 2)

- Mehrheit der Bewertungen:
  Logisch für eine Norm wäre ebenfalls, dass man von einer Mehrheit der Bewertungen spricht, die diese Norm erfüllen. Das bedeutet, dass es sich bei kognitiven Effekten um eine kleinere Anzahl der Bewertungen handelt, die fehlerbehaftet sind. Auf dem Gebiet der robusten Statistik werden Verfahren des Umgangs mit Ausreißern entwickelt. Als Ausreißer werden dabei die Werte verstanden, durch die statistische Analysen ‚gestört' werden und die stark von der Masse der Daten

[29] Auf die Studien ausgewählter kognitiver Effekte wird im Kapitel 3, Abschnitt 3.1.1.1 eingegangen.

[30] Auf die Polarität wird im Laufe der Arbeit in mehreren Abschnitten eingegangen (s. Abschnitte 2.1.1.3.1, Seite 23; 2.2.2.2.2, Seite 47).

> abweichen (Buttler, 1996, S. 2). Genauso wie kognitive Effekte können einige Ausreißer zu Fehlern gezählt werden, die bei den Analysen empirischer Daten entweder korrigiert oder eliminiert werden (ebd., S. 4f.).

Wie stellen sich Abweichungen von der eben festgelegten Norm einer Bewertung dar? Wie kann man diese automatisch identifizieren? Logisch ist, dass eine automatische Textverarbeitung anhand der Möglichkeiten erfolgen kann, die der Text und evtl. weitere Metadaten bieten. Auf den Arztbewertungsportalen sind es wertende Elemente der im Kapitel 1, Abschnitt 1.2 vorgestellten Komponenten einer Bewertung.

2.1.1.2.3 Allgemeine Definition kognitiver Effekte

Aus den bisherigen Ausführungen lässt sich die folgende auf die Domäne der Arztpraxen bezogene Definition kognitiver Effekte aufstellen:

Als **kognitive Effekte im Kontext der Online-Bewertungen von Arztpraxen** werden die Bewertungsfehler – was als Abweichung von der Erwartungsnorm zu interpretieren ist – verstanden, die unbewusst, aber nicht zufällig von den Autoren der Bewertungen begangen werden. Diese Bewertungsfehler führen zu einer Art Verzerrung, die sich aufgrund des Bewertungsverhaltens anhand wertender Elemente entsprechend interpretieren und klassifizieren lässt. Zu jedem kognitiven Effekt müssen Gegenbeispiele auffindbar sein, die, ihrer Menge entsprechend, die Norm einer fehlerfreien Bewertung vertreten.

Die aufgestellte Definition ist breit gefasst, was bedeutet, dass weitere Konkretisierungen nötig sind. Diese Konkretisierungen betreffen die genannten wertenden Elemente, die zur Aufstellung der Kriterienkataloge pro identifizierbarer Effekt im Kapitel 4 (Abschnitt 4.3) verwendet werden. Das heißt: Jeder Effekt bedarf einer eigenen Definition, die anhand der aufzustellenden Kriterien, die zur automatischen Identifikation kognitiver Effekte nötig sind, konkretisiert und präzisiert wird.

2.1.1.3 Wie kommen kognitive Effekte zustande

2.1.1.3.1 Einstellungen

Die Sozialpsychologie beschäftigt sich mit der Erforschung der Einstellungen[31] von Individuen (Stürmer, 2009, S. 69). Die Grundannahme der Auseinandersetzungen mit diesen besteht darin, dass von ihnen menschliche Handlungen und Entscheidungen bezüglich des eigenen Verhaltens geleitet werden (ebd.; Schöberl, 2012, S. 36; Zick, 2004, S. 129). „Die Einstellung einer Person zu einem Objekt ist die subjektive Bewertung dieses Objekts“ (Stürmer, 2009, S. 70). Unter einem Einstellungsobjekt kann man alles verstehen, woran man denken und was man unterscheiden kann (Zick, 2004, S. 130). Es können nichtsoziale oder soziale Stimuli, konkrete, abstrakte oder leblose Dinge sein (Stürmer, 2009, S. 70; Zick, 2004, S. 130). Produkte, Personen, Rauchen, Pizza, Sportautos, politisches Engagement, Redefreiheit, Flaggen, Embleme etc. können zu den Einstellungsobjekten gezählt werden (ebd.; ebd.). In Bezug auf die Domäne der Arztbewertungen bilden die z. B. im Kapitel 1, Seite 8 oder im Abschnitt 2.1.1.1.3 anhand der Tabelle 2.1 erläuterten Bewertungsdimensionen die Einstellungsobjekte.
Ihrer Struktur nach, weisen Meinungen eine kognitive, eine affektive und eine konative Komponente[32] auf (ebd.; ebd., S. 131):

- *kognitiv*: Überzeugungen einer Person zu einem Einstellungsobjekt (z. B. Kenntnis seiner positiven, negativen Seiten) (Stürmer, 2009, S. 71).

- *affektiv*: „Gefühle oder Emotionen [...], die eine Person mit einem Einstellungsobjekt assoziiert“ (ebd.).

[31] Die Einstellungen werden in weiteren Ausführungen der vorliegenden Arbeit als Meinungen bezeichnet, um die Einheitlichkeit der Begriffsverwendung im computerlinguistischen Kontext zu gewährleisten. Zwischen beiden Begriffen und deren Charakteristika können eindeutige Parallelen gezogen werden, so dass diese von vorne herein als äquivalent zueinander betrachtet werden können.

[32] Die konative Komponente „[...] bezieht sich auf Informationen bzgl. des Einstellungsobjekts, die aus dem eigenen Verhalten im Umgang mit diesem Objekt abgeleitet werden“ (Stürmer, 2009, S. 73). Sie lässt sich schwer bzw. gar nicht in den Bewertungstexten erkennen. Es ist schwer vorstellbar, dass man eigenes Verhalten zu Bewertungsdimensionen schriftlich reflektieren und dieses in einer Bewertung formulieren würde. Aus diesem Grund werden hier die Ausführungen dazu ausgelassen.

Es ist umstritten, ob alle Komponenten eine Meinung gleichzeitig beeinflussen müssen (Schöberl, 2012, S. 37; Zick, 2004, S. 131). Eine der zwei Dimensionen, anhand derer man Meinungen charakterisieren kann, ist die Valenz, die sich in zwei Ausprägungen sichtbar macht: positiv und negativ (Stürmer, 2009, S. 70); eine andere Dimension ist Stärke, die man daran beobachten kann, „wie schnell ein Einstellungsobjekt eine wertende Reaktion auslöst“ (ebd.)[33].
Die kognitiven und affektiven Komponenten sowie Polaritätsausprägungen lassen sich anhand wertender Äußerungen der Patienten in Bewertungstexten wiederfinden. Was die Intensität einer Meinung betrifft, kann man diese anhand bestimmter Signalwörter identifizieren. „Sehr gut“ ist z. B. stärker bzw. intensiver als „gut“. In der vorliegenden Arbeit wird die feinkörnige Analyse der Intensität von Äußerungen nicht durchgeführt. Die o. g. Reaktionszeit kann man lediglich aus Spontanreaktionen mündlicher Äußerungen erschließen, was für die schriftlichen Texte irrelevant ist.

2.1.1.3.2 Wahrnehmung und kognitive Informationsverarbeitung

Der Begriff Wahrnehmung wird teilweise synonym zur Kognition verwendet und bedeutet aktiver, selektiver und subjektiver Verarbeitungsprozess, der zu einer Wissensrekonstruktion führt (Schneider, 2013, S. 7):

- *Aktiv*, weil kognitive Systeme nicht nur die Informationen rezeptiv aufnehmen, sondern diese auch verarbeiten, woraus bestimmte Rekonstruktionen entstehen
- *Selektiv*, weil kognitive Systeme gezielt bestimmte Informationen auswählen (selektieren), die verarbeitet werden
- *Subjektiv*, weil unterschiedliche kognitive Systeme Informationen individuell verschieden verarbeiten können, was wiederum zu unterschiedlichen Rekonstruktionen oder Ergebnissen führen kann (ebd.).

Aktive Verarbeitungsprozesse laufen bei jedem Patienten bei einem Arztbesuch individuell und unterschiedlich ab (Subjektivität). Je nachdem, mit

[33] Valenz und Stärke werden in weiteren Ausführungen der vorliegenden Arbeit als Polarität und Intensität bezeichnet, um die Einheitlichkeit der Begriffsverwendung im computerlinguistischen Kontext zu gewährleisten.

welchen Zielen und Erwartungen ein Patient zum Arzt kommt, wie er seine Prioritäten setzt, werden ebenfalls Prioritäten bei der Informationsverarbeitung gesetzt (Selektion). Um die Aktivität der Informationsverarbeitung nachzuvollziehen, muss sich mit einigen kognitiven Prozessen auseinandergesetzt werden.

Die kognitive Psychologie betrachtet den Menschen als ein informationsverarbeitendes System und beschäftigt sich mit kognitiven Komponenten, deren Speicherung, Abrufbarkeit, Reorganisation und Zuwachs etc. (Wolff, 1993, S. 29; Zick, 2004, S: 131). Die aufgezählten Prozesse werden kognitiv genannt (Kühne, 2013, S. 7).
Was passiert im Einzelnen z. B. bei der Speicherung und Abrufbarkeit? Eine der Theorien der Wissensspeicherung ist die Skripttheorie. Sie besagt, dass die gespeicherten Informationen stereotype Handlungssequenzen enthalten (Prestin, 2003, S. 496). Auf den Arztbesuch bezogen, könnte man folgende abgespeicherte Handlungssequenzen aufführen: sich an der Rezeption anmelden, ggf. im Warteraum Platz nehmen und Zeitschriften lesen, zum Arzt aufgerufen werden, in den Behandlungsraum eintreten u. ä.[34]. Die grundlegende Theorie der Gedächtnisforschung nimmt die Existenz eines Kurzzeitgedächtnisses (KZG) an, in dem sensorisch aufgenommene Informationen behalten werden, aber schnell verloren gehen, wenn diesen keine weitere Aufmerksamkeit geschenkt wird. Damit Informationen in ein relativ andauerndes Langzeitgedächtnis (LZG) gelangen können, müssen diese wiederholt bzw. memoriert werden (Anderson, 2001, S. 175f.). Die aufgezählten Handlungssequenzen sind für einen Patienten aus seinem LZG leicht abrufbar. Diese Handlungssequenzen sind nichts anderes als Erfahrungen des Patienten, der sie in die Bildung seiner Meinungen einbezieht. Sensorisch wahrgenommene (KZG) oder in Vergessenheit geratene Informationen können bei der Meinungsbildung aus Mangel an Informationen zu falschen Ergebnissen führen (Schneider, 2013, S. 14; s. Abschnitt 2.1.1.2.1, Seite 18). Wenn einem Menschen z. B. ein Vorgang vertraut ist, so kann dieser das entsprechende Wissen zum „Arztbesuch-Skript“ aktivieren und abrufen. In einem anderen Fall ist die besagte Situation noch unbekannt, so dass sich die Meinung zu den Abläufen ebenfalls neu bildet und – aufgrund der individuellen Verar-

[34]Neben den Skripttheorien existieren zahlreiche andere Theorien wie z. B. Schema- und Szenariotheorien oder mentale Modelle (Prestin, 2003, S. 496), die die Wissensorganisation und -speicherung tiefer differenzieren. Im Rahmen der vorliegenden Arbeit wird auf diese nicht weiter eingegangen.

beitung (s. o. ‚subjektiver Verarbeitungsprozess') – anders ausfallen kann als bei Menschen, denen eine Situation vertraut ist.
Kognitive Wissensstrukturen werden in deklaratives und prozedurales Wissen eingeteilt. Unter deklarativem Wissen wird „[...] das Wissen über Fakten, Zustände und Geschehnisse der wirklichen oder erdachten Welt, das sich der Informationsverarbeiter [Mensch] im Verlauf seines Lebens in der Auseinandersetzung mit der Umwelt erworben hat [...]" (Wolff, 1993, S. 32), verstanden. Wenn man deklaratives Wissen als ‚Was-Wissen' bezeichnet, ist prozedurales Wissen dagegen als ‚Wie-Wissen' zu betrachten (ebd.). Das heißt, dass dies das Wissen über die Prozeduren ist, die der Mensch durchführt, um die eingegangenen Informationen zu verarbeiten. Prozedurales Wissen steuert den Erwerb und den Einsatz des deklarativen Wissens.
In Anlehnung an ein Sprachverarbeitungsmodell, das im Kontext der Arbeit mit Filmen im DaF[35]-Unterricht von Biechele (2006, S. 320) entwickelt wurde, könnte man die Informationsverarbeitung im Kontext der wahrgenommenen Situationen bei einem Arztbesuch wie auf der Abbildung 2.2 darstellen. Zum deklarativen Wissen gehören generelles Weltwissen, narratives und formales Wissen zu Geschehnissen und Vorgängen in einer Arztpraxis (ebd., S. 315f.). Mit Hilfe des generellen Wissens eines Patienten muss dieser aus einer konkret entstandenen Situation Schlussfolgerungen ziehen, die sich dann auf noch nicht Geschehene beziehen. Was bedeutet, dass er antizipieren (Vorhersagen treffen, Interpretationen und Wertungen abgeben), inferieren (Schlussfolgerungen oder Konklusionen ziehen, s.o.) etc. (ebd., S. 321) muss (ebd., S. 316). Das narrative Wissen umfasst typische Abläufe, Rollen (Arzt, Empfangsdame), Handlungssituationen u.ä. (ebd.). Formales Wissen impliziert ein fachspezifisches Wissen über z. B. bestimmte Untersuchungen, Prozedur der Blutabnahme oder Zusammenhänge bei Impfungen bzw. Einnahmen von Medikamenten. Dieser Wissensbestand schafft die Grundlage für das Verstehen und das tiefe Verarbeiten (elaborieren) arztpraxisbezogener Narration sowie für die Wertung der entstandenen Situation (ebd.). Die eingehenden Informationen werden wahrgenommen, mit dem vorhandenen Wissen in Einklang gebracht und in entsprechende kognitive Strukturen integriert (Wolff, 2002, S. 33) (top-down). Zur Verarbeitung unbekannter Vorgänge werden aufgezählte prozedurale Prozesse eingeschaltet (bottom-up) und die existierenden kognitiven Strukturen entsprechend verändert (Schneider, 2013, S. 7). Die neu gebildeten bzw. ergänzten kognitiven Strukturen

[35]Deutsch als Fremdsprache.

beeinflussen wiederum die Meinungen und weitere Handlungen der Individuen (Kühne, 2013, S. 3). „Die Grundannahme ist, dass Individuen für die Urteilsbildung diejenigen Informationen verwenden, welche aus dem Langzeitgedächtnis abgerufen werden können“ (ebd., S. 7).

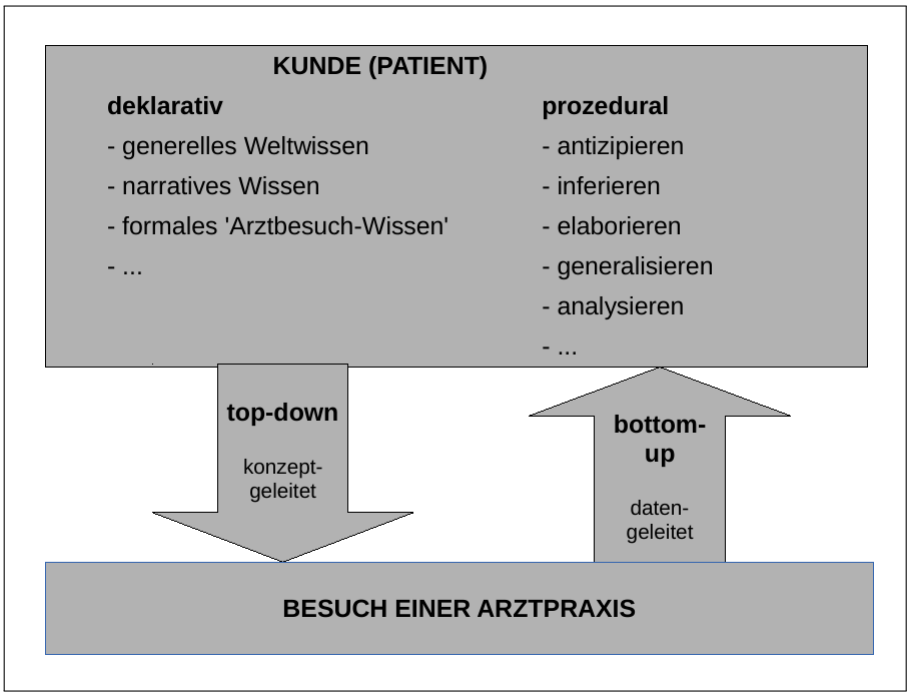

Abbildung 2.2: Informationsverarbeitung beim Besuch einer Arztpraxis (vgl. Biechele, 2006, S. 320)

2.1.1.3.3 Interpretationsspielräume in der natürlichen Sprache

Die Kontexte, in denen kognitive Effekte auftreten, haben etwas mit Meinungsbildung zu tun. Bedeutsam ist ihr Auftreten in Arztbewertungen. Es ist denkbar, dass die Verzerrungen eines Bewertungsbildes zu einem Arzt / einer Arztpraxis, die u. a. durch kognitive Effekte verursacht wurden, zu einem Reputationsverlust oder Rufmord, Abwerbung der Patienten etc. zur Folge haben können. Immer häufiger treten rechtliche Aspekte von Inter-

netbewertungen[36,37] und das Recht auf freie Meinungsäußerung sowie die Überschreitung der Grenzen zur Schmähkritik in den Fokus der Öffentlichkeit[38]. In den Beispielen mit der Leistungsschätzung der Schüler und den von Senioren besetzten Sitzplätzen im Abschnitt 2.1.1.2 (Seite 18) fällt auf, dass die resultierenden Fehler die Ergebnisse von im vorigen Abschnitt beschriebenen Prozessen des Denkens und Urteilens darstellen bzw. aus den Zugriffen auf bestimmte Heuristiken resultieren. So geht es bei Arztbewertungen um nichts anderes als diese Prozesse und deren Ergebnisse, die bei den Bewertung schreibenden Patienten ablaufen. Fest steht: auch hier treten die Denkfehler auf, die man anhand der niedergeschriebenen wertenden Texte und den vergebenen Noten erkennen könnte. Die Anonymität der Bewertenden macht es unmöglich, sich mit konkreten Personen und ihren Denkweisen anhand der verfassten Texte auseinanderzusetzen, wie man dies in speziellen Studien zu konkreten kognitiven Effekten durch Simulationen entsprechender Situationen untersucht (s. Kapitel 3, Abschnitt 3.1.1.1). Durch gewinnende Popularität der Bewertungsportale und die permanent wachsende Anzahl an Arztbewertungen (Geierhos et al., 2015b, S. 1) wird daher die Notwendigkeit der Ausarbeitung eines automatischen Identifikationsverfahrens deutlich. Es sollte dabei u. a. nach bestimmten sprachlichen Ausdrücken gesucht werden, die auf diesen oder jenen Effekt hindeuten würden. Im erwähnten Beispiel der Senioren könnten z. B. solche expliziten Äußerungen wie „Schon wieder diese Senioren! Man kann sich wegen ihnen nirgendwo hinsetzen!“ als Hinweise auf den beschriebenen Denkfehler interpretieren.
Welche kognitiven Effekte in Arztbewertungen auffindbar sind und wie man diese anhand sprachlicher Muster bzw. Musterkombinationen, die als graphisch erkennbare Ergebnisse stattgefundener Informationsverarbeitungsprozesse (s. o.) fungieren, interpretieren und klassifizieren kann, ist eine der wichtigsten Fragen der vorliegenden Arbeit. Um dieser Frage näher zu kommen, werden Effekte betrachtet und ihre Auffindbarkeit in der gewählten Domäne diskutiert.

[36] http://www.anwalt.de/rechtstipps/wenn-ihre-meinung-gefragt-ist-bewertungen-im-internet-und-die-rechtlichen-aspekte_005556.html (04.03.2016).

[37] http://www.zwp-online.info/de/zwpnews/wirtschaft-und-recht/marketing/stressfreies-marketing-mit-arztbewertungen-bei-jameda-und-co (04.03.2016).

[38] http://www.gesundheitsstadt-berlin.de/patienten-duerfen-negative-arztbewertungen-abgeben-3840/ (04.03.2016).

2.1.2 Klassifikation von Effekten

Es gibt eine Reihe kognitiver Effekte, z. B. „Halo-Effekt“, „Framing-Effekt“, „Bestätigungsfehler“. Bezeichnungen dafür sind in der Literatur unterschiedlich. Außer wissenschaftlichen Studien existieren zahlreiche Publikationen psychologischer Natur, aber auch Ratgeber, die Erklärungsversuche zur beschriebenen Problematik unternehmen oder Tipps und Hilfen zum Umgang mit sogenannten Denkfehlern anbieten[39,40,41,42]. Alle diese Publikationen haben eine Gemeinsamkeit: Es geht immer um ‚typische‘ Denkfehler oder um eine bestimmte Anzahl dieser. Nie um alle! Man spricht von ausgewählten, rubrikartigen Überblicken. Es gibt keine einheitliche Klassifikation oder Terminologie. Offensichtlich ist die Tatsache, dass es unmöglich zu sein scheint, einen Überblick aller Denkfehler zu kreieren. Einen tabellarischen Überblick ausgewählter kognitiver Effekte, deren Klassifikation und Erläuterung bietet u. a. Schneider (2013, S. 15ff.).
Um sich vorstellen zu können, was konkrete kognitive Effekte sind, wie diese automatisch identifiziert und klassifiziert werden könnten, werden in diesem Abschnitt einige Effekte vorgestellt. Aus der vorgestellten Menge werden diese auf ihre Identifizierbarkeit in der gegebenen Domäne (laut Selektionskriterien, s. u.) überprüft, analysiert und diskutiert. Außerdem werden domänenspezifische Denkfehler empirisch ermittelt. In Folge werden identifizierbare und somit für diese Arbeit relevante Effekte zusammengefasst und domänenspezifisch definiert.

2.1.2.1 Selektionskriterien

Eine der Kernfragen der vorliegenden Arbeit ist die automatische Auffindbarkeit kognitiver Effekte in gegebenen Bewertungstexten (s. Seite 6). Wie festgestellt, entstehen kognitive Effekte bei der Meinungs- bzw. Urteilsbildung. Für ihre maschinelle Identifikation in Arztbewertungen benötigt man bestimmte graphische Muster, aus denen hervorgeht, dass dieser oder jener Effekt stattgefunden hat. Die graphischen Muster, die auf den Arztbewer-

[39] `http://www.faz.net/aktuell/finanzen/meine-finanzen/denkfehler-die-uns-geld-kosten/denkfehler-die-uns-geld-kosten-60-trinkgeld-und-andere-fehler-12148100.html` (31.03.2016).

[40] `http://www.skeptiker.ch/themen/kognitive-verzerrungen/` (04.03.2016).

[41] `http://www.psychotherapie-davos.ch/Kontakt/Service/Download_Materialien/Checkliste_kognitiver_Verzerrungen.pdf` (31.03.2016).

[42] `https://www.youtube.com/watch?v=woug36Y4_y8` (04.03.2016).

tungsportalen zu finden sind und die die o. g. Meinungsbildung betreffen, sind frei formulierte Texte und vergebene nummerische Werte (z. B. Noten) zu den von jedem Bewertungsportal vordefinierten Bewertungsdimensionen (u. a. wertende Elemente, s. Abbildung 1.1). Das Zusammenspiel dieser wertenden Elemente, der Vergleich deren Polarität mit den Polaritäten von anderen Bewertungen, die Möglichkeit einer Suche nach Ausreißern etc. sollten einige Selektionskriterien darstellen, deren Kombination auf einen bestimmten kognitiven Effekt hindeutet. Für die getroffene Auswahl an kognitiven Effekten wird überprüft, ob einige diese aufgestellten Kriterien erfüllen und somit in den Arztbewertungen identifizierbar sind. Die Effekte, die dieses nicht tun, werden verworfen, da sie im Sinne einer automatischen Identifikation irrelevant wären. Folgende drei Selektionskriterien können auf das Vorhandensein eines kognitiven Effekts hindeuten, wovon mindestens zwei gleichzeitig pro eine Meinung identifizierbar sein sollten, um einen Effekt als bestätigt zu interpretieren:

(a) Ausdrucksmöglichkeiten in der natürlichen Sprache

(b) Vergleich und Verifizierbarkeit textueller und nummerischer Bewertungen

(c) Ausreißerbestimmung

2.1.2.1.1 Ausdrucksmöglichkeiten in der natürlichen Sprache

Um eine ‚falsche Meinung' in einem Text automatisch identifizieren zu können, muss diese Meinung in der natürlichen Sprache formuliert und niedergeschrieben werden. In Arztbewertungen können diese Meinungen zu den vordefinierten Bewertungsdimensionen wie z. B. „die Wartezeit war zu lang" oder zu anderen Bewertungsobjekten („der Arzt ist zu alt") formuliert werden. Im Sinne des automatischen Identifikationsverfahrens müssen aus den Meinungsäußerungen die Bewertungsgegenstände (hier: Dimensionen) und die Polarität der Meinung herausgefunden werden. Diese zwei Indikatoren werden benötigt, um die Entscheidung zu treffen, ob die ausgedrückte Meinung als Ergebnis falsch abgelaufener kognitiver Prozesse zu werten wäre (kognitiver Effekt) oder ob diese zur überwiegenden Mehrheit, zur üblichen Meinung, die diese oder jene Praxis betrifft, gehört (kein kognitiver Effekt).

2.1.2.1.2 Textuelle und nummerische Bewertungen

Wie bereits gezeigt, bieten die Strukturen der Bewertungsportale die Möglichkeit, die Polaritäten der in freien Textfeldern ausgedrückten Meinungen mit den gleichzeitig vergebenen nummerischen Werten zu vergleichen. Dies gilt jedoch nur, wenn eine Meinung formuliert und eine nummerische Bewertung zu einem und demselben Gegenstand getätigt wurde. Die Gegenstände, die man nummerisch bewerten kann, sind die vordefinierten Dimensionen des jeweiligen Portals. Das bedeutet, dass die kognitiven Effekte, bei denen die formulierte Meinung nicht die Bewertungsdimensionen betrifft, dieses Kriterium nicht erfüllen und somit auch nicht verifiziert werden können. Die angedeutete Verifikation kann entweder die Polarität einer ausgedrückten Meinung bestätigen (z. B. positive Meinung – positive Note) oder diese widerlegen (z. B. positive Meinung – negative Note), was je nach Effekt entsprechend interpretiert werden kann.

2.1.2.1.3 Ausreißer

Im Abschnitt 2.1.1.2.2 (Seite 20) wurden Ausreißer als Werte definiert, die stark von der Masse der Daten abweichen. In Bewertungstexten können Ausreißer anhand vergebener nummerischer Werte zu Dimensionen identifiziert werden. ‚Die Masse der Daten' ist als objektive Norm zu interpretieren, die durch diese nummerischen Werte, von der Mehrheit der Bewertenden vergeben, charakterisiert werden kann. Anhand der Anzahl von festgestellten Ausreißern bzw. anhand der Bewertungsdimensionen, die als Ausreißer identifiziert wurden, kann durch die Auffindbarkeit anderer Indikatoren (z. B. Polarität der textuellen Bewertung zur gleichen Dimension) ein konkreter kognitiver Effekt festgestellt werden.

2.1.2.2 Überblick und domänenbezogene Diskussion

Im Abschnitt 1.1 der Arbeit wurde der Begriff der Heuristiken erläutert. Heuristiken benutzen Menschen oft, um schnell Urteile zu fällen, was zu kognitiven Effekten führen kann. Bei dem weiter folgenden Überblick kognitiver Effekte wird ein Versuch unternommen, jedem eine zugrunde liegende Heuristik zuzuordnen.

2.1.2.2.1 Halo-Effekt

Der Halo-Effekt impliziert eine Tendenz, unabhängige oder bedingt abhängige Eigenschaften von Personen oder Sachen fehlerhaft als zusammenhängend zu interpretieren. „The halo effect fallacy is based on the „halo effect“, a psychological tendency many people have in judging others based on one trait that they approve of and concluding that the person must have other attractive traits“ (Grcic, 2008, S. 1). „Für das Produkt und dessen Wahrnehmung bedeutet dieser Effekt, das man bei Produkten, die man mag und schätzt, auch alle Eigenschaften dieses Produktes für gut hält. Grundsätzlich positiv unterstellte Eigenschaften überstrahlen also andere Eigenschaften“[43].
Annahme: Als Produkt wäre in der hier behandelten Domäne die Dienstleistung in einer Arztpraxis zu verstehen. Die Eigenschaften eines Produktes ist die Unterteilung dieser Dienstleistung in die Teilleistungen, also Bewertungsdimensionen. Z. B. ein Patient, der mit der „Behandlung“ eines Arztes zufrieden ist, könnte automatisch eine positive Note und oder Bewertung z. B. für die Dimension „Parkmöglichkeiten“ abgeben, weil die Parkmöglichkeiten diesem Patienten möglicherweise unwichtig sind o. ä., obwohl die tatsächliche Parksituation gerade bei dieser Praxis eher schlecht zu bewerten wäre. Die Urteilsheuristik, durch die dieser Effekt zustande kommen könnte, ist ‚Anker und Anpassung‘. „In many situations, people make estimates by starting from an initial value that is adjusted to yield the final answer. The initial value, or starting point, may be suggested by the formulation of the problem, or it may be the result of a partial computation“ (Tversky und Kahneman, 1974, S. 1128). Als ‚starting point‘ oder ‚Ankerdimension‘ ist in dem genannten Beispiel die „Behandlung“, und der eigentliche Fehler findet durch die falsche Bewertung der Dimension „Parkmöglichkeiten“, die als ‚Anpassungsdimension‘ zu interpretieren ist. Auf die Selektionskriterien bezogen, kann dieser Effekt in Arztbewertungen alle drei Punkte erfüllen[44].

2.1.2.2.2 Framing-Effekt

„Der durch den Framing-Effekt angesprochene Defekt ist, dass Personen unterschiedlich repräsentierte, aber inhaltsgleiche Informationen unterschiedlich

[43] `http://marketing_lexikon.deacademic.com/59/Halo-Effekt` (29.09.2014).

[44] Die im Abschnitt 2.1.2.1 aufgestellten Selektionskriterien werden ausführlicher im Kapitel 4, Abschnitt 4.3 in Bezug auf jeden identifizierbaren Effekt diskutiert und im Kapitel 5, Abschnitt 5.3.2 praktisch realisiert.

bewerten“ (Schneider, 2013, S. 16).
Annahme: Unterschiedlich repräsentierte Informationen können sich in Bezug auf Arztbewertungen als konkrete Situationen, die bei einem Arztbesuch erfolgen, vorgestellt werden. Es können z. B. zwei verschiedene sprachliche Formulierungen eines Arztes zu einer Diagnose sein, durch die Bewertungen der betroffenen Patienten zu diesem Arzt unterschiedlich beeinflusst werden. Allerdings müsste man solche Situationen simulieren (s. Kapitel 3, Abschnitt 3.1.1.1, Seite 54f.), um diesen Effekt in Arztbewertungen nachzuweisen[45].

2.1.2.2.3 Bestätigungsfehler

Wenn beim Urteilen über Sachverhalte nach einer Bestätigung eigener Überzeugungen gesucht wird, können sogenannte Bestätigungsfehler, auch ‚Confirmation Bias‘ genannt, auftreten. „People hold onto their beliefs strongly. Changing beliefs takes time and effort, and it is often easier to disregard alternative perspectives rather than to adapt exisiting beliefs“ (Hernandez und Preston, 2012, S. 1). „Personen tendieren dazu, Informationen zu generieren und stärker zu gewichten, welche die eigenen Einstellungen bestätigen oder Erwartungen erfüllen als diejenigen Informationen, welche die eigenen Erkenntnisse widerlegen“ (Schneider, 2013, S. 18). Dieser Effekt kommt zustande, wenn Menschen nach Bestätigung eigener Hypothesen suchen, was in der wissenschaftlichen Literatur unter der Bezeichnung ‚Positive Teststrategie‘ bekannt ist[46]. „Positive Teststrategie bezeichnet das Phänomen einer einseitigen Suche nach bestätigender Information, denn Menschen als Informationssucher suchen die Umwelt nach bekannten und unbekannten Hinweisen ab, wobei die Informationssuche in der Regel zielgerichtet läuft, d. h. es wird eine Hypothese konstruier[t] und dann nach Hinweisen gesucht, die die Hypothese unterstützen. Objektiv wissenschaftlich wäre es, gleichermaßen nach bestätigenden und widerlegenden Hinweisen zu suchen. Durch diese einseitige Informationssuche wird der Mensch tendenziell zu einer Bestätigung seiner Vermutung kommen [...]“[47].
Annahme: Automatisch erkennen könnte man diesen Effekt an den typischen Phrasen, die Erfahrungen der Patienten ausdrücken würden (s. Bei-

[45] Aufgrund dessen wird an dieser Stelle nicht weiter auf den Framing-Effekt eingegangen, was dessen Verwerfung in Bezug auf die vorliegende Arbeit impliziert.
[46] `http://lexikon.stangl.eu/1546/positive-teststrategie/` (02.04.2016).
[47] ebd.

spiele (2.1)). In der Kombination mit dem Selektionskriterium ‚Ausreißer' (s. Abschnitt 2.1.2.1.3, Seite 30) würde die Wahrscheinlichkeit der Identifikation von einem Bestätigungsfehler ansteigen.

(2.1) (a) Phrase ‚Vermutung bestätigt':
„Dann kam eine sehr schnelle Untersuchung, wo meine Vermutung bestätigt wurde und mir einfach irgendwas verschrieben wurde"

(b) Phrase ‚immer so':
„Ärzte immer so hopphopp"
„Ärzte immer so besserwisserisch"

(c) Phrase ‚immer schon':
„ich finde es immer schon nicht so gut, wenn die anmeldung im wartezimmer liegt, erst recht nicht, wenn die sprechstundenhilfe sich lauthals über nicht anwesende personen austauscht"

(d) Phrase: ‚wie gewohnt':
„Frau Dr. Winter und die Helferinnnen wie gewohnt kompetent und freundlich"

(e) Phrase: ‚wie immer schon':
„Das Team ist sehr freundlich, wie immer schon"

2.1.2.2.4 MUM-Effekt

Als MUM-Effekt bezeichnet man Situationen, in denen Menschen aus Angst oder dem Gedanken heraus, jemanden zu verletzen bzw. sich selbst nicht in einem schlechteren Licht darstellen zu wollen, etwas mit Absicht verschweigen bzw. nicht die Wahrheit sagen. „[...] MUM-Effekt, der besagt, dass der Mensch versucht, sich von schlechten Nachrichten zu distanzieren. Niemand will derjenige sein, der eine schlechte Nachricht überbringen muss, da jeder fürchtet, dass er dann auch gleichzeitig als Schuldiger betrachtet wird[48].
Annahme: In Arztbewertungen ist es schwierig bzw. unmöglich, diesen Effekt automatisch zu identifizieren. Durch die Anonymität der Bewertungstexte haben Patienten in der Regel keine Bedenken, in der sozialen Gemeinschaft als „Schuldiger" betrachtet zu werden. Dementsprechend werden selbst intimste Erfahrungen öffentlich dargeboten. Selbst wenn etwas verschwiegen

[48] `http://www.wirtschaft48.info/a/Wie_Angst_und_Misstrauen_zu_Umsetzungsproblemen_f%FChren-578960.html` (29.09.2014).

werden sollte, könnte man es nicht untersuchen, da dies grafisch nicht sichtbar wäre. Was allerdings möglich wäre, sind beispielsweise die theoretisch bewusst nicht vergebenen schlechten Noten mit der Absicht, das Gesamtbild bzw. die Gesamtnote einer Arztpraxis nicht „verschlechtern“ zu wollen. Gegen die Möglichkeit, MUM-Effekte zu erkennen, spricht jedoch ein zu breit gefasster Interpretationsrahmen der Nichtvergabe nummerischer Bewertungen zu konkreten Dimensionen, da allzu viele Gründe dafür denkbar wären. Einer der trivialsten Gründe wäre z. B. der Zeitmangel. Statistisch gesehen, bestätigt sich die Annahme der Vielfalt an möglichen Gründen dadurch, dass rund 40% aller vergebenen Noten in den Bewertungen von Jameda nicht angegeben (n/a), d. h. nicht vergeben sind[49]. Die hier durchgeführte Diskussion veranlasst dazu, den MUM-Effekt im Sinne einer automatischen Identifikation zu verwerfen.

2.1.2.2.5 Dunning-Kruger-Effekt

Der Dunning-Kruger-Effekt bezeichnet eine Selbstüber- bzw. Selbstunterschätzung. „Personen neigen dazu, ihr eigenes Wissen und auch Können zu überschätzen (vor allem Männer) bzw. zu unterschätzen (vor allem Frauen)“ (Schneider, 2013, S. 18). „Menschen neigen bei vielen gesellschaftlichen und intellektuellen Fragen dazu, ihre Fähigkeiten zu hoch einzuschätzen“ (Droste, 2013, S. 2). Erklärt wird dieses Phänomen dadurch, dass Menschen – aufgrund ihrer mangelnden Qualifizierung – falsche Schlussfolgerungen ziehen und nicht in der Lage sind, sich ihren eigenen Fehlern bewusst zu werden (ebd.).
Annahme: Auch dieser Effekt bleibt in Arztbewertungen nicht identifizierbar. Zum einen kann man aus den persönlichen und freiwillig gemachten Angaben der Patienten weder ihren Beruf noch den Wissenstand erschließen und einschätzen. Zum anderen sind die Bewertungen, wie bereits an mehreren Stellen angedeutet, anonym, so dass man persönliche Angaben nicht konkreten Personen zuordnen kann. Es existieren zwar BewertungsIDs (s. Kapitel 5, Tabelle 5.2, Seite 131), was die Nummer einer neu abgegebenen Bewertung impliziert, ohne jedoch zu unterscheiden, welcher Patient unter der Nummer geführt wird und ob z. B. zwei unterschiedliche Bewertungen mit zwei unterschiedlichen IDs von verschiedenen oder von einem Patienten verfasst wurden usw. Schließlich wäre die Kompetenz – selbst wenn man eigene Darstellungen

[49] Stand der Berechnung: Oktober 2013

von Patienten in Texten erkennen würde („Ich bin selbst Chirurg.“) – durch die Subjektivität und Individualität deren Beschreibung für eine automatische Identifikation zu komplex und schwer zu messen. Aus diesen Gründen muss der Dunning-Kruger-Effekt ebenfalls verworfen werden.

2.1.2.3 Domänenspezifische kognitive Effekte

Bei der Untersuchung der Bewertungstexte fällt auf, dass außer einigen oben genannten Effekten andere Anzeichen vorhanden sind, durch die Bewertungen als fehlerhaft charakterisiert werden können. In der vorliegenden Arbeit wird die Untersuchung zusätzlicher domänenspezifischen und empirisch ermittelten Effekte auf sogenannte Überbewertungen und Diskriminierungen eingeschränkt[50].

2.1.2.3.1 Überbewertung einer Dimension durch die Hervorhebung einer anderen

Bei der sogenannten „Überbewertung einer Dimension durch die Hervorhebung einer anderen“ geht es um Dimensionen, deren nummerische Bewertung möglicherweise positiver ausfällt als es in der Wirklichkeit der Fall ist.

(2.2) „Herr Dr. Brachvogel nimmt sich für seine Patienten ausreichend Zeit, was gelegentlich auch dazu führen könnte, dass man länger warten muss!“.

In der Bewertung (2.2) war die Note der Dimension „Wartezeit (Praxis)“ 1.0. Es ist auffällig, dass die Wartezeit zwar als „lang“ wahrgenommen wird, jedoch wird vorher eine andere Bewertungsdimension (in diesem Fall „Genommene Zeit“) positiv hervorgehoben. So lässt sich bei diesem Phänomen beobachten, dass entweder im gleichen Satz oder in unmittelbarer Nähe von dem betreffenden negativen Pattern einer Dimension eine andere Dimension (oft „Genommene Zeit“, „Behandlung“ oder „Freundlichkeit“) positiv hervorgehoben wurde. In anderen Fällen wurden positive Äußerungen zu mehreren Dimensionen aufgelistet und damit zum Schluss z. B. die langen Wartezeiten begründet bzw. ‚entschuldigt‘. Selbstverständlich kann man die dargestellte Beobachtung als rein interpretativ bezeichnen. Obwohl aus dem

[50] Diese Einschränkung ist im Rahmen der Dissertation notwendig, schließt jedoch andere mögliche domänenspezifische Phänomene nicht aus.

o. g. Beispiel die angesprochene Wahrnehmung der Wartezeitdauer und somit die Polarität (s. Abschnitt 2.2.2.2.2, Seite 47) offensichtlich zu sein scheint, könnte man ja schlichtweg behaupten, dass diese für die Dimension „Wartezeit (Praxis)“ in dem aufgeführten Kontext nicht als negativ, sondern als positiv zu verzeichnen wäre. Es würde ebenfalls die Existenz des hier angesprochenen Effekts infrage gestellt werden. Als Gegenargument sei allerdings angemerkt, dass es bei der empirischen Analyse des Korpus auffällt, dass es eine Menge, ja die Mehrheit von Äußerungen ähnlicher Art gibt, bei denen die Patienten lange Wartezeiten durch gut bewertete Fachkompetenz, Freundlichkeit etc. gern in Anspruch nehmen, die Dimension selbst jedoch negativ benotet wird. Laut der für eine Stichprobe des Korpus erstellten Statistik gehören etwa 22% der Bewertungen zu kognitiven Effekten. 78% sind dementsprechend keine (s. Beispiel (2.3)).

(2.3) „Immer freundlich, kompetent, empathisch</column> <column name=„Bewertung“>Die langen wartezeiten nimmt man da gerne in Kauf. [...] <column name=„b_WartezeitPraxis“>5.0</column>“.

Da bei den kognitiven Effekten von einer kleineren Anzahl von fehlerhaften Bewertungen (s. Abschnitt 2.1.1.2.2, Seite 20f.) ausgegangen wird, müssen die Äußerungen im Beispiel (2.2) eindeutig zu den kognitiven Effekten zählen. Alle drei Selektionskriterien können bei diesem Effekt erfüllt werden. Ähnlich dem Halo-Effekt wird hier die Heuristik ‚Anker und Anpassung‘ verwendet (s. Abschnitt 2.1.2.2.1, Seite 31).

2.1.2.3.2 Diskriminierung

Diskriminierung bezeichnet eine Benachteiligung aufgrund bestimmter Merkmale wie ethnische Herkunft, Religion, Weltanschauung, Behinderung, Alter, sexuelle Identität etc. (Elspass und Maitz, 2011, S. 2; Schneider und Bauhoff, 2013, S. 15). Diskriminierungen finden immer in verschiedenen Situationen des Alltagslebens statt und erfolgen im Prozess einer Bewertung anderer. Eine soziale Benachteiligung beruht „auf der kategorialen Behandlung und einer damit verbundenen negativen Bewertung einer Person“ (Elspass und Maitz, 2011, S. 2). Gründe für stattfindende Diskriminierungen lassen sich auf die Verfügbarkeitsheuristik zurückführen, die besagt, dass Menschen beim Urteilen die Informationen verwenden, die sie einfacher aus dem Gedächtnis abrufen können (Tversky und Kahneman, 1974, S. 1127; s. auch Abschnitt 2.1.1.3.2, Seite 23f.). „Assoziiert man aufgrund solcher gedanklich

leicht greifbaren Schemata und Konzepte mit einer bestimmten Personengruppe vorrangig negative Erzählungen oder schlechte Erfahrungen, ist es wahrscheinlich, dass man diese unbewusst auf ein Mitglied dieser Gruppe überträgt. Man schreibt der betreffenden Person dann mit hoher Wahrscheinlichkeit die gleichen Eigenschaften zu, ohne diese Vorannahmen an der Realität zu überprüfen“ (Schneider et al., 2014, S. 30).
Obwohl in der wissenschaftlichen Literatur Diskriminierungen nicht direkt als kognitiver Effekt definiert sind, ist deren Zugehörigkeit zu diesen Phänomenen offensichtlich. Wenn die allgemein aufgestellte Definition kognitiver Effekte im Abschnitt 2.1.1.2.3 (Seite 21) betrachtet wird, kann dort die unbewusste, jedoch nicht zufällige Benachteiligung der Personen in einer Arztpraxis zu Unterbewertungen deren Leistungen führen, die automatisch anhand typischer Phrasen sowie der Ausreißer (s. Abschnitt 2.1.2.1.3, Seite 20), erkannt werden können. Allerdings muss man sich darüber im Klaren sein, dass eine automatische Extraktion – aufgrund der nötigen Einschränkungen sowie domänenspezifischer Verfügbarkeit solcher Aussagen – nicht für alle Merkmale von Diskriminierungen möglich ist. Zunächst sollten die impliziten und expliziten Diskriminierungen unterschieden werden, wobei sich diese Arbeit auf die automatische Extraktion der expliziten sprachlichen Diskriminierungen – aufgrund der sonst zu hohen semantisch bedingten Komplexität – beschränkt. „Unter sprachlicher Diskriminierung wird eine soziale Diskriminierung verstanden, die mittels Sprache realisiert wird. [...]. Kommt die Bewertung in der Wortwahl oder im Inhalt einer Äußerung zum Vorschein, handelt es sich um eine explizite Diskriminierung, geht die Bewertung lediglich aus dem Kontext hervor, handelt es sich um eine implizite Diskriminierung“ (Höer et al., 1996, S. 2). Somit kann eine explizite Diskriminierung in Bewertungstexten „anhand einer aus einem oder mehreren Wörtern bestehenden sprachlichen Einheit [...] oder aus dem Inhalt des ganzes[n] Satzes“ (ebd., S. 4) erkannt werden.
Auf welche Art werden Diskriminierungen in Bewertungstexten der Arztpraxen sichtbar?
Nach einer empirisch erfolgten Analyse der Arztbewertungstexte lässt sich eine interessante Tatsache beobachten, die der Domänenspezifik zugeschrieben wird. So gibt es zwei Arten von Diskriminierungen:

- Patienten diskriminieren Ärzte oder das Praxispersonal aufgrund bestimmter Merkmale:

(2.4) (a) Alter: „weil Dr. Bach zu alt ist“

(b) Nationalität: „eine unmotivierte arrogante Asiatische Ärztin“

- Patienten fühlen sich aufgrund bestimmter Merkmale von dem Arzt bzw. dem Praxispersonal diskriminiert:

(2.5) (a) Nationalität: „Türken bezahlen fürs Augenlasern weniger als Deutsche“

(b) Aussehen: „Dicke werden in der Praxis diskriminiert“

In beiden Fällen ist zu vermuten, dass eine Abwertung der betreffenden Praxis stattfindet, die man anhand der vergebenen nummerischen Bewertungen zu den Dimensionen ablesen kann.

2.1.3 Überblick und Definitionen relevanter Effekte

Aus der vorgestellten Auswahl kognitiver Effekte werden die Ergebnisse in Bezug auf deren Ermittlungsart (aus der Literatur / empirisch) und automatische Erkennbarkeit (ja / nein) in der Tabelle 2.2 zusammengefasst. Somit werden vier für die vorliegende Arbeit relevante identifizierbare Effekte festgelegt, deren domänenspezifische Definitionen in nachfolgenden Abschnitten aufgeführt werden:

- Halo-Effekt
- Bestätigungsfehler
- Überbewertung
- Diskriminierung

2.1.3.1 Definition: Halo-Effekt

Unter *Halo-Effekt* versteht man die Phänomene, bei denen durch eine positiv / negativ angesprochene und bewertete Dimension eine andere bzw. mehrere andere ebenfalls positiv / negativ bewertet werden. Dies bedeutet, dass die Bewertung dieser ersten Dimension als „Leitdimension“ (Ankerdimension) zu verstehen ist und die andere(n) Dimensionen (Anpassungsdimension(en)) durch sie über- bzw. unterbewertet werden. Die dadurch über- / unterbewerteten Dimensionen sollen sich im Allgemeinen von den durch die meisten

Ermittlungsart		Erkennbarkeit	
aus der Literatur	**empirisch ermittelt**	**ja**	**nein**
Halo		x	
Framing			x
Bestätigungsfehler		x	
MUM			x
Dunning-Kruger			x
	Überbewertung	x	
	Diskriminierung	x	

Tabelle 2.2: Erkennbarkeit kognitiver Effekte in Arztbewertungen

Patientenbewertungen (‚objektive Norm') derselben Dimensionen deutlich unterscheiden (Ausreißer).

2.1.3.2 Definition: Bestätigungsfehler

Bestätigungsfehler sind Tendenzen, nach den Bestätigungen eigener Hypothesen einseitig zu suchen. Hinweise auf diese Einseitigkeit in Arztbewertungen bilden typisch und explizit formulierte Phrasen ab. Gleichzeitig müssen innerhalb der betreffenden Bewertungen Ausreißer identifiziert werden können.

2.1.3.3 Definition: Überbewertung

Als *Überbewertungen* werden diejenigen Phänomene bezeichnet, bei denen eine negative Äußerung zu einer Dimension (Anpassungsdimension) stattfindet, diese jedoch positiv benotet wird. Gleichzeitig wird auf andere(n) Dimension(en) (Ankerdimension(en)) positiv verwiesen. Die nummerische(n) Bewertung(en) der Ankerdimension(en) hat / haben ebenfalls eine positive Polarität. Die Anpassungsdimension wird durch das Vorhandensein der Ankerdimension(en) auf irgendeine Weise „gerechtfertigt" und trotz des negativen Empfinden von derselben positiv benotet bzw. überbewertet, wodurch die Verzerrung entsteht.

2.1.3.4 Definition: Diskriminierung

Diskriminierungen in Arztbewertungen sind mittels einer Sprache explizit realisierte soziale Benachteiligungen, die sich auf einen oder mehrere Merkmale beziehen. Mit Merkmalen sind beispielsweise Geschlecht, Alter, sexuelle Identität, ethnische Herkunft, Religion, Behinderung etc. gemeint. Die Diskriminierungen können sich einerseits gegenüber den Personen in einer Praxis richten, andererseits können sich die Bewertenden selbst diskriminiert vorkommen. Durch eine erfolgte Diskriminierung ist zu vermuten, dass eine Abwertung ärztlicher Leistungen stattfindet, die in der Kombination von diskriminierenden Äußerungen und nummerischen Bewertungen zu Dimensionen automatisch erkannt wird.

2.2 Maschinelle Erkennung kognitiver Effekte

Die den gebildeten und geäußerten Meinungen vorausgegangenen kognitiven Verarbeitungsprozesse kann man allgemein nicht in den niedergeschriebenen Texten erkennen. Vorstellbar ist jedoch die Erkennung der Ergebnisse stattgefundener Prozesse, also der Meinungen selbst.
Durch die im vorigen Abschnitt beschriebenen und definierten Effekte wird deutlich, dass jeder Effekt mit Interpretationen, Bewertungen, Informationsgewichtungen, Über- und Unterschätzungen etc. zu tun hat. Das Erkenntnisinteresse der vorliegenden Arbeit betrifft die automatische Identifikation kognitiver Effekte, was impliziert, dass zunächst automatisch wertende Äußerungen, also Meinungen in Texten erkannt werden müssen, was die Auseinandersetzung mit den am Anfang dieses Kapitels erwähnten computerlinguistischen Disziplinen erforderlich macht.

2.2.1 Informationsextraktion und Domänenspezifik

2.2.1.1 Definition von Informationsextraktion

Die *Informationsextraktion (IE)* beschäftigt sich mit automatischer Textanalyse. Der Begriff der IE ist in der Literatur breit gefasst, worauf z. B. die nachfolgende Definition hindeutet: „Information Extraction (IE) is the name given to any process which selectively structures and combines data

which is found, explicitly stated or implied, in one or more texts. The final output of the extraction process varies, in every case, however, it can be transformed so as to populate some type of database“ (Cowie und Wilks, 2000)[51]. Um dieses für die vorliegende Arbeit zu präzisieren, kann man eine traditionelle domänenspezifische und eine domänenoffene IE unterscheiden (Geierhos, 2010, S. 17ff.). „Im Gegensatz zur domänenoffenen Informationsextraktion steht das Festlegen einer bestimmten Domäne hier an erster Stelle, bevor die IE-Aufgabe überhaupt spezifiziert werden kann“ (ebd., S. 19). Die relevanten Informationen sollen also domänenspezifisch aus freien Texten aufgespürt und strukturiert werden, während die irrelevanten gleichzeitig „überlesen“ werden (ebd., S. 17). „IE-Systeme [...] sollen nur die Textpassagen analysieren bzw. „verstehen“, die relevante Informationen beinhalten. Was als relevant gilt, wird dabei durch vordefinierte domänenspezifische Lexikoneinträge oder Regeln dem System fest vorgegeben“ (Neumann, 2004, S. 502).

2.2.1.2 Aufgabenspezifizierung

Durch die Entscheidung, sich auf die Domäne von Arztbewertungen zu konzentrieren, ist die Relevanz der domänenspezifischen IE offensichtlich. Die Extraktion der genannten relevanten Informationen sollte dabei nicht auf lokales Textverständnis ausgerichtet sein, da zum einen IE diese Aufgabe nicht bewältigen kann (Geierhos, 2010, S. 30). Zum anderen sind die hier interessierenden Informationen auf wertende Phrasen zu Bewertungsdimensionen und auf die für ausgewählte Effekte spezifischen Phrasen eingeschränkt. Das Ziel ist es hier, keine Templates für z. B. Datenbanken oder Frage-Antwort-Systemen anzufertigen, wobei Informationen dafür üblicherweise aus meistens gut strukturierten Texten, die bestimmte Tatsachen zu Sachverhalten darlegen, extrahiert werden (vgl. Geierhos, 2010; vgl. Stotz, 2018). Die domänenspezifische IE soll im Rahmen dieser Arbeit als ein Mittel verstanden werden, um nötige Muster zur Weiterverarbeitung (hier: Identifikation und Klassifikation kognitiver Effekte) zu extrahieren. Die Informationsextraktion kann hierbei als zweistufiger Prozess betrachtet werden (Grishman, 1997, zitiert in Geierhos (2010, S. 27)), bei dem „lokale Textanalyse zur Extraktion relevanter Textpassagen und Faktenbestimmung“ und „Diskursanalyse zur Vervollständigung der extrahierten Fakten“ (ebd.) im Mittelpunkt

[51] `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.6480&rep=rep1&type=pdf` (19.04.2015).

stehen. Die erwähnten Muster, die in Arztbewertungen extrahiert werden müssen, also, ‚die relevanten Textpassagen' (s. o.) sind auf Meinungen und deren Polaritäten (s. Abschnitt 2.2.2.2.1) in Bezug auf die Bewertungsdimensionen eingeschränkt. Die lokale Textanalyse zur Extraktion dieser Muster wird – semantisch gesehen – in Bezug auf Entitäten oder Objekte (s. Abschnitt 2.2.1.2.2) – Bewertungsdimensionen, Polaritätswörter, bestimmte wertende Phrasen – erfolgen, die in relevanten Kontexten in entsprechenden Phrasen / Sätzen (syntaktisch gesehen) vorkommen (Diskursanalyse). „Depending of the granularity of analysis, a sentiment target may refer to a concrete entity or to a more abstract topic. For instance, in aspect-oriented review mining we are interested in determining the reviewers' evaluations of very concrete aspects [Bewertungsdimensionen]. Such targets typically become manifest at phrase or sentence level (e.g., „I really like the picture quality."). In this case, the task ist primarily regarded as an information extraction task" (Broß, 2013, S. 14).

2.2.1.2.1 Bewertungsmuster als Subsprache

Die Spezifik der Extraktionsmuster kann als eine Subsprache betrachtet werden. „Certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it" (Harris, 1979, S. 152). Es ist selbstverständlich nicht einfach, die Grenzen von solchen Subsprachen zu definieren, da selbst natürliche Sprachen unscharf abgegrenzt sind, wofür z. B. eine Menge von Dialekten bezeichnend ist (Sager, 1986, S. 2). Einige Eigenschaften der Subsprachen (Geierhos, 2010, S. 28f.), die auf die beschränkte Auswahl der zu extrahierenden Ausdrücke in Arztbewertungen ebenfalls zutreffen, sind:

- thematische Begrenztheit
- lexikalische, syntaktische und semantische Einschränkungen
- andere grammatikalische Eigenschaften als in Allgemeinsprache
- Wiederholung gewisser lexikalischer Strukturen
- Verwendung eigener Struktur und Symbolik (ebd.)

2.2.1.2.2 Entitäten vs. Objekte

Seit 1991 ist eines der Ziele der *Message Understanding Conferences (MUC)*, die strukturierte Repräsentation der wichtigsten Textinformationen zu realisieren. Der Hauptakzent liegt auf Entitäten und deren Relationen (Geierhos et al., 2015a, S. 17). „Entities are typically noun phrases and comprise of one to a few tokens in the unstructured text. The most popular form of entities is named entities like names of persons, locations and companies [...]“ (Sarawagi, 2008, S. 269). Die Extraktion von benannten Entitäten (named entity recognition – NER) wurde zum ersten Mal in der 6 MUC 1995 als Aufgabe der IE definiert, was in drei Bereiche aufgeteilt wurde:

- Eigennamen und Akronyme von Personen, Orten und Organisationen (ENAMEX)
- absolute Zeitangaben (TIMEX)
- nummerische Ausdrücke (NUMEX) (vgl. ebd.).

Obwohl lediglich Bezeichner für Personen, Organisationen und Orte zu benannten Entitäten gezählt werden (Geierhos et al., 2015a, S. 19), können trotzdem alle möglichen Erkennungsobjekte unter dem Begriff Entitäten zusammengefasst werden. „Now the term entities is expanded to also include generics like disease names, protein names, paper titles, and journal names. The ACE [Automatic Content Extraction] competition for entity relationship extraction from natural language text lists more than 100 different entity types“ (Sarawagi, 2008, S. 269). Die aus dem Forschungsfeld Informationsextraktion angesprochene Zusammenfassung aller denkbaren Erkennungsobjekte unter dem Begriff „Entitäten“ ist in der Stimmungsanalyse unter der Bezeichnung „Objekte“ bekannt: „An *object o* is an entity which can be a product, person, event, organization, or topic“ (Liu, 2010, S. 3). Außer den im Abschnitt 2.1.1.3.1 (Seite 22f.) definierten Einstellungsobjekten (hier: Bewertungsdimensionen) sind für die vorliegende Arbeit z. B. Personennamen insofern interessant, dass man diese von anderen Bewertungsobjekten wie ‚Praxis‘, ‚Parkplätze‘, ‚Ausstattung‘ usw. unterscheiden kann.
Die Repräsentation eines Objekts kann man sich als eine endliche Menge von *Eigenschaften (Features)*, F = {f1, f2, ..., fn} von diesem Objekt vorstellen, wobei Objekt selbst als eine Eigenschaft in der Menge F enthalten ist (ebd.,

S. 5). Eigenschaften können *explizit* und *implizit* ausgedrückt werden (s. Seite 37): „If a feature f or any of its synonyms appears in a sentence s, f is called an *explicit feature* in s. If neither f nor any of its synonyms appear in s but f is implied, then f is called an *implicit feature* in s“ (ebd., S. 4). Bei expliziten Eigenschaften handelt es sich um Nomen und Nominalphrasen (Liu, 2012, S. 22). So kann man sich die Dienstleistung einer Praxis als ein Bewertungsobjekt vorstellen, dessen Eigenschaften Bewertungsdimensionen sind. Die Bewertungsdimensionen können wiederum gleichzeitig als Objekte und Eigenschaften definiert werden. Als Beispiel einer expliziten Eigenschaft, die gleichzeitig als Objekt betrachtet werden kann, ist das Wort bzw. die Dimension „Wartezeit“ im folgenden Satz zu benennen (Beispiel (2.6)):

(2.6) „90 Minuten Wartezeit im Wartezimmer [...].“

Im Beispiel (2.7) dagegen wird die Eigenschaft bzw. das Objekt selbst nicht genannt, jedoch wird ebenfalls hier die Wartezeit im Wartezimmer angesprochen. In solchen Fällen handelt es sich um *implizite* Eigenschaften:

(2.7) „Musste zwar lange im Wartezimmer sitzen [...].“

2.2.1.2.3 Relationen

Relationen bezeichnen Beziehungen zwischen den oben aufgeführten Entitäten oder Objekten. „Relationships are defined over two or more entities related in a predefined way. Examples are „is employee of“ relationship between a person and an organization, „is acquired by“ relationship between pairs of companies, „location of outbreak“ relationship between a disease and a location and „is price of“ relationship between a product name and a currency amount on a web-page“ (Sarawagi, 2008, S. 270f.). Geierhos (2010), die sich mit der Klassifikation und Extraktion karrierespezifischer Informationen beschäftigte, definierte „biographische Relation als eine Prädikat-Argument-Struktur (PAS), bei der mindestens ein Argument mit einer Instanz aus der Objektklasse <Person> belegt sein muss“ (ebd., S. 11), wobei Objektklassen die Argumentpositionen eines Verbs, Adjektivs oder Nomens belegen können (ebd., S. 10). Einige der Relationen, die sie in ihrer Arbeit extrahierte, sind: „to study“, „to be educated“, „to finish school“, „to work“, „to teach“, „to be engaged“, „to be nominated“, „to set up a company“, „successor of“ usw. (Geierhos, 2010, S. 246ff.). Auf diese Arbeit bezogen, haben die Relationen

bei der Identifikation kognitiver Effekte einen anderen Charakter. Innerhalb der aufgestellten Identifikationskriterien (s. Kapitel 4, Abschnitt 4.3) sind z. B. bei den linguistischen Mustern einerseits die Beziehungen zwischen den expliziten Benennungen vordefinierter Dimensionen (Objekte) und den Polaritätswörtern interessant („Die **Behandlung** ist **gut**."). Ein anderes Beispiel für Relationen wären die (In)Konsistenzen in der Polarität, d. h. die Relationen zwischen den Aussagen in Bewertungstexten und der Vergabe nummerischer Werte bezüglich einer Dimension. Andererseits kann man bestimmte Kombinationen von denselben Kriterien als Relationen verstehen, durch die Effekte definiert sind.

2.2.2 Stimmungsanalyse

2.2.2.1 Definition und Bedarf

Die *Stimmungsanalyse (SA)*, ebenfalls unter Begriffen *Sentiment Analysis* (englische Übersetzung), *Opinion Mining* oder *Subjectivity Analysis* (Bornebusch und Cancino, 2014, S. 2389; Liu, 2010, S. 1; Wolfgruber, 2015, S. 17) bekannt, beschäftigt sich mit der Untersuchung und der automatischen Verarbeitung von Meinungen, Gefühlen und Emotionen, die in einem Text ausgedrückt werden (Liu, 2010, S. 3; Bornebusch und Cancino, 2014, S. 2389). Das Interesse an der SA seitens der Wirtschaft und Wissenschaft ist immens groß (Liu, 2010, S. 1). „SA [...] has recently become the focus of many researchers, because analysis of online text is beneficial and demanded for market research, scientific surveys from psychological and sociological perspective, political polls, business intelligence, enhancement of online shopping infrastructures, etc. Nowadays if one wants to buy a consumer product one prefer[s] user reviews and discussion in public forums on web about the product. As a result opinion mining has gained importance" (Shailesh, 2015, S. 113). Im Vergleich zur IE, die sich mehr auf die Extraktion der Entitäten konzentriert (s. o.), was als „factual information" (ebd.) interpretiert werden kann, hat sich die SA zur Aufgabe gemacht, subjektive Informationen – Meinungen (s. Abschnitt 2.2.2.2, Seite 46) – aus dem UGC (s. Seite 4) zu extrahieren. UGC kann man als „opinionated text in the form of reviews, blog posts, social media comments and more recently, tweets" (ebd.) zusammenfassen. „In many cases, opinions are hidden in long forum posts and blogs. It is difficult for a human reader to find relevant sources, extract related sentences with opinions, read them, summarize them, and organize them into usable forms.

Thus, automated opinion discovery and summarization systems are needed. *Sentiment analysis*, also known as *opinion mining*, grows out of this need“ (Liu, 2010, S. 1). Die wichtigsten Aufgaben der Stimmungsanalyse sind also die *Extraktion* subjektiver Informationen aus Texten und Klassifikation dieser Informationen entsprechend ihrer Polarität (Wolfgruber, 2015, S. 18; Chen, 2012, S. 173).

2.2.2.2 Meinungen

2.2.2.2.1 Definition

Unter einer Meinung wird eine subjektive positive oder negative Auffassung, Einstellung, Emotion oder Bewertung zu Entitäten, Ereignissen und deren Eigenschaften verstanden, die von einem Meinungsträger gehalten wird (Liu, 2010, S. 4; Liu, 2012, S. 17; Wolfgruber, 2015, S. 12). Die Verbindung zum sozialpsychologischen Kontext bilden die im Abschnitt 2.1.1.3.1, 22 definierten Einstellungen, die ein Äquivalent zu Meinungen bilden und innerhalb derer nach Hinweisen auf kognitive Effekte gesucht werden soll. Der Meinungsträger ist diejenige Person, die die entsprechende Meinung ausdrückt (Liu, 2010, S. 4). In Bewertungstexten der Bewertungsportale sind Meinungsträger die Autoren der Bewertungen selbst (Broß, 2013, s. 13), also die Patienten oder deren Angehörige.

2.2.2.2.2 Klassifikationskriterien

a) Subjektivität vs. Objektivität

Das Problem der Klassifikation von Aussagen ihrer Subjektivität entsprechend ist als eine binäre Aufgabe zu verstehen, die objektiven von den subjektiven Informationen zu unterscheiden (ebd.). „An objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs“ (Liu, 2010, S. 7). Charakteristisch für Bewertungsportale ist, dass die Bewertungstexte meistens subjektiv sind (Broß, 2013, S. 18; Wolfgruber, 2015, S. 21). Eine genaue Definition der Subjektivität ist allerdings kontext- und domänenabhängig (Broß, 2013, S. 13), was man im Beispiel (2.8) sehen kann:

(2.8) (a) „Es wird sich zu wenig Zeit genommen“

(b) „Die Ärztin hat sich dann ca. 20 Minuten Zeit genommen“

Während im Beispiel (2.8)(a) die Aussage eindeutig subjektiven Charakter hat, ist dies im Beispiel (2.8)(b) nicht eindeutig klar, da eine genaue Zeitangabe individuell in Bezug auf ihre Positivität oder Negativität empfunden wird.

b) Implizitheit vs. Explizitheit

Der Begriff der Im- und Explizitheit wurde in der vorliegenden Arbeit bereits im Kontext von Diskriminierungen (s. Abschnitt 2.1.2.3.2, Seite 37f.) und Eigenschaften von Objekten (s. Abschnitt 2.2.1.2.2, Seite 44ff.) angesprochen. In diesem Sinne können ebenfalls Meinungen implizit oder explizit ausgedrückt werden (Wolfgruber, 2015, S. 12). Im ersten Fall könnten diese z. B. in Form einer rhetorischen Frage ausgedrückt werden (ebd; s. Beispiel (2.9)), die eine negative Meinung zu genannten Aspekten suggeriert. Im zweiten Fall wird eine Meinung direkt in einem subjektiven Satz mitgeteilt (s. z. B. Beispiel (2.8)(a)).

(2.9) „Sollte der Arzt sich nicht Zeit nehmen und interesse zeigen?“

c) Polarität und Intensität

Die Polarität einer Meinung gibt an, ob diese positiv, negativ oder neutral ist (Liu, 2010, S. 5). Auf der Satzebene wird die Polarität des ganzen Satzes analysiert, während auf der Aspektebene (s. Abschnitt 2.2.2.3.2) die Polaritäten einzelner Aspekte bestimmt werden (Bornebusch und Cancino, 2014, S. 2389). Bei feinkörnigeren Analysen kann die Differenzierung verschiedener Grade der Positivität oder Negativität einer Meinung, eine sogenannte Intensität (s. „Stärke“, Seite 23), vorgenommen werden (Broß, 2013, S. 13). Im Kontext der sozialpsychologischen Auseinandersetzung mit Einstellungen (s. Abschnitt 2.1.1.3.1, Seite 22ff.) wurden Polarität und Intensität als Valenz (s. Seite 23) und Stärke (s. o.) bezeichnet, jedoch liegt hier der Fokus auf einer automatischen Klassifizierung der Meinungen nach Kriterien ‚positiv‘ und ‚negativ‘ in schriftlichen Texten. Diese Aufgabe ist nicht immer trivial und zum Teil stark domänenabhängig (ebd.). Aufgrund einer individuellen Situationswahrnehmung (s. Abschnitt 2.1.1.3.2, Seite 23) kann z. B. die Wartezeit in Minuten in der Arztpraxis bei einigen Patienten positiv, bei anderen dagegen negativ wahrgenommen werden (s. (2.10)).

(2.10) (a) Positiv:
„30 Minuten Wartezeit waren für mich erträglich“

(b) Negativ:
”30 Minuten Wartezeit trotz Termin“

Im Beispiel (2.11)(a) kann die Teilaussage „[...] hat sich viel Zeit genommen [...]“ negativ empfunden werden, im Kontext der Arztbewertungen jedoch ist dies eindeutig als eine positive Bewertung zu interpretieren.

(2.11) (a) „Er hat sich viel Zeit genommen und alles gründlich gecheckt“

(b) „Fachlich sehr gute Behandlung, aufgrund des sehr hohen Patientenaufkommens oft lange Wartezeiten.“

Eine eindeutige Einteilung polaritätsbestimmender Phrasen (s. nächsten Abschnitt) in positive und negative Kategorien kann innerhalb einer Domäne ebenfalls problematisch werden. Der Ausdruck „viel Zeit“ ist bei der Bewertungsdimension „Wartezeit (Praxis)“ eindeutig als negativ und bei der „Genommenen Zeit“ als positiv zu klassifizieren.

d) Ausgewählte Indikatoren und Modifikatoren

Mit Indikatoren sind polaritätsgebende Wörter, Wortkombinationen oder andere syntaktische Phänomene gemeint (Wolfgruber, 2015, S. 24). „*Opinion words* are words that are commonly used to express positive or negative sentiments. For example, *beautiful*, *wonderful*, *good*, and *amazing* are positive opinion words, and *bad*, *poor*, and *terrible* are negative opinion words“ (Liu, 2010, S. 11). Die Wortarten wie Adjektive, Adverbien, Nomen, Verben können auf die Polarität einer Meinung hindeuten (Wolfgruber, 2015, S. 24ff.). Außerdem können die ganzen Phrasen, feste Wendungen (Liu, 2010, S. 11; Wolfgruber, 2015, S. 26f.), aber auch ironische Ausdrücke (Wolfgruber, 2015, S. 27ff.) polaritätsbestimmend sein. Weitere graphemische Indikatoren sind die Emoticons, Anführungszeichen, Großbuchstaben usw., auf die hier nicht ausführlicher eingegangen wird (vgl. Wolfgruber, 2015; vgl. Schieber et al., 2012). Auf morphologischer und lexikalischer Ebene können Phrasen oder Sätze verstärkende, abschwächende oder inverse Modifikationen bezüglich ihrer Polaritäten erhalten (Wolfgruber, 2015, S. 32ff.). Diese erfolgen mit Wirkung sogenannter Modifikatoren wie z. B. Negation, die die Polarität einer Meinung umkehrt (ebd., S. 32f.; s. Beispiel (2.12)):

(2.12) (a) „nicht gut behandelt“

(b) „unübersichtliche Rechnung“

Lexikalische Modifikatoren sind z. B. folgende Wörter wie „nein“, „nie(mals)“, „kein“, „nicht(s)“, „keinerlei“, „ohne“, „außer“ etc. und zu morphologischen zählen wortbildende Präfixe wie „un-“, „nicht-“, „ent-“, „non-“, „dis-“, „a-“, „in-“ und Suffixe wie „-los“, „-frei“ (ebd., S. 33.). Die verstärkend oder abschwächend wirkenden Modifikatoren sind z. B. folgende Partikel wie „kaum“, „sehr“, „total“, aber auch andere Wortarten und Ausdrücke, die Polarität einer Aussage in entsprechende Richtung intensivieren können (ebd., S. 34). Eine ausführliche Auseinandersetzung mit den Modifikatoren kann z. B. im Rahmen der Intensitätsbestimmung einer Meinung stattfinden, was in der vorliegender Arbeit irrelevant ist und daher nur am Rande erwähnt wird.

2.2.2.3 Ablauf

2.2.2.3.1 Prozess der Stimmungsanalyse

Der Prozess der Stimmungsanalyse kann auf folgende Weise erfolgen:
UGC bzw. die Bewertungen werden zunächst auf das Vorhandensein subjektiver Aussagen untersucht (*Sentiment Identification*[52]) (Wolfgruber, 2015, S. 19). Die Objekte und deren Eigenschaften (s. Abschnitt 2.2.1.2.2, Seite 43ff.) werden extrahiert (*Feature or Aspect Selection*[53]) und entsprechend ihrer Polarität klassifiziert (*Sentiment Classification und Sentiment Polarity*[54]) (Shailesh, 2015, S. 115).
Für die vorliegende Arbeit sind nicht alle Phrasen, die Meinungen beinhalten, interessant. Die Stimmungsanalyse wird hier daher auf wertende Aussagen zu den vordefinierten Dimensionen und zu ausgewählten Aspekten, die einige Phrasen zu weiteren kognitiven Effekten betreffen, eingeschränkt. D. h. die Dimensionen sind nur innerhalb der wertenden Aussagen relevant. Nach ihrer Extraktion (Bewertungsobjekte) erfolgen „Sentiment Identification“, „Sentiment Classification“ und „Sentiment Polarity“ anhand wertender Phrasen, in denen die Bewertungsobjekte vorkommen. Findet man die Bewertungen zu

[52] Stimmungsidentifikation

[53] Auswahl von Objekten und deren Eigenschaften (hier: Extraktion der Bewertungsobjekte)

[54] Stimmungsklassifikation und -polarität (hier: Stimmungsidentifikation und -klassifikation (Polarität))

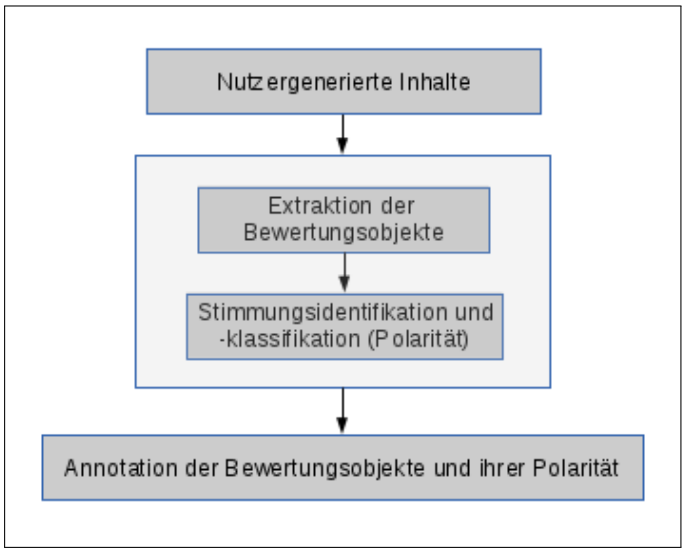

Abbildung 2.3: Prozess der Stimmungsanalyse

Dimensionen, werden diese zur automatischen Weiterverarbeitung im Sinne der Identifikation kognitiver Effekte annotiert. Der Prozess der Stimmungsanalyse kann dementsprechend wie auf der Abbildung 2.3 schematisch dargestellt werden.

2.2.2.3.2 Klassifikationsebenen

Der dargestellte Prozess kann auf unterschiedlichen Ebenen erfolgen.
Grundsätzlich werden drei Ebenen differenziert (Shailesh, 2015, S. 115ff.; Wolfgruber, 2015, S. 18):

- Dokument-Ebene: „Given a set of opinionated documents D, it determines whether each document $d \in D$ expresses a positive or negative opinion (or sentiment) on an object“ (Liu, 2010, S. 10)
- Satz-Ebene: In einem gegebenen als subjektiv identifizierten Satz wird ermittelt, ob die Meinung positiv oder negativ ist (Wolfgruber, 2015, S. 18; Liu, 2010, S. 13)
- Aspekt-Ebene: Hier wird nach positiven und negativen Aspekten einer Entität klassifiziert (ebd.; Shailesh, 2015, S. 117)

Für die Arbeit ist die Stimmungsanalyse auf der Aspekt-Ebene innerhalb der Phrasen oder Sätze interessant. „Das Ziel [...] [einer aspektbasierten Stimmungsanalyse] ist es, eine Sentiment Analysis durchzuführen, die darauf basiert, dass zunächst die Polaritäten von einzelnen Aspekten eines Satzes oder Absatzes bestimmt werden“ (Bornebusch und Cancino, 2014, S. 2389). Als Aspekte ist, wie auch auf Bewertungsportalen mit individuellen Produktaspekten (Broß, 2013, S. 17), eine endliche Menge vordefinierter Bewertungsdimensionen zu verstehen. Zum Zweck einer automatischen Identifikation kognitiver Effekte benötigt man, wie bereits mehrmals festgestellt, Polaritäten wertender Phrasen zu konkreten Dimensionen, um diese dann mit nummerischen Bewertungen zu vergleichen und mögliche Unstimmigkeiten zu erkennen. Anders als bei einer ‚absatz- oder dokumentenbasierten‘ Stimmungsanalyse, die im Ergebnis allgemeine Informationen zur z. B. Anzahl der zufriedenen und unzufriedenen Patienten mit ärztlichen Leistungen (ebd., S. 21) liefert oder eine Bewertung bzw. einen Satz als insgesamt ‚positiv‘ oder ‚negativ‘ klassifiziert (ebd., S. 12; Shailesh, 2015, S. 117; Bornebusch und Cancino, 2014, S. 2395), ist in dieser Arbeit eine phrasenbasierte Stimmungsanalyse interessant. „At the phrase level we are interested in extracting individual mentions of aspects and related expressions of sentiment“ (Broß, 2013, S. 22). Eine Wertung zu einer Dimension kann dabei innerhalb eines Satzes stehen, wie dies in Beispielen (2.8) zu sehen ist, oder innerhalb einer Phrase in einem Satz (s. Beispiel (2.11)(a)). Es können ebenfalls Sätze vorkommen, die Wertungen zu mehreren Dimensionen mit diversen Polaritäten enthalten (s. Beispiel (2.11)(b).).

Kapitel 3

Stand der Wissenschaft

Nach der Vorstellung kognitiver Effekte und der Aufstellung der Grundlagen zu ihrer automatischen Identifikation wird in diesem Zusammenhang der Stand der Wissenschaft betrachtet. Zunächst werden Vorgehensweisen sozialpsychologischer Experimente und Verfahren aspektbasierter Stimmungsanalyse gegenübergestellt, was die Herausforderungen einer automatischen Identifikation kognitiver Effekte in Arztbewertungen verdeutlicht. Dabei sind die Fragen vorrangig interessant, welche Möglichkeiten die Verfahren beider Wissensbereiche bieten und wie diese ablaufen, um ein eigenes Vorgehen zu bestimmen und eine eigene Verfahrensauswahl zu treffen.

Da die automatische Identifikation von kognitiven Effekten noch nie durchgeführt wurde, werden andere sozialpsychologische Phänomene betrachtet, die bereits durch die Aufstellung von typischen Indikatoren und den Einsatz automatischer Verfahren identifiziert wurden. Das Augenmerk wird hier wiederum auf die jeweiligen Indikatoren und deren Hintergründe gelegt. Bevor man allerdings automatische Verfahren zur Identifikation kognitiver Effekte einsetzt, wird man u. a. mit Problemen wie subjektive Pattern-Definitionen oder Indikatoren, die auf der Ebene der Gesamtdaten definiert und entsprechend aufbereitet werden müssen, konfrontiert. Dies impliziert, dass eine Reihe von Vorarbeiten geleistet werden muss, auf die zum Ende des aktuellen Kapitels eingegangen wird. Auf diese Weise entsteht ein differenziertes Bild von Indikatoren und automatischen Verfahren, mit denen sie identifiziert werden können. Die Ausführungen dieses Kapitels dienen daher als Basis zur Ausarbeitung einer methodischen Konzeption, die u. a. die Aufstellung der Kriterien / Indikatoren für identifizierbare Effekte einschließt und im Kapi-

tel 4 beschrieben wird.

3.1 Zu aktuellen Verfahren für Identifikation kognitiver Effekte

In diesem Abschnitt wird zunächst die Spezifik automatischer Erkennung kognitiver Effekte deutlich gemacht. Als Erstes werden Methoden aufgeführt, die zur Bestimmung von Effekten mittels sozialpsychologischer Experimente herangezogen wurden. Im nächsten Schritt werden aktuelle Verfahren aspektbasierter Stimmungsanalyse dargestellt, da kognitive Effekte mit Äußerungen von Meinungen verbunden sind (s. Kapitel 2, Abschnitt 2.1.1.2.1, Seite 19). Dieses Vorgehen dient der Findung eines eigenen Ansatzes zur automatischen Identifikation solcher Phänomene.

3.1.1 Spezifik automatischer Effekte-Erkennung

Bevor man Verfahren der automatischen Textanalyse betrachtet, werden hier einige Methoden aus aktuellen Studien beschrieben, anhand derer kognitive Effekte identifiziert wurden. Dies ist wichtig, um Kontraste zur methodischen Vorgehensweise bei der automatischen Identifikation der Effekte aufzuzeigen sowie die Richtung computerlinguistischer Verfahrensauswahl zu bestimmen und zu begründen.

3.1.1.1 Sozialpsychologische Experimente

Einige Beispiele kognitiver Effekte wurden bereits im Kapitel 2 (Abschnitt 2.1.1.2, Seite 18ff.) beschrieben. In diesem Abschnitt jedoch ist die Frage interessant, mit welchen Methoden sich diese Denkfehler bestimmen lassen. Wie auf der Seite 31 in Bezug auf den **Framing-Effekt** angedeutet, wird dieser mit Hilfe von Darstellungen entsprechender Situationen identifiziert. Ein Beispiel dafür lieferten die Psychologen Kahneman und Tversky (1984), indem sie zwei Probandengruppen Programme zur Bekämpfung einer Krankheit auf unterschiedliche Weisen vorstellten. Dabei mussten sich beide Gruppen jeweils für eines der zwei zur Verfügung gestellten Programme entscheiden. Obwohl beide Programme für die jeweilige Probandengruppe vollkommen identisch waren, entschieden sich diese für unterschiedliche Programme.

Der Grund dafür liegt in den unterschiedlichen Formulierungen. ‚Freundlicher' formulierte Aussagen wurden bevorzugt und somit die Entscheidungen beeinflusst (Kahneman und Tversky, 1984, S. 343). Auch neuere Studien zum Framing-Effekt untersuchen dieses Phänomen mittels der Situationen, die den Probanden auf unterschiedliche Weisen dargestellt werden. So stellten Gehrig und Breu (2013, S. 55) fest, dass von der Informationsdarbietung die Entscheidung abhängt, ob ein Projekt angenommen oder abgelehnt wird. „Wenn eine Investitionsrechnung mit einer Erfolgschance von 50% errechnet und präsentiert wird, so wird das Projekt mit einer höheren Wahrscheinlichkeit angenommen, als wenn die Angabe einer Misserfolgschance von 50% aufgeführt wird" (ebd., S. 50). Schweizer (2005) untersuchte u. a. den Framing-Effekt („Darstellungseffekt"), indem er den Richterinnen und den Richtern die Fragestellung eines bestimmten Sachverhalts aus der Sicht der Klägerin oder der Beklagten darlegte. In einem Prozess zwischen zwei Baufirmen, in dem eine Firma eine andere auf die Zahlung eines Gewinnanteils aus einem gemeinsamen Bauprojekt verklagte, wurden Richter, die den Fall aus der Sicht der Klägerin (Gewinn-Gruppe) beurteilen sollten, gefragt, ob die Klägerin ein Drittel der verlangten Summe von der Beklagten akzeptieren und dafür die Klage zurückziehen sollte (ebd., S. 95). Die andere Gruppe der Richter (Verlust-Gruppe) wurde gefragt, ob die Beklagte der Klägerin zwei Drittel der Summe zahlen sollte, wodurch der Prozess erledigt wäre (ebd.). In dieser Situation ist es „wichtig zu sehen, dass die Alternativen wirtschaftlich identisch sind. Die Parteien streiten sich über die Aufteilung eines fixen „Kuchens", nämlich des Gewinns aus dem gemeinsamen Projekt" (ebd., S. 96), wobei in beiden vorgestellten Fällen den beiden Parteien nach Abzug der Gerichtskosten jeweils die gleiche Summe bliebe (ebd.). Die Ergebnisse dieser Studie haben gezeigt, dass die dem Framing-Effekt unterliegenden Richter eher der Klägerin zum Vergleich rieten als der Beklagten, was sich in der Mehrheit der Fälle bestätigte (ebd., S. 97).

Um die Existenz der in Bezug auf den **Halo-Effekt** beschriebenen Ankerheuristik (s. Kapitel 2, Abschnitt 2.1.2.2.1, Seite 31) nachzuweisen, wurden in einem Experiment die Probanden in zwei Gruppen aufgeteilt und gefragt, wie hoch diese den Anteil der afrikanischen Staaten unter den Mitgliedern der UNO schätzen würden (Schneider, 2013, S. 20). Per Zufall wurden für beide Gruppen Werte bestimmt, an die sie sich insofern orientieren sollten, dass sie zuerst angeben mussten, ob sich der gesuchte Anteil unter oder über der gegebenen Zahl befand und sich dann auf einen konkreten Wert festlegen (ebd., S. 20f.). Pollock (2012) wies den Halo-Effekt nach, indem den 25 männlichen

Probanden 30 Facebook-Bilder von Frauen gezeigt wurden, wobei zu jedem Bild fünf Fragen beantwortet werden mussten. Anhand der gestellten Fragen wurde der Zusammenhang der Eigenschaften „Attraktivität“ und „Promiskuität“ festgestellt.
Ein Beispiel mit Einschätzungen als eine mögliche Form der Identifikation kognitiver Effekte wurde bereits bei der Aufstellung ihrer Definition (Kapitel 2, Abschnitt 2.1.1.2, Seite 18) aufgezeigt, indem die Schüler ihre eigenen Leistungen einschätzen sollten (**Dunning-Kruger-Effekt**).
Nicht nur Beschreibungen der Sachverhalte, sondern auch das Bild- und Videomaterial sowie die Simulationen authentischer Situationen dienen als Input für die Probanden, um die Reaktionen auf die zu untersuchenden Effekte hervorzurufen. „For example, when children in a video were given a label of high or low socioeconomic status, people used that label to make judgements of future academic ability, disregarding other relevant information“ (Darley & Gross, 1983, zitiert in Hernandez und Preston (2012, S. 1)) (**Bestätigungsfehler**).
Auch für empirisch ermittelte Effekte wie **Diskriminierung** sind einige Forschungsarbeiten auffindbar. Im Rahmen einer Studie im Forschungsbereich für Integration und Migration wurden Diskriminierungen der Jugendlichen mit Migrationshintergrund im Vergleich zu deutschen Jugendlichen in der Bewerbungsphase untersucht (vgl. Schneider et al., 2014). In einem Test-Verfahren, bei dem diskriminierungsrelevante Situationen simuliert wurden, überprüfte man, ob bei gleichen Voraussetzungen der Testpersonen eine Ungleichbehandlung bzw. eine Diskriminierung vorlag (ebd., S. 14). Ein Beispiel eines solchen Verfahrens waren jeweils zwei konkurrierende Bewerber auf eine gleiche Stelle, die im Hinblick auf ihre Sprachkenntnisse, Qualifikationen etc. eine absolut gleiche Eignung für die entsprechenden Stellen aufwiesen, die Unterschiede bestanden lediglich in den Namen der Bewerber und ihrer Staatsangehörigkeiten (deutsch bzw. türkisch) (ebd.). Durch die Ergebnisse dieser Studie wurden erstmals Diskriminierungen anhand der Differenz zwischen den Rückmeldungen der Arbeitgeber für die Bewerber mit einem deutschen Namen und denen für die Bewerber mit einem türkischen Namen belegt (ebd., S. 20). Türkische Bewerber erhielten seltener Rückmeldungen potentieller Arbeitgeber, sie wurden auch seltener zu einem Vorstellungsgespräch eingeladen und bekamen mehr Absagen als ihre deutschen Mitstreiter (ebd., S. 24). Auch die Häufigkeit und die Form der Rückmeldungen sowie einige andere Faktoren wurden zur Interpretation der Ergebnisse herangezogen (ebd., S. 25ff.). Einer von diesen Faktoren waren die Interpretationen

der Anreden schriftlicher Rückmeldungen (ebd.). „Während sie die Bewerber mit deutschen Namen bei der Rückmeldung auf ihre Bewerbung [...] eher mit Nachnamen adressierten, wurden Bewerber mit einem türkischen Namen häufiger geduzt und mit Vornamen angeredet“ (ebd., S. 26). Auf die Methode der Textinterpretationen griffen ebenfalls Schneider und Bauhoff (2013) zurück, indem sie 332 Stellenanzeigen der *Frankfurter Allgemeine Zeitung* auf die Diskriminierungen wegen des Alters und Geschlechts untersuchten. In der Datenauswertung hielten sie sich streng an die Formulierungen aus Rechtsprechung und Kommentarliteratur (ebd., S. 63). Einige solcher Formulierungen sind in Beispielen (3.1) und (3.2) aufgeführt.

(3.1) Geschlecht (nicht diskriminierende Ausdrücke):

(a) „Sachbearbeiter / Sachbearbeiterin“

(b) „Krankenschwester (m / w)“

(c) „Reinigungskraft“

(3.2) Alter (diskriminierende Ausdrücke):

(a) „Sie sind zwischen ... Jahre und ... Jahre alt“

(b) „Wir bevorzugen Mitarbeiter bis ... Jahre“

(c) „junge dynamische Führungskraft“

(d) „erfahrener alter Hase“

3.1.1.2 Besonderheiten computerlinguistischen Vorgehens

Durch die Darstellung einiger sozialpsychologischer Experimente im vorigen Abschnitt kristallisieren sich bei der Identifikation kognitiver Effekte die Vorgehensweisen heraus, die man auf folgende Weise zusammenfassen kann:

- Einschätzungen anhand dargebotenen Informationen
- Simulationen authentischer Situationen
- Textinterpretationen

Im Sinne der obigen Zusammenfassung und im Hinblick auf die im Abschnitt 1.2 formulierten Ziele der vorliegenden Arbeit kann man die Vorgehensweise als eine experimentelle Studie begreifen, die auf eine automatische Identifikation mehrerer kognitiver Effekte in einer konkreten Situation (und Domäne) anhand der schriftlich formulierten Bewertungstexte abzielt. Wenn man die ersten beiden formulierten Vorgehensweisen sozialpsychologischer Experimente betrachtet, so haben diese Möglichkeiten, die beim computerlinguistischen Vorgehen ausgeschlossen sind. Die den Probanden anhand der im Voraus formulierten Texte, ausgewählten Bilder oder Videos dargebotenen Stimuli mit dem Ziel, eine kognitive und emotionale Informationsverarbeitung (s. Abschnitt 2.1.1.3, Seite 22f.) anzuregen, um so zu erwartbare Reaktionen in Form von Meinungen, Einstellungen, Einschätzungen hervorzurufen, können im Kontext der automatischen Textverarbeitung nicht gewährleistet werden. Die Möglichkeiten einer kontrollierten Vorarbeit sind hier begrenzt. Hinweise auf kognitive Effekte können nur aus den vorliegenden Daten, d. h. aus den bereits erfolgten Meinungen automatisch herausgelesen werden. Was bedeutet, dass man sich bei der automatischen Identifikation der Effekte auch zeitlich im Stadium der Textinterpretationen befindet, wie diese z. B. bei der Auswertung der Rückmeldungen von potentiellen Arbeitgebern im Vorgehen mit den Diskriminierungen erfolgten. Das Hervorrufen der Reaktionen, auf die die Formulierungen von Meinungen resultieren, die Stimuli erfolgen in einer authentischen Situation, die jedoch dem Computerlinguisten verborgen bleibt. Eine der Aufgaben eines computerlinguistischen Vorgehens besteht in diesem Sinne in der Interpretation der Ergebnisse, die in Form der Patientenbewertungen vorliegen. Anhand der im Abschnitt 2.1.2.1 (Seite 28) genannten Selektionskriterien und deren konkreten graphischen Erscheinungen (s. auch Kriterienkataloge im Kapitel 4, Abschnitt 4.3, Seite 118) in den Bewertungstexten sollen Effekte klassifiziert werden, was als zweite computerlinguistische Aufgabe zu betrachten ist. Zur Verdeutlichung der Kontraste bei entsprechenden Ausgangssituationen sowie bei Identifikationsmöglichkeiten kognitiver Effekte in sozialpsychologischen Studien und im Rahmen computerlinguistischer Verfahren, welches in den nächsten Abschnitten näher betrachtet wird, werden einige Punkte dazu in der Tabelle 3.1 zusammengefasst[55].

[55]Die ausgewählten Effekte der Studien werden dabei in Klammern aufgeführt.

	sozialpsycholog. Studien	**computerling. Verfahren**
1. Ausgangssituation: Meinungsbildung	• Bekämpfung einer Krankheit (Framing) • Streitprozess beim Gericht (Framing) • Eigenschaften der Frauen (Halo) • Bewerbungen der Jugendlichen mit Migrationshintergrund (Diskriminierung)	• Besuch einer Arztpraxis
2. Probanden	• männliche Probanden (Halo) • Jugendliche mit Migrationshintergrund (Diskriminierung)	• Patienten
3. Das auszuwertende Material	• vorgegebene Fragen nach einer Skala (Halo) • Texte mit Stellenanzeigen (Diskriminierung) • Rückmeldungen von Arbeitgebern (Diskriminierung)	• frei formulierte Bewertungstexte
4. Erkennungsmöglichkeiten	• Interpretationen der Lösungen von Probanden (Framing, Halo) • Linguistische Muster (Diskriminierung)	• Linguistische Muster • Metadaten der Textsorte

Tabelle 3.1: Ausgewählte Aspekte der Ausgangssituationen in sozialpsychologischen Studien und bei computerlinguistischen Verfahren im Vergleich

3.1.2 Verfahren aspektbasierter Stimmungsanalyse und domänenspezifischer Informationsextraktion

Wie in Tabelle 3.1 sichtbar, wird man vor dem Eintritt der Effekte mit einer Situation konfrontiert, die den Prozess einer Meinungsbildung auslöst. Anhand einiger Kriterien kann man die bei der Meinungsbildung produzierten kognitiven Effekte erkennen. Daraus folgt, dass die zentrale Aufgabe der Identifikation eine automatische Extraktion von Meinungen bzw. Stimmungen zu vordefinierten Objekten (hier z. B. Bewertungsdimensionen) ist. Aus diesem Grund beschäftigt sich dieser Abschnitt mit den Verfahren automatischer Textverarbeitung, wobei das Augenmerk auf die Stimmungsanalyse und Informationsextraktion gelegt wird.
Der Prozess der Stimmungsanalyse wurde im Kapitel 2, Abschnitt 2.2.2.3, Seite 49f. beschrieben, wobei in Bezug auf die vorliegende Arbeit festgestellt wurde, dass in deren Rahmen eine aspekt- bzw. phrasenbasierte Stimmungsanalyse relevant ist. Die Phrasen innerhalb der Sätze beziehen sich dabei einerseits auf die Aspekte (Bewertungsdimensionen), zu denen Stimmungen bzw. Meinungen ausgedrückt werden. Andererseits sind es typische Phrasen, die sprachliche Muster zu zwei konkreten Effekten (Diskriminierungen und Bestätigungsfehler) darstellen. Aufgrund der eben angedeuteten Tatsache, dass für die vorliegende Arbeit die Meinungen zu einer begrenzten Anzahl der voneinander unterscheidbaren Bewertungsobjekte oder Entitäten (s. auch Begründung eingeführter Einschränkungen in der Einleitung (Abschnitt 1.2, Seite 6) interessant sind, muss ebenfalls auf die hier relevanten Verfahren der Informationsextraktion eingegangen werden. Einen ausführlichen Überblick von aktuellen Methoden und Arbeiten der Stimmungsanalyse gibt u. a. Medhat et al. (2014, S. 1095ff.). Die Verfahren der Stimmungsanalyse werden in der wissenschaftlichen Literatur unterschiedlich klassifiziert, wobei die wesentlichen von ihnen die lexikonbasierten, die maschinellen und die Kombination aus beiden genannten, die hybriden Verfahren sind (ebd., S. 1098; Broß, 2013, S. 22f.; Shailesh, 2015, S. 117; Bretschneider, 2015, S. 3f.). Was die domänenspezifische Informationsextraktion betrifft, so existieren hier dieselben Verfahren, die entsprechend der Aufgabenspezifizierung (s. Kapitel 2, Abschnitt 2.2.1.2, Seite 41f.) eingesetzt werden.

3.1.2.1 Lexikonbasierte Verfahren

Was mit der Polarität einer Meinung gemeint ist, wurde im Kapitel 2, Abschnitt 2.2.2.2.2, Seite 47 erläutert. ‚Opinion words' (Polaritätswörter) als mögliche Indikatoren der Polarität wurden im Abschnitt 2.2.2.2.2 (Seite 48) definiert. Da ‚opinion words' bei vielen Stimmungsklassifikationsaufgaben herangezogen werden (Liu, 2010, S. 14), behandelt dieser Abschnitt lexikonbasierte Verfahren und setzt sich mit den Ansätzen der Lexika-Erstellung auseinander, die den o. g. Verfahren zugrunde liegt. „*Lexikonbasierte Verfahren* schlagen alle oder nur bestimmte Worte eines Eingabetextes in einem vorher angelegten Lexikon nach" (Bretschneider, 2015, S. 3). Anhand der Polarität der im Lexikon kodierten ‚opinion words' kann man auf die Polarität einer Meinung schließen (s. Kapitel 2, Abschnitt 2.2.2.2.2, Seite 48) bzw. auf die semantische Orientierung eines Textes (Wolfgruber, 2015, S. 40). Was die Entitäten betrifft, so bezieht sich dies in der vorliegenden Arbeit auf die Extraktion der Ärzte-Namen und Fachvokabulars sowie die Anfertigung eines externen Lexikons mit Wörtern zur Nationalität. Lexikonbasierte Verfahren kann man in manuelle, wörterbuchbasierte und korpusbasierte Verfahren aufteilen (ebd.).

3.1.2.1.1 Manuelle Verfahren

Die manuelle Klassifikation der Polaritätswörter ist sehr aufwendig (Bretschneider, 2015, S. 4; Wolfgruber, 2015, S. 40). Als eine Erweiterung eines domänenspezifischen Lexikons und in einer zu anderen Verfahren unterstützenden Funktion können jedoch manuelle Verfahren zu einer Verbesserung der Ergebnisse beitragen (Wolfgruber, 2015, S. 40).

3.1.2.1.2 Wörterbuchbasierte Verfahren

Diese Verfahren basieren auf der „Bootstrapping"-Methode, indem von einer kleinen Seed-Liste der bekannten Polaritätswörter ausgegangen wird, die mittels Online-Wörterbüchern (z. B. WordNet) um ihre Synonyme und Antonyme erweitert werden (Liu, 2010, S. 15; Wolfgruber, 2015, S. 40f.). Die Methode „Bootstrapping" impliziert dabei, dass auf die beschriebene Weise so lange vorgegangen wird, bis keine neuen Wörter gefunden werden (Wolfgruber, 2015, S. 41). Auf diese Methode wird zusätzlich im Rahmen des in der vorliegenden Arbeit gewählten Verfahrens im Kapitel 4 (Abschnitt 4.2.1.3,

Seite 110) eingegangen. Der Einsatz von „Bootstrapping“ in der vorliegenden Arbeit wird im o. g. Kapitel konzipiert und im Kapitel 5 (Abschnitt 5.1.2.2.1, Seite 137) ausführlicher erläutert. Aufgrund des Nachteils der wörterbuchbasierten Verfahren, keine domänenspezifischen Polaritätswörter finden zu können (Liu, 2010, S. 15; Wolfgruber, 2015, S. 41), wird in der vorliegenden Arbeit auf folgende Weise vorgegangen: statt einer kleinen Seed-Liste wird eine Ressource (*SentiWS*) verwendet, auf deren Basis korpusbasierte Akquise angestrebt wird, um den o. g. Nachteil ausgleichen zu können (Liu, 2010, S. 15). Im aktuellen Abschnitt werden daher zwei in der vorliegenden Arbeit verwendete Lexika beschrieben. Das erste umfassende Lexikon *CISLEX* wurde für diese Arbeit zur Verfügung gestellt und wird mit Hilfe von dem zweiten Lexikon *SentiWS* um Polaritätswörter erweitert (s. Kapitel 5, Abschnitt 5.1.2.2.1, Seite 137). Zur Bestimmung relevanter Kontexte der Bewertungsobjekte, die zum Aufbau wertender Phrasen verwendet werden, wird „Bootstrapping“-Methode ebenfalls eingesetzt, worauf im Kapitel 4, Abschnitt 4.2.1.3 kurz eingegangen wird.

a) CISLEX

CISLEX ist ein Wörterbuch, das am Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwig-Maximilians-Universität München (LMU) entwickelt wurde. Mit dem Aufbausystem von CISLEX wurde versucht, ein vollständiges theorieneutrales elektronisches Wörterbuch der deutschen Sprache zu erstellen (Guenthner und Maier, 1994, S. 1). Den Prinzipien der Vollständigkeit, der Abgeschlossenheit und der Korrektheit folgend, wurde eine Klassifizierung beobachtbarer lexikalischer Entitäten in Objektklassen vorgenommen, wobei man von einer wichtigen Grundmenge der einfachen Formen ausging, für die flektierte Formen, Komposita und andere morphologische pragmatisch relevante Regularitäten ausgearbeitet wurden (ebd., S. 1f.). Die Einträge des Lexikons enthalten einfache und komplexe Formen (ca. 150 000), Eigennamen, Fremd- und Fachwörter, Kurzformen (ebd., S. 4). Die einfachen Formen sind nach morphologischen (Flexionsverhalten) und syntaktischen (Wortarten) Kriterien klassifiziert (ebd., S. 7). Flektierte Wortformen wurden aus Grundformen mittels eines Prologprogramms (funktionale Programmierung) ermittelt (ebd., S. 4).
Außer Tagging ist eine der typischen Anwendungen von CISLEX die maschinelle Übersetzung, so dass die Kodierung des Wörterbuchs „bewusst in Übereinstimmung mit den entsprechenden Wörterbüchern fürs Französi-

sche, Englische, Italienische, Spanische, Portugiesische usw.“ (ebd., S. 13) erfolgte, „die entweder direkt am LADL in Paris oder in den entsprechenden Ländern unter Regie des LADL entwickelt wurden“ (ebd.). Was die Wortarten-Klassifizierung angeht, so sind die häufigsten grammatischen (syntaktischen) Kategorien in der Tabelle 4.1 (Kapitel 4) aufgeführt.

b) SentiWS

SentimentWortschatz oder SentiWS ist ein deutschsprachiges öffentlich zugängliches Wörterbuch, das im Rahmen eines Projekts der Universität Leipzig entwickelt wurde (Remus et al., 2010, S. 1168; Remus und Ahmad, 2010, S. 27). Das Wörterbuch enthält 1 650 positive und 1 818 negative Grundformen der Wörter[56], was mit flektierten Formen 16 406 positive und 16 328 negative Wörter ausmacht (ebd. Remus et al., 2010). Ein Wörterbuch-Eintrag enthält somit ein Wort, deren syntaktische Kategorie, Flexionsformen, Polarität und deren Stärke (ebd.). Die Wörter „wurden per Google translate (`http://translate.google.com`) halb automatisch, halb manuell ins Deutsche übersetzt, überprüft und anschließend um ihre Flexionsformen erweitert“ (Remus und Ahmad, 2010, S. 25).
Die Polaritäten der Wörter wurden mittels *Pointwise Mutual Information (PMI)* ermittelt (vgl. Turney, 2002), wobei bei dieser Methode die semantische Orientierung (SO) eines Wortes aus der *semantischen Assoziation* abgeleitet wird (Remus et al., 2010, S. 1169). D. h.: Die (SO) eines gegebenen Wortes w wird mittels der Differenz der Stärke seiner Assoziation (A) mit dem manuell zusammengestellten Set von positiven Wörtern (P) und der Stärke seiner Assoziation mit dem Set von negativen Wörtern (N) berechnet (ebd.) (s. Formel 3.1). Fällt diese Differenz (Stärke der semantischen Orientierung des Wortes) positiv aus, hat das Wort positive SO und umgekehrt. (ebd.).

$$\text{SO-A(w)} = \sum_{p \in P} A(w,p) - \sum_{n \in N} A(w,n) \tag{3.1}$$

Die semantischen Assoziationen *A(w,p)* und *A(w,n)* werden anhand von PMI berechnet (s. Formel 3.2). *P(w)* bedeutet dabei die Wahrscheinlichkeit des Auftretens eines Wortes w und *P(w_1,w_2)* die Wahrscheinlichkeit des gemeinsamen Auftretens der Wörter w_1 und w_2 (Wolfgruber, 2015, S. 42). Wenn

[56] `http://asv.informatik.uni-leipzig.de/download/sentiws.html` (04.10.2015)

Wörter statistisch unabhängig sind, dann ist die Wahrscheinlichkeit ihres gemeinsamen Auftretens durch das Produkt deren Einzelwahrscheinlichkeiten $P(w_1) \cdot P(w_2)$ gegeben (Turney, 2002, S. 419). Das Verhältnis zwischen $P(w_1,w_2)$ und $P(w_1) \cdot P(w_2)$ stellt daher ein Maß des statistischen Abhängigkeitsgrades zwischen den Wörtern dar (ebd.).
Die Gewichte der Wortpolaritäten liegen im Interval [-1;1] (Remus et al., 2010, S. 1168), wobei -1 absolut negative und 1 absolut positive Werte implizieren (ebd., S. 1170). Als Beispiele für stark positive Wörter kann man „Freude“ (Gewicht 0,6502) und „perfekt“ (Gewicht 0,7299) und für stark negative „betrügen“ (Gewicht -0,743) und „schädlich“ (Gewicht -0,9269) benennen (ebd.).

$$PMI(w_1, w_2) = \log_2 \left(\frac{P(w_1, w_2)}{P(w_1) \cdot P(w_2)} \right) \tag{3.2}$$

3.1.2.1.3 Korpusbasierte Verfahren

Die Idee der korpusbasierten Verfahren beruht auf syntaktischen Mustern, deren Ausgangspunkt eine kleine Seed-Liste (s. Abschnitt 3.1.2.1.2, Seite 61) von z. B. Polaritätswörtern ist, mit dem Ziel, andere Polaritätswörter in einem großen Korpus zu finden (Liu, 2010, S. 15), ohne dabei externe Ressourcen zu nutzen (Vázquez et al., 2012, S. 1272). Auch neue Entitäten wie z. B. Namen können auf diese Weise korpusbasiert gefunden werden, wenn man von einer kleinen Liste der bereits bekannten Namen ausgeht und deren relationalen Beziehungen berücksichtigt. In dieser Arbeit gelten die Bewertungsdimensionen als Entitäten / Objekte (s. Kapitel 2, Abschnitt 2.2.1.2.2), deren explizite Benennungen manuell erstellt werden. In der Stimmungsanalyse der Hotelbewertungen wendet Wolfgruber (2015, S. 66ff.) ein korpusbasiertes Verfahren zur Akquise der Adjektive und anderer Wortarten an, woraufhin deren manuelle Analyse erfolgt. Dabei verzichtet sie auf Lexika mit Polaritätswörtern aufgrund der orthographischen Fehler der gewählten Textsorte (ebd., S. 63). Die Adjektive wurden „nach dem Muster <ADJ> (= alle Adjektive im Text)“ (ebd.) extrahiert, manuell analysiert und in zwei Gruppen (positiv und negativ) aufgeteilt (ebd.). In der vorliegenden Arbeit wird ebenfalls eine korpusbasierte Akquise von Adjektiven durchgeführt. Als ‚Start-Set‘ der Adjektive wird allerdings nicht die manuell erstellte Liste polaritätsspezifischer Adjektive verwendet, sondern die positiven und die negativen Adjektive des Lexikons SentiWS (s. Abschnitt 3.1.2.1.2, Seite 63). Wie

die Gewinnung weiterer neuer Adjektive aus dem Korpus erfolgt und welche Ergebnisse dabei erzielt wurden, wird im Kapitel 5 (Abschnitt 5.1.2.2.1, Seite 137ff.) beschrieben.
Eine ältere, jedoch als Grundlage für eine Reihe von aktuellen Forschungen verwendete Arbeit, die die Akquise von Adjektiven betrifft, wurde von Hatzivassiloglou und McKeown (1997) durchgeführt, wobei sie im Ergebnis über 90% Genauigkeit erreichten (ebd., S. 174). Mit dem anfänglich konstruierten Set von Adjektiven verschiedener Polarität wurde nach weiteren neuen Adjektiven in einem aus 21 Millionen Wörter bestehenden Korpus anhand der verbindenden Konjunktionen gesucht (ebd., S. 176), was aus der Sicht der Informationsextraktion als Relationen zwischen den Adjektiven (Entitäten) interpretiert werden kann. Einige der Annahmen dieser kontextgebundenen Suche implizierten z. B., dass die Adjektive, die mit Konjunktion AND miteinander verbunden waren, die gleiche semantische Orientierung, während diejenigen, die mit BUT kombiniert wurden, die gegensätzlichen Polaritäten aufwiesen (ebd.). Mit dieser Vorgehensweise und der Verwendung weiterer Konjunktionen wie OR, EITHER-OR, NEITHER-NOR erstellten sie ein Set von 657 positiven und 679 negativen Adjektiven (ebd., S. 175). Wie bei wörterbuchbasierten, so auch bei korpusbasierten Verfahren wird die Methode „Bootstrapping“ verwendet (Wolfgruber, 2015, S. 46f.). Auf der eben beschriebenen Arbeit aufbauend, verwendeten Vázquez et al. (2012) die „Bootstrapping“, um neue domänenspezifische Adjektive zu gewinnen. Ausgehend von einer kleinen Seed-Liste mit 28 positiven und 7 negativen Adjektiven suchten sie nach deren Verbindungen mit anderen Adjektiven mit Konjunktionen Y (‚und‘) und PERO (‚aber‘) (Vázquez et al., 2012, S. 1275f.). Trotz der erzielten hohen Genauigkeit der Ergebnisse für positive Adjektive (97,6%), betrug die Abdeckung der Adjektive lediglich 67% (ebd., S. 1278). Außerdem wurden sowohl Hatzivassiloglou und McKeown (1997) als auch Vázquez et al. (2012) mit dem Problem der ambigen Adjektive konfrontiert. Im Beitrag zu Polaritätswörtern in Lexika stellten Vázquez und Bel (2013) fest, dass mehr als die Hälfte aller Adjektive (67,32%) eine domänenabhängige Polarität haben. Dabei setzten sie sich mit 514 Adjektiven aus drei domänenspezifischen Korpora für Autos, Mobiltelefone und Filme auseinander (ebd., S. 3557). Am Beispiel des Adjektivs „small“ kann man nicht nur Domanenspezifik der Adjektivpolarität, sondern auch die Individualität der Sichtweise verschiedener Menschen, was auch innerhalb einer Domäne erfolgen kann, nachvollziehen: „It will be considered by a majority of people as a good characteristic for mobile phones, however if someone ask if having a „small car“ is good, so-

me people will answer yes and others will answer no. Moreover, if we also ask about a „small film“ probably people won't know how to answer to our question“ (ebd.). Das Problem der Domänen- und Aspektspezifik ist auch in der vorliegenden Arbeit aktuell. Wie mit diesem umgegangen wird, wird im Kapitel 5 (Abschnitt 5.2.1.2, Seite 154f.) thematisiert.
Die vorgeschlagenen relationalen Konjunktionsverbindungen zwischen den Adjektiven sind nicht in allen Phrasen konsistent (Liu, 2010, S. 15; Ding et al., 2008, S. 6). In Bezug auf die Bewertung einer Ärztin und der Wartezeiten im gleichen Satz (s. Beispiel (3.3)) haben zwei mit der Konjunkiton ‚und‘ verbundene Adjektive unterschiedliche Polaritäten.

(3.3) „Sie ist grundlegend sehr gut und lange Wartezeiten bei Frauenärzten sind ja überall normal.“

Bei der Erweiterung des oben beschriebenen Konzepts von Hatzivassiloglou und McKeown (1997) auf der Satzebene schlugen Ding et al. (2008) die Erfassung der Polaritätswörter zusammen mit Objekteigenschafen vor. Sie durchsuchten jeden Satz nach allen Bewertungsobjekten und kalkulierten für jedes Objekt einen Orientierungswert (semantische Orientierung) anhand der im gleichen Satz vorhandenen Objekteigenschaften, wobei der Abstand zwischen jedem Objekt und jeder Eigenschaft berücksichtigt wurde (ebd., S. 4f.). Geierhos et al. (2015b, S. 7) arbeiteten ebenfalls mit Kombinationen aus den Objekten und Objekteigenschaften und entwickelten in Arztbewertungen eine Reihe von syntaktischen Mustern wie ‚<A> Praxis‘, wobei <A> ein Adjektiv wie z. B. „unordentlich“ impliziert (ebd.). Wie bereits mehrmals erwähnt, ist man auf gewisse Einschränkungen angewiesen, um den gestellten Zielen gerecht zu werden. Wie bekannt, beschränkt sich die Extraktion der Bewertungsobjekte auf die vordefinierte Dimensionen des Jameda-Portals. Die Extraktion von diesen Dimensionen erfolgt, wie in den letzten genannten Arbeiten dieses Abschnitts beschrieben, nicht isoliert, sondern in Bezug auf den Kontext, was ausführlicher im Rahmen der Entwicklung lokaler Grammatiken im Kapitel 5 (Abschnitt 5.2, Seite 148ff.) erläutert wird. Bei der Akquise der Polaritätswörter muss das Vorgehen nicht ausschließlich auf der syntaktischen Ebene erfolgen, indem man sich mit den zusammen auftretenden Wörtern und deren Verbindungen auseinander setzt. So kann man auch auf der morphologischen Ebene mittels wortbildenden Elemente korpusbasiert nach neuen Wörtern suchen. Mikheev (1996) hat die Technik von „guessing strategies“ angewandt, um neue Wörter für Part-of-Speech-Tagging zu

erweitern. Dabei wurden „guessing rules“ für Präfixe, Suffixe und Endungen der Wörter entwickelt, wobei von den bekannten Wörtern im Lexikon ausgegangen wurde. So impliziert z. B. die Regel

$$A^p : [un\ (VBD \quad VBN) \quad (JJ)], \tag{3.3}$$

dass ein unbekanntes Wort, das Präfix „un“ in einem aus dem Lexikon als „a past verb and participle“ bekannten Verb aufweist, als Adjektiv interpretiert wird (ebd., S. 770). Außer dass die beschriebene Regelbildung ziemlich gute Ergebnisse erzielte (ebd., S. 774), zeigt auch die Studie in einem anderen Kontext, dass der Lernprozess der unbekannten Wörter mit einer Kontextumgebung bei Menschen erfolgreicher als das Lernen isolierter Wörter ist (vgl. Vakili Samiyan, 2014).
Die korpusbasierte Akquise der Polaritätswörter erfolgt auf der syntaktisch-semantischen und morphologischen Ebene auch in der vorliegenden Arbeit und wird im Kapitel 5, Abschnitt 5.1.2.2.1, Seite 137ff. beschrieben.

3.1.2.2 Maschinelle Lernverfahren

Die Methoden der Textklassifikation, die auf maschinellen Lernverfahren basieren, kann man grob in überwachtes und unüberwachtes Lernen einteilen (Shailesh, 2015, S. 118). Überwachtes Lernen, „because a supervisor (the human who defines the classes and labels training documents) serves as a teacher directing the learning process“ (Manning et al., 2009, S. 256). Die überwachten Lernverfahren arbeiten mit einer großen Menge an annotierten Trainingsdaten, während die unüberwachten dann verwendet werden, wenn solche Daten schwer zu finden sind (Shailesh, 2015, S. 118). „No supervision means that there is no human expert who has assigned documents to classes“ (Manning et al., 2009, S. 349). Im Kontext der Stimmungsanalyse stammen die Trainings- und Testdaten üblicherweise aus dem UGC (Shailesh, 2015, S. 115; Bretschneider, 2015, S. 4). Auf der Dokument-Ebene (s. Kapitel 2, Abschnitt 2.2.2.3.2, Seite 50) kann man die Textklassifikation als binäre Aufgabe mit z. B. zwei Kategorien positiv und negativ betrachten (Wolfgruber, 2015, S. 47; Liu, 2010, S. 10). Bei der Relationsextraktion werden in diesem Zusammenhang die Entitäten danach klassifiziert, ob sie eine Instanz einer gesuchten Relation bilden (Stotz, 2018, S. 53). Unter den überwachten Lernverfahren haben sich Klassifikatoren wie Support Vector Machines (SVM), Naive Bayes (NB) und Maximum Entropy (ME) durchgesetzt (Bretschneider, 2015, S. 4). Zu den unüberwachten Verfahren zählen z. B. Hidden Markov

Models (HMM) und Conditional Random Fields (CRF) (Wolfgruber, 2015, S. 50). Trotz deren großen Verbreitung auf dem Gebiet der Stimmungsanalyse (Bretschneider, 2015, S. 3), besteht einer der Nachteile maschineller bzw. statistischer Lernverfahren darin, dass sie nicht „im Detail analysieren, welches Objekt zu welchem Sentiment gehört“ (Wolfgruber, 2015, S. 48). Da in der vorliegenden Arbeit die Stimmungsklassifikation auf der Aspekt-Ebene erfolgen soll (s. Kapitel 2, Abschnitt 2.2.2.3.2, Seite 50), wird in weiteren Abschnitten im genannten Kontext hauptsächlich aspektbasiert auf überwachte und unüberwachte Lernverfahren eingegangen.

3.1.2.2.1 Überwachte Lernverfahren

Der NB-Klassifikator bestimmt die Wahrscheinlichkeiten dafür, dass ein Dokument (d) zu einer bestimmten Klasse (c) gehört (Wolfgruber, 2015, S. 49; Linke, 2003, S. 151). Das Dokument oder der Text wird dann der Klasse mit der höchsten Wahrscheinlichkeit zugeordnet (Linke, 2003, S. 151). Diesem Klassifikator liegt die Bayes-Formel (Formel 3.4) zugrunde, die sich mit bedingten Wahrscheinlichkeiten befasst (ebd., S. 152).

$$P(c \mid d) := \frac{P(c \wedge d)}{P(d)} \tag{3.4}$$

Dabei bedeutet $P(c \mid d)$ die Wahrscheinlichkeit dafür, dass die Klasse c dem Dokument d zugeordnet werden kann. Mit anderen Worten: Wie wahrscheinlich ist es, dass wenn d wahr ist, dass auch c wahr ist. $P(d)$ ist die Wahrscheinlichkeit für d und $P(c \wedge d)$ die Wahrscheinlichkeit für gemeinsames Auftreten von c und d (ebd.). Da $P(c \wedge d) = P(d \mid c)$, lässt sich aus der Formel der bedingten Wahrscheinlichkeit 3.4 die Bayes-Formel ableiten (ebd.) (Formel 3.5). Da beim NB-Klassifikator angenommen wird, dass z. B. die Wörter im Text unabhängig voneinander auftreten, nennt man diesen ‚naiv‘ (Wolfgruber, 2015, S. 49).

$$P(c \mid d) := \frac{P(d \mid c) \cdot P(c)}{P(d)} \tag{3.5}$$

Der SVM-Klassifikator, der ebenfalls in der traditionellen Textklassifikation angewandt wird (ebd., S. 48), „is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data“ (Manning et al., 2009,

S. 319). Die Suche nach dieser sogenannten ‚Hyperebene', die in Form eines Vektors $\vec{w}$ dargestellt wird, ist mit einem Optimierungsproblem verbunden, bei dessen Lösung nach *Lagrange-Multiplikatoren* α_j's ≥ 0 (s. ausführlicher ebd., S. 324f.) gesucht wird, durch die Support-Vektoren $\vec{d}$'s bestimmt werden (Pang et al., 2002, S. 82). Die ‚Hyperebene', die die Dokumente mit möglichst größtem Abstand in zwei Klassen ($c_j \in \{1, -1\}$: positiv und negativ) voneinander trennt, ist dann auf folgende Weise darstellbar (ebd.):

$$\vec{w} := \sum_j \alpha_j c_j \vec{d_j}, \quad \alpha_j \geq 0 \tag{3.6}$$

Anders als bei der oben dargestellten Testklassifikation mit den vordefinierten Klassen wie z. B. Politik, Wissenschaft, Sport etc., sind in der Stimmungsklassifikation die ‚opinion words' relevant (Liu, 2010, S. 11). In diesem Kontext haben Pang et al. (2002) u. a. oben beschriebene überwachte Verfahren zur Klassifikation von Filmbewertungen in zwei Klassen (positiv und negativ) eingesetzt. Sie zeigten, dass bei der Abdeckungsrate zwischen 50% und 69% eine gute Genauigkeit der Ergebnisse mit dem Einsatz von Unigrammen erzielt werden kann. Jedoch mussten sie feststellen, dass die Klassifikation im Kontext der Stimmungsanalyse im Vergleich zur traditionellen Textklassifikation schwieriger zu bewältigen ist. Die vordefinierten Themen der Textklassifikation (s. o.) kann man oft allein durch die dazu gehörigen Schlüsselwörter erschließen, was bei den Stimmungen anders ist (ebd., S. 79). Im Beispiel (3.4) ist eine negative Bewertung dargestellt, die kein einziges negatives Wort enthält (ebd.).

(3.4) „How could anyone sit through this movie?"

Auch unterschiedliche Ebenen (s. Kapitel 2, Abschnitt 2.2.2.3.2, Seite 50) zeigen unterschiedliche Ergebnisse bei den überwachten Verfahren. Aufgrund einer kleinen Menge an Wörtern innerhalb eines Satzes enthalten die resultierenden Vektoren deutlich weniger Wörter als bei der Klassifikation auf der Dokumenten-Ebene (Wiegand und Klakow, 2009, S. 296). Die Ergebnisse können jedoch durch die zusätzliche Berücksichtigung verschiedener linguistischer Eigenschaften wie Informationen zu Wortarten, Hypernymie (Wortebene), gegenseitige Beeinflussung von Polaritätsphrasen (Satzebene), Negation u. ä. deutlich verbessert werden (ebd., S. 296ff.). Bei der aspektbasierten Stimmungsanalyse von Laptop- und Restaurantbewertungen auf der Satzebene unterteilten Bornebusch und Cancino (2014, S. 2389) das Klassifikationsproblem in vier Teilaufgaben:

1. Extraktion der Aspekte aus einem Satz
2. Bestimmung von Polaritäten der Aspekte
3. Ermittlung der Kategorien einzelner Sätze
4. Bestimmung der Polaritäten zu den ermittelten Kategorien

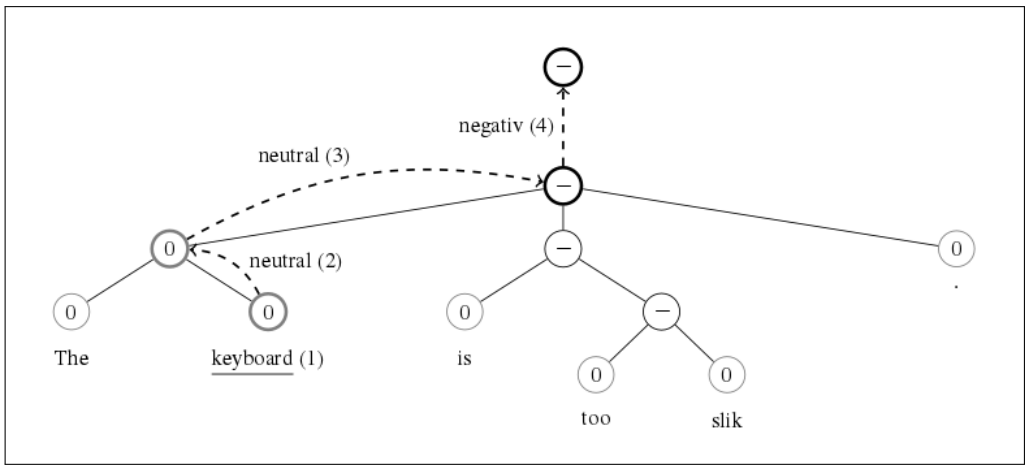

Abbildung 3.1: Beispiel des Sentiment-Tree-Algorithmus für den Satz "The keyboard is too slik." (Bornebusch und Cancino, 2014, S. 2395)

Dabei wurden unterschiedliche Klassifizierer überwachter Lernverfahren eingesetzt und die in Python implementierte Bibliothek *Natural Language Toolkit (NLTK)* verwendet (ebd., S. 2390). Die Bestimmung von Polaritäten der Aspekte z. B. erfolgte mit Hilfe von *Sentiment Trees*, die auf der Satzebene arbeiteten (ebd., S. 2395). Der Satz „The keyboard is too slik." wurde, wie auf der Abbildung 3.1 dargestellt, ausgewertet. Ausgehend von dem Hauptwort des Aspekts (keyboard (1)), wurde nach einer Polarität gesucht (2) (ebd.). Da sein Elternknoten eine neutrale Polarität aufwies (2), wurde die Polarität des Wurzelknotens untersucht (3) (ebd.). Aufgrund der negativen Polarität des Wurzelknotens (4) bekam der Aspekt ‚keyboard' eine negative Polarität zugeordnet (ebd.). Mit diesem Vorgehen wurden für die positiven Polaritäten beider Produkte etwas bessere Ergebnisse als für die negativen erreicht (F_1-Scores[57] für Laptop: 0,67 und 0,60; für Restaurant: 0,74 und 0,50).

[57] s. Kapitel 6, Abschnitt 6.1.3, Seite 187.

Durch die schwer zu beschaffenden annotierten Trainingskorpora existieren generell nur wenig Arbeiten, die auf den überwachten Lernverfahren basieren (Broß, 2013, S. 208). „Most approaches to clause or phrase level polarity detection are rule-based, using a sentiment lexicon“ (ebd.). Auf die regelbasierten Verfahren wird im Abschnitt 3.1.2.4 eingegangen, diese werden auch in der vorliegenden Arbeit eingesetzt und im Kapitel 5 (Abschnitt 5.2, Seite 148ff.) präsentiert. Die Stimmungsanalyse, die als Extraktion wertender Muster zu begreifen ist, wird dabei, ähnlich der Arbeit von Bornebusch und Cancino (2014) in mehreren Teilaufgaben aufgeteilt.

3.1.2.2.2 Unüberwachte Lernverfahren

Unüberwachte Lernverfahren stützen sich meistens auf die Lexika, die stimmungsrelevante Wörter und Phrasen beinhalten (ebd., S. 23). Einer der Nachteile der unüberwachten gegenüber den überwachten Lernverfahren ist ihre begrenzte Genauigkeit (ebd., S. 22; Wolfgruber, 2015, S. 50). Nicht immer verwenden unüberwachte Verfahren die Stimmungslexika (Broß, 2013, S. 24). Turney (2002) stellte ein unüberwachtes Lernverfahren zur Klassifikation der Bewertungen als ‚empfehlenswert‘ oder ‚nicht empfehlenswert‘ vor, wobei der Algorithmus in drei Schritten arbeitete:

- Zunächst wurden die Phrasen mittels eines Part-of-Speech-Taggers (POS-Tagger) extrahiert, die Adjektive oder Adverbien beinhalteten, weil sich diese Wortarten als gute Indikatoren für subjektive Sätze erwiesen haben (ebd., S. 417f.).

	First word	Second word	Third word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Tabelle 3.2: Muster für einen POS-Tagger zur Phrasenextraktion (Turney, 2002, S. 418)

Dabei wurde nicht mit isolierten Wörtern (z. B. Adjektiven wie bei Hatzivassiloglou und McKeown (1997)) gearbeitet, sondern mit Phrasen, die den ‚plausiblen' Kontext zu ‚opinion words' lieferten. In der Tabelle 3.2 sind POS-Tags aufgelistet, anhand derer die eben beschriebenen Phrasen gefunden werden sollten. Z. B. impliziert das Muster in der zweiten Reihe, dass zwei aufeinander folgende Wörter zu extrahieren sind, wenn das erste Wort ein Adverb und das zweite ein Adjektiv ist, das dritte Wort jedoch kein Nomen sein darf (ebd., S. 418).

- Im zweiten Schritt wurde die semantische Orientierung jeder extrahierten Phrase berechnet (ebd.). Dafür wurde der Algorithmus PMI-IR (Pointwise Mutual Information und Information Retrieval) zur Messung der Stärke von der semantischen Assoziation zwischen zwei Wörtern (ebd.) eingesetzt (s. auch Abschnitt 3.1.2.1.2, Seite 63). Die semantische Orientierung einer gegebenen Phrase (SO(phrase): s. Formel 3.7) ergab sich dann aus der Differenz des PMI-Wertes dieser Phrase mit dem Wort „excellent" und des mit dem Wort „poor" (ebd., S. 419), wobei diese ‚Referenzwörter' zwei Extreme der gegebenen Skala darstellten (ebd.).

$$SO(phrase) = PMI(phrase, „excellent") - PMI(phrase, „poor") \tag{3.7}$$

PMI-IR berechnet PMI, indem die Anfragen (queries) an die Suchmaschine aufgestellt und die Anzahl der gefundenen Dokumente (hits) verzeichnet werden (ebd.). Wenn man nach zwei Wörtern zusammen und getrennt sucht, kann man die Wahrscheinlichkeiten mit der Formel 3.2 ausrechnen (Liu, 2010, S. 12). Turney (2002) benutzte die AltaVista-Suchmaschine, weil diese über einen NEAR-Operator verfügt. Wenn hits(query) die Anzahl gefundener hits bei der Suchanfrage „query" ist, so kann man, ausgehend von der Formel 3.2, die Formel 3.7 auf folgende Weise umschreiben (ebd.):

$$SO(phrase) = \log_2 \left[\frac{hits(phrase\ NEAR\ „excellent")hits(„poor")}{hits(phrase\ NEAR\ „poor")hits(„excellent")} \right] \tag{3.8}$$

- Im letzten Schritt wurde der Durchschnitt der semantischen Orientierung von Phrasen in einer gegebenen Textbewertung ausgerechnet und die Bewertung als ‚empfehlenswert' klassifiziert, wenn der Wert positiv bzw. als ‚nicht empfehlenswert', wenn der Wert negativ ausgefallen ist (ebd.).

Die erzielten Ergebnisse variieren zum einen in Abhängigkeit von der Domäne, was z. B. auf die Beschreibungen von negativen Szenen in einer positiven Filmbewertung schließen lässt (ebd., S. 422). Während die Genauigkeit von lediglich 66% bei dem beschriebenen Vorgehen in der Domäne der Filmbewertungen erreicht wurde, variierte diese bei den Bank- und Fahrzeugbewertungen von 80% bis 84% (ebd., S. 424).
Auf der Aspektebene schlugen Bagheri et al. (2013) ein unüberwachtes Lernverfahren vor, das in vier Schritten die aus mehreren Wörtern bestehenden Aspekte identifizierte, und erreichten damit im Durchschnitt 84,1% Genauigkeit und 66,2% Abdeckung der Daten bei der Analyse. Ähnlich der eben vorgestellten Arbeit wurde hier POS-Tagger benutzt, um Aspekte zu identifizieren (ebd., S. 143), woraufhin die Korrektur identifizierter Aspekte mit folgenden zwei Heuristiken (s. Kapitel 1 Abschnitt 1.1) folgte (ebd., S. 144):

- „Rule #1: Remove aspects which there are no opinion words in a sentence."

- „Rule #2: Remove aspects that contain stop words."

Eine kleine Seed-Liste häufigster Aspekte wurde verwendet, um, basierend auf Mutual Information, die inter-relationalen Informationen zwischen den Wörtern zu berechnen (A-Score) (ebd.). Mit der „Bootstrapping"-Methode wurde iterativ die endgültige Liste der Aspekte zusammengestellt (ebd.). Der „Bootstrapping"-Algorithmus ist auf der Abbildung 3.2 zu sehen. Am Ende wurden zwei Methoden (Subset und Superset-Support-Pruning) eingesetzt, um redundante Aspekte zu löschen (ebd., S. 145f.). Ein Beispiel für die ‚Subset-Redundanz' sind „free speakerphone" oder „rental dvd player", da diese solche Aspekte wie „speakerphone" und „dvd player" unnötig differenzieren (ebd., S. 145). Mit Superset-Support-Pruning wurden redundante Aspekte gelöscht, die aus einem Wort bestanden (ebd., S. 146). „Suite" oder „life" sind Beispiele für solche Aspekte, deren sinnvolle Supersets „PC Suite" oder „battery life" sind.

Algorithm: Iterative Bootstrapping for Detecting Aspects
Input: Seed Aspects, Candidate Aspects
Method:
FOR each candidate aspect
Calculate A-Score
Add the Aspect with Maximum A-Score to the Seed Aspects
END FOR
Copy Seed Aspects to Final Aspects
Output: Final Aspects

Abbildung 3.2: Iterativer „Bootstrapping"-Algorithmus zur Identifikation von Aspekten (Bagheri et al., 2013, S. 145)

Da in der vorliegenden Arbeit eine Musterextraktion erfolgt, die sich auf die zu vordefinierten Objekten wertenden Phrasen beschränkt, wird weder nach neuen Bewertungsobjekten gesucht, noch sich an POS-Tags zur Extraktion kompletter Phrasen orientiert. Daher sind unüberwachte Lernverfahren für die hier gestellten Ziele ungeeignet.

3.1.2.3 Semiüberwachte / hybride Lernverfahren

Semiüberwachte und hybride Lernverfahren werden in der vorliegenden Arbeit synonym verstanden. Die Idee dieser Verfahren besteht in einer Kombination überwachter und unüberwachter Lernverfahren (Archak et al., 2011, S. 1489), wobei die Vorteile der beiden zum Erreichen besserer Ergebnisse ausgenutzt werden. „Halfway between supervised and unsupervised learning, it adresses the problem of learning in the presence of both, labeled and unlabeled data. The underlying assumption is that the required number of labeled samples can be reduced by taking advantage of large amounts of unlabeled data (which is typically available at no costs)" (Broß, 2013, S. 34). Die im Kapitel 4 (Abschnitt 4.2.1.3, Seite 110) in Bezug auf lokale Grammatiken vorgestellte Methode „Bootstrapping" gehört als eine der ersten Ideen in diesem Kontext zu hybriden Lernverfahren. Archak et al. (2011) benutzte eine „Crowdsourcing-Based Technique", um Produkteigenschaften (Aspekte) zu identifizieren. Das genannte System heißt *Amazon Mechanical Turk*, bei dem die Annotatoren (s. Abschnitt 3.3.1) für verschiedene Aufgaben eingesetzt werden, weil diese nicht voll automatisch erledigt werden können (ebd.,

S. 1489). Solche Aufgaben sind unter der Bezeichnung ‚human intelligence tasks (HITs)' bekannt (ebd.). Bei der Identifikation der Produktaspekte wurden drei unabhängige Annotatoren eingesetzt, die aus jeder Bewertung die besagten Aspekte in freier Form beschreiben sollten (ebd., S. 1489f.). Wenn zwei Annotatoren gleiche Aspekte aus derselben Textbewertung extrahiert haben, so wurden diese für zuverlässig erklärt (ebd., S. 1489). Im Vergleich zu voll automatischen Lernverfahren (POS-Tagging, Erweiterung extrahierter Phrasen zu einer Menge ähnlicher Nomen und Nominalphrasen sowie deren Gruppierung mit Berücksichtigung des Kontextfensters von vier Wörtern um sie herum) kann man mit dem ‚Mechanical Turk' zum Teil etwas bessere Ergebnisse erreichen (ebd., S. 1490).
Sowohl „Bootstrapping" als auch die Leistungen von Annotatoren werden in der vorliegenden Arbeit verwendet. Das zuerst genannte Verfahren wird zur Akquise der Adjektive sowie Filterung relevanter Kontexte von wertenden Phrasen herangezogen, worauf im Kapitel 5 (Abschnitte 5.1.2.2.1, Seite 137ff. und 5.2.1.1.3, Seite 152) ausführlicher eingegangen wird. Die Problematiken, die durch Annotatoren im Rahmen der vorliegenden Arbeit gelöst werden, werden im Abschnitt 3.3.1 thematisiert und im Kapitel 5 (Abschnitt 5.3.1, Seite 167ff.) behandelt. Die Ergebnisse des Annotatoreneinsatzes werden im Kapitel 6 (Abschnitt 6.2, Seite 187ff.) präsentiert.

3.1.2.4 Muster- und regelbasierte Lernverfahren

Laut Medhat et al. (2014, S. 1095) gehören regelbasierte zu den überwachten Lernverfahren. Da die üblichen überwachten Verfahren auf einer niedrigeren Ebene wie z. B. Satzebene schlechtere Ergebnisse erzielen (s. Abschnitt 3.1.2.2.1), setzen diesbezüglich muster- und regelbasierte Verfahren in Bereichen wie Informationsextraktion und Stimmungsanalyse eine genauere empirische Auseinandersetzung mit den semantischen und syntaktischen Strukturen der Sätze voraus, was für die vorliegende Arbeit relevant ist. Daher werden die genannten Verfahren gesondert als eine eigene Klasse behandelt.
Auf der Dokument-Ebene können allgemeine Statistiken wie z. B. Anzahl der zufriedenen oder unzufriedenen Kunden erstellt werden. Diese liefern jedoch keine differenzierteren Informationen zu konkreten Aspekten der Produkte bzw. Dienstleistungen oder zu den Gründen der Kundenzufriedenheit (Broß, 2013, S. 21). Diesen notwendigen Schritt geht die aspektbasierte Stimmungsanalyse, indem sie die Meinungen in Bezug auf die individuellen Produkt-

aspekte analysiert (ebd.). „Whereas review classification considers only a single dimension (namely „sentiment polarity“), aspect-oriented review mining involves the joined analysis of two dimensions. On one dimension we want to discover all relevant product aspects and on a second dimension we want to identify related expressions of sentiment [s. Kapitel 2, Abschnitt 2.2.1.2.3] and determine their polarity. In contrast to review classification, the task is better characterized as a problem in information extraction [s. Kapitel 2, Abschnitt 2.2.1] than a problem in text categorization“ (ebd.). Die in der vorliegenden Arbeit zu extrahierenden Produkt-Aspekte (vordefinierte Jameda-Dimensionen) bilden zusammen mit wertenden Ausdrücken die eben angedeuteten Relationen (s. auch Kapitel 2, Abschnitt 2.2.1.2.3, Seite 44). Auf diese Informationen zielt die Extraktion wertender Aussagen in dieser Arbeit ab. Syntaktische Konstruktionen weisen bestimmte Muster auf, auf deren Basis die Regeln für die Extraktionssysteme entwickelt werden. Auf der Abbildung 3.3 ist die Transformation der unstrukturierten Informationen aus einem Bewertungstext in ein strukturiertes Format entsprechend der Aufgabe von Informationsextraktion (s. o.) dargestellt. Trotz der teilweisen Zuordnung der Identifikation kognitiver Effekte dem Gebiet der Informationsextraktion, bleibt das Ziel der vorliegenden Arbeit allerdings keine Präsentation der gewonnenen Informationen in einem strukturierten Format, sondern die Erkenntnis über die Möglichkeiten der Identifikation, Klassifikation und das Ausmaß des Vorhandenseins kognitiver Effekte in den Bewertungstexten.
Im Bereich der Informationsextraktion entwickelte Stotz (2018) ein regelbasiertes System zur Erkennung von Unternehmenszusammenschlüssen. Das System basiert auf Lexikongrammatiken, die „Interaktion zwischen Wörtern aus dem Lexikon und grammatischen Regeln beschreiben“ (ebd., S. 39) und in Form von Verbtafeln dargestellt werden (ebd.). Auf der Grundlage dieser Verbtafeln und der Auseinandersetzung mit den distributionellen und strukturellen Eigenschaften der Verben wurden lokale Grammatiken[58] entwickelt, „die den großen Vorteil bieten, die Interaktion syntaktischer und semantischer Informationen abbilden zu können“ (ebd., S. 48). Da das Korpus, auf dem das genannte Extraktionssystem entwickelt wurde, hauptsächlich aus zusammengestellten Zeitungsartikeln bestand (ebd., S. 139f.), kann man zumindest in den meisten Fällen von syntaktisch korrekten Satzkonstruktionen ausgehen, was die Sprachstile der Autoren betrifft. Dadurch können mit der

[58] s. Kapitel 4, Abschnitt 4.2.1

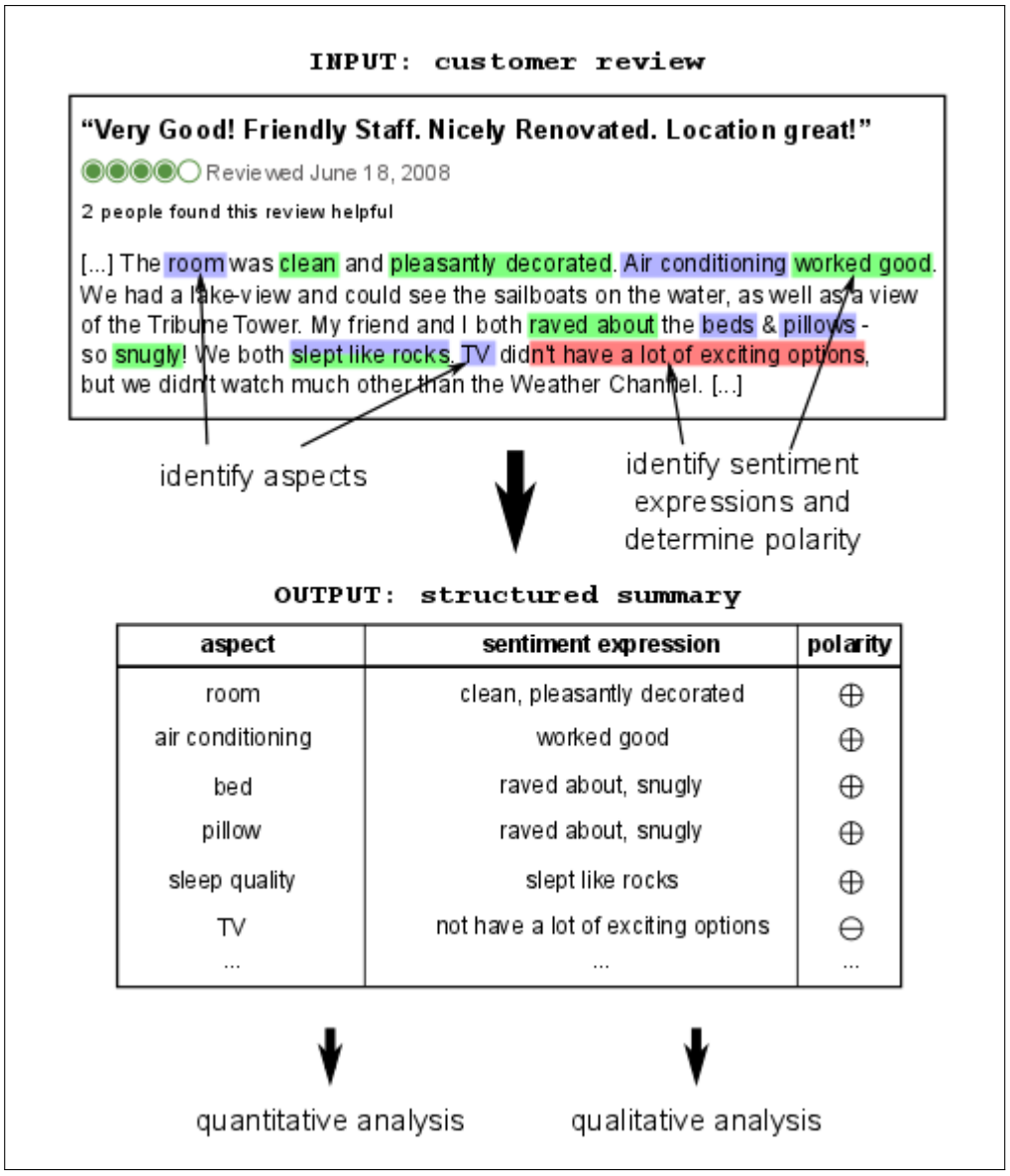

Abbildung 3.3: Strukturierter Überblick extrahierter Relationen (Produktaspekte und Stimmungen) aus den unstrukturierten Bewertungstexten der Kunden (Broß, 2013, S. 21)

vorgeschlagenen Methode jedoch nur die Sätze berücksichtigt werden, die die Verben beinhalten (ebd., S. 115). Für das in der vorliegenden Arbeit verwendete Korpus sind elliptische Sätze mit Auslassungen der Verben und anderer Wortarten charakteristisch. Der Aufbau von Verbtafeln für die Konstruktion entsprechender syntaktischer Regeln wäre daher im Rahmen dieser Arbeit nicht sinnvoll.

Wie kann der regelbasierte Ansatz im Rahmen aspektbasierter Stimmungsanalyse und domänenspezifischer Informationsextraktion umgesetzt werden?

Ding et al. (2008, S. 235) schlugen z. B. die Regeln für die Negationswörter oder -phrasen vor, die die Polaritäten der ausgedrückten Meinungen gewöhnlich umkehren (Beispiel (3.5)). Die Regel (3.5)(a) impliziert z. B. einen positiven sprachlichen Ausdruck bei einer Verneinung eines negativen Wortes. Außer der einfachen Negationswörter wie „no“, „not“, „never“ gibt es musterbasierte Verneinungen wie „stop“+„vb-ing“ oder „cease“+„to vb“ (z. B. „stopped working“), die den Aussagen bei der Anwendung der Regel (3.5)(c) eine negative Polarität verleihen (ebd.).

(3.5) (a) Negation Negative → Positive // e.g., „no problem“
(b) Negation Positive → Negative // e.g., „not good“
(c) Negation Neutral → Negative // e.g., „does not work“, where „work“ is a neutral verb

Wolfgruber (2015) behandelte u. a. die Negation, wobei lexikalische und morphologische Negationsträger unterschieden wurden (ebd., S. 31f.). Bei ihrem Ansatz zur Extraktion der Stimmungen in Hotelbewertungen entwickelte sie Graphen, die verneinende Modifikatoren wie „kaum“, „kein“, „nicht“, „nichtmal“, „niemals“ etc. (ebd., S. 84) erkennen, wodurch die Polarität eines Satzes bzw. einer Phrase umgekehrt wird. Im Rahmen ihrer Dissertation realisierte sie das Vorhaben, „mit lokalen Grammatiken eine umfangreiche und flexible Erkennung von Sentiments im domänenspezifischen Kontext der Hotelbewertungen zu erreichen“ (ebd., S. 1), wobei die Genauigkeit von 97,9% und die Abdeckung des Korpus von 91,1% erzielt wurden (ebd., S. 109). Die Erkennung von Sentiments erfolgte auf dem Korpus, das von dem Bewertungsportal `www.holidaycheck.de` stammte (ebd., S. 62). Die korpus-

basierte semiautomatische Akquise lexikalischer Ressourcen wurde nach den Mustern zur Wortarten-Suche und auf der Basis der in das Korpusverarbeitungssystem *UNITEX* (s. Kapitel 4 Abschnitt 4.2.2, Seite 111ff.) eingebetteten Wörterbücher realisiert (ebd., S. 62ff.). Die aus dem Korpus extrahierten wertenden Adjektive, Nomen, Verben etc. wurden manuell analysiert und „entsprechend ihrer Polarität in separate Graphen [lokale Grammatiken (s. Kapitel 4, Abschnitt 4.2.1)] aufgeteilt“ (ebd., S. 63). Zusätzlich wurden externe Quellen wie Duden oder Internet zur Gewinnung der wertenden festen Wendungen oder Emoticons herangezogen (ebd., S. 63ff.). Die bewertenden Objekte, die sogenannten Features (s. Kapitel 2 Abschnitt 2.2.1.2.2, Seite 43), wurden anhand der häufigsten Nomen oder Nominalphrasen identifiziert. Nach der beschriebenen Akquise wurden Regeln zur Phrasenextraktion aufgestellt und auf das genannte Korpus in Form von Mastergraphen (s. Kapitel 5, Abschnitt 5.2.2.4, Seite 160ff.) angewandt. Die Metainformationen bezüglich der Polarität, der Wortart und der Intensität einer Aussage sowie deren eindeutige Zuordnung zu den bewertenden Objekten wurden in einer an XML angelehnten Notation in entsprechenden Tags produziert (siehe Beispiel (3.6)). Die Qualität erzielter Ergebnisse in der Dissertation von Wolfgruber (2015), zahlreiche Vorteile und Möglichkeiten, die die Methode lokaler Grammatiken (Kapitel 4, Abschnitt 4.2.1, Seite 107ff.) bietet, was die Arbeit auf der Satz- und Phrasenebene betrifft (Kap. 2, Abschnitt 2.2.2.3.2, Seite 50), veranlassen dazu, sich mit diesen an entsprechenden Stellen (s. o.) genauer auseinanderzusetzen. Die Entwicklung eines muster- und regelbasierten IE-Systems von Wolfgruber (2015) stellt daher für die vorliegende Arbeit eine Grundlage zur Erkennung wertender Muster zu vordefinierten Dimensionen dar. Im Laufe der Entwicklung des Identifikationssystems für kognitive Effekte wird an mehreren Punkten an die eben beschriebene Dissertation angeknüpft, worauf an entsprechenden Stellen verwiesen wird und die Zusammenhänge erläutert werden.

(3.6) „<POS><O type=“N“>Die <id=“Zimmer“>Zimmer</O> waren<A pol=“pos“ type=“A“> <id=“sauber“>sauber</A></POS>“

3.2 Automatisierte Textinterpretationen

Einige Textinterpretationen sozialpsychologischer Experimente wurden am Ende des Abschnitts 3.1.1.1 aufgeführt. Diese Interpretationen bezogen sich auf die Phrasen, linguistische Indikatoren, wie sie z. B. im Beispiel (3.2) gezeigt wurden. Im aktuellen Abschnitt werden wissenschaftliche Arbeiten dargestellt, die sozialpsychologische Phänomene in Texten anhand bestimmter Indikatoren interpretieren und automatisch identifizieren. Die Darstellung der Vorgehensweisen bei Textinterpretationen dient der Findung von Indikatoren für ausgewählte automatisch erkennbare Effekte in der vorliegenden Arbeit. Bei den Ausführungen der nachfolgenden Abschnitte wird eine Aufteilung in rhetorische Stilmittel und weitere sozialpsychologische Phänomene vorgenommen. Während die ersten dominant auf linguistischen Mustern basieren, durch die Meinungen ausgedrückt werden, sind die zweiten – neben den linguistischen Musteranalysen – in einem höheren Maß auf die Meta-Daten jeweiliger Online-Dienste angewiesen, um mögliche Indikatoren / Features zur Identifikation entsprechender Phänomene zu bestimmen.

3.2.1 Rhetorische Stilmittel

3.2.1.1 Ironie und Sarkasmus

Ironie und Sarkasmus werden in der wissenschaftlichen Literatur unterschiedlich definiert. Schieber et al. (2012, S. 5) verstehen unter Ironie eine Möglichkeit, „Einstellung oder Gefühle auszudrücken, die man nicht hat, und gleichzeitig zu verstehen zu geben, dass man sie nicht hat“ (Lapp, 1992, S. 141, zitiert in ebd.). Sarkasmus wird als eine Ironisierung von Unrecht und Leid verstanden (Lapp, 1992, S. 110ff., zitiert in ebd.) und ist somit als eine spezielle Form der Ironie zu deuten (ebd.). In ihrer Arbeit benutzten Hallmann et al. (2016, S. 2) folgende Definition von Ironie: „Irony is an utterance with „a literal evaluation that is implicitly contrary to its intended evaluation“ (Burgers et. al., 2011, p. 190)“ (ebd.) und behandelten Sarkasmus als Synonym zur Ironie (ebd.). In der vorliegenden Arbeit werden die beiden Begriffe ebenfalls synonym und, in Anlehnung an beide obigen Definitionen, als rhetorisches Stilmittel zum Ausdruck einer Meinung in Form von einer Äußerung des Gegenteils dieser Meinung verstanden.
Wissenschaftliche Arbeiten zur automatischen Identifikation von Ironie bedienten sich meistens den expliziten (s. Kapitel 2, Abschnitte 2.1.2.3.2 (Sei-

te 36), 2.2.2.2.2 (Seite 47)) Markern wie (#not) #sarcasme (#sarcasm), Anführungszeichen, Ausrufezeichen, Großbuchstaben, Emoticons u. ä. (vgl. ebd.; vgl. Schieber et al., 2012; vgl. Wolfgruber, 2015). Bamman und Smith (2015) zeigten, dass Ironie ein komplexes Phänomen ist, dessen automatische Identifikation nicht allein aus dem lokalen Kontext der Mitteilungen selbst, sondern durch die Betrachtung zusätzlicher linguistischer Informationen wie die zum Autor der Mitteilungen selbst, zu seinen Beziehungen zu Adressaten der Mitteilungen etc. erfolgen sollte (ebd., S. 574). Sie arbeiteten dabei mit Tweeter-Daten und definierten vier Klassen von Features (Tweet Features, Author Features, Audience Features und Response Features) (ebd., S. 575). Einige Beispiele von diesen sind:

- Tweet Features: ‚Tweet whole sentiment', ‚Intensifiers' (so, too, very, really), ‚Tweet word sentiment'
- Author Features: ‚Author historical topics', ‚Profile information' (number of friends, followers)
- Audience Features: ‚Author / Addressee interactional topics'
- etc. (ebd., S. 575f.)

Bei ihrem Vorgehen zeigten die Autoren, dass je mehr Features sie zur Analyse hinzuzogen, desto höher die Genauigkeit der Ironie-Identifikation in den Tweeter-Daten wurde (s. Abbildung 3.4). Generell wurde außerdem festgestellt, dass ein solches Hashtag wie #sarcasm nicht zu einem natürlichen Indikator von Ironie zwischen den Freunden gehörte, sondern eine kommunikative Funktion erfüllte und eher die Absicht eines Autors implizierte, den Lesern gegenüber den Ausdruck von Sarkasmus zu symbolisieren, die, anders als Freunde, dies sonst nicht hätten verstehen können (ebd., S. 577). Mit Ironie-Markern in diesem Kontext beschäftigten sich ebenfalls Hallmann et al. (2016) und stellten fest, dass diese in Abhängigkeit von Tweeter-Gruppen unterschiedlich benutzt wurden (ebd., S. 13f.). Dabei teilten sie die Tweeter in zwei Gruppen:

- Adressierte Tweets: diejenigen, die an andere Personen gerichtet wurden (Suche nach Mustern, die mit „@" begannen)
- Nicht adressierte Tweets: diejenigen, die sich an keine Personen richteten (ebd., S. 7).

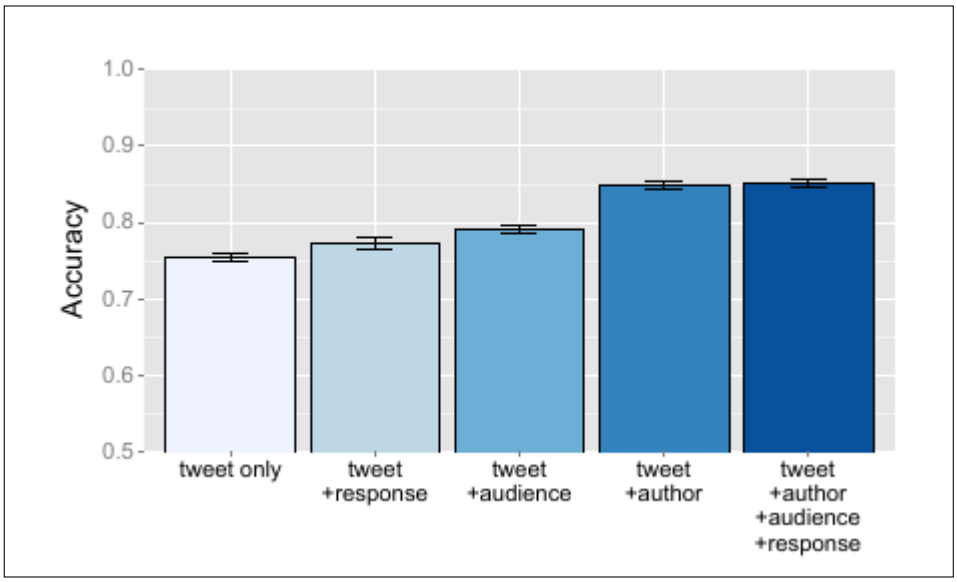

Abbildung 3.4: Genauigkeit der Sarkasmus-Identifikation bzgl verschiedener Feature Sets (Bamman und Smith, 2015, S. 576)

Für 10 Kategorien von Ironie-Elementen überprüften sie, wie präsent diese in beiden Gruppen von Tweets waren (ebd., S. 10ff.). Einige dieser Kategorien sind, wie folgt, aufgeführt:

- Polarität (positive oder negative Bewertung)
- Ambiguität (Elemente, die mehrere mögliche Interpretationen zulassen)
- Hyperbeln (Wörter, die stark vom semantischen Durchschnitt abweichen, z. B. „fantastic“ ist eine Hyperbel, während „nice“ keine ist)
- Wiederholungen der Buchstaben (z. B. „grrrrrreat“)
- Hashtags (vor den Wörtern, z. B. „#fun“)
- Emoticons (Simulation der Gesichtsausdrücke anhand der Punktuationszeichen)
- etc.

Die Ergebnisse dieser Untersuchung haben gezeigt, dass das Benutzen von verschiedenen Ironie-Markern signifikante Unterschiede in Bezug auf adressierte und nicht adressierte Tweets aufwiesen (ebd., S. 15). So enthielten adressierte Tweets z. B. deutlich weniger Wiederholungen der Buchstaben und Hashtags, während die Emoticons in denselben mehr vertreten waren (ebd., S. 16). Generell ließ sich feststellen, dass Ironie-Marker in nicht adressierten Tweets öfter benutzt wurden, was darauf schließen lässt, dass die einander kennenden Personen das Wissen gegenseitig voraussetzen und das Angedeutete auch inferieren können (s. Kapitel 2, Abschnitt 2.1.1.3.2, Seite 23) (ebd.). Die beiden beschriebenen Untersuchungen zu Ironie-Markern erfolgten auf Tweeter. Aufgrund der Möglichkeiten dieses Online-Dienstes kann man z. B. die Informationen zum Autor der Tweets, zu seinen Kontakten, Anzahl von Freunden, Themen, zu denen dieser sich äußert u. a. verfolgen. Diese Informationen haben einen Mehrwert im Sinne der Suche nach Indikatoren und der Aufstellung der Kriterien zur automatischen Identifikation von Ironie. Diese Möglichkeiten sind bei Kundenbewertungsportalen eingeschränkt. „Das Kommunikationsmodell der ironischen Äußerung besteht aus einem Sprecher, einem Gegenspieler und einem Publikum“ (Hartung, 2002, S. 182, zitiert in Wolfgruber (2015, S. 28)). „Diese Konstellation ist bei einer Hotelbewertung nur eingeschränkt gegeben. Man könnte die bewertende Person [Patient] als Sprecher, das Hotel [Arztpraxis] als Gegenspieler und die Leser der Bewertung als Publikum sehen, wobei die bewertende Person eher direkt für den Leser schreibt und sich nicht in Zwiegespräch mit dem Hotel [Arztpraxis] befindet“ (Wolfgruber, 2015, S. 28). Aufgrund dessen und der bekannten Tatsache der Anonymität von Bewertenden (s. Kapitel 2, Abschnitte 2.1.1.1.1 (Seite 14) und 2.1.1.3.3 (Seite 26)) kann man keine adressierten Bewertungen auf den Bewertungsportalen erwarten, so dass man das Netzwerk der Patienten als einen potentiellen Indikator für Ironie o. a. nicht verfolgen kann. In ihrer Dissertation versuchte Wolfgruber (2015) Ironie mittels lokaler Grammatiken (s. Kapitel 4, Abschnitt 4.2.1 (Seite 107)) zu erkennen mit der Argumentation, dass dieses Phänomen nicht anhand einzelner Wörter, sondern durch die Berücksichtigung semantischer Kriterien erkannt werden kann, was statistische Methoden in diesem Zusammenhang scheitern lässt (ebd., S. 28). Sie bediente sich dabei den Ironie-Markern wie polaren Adjektiven und Nomen, sarkastischen Emoticons sowie den Anführungszeichen (z. B. „tolle“ Aussicht) aus dem Trainingskorpus (ebd., S. 68f.) und baute aus diesen entsprechende semantische Muster auf, die mit Hilfe von einem Graphen auf den Text angewandt wurden (ebd., S. 87). Die erzielten Ergeb-

nisse konnten auf dem Testkorpus nicht evaluiert werden, da dort lediglich eine ironische Aussage auffindbar war, die durch lokale Grammatiken nicht erkannt wurde (ebd., S. 110). Dies bestätigt die Erkenntnis aus anderen wissenschaftlichen Untersuchungen darüber, dass Ironie ein seltenes Phänomen ist (Hallmann et al., 2016, S. 6), was kognitive Effekte ebenfalls betrifft. Die Genauigkeit der Ironie-Erkennung im Trainingskorpus mit lokalen Grammatiken betrug 72,5% (Wolfgruber, 2015, S. 120). Auch Schieber et al. (2012) erkannten die Problematik der Ironie-Identifikation in Kundenrezensionen. Zur Erkennung dieses Phänomens bedienten sich die Autoren eines Kommunikationsmodells von Schulz Thun (2008, S. 25ff., beschrieben in Schieber et al. (2012, S. 5)), welches besagt, dass „die Aussage einer Person vier Botschaften enthält: die Sachebene (Informationen, Daten, Sachverhalte), die Selbstkundlage (Persönlichkeit des Sprechers), den Beziehungshinweis (Verhältnis der kommunizierenden Personen) sowie den Appell (Einflussnahme auf den Empfänger)“. Dabei stellten sie fest, dass in den niedergeschriebenen Texten lediglich die Sachebene und der Appell extrahiert werden können (ebd.). Weiterhin benutzten die Autoren empirische Untersuchungen von Winner (1988, S. 147ff., beschrieben in Schieber et al. (2012, S. 6ff.)), laut denen die Erkennung von Ironie nach drei wesentlichen Punkten möglich ist:

- Fakten (die Wahrheit oder die Unwahrheit einer Aussage anhand der gegebenen Informationen)
- Glauben (die Absicht oder die Aufrichtigkeit einer falsch gemachten Äußerung)
- Absicht des Sprechers (die Täuschung oder die Ironie der bewusst falsch gemachten Äußerung)

Dem Prozess zur automatischen Erkennung von Ironie und Sarkasmus liegt ein halbüberwachtes Klassifikationsverfahren zugrunde (ebd., S. 6). Im ersten Schritt wurden die vordefinierten Muster ausfindig gemacht (ebd.). Im zweiten Schritt wurden erkannte Sätze in einem Wertebereich von 1 (keine ironische Aussage) bis 5 (höchst wahrscheinlich eine ironische Aussage) klassifiziert (ebd., S. 6ff.). Die Erkennung sarkastischer Sätze erfolgte in den Kundenreviews zum Produkt „iPad“ des Herstellers Apple Inc (ebd., S. 10). Die Inhalte der Dokumente wurden featurebasiert auf der Satzebene analysiert (ebd., S. 4). In erster Linie ging es um die Erkennung widersprüchlicher Aussagen, wobei Produkteigenschaften und die dazu gehörigen Meinungswörter

auf ihre Zusammengehörigkeit überprüft wurden (ebd., S. 7). Dann wurde nach Kontextwörtern gesucht, die keine Korrelation mit dem jeweiligen Produkt hatten, wodurch festgestellt wurde, ob der Autor tatsächlich das gegebene Produkt bewertete (ebd.). Bei der Identifizierung der Aussagen solcher Art wurden diese mit Stufe 3 bewertet (ebd., S. 8). Im Beispiel (3.7) wird deutlich, dass die positive Aussage zu einem Spiegel in der Bewertung eines Monitors als Ironie interpretiert werden kann (ebd.).

(3.7) „Der Monitor ist ein super Spiegel!“

Bei der Signalanalyse ging es um Zeichen- (Anführungszeichen, Ausrufezeichen, Großbuchstaben), Emotions- und Superlativanalyse (ebd.). Bei der Markierung von Meinungswörtern mit Anführungszeichen z. B. erhielt die Aussage die Ironiestufe 4, bei Wortgruppen mit mehr als drei Wörtern – Stufe 2, da die Verwendung von Zitaten angenommen wurde (ebd.). Bei der Verwendung der Großbuchstaben wurde der Ironiewert ebenfalls auf 4 gesetzt, wenn das großgeschriebene Wort ein Meinungswort war (ebd.). Die Vergabe der Ironiewerte bei Ausrufezeichen hing davon ab, ob diese innerhalb eines Satzes oder an seinem Ende gebraucht wurden sowie von deren Anzahl (ebd.). Die Emotionen können durch Smileys, Abkürzungen und Akronyme ausgedrückt werden (ebd.). Im Beispiel (3.8) stellt die Aussage „einen Widerspruch zwischen der Aussage und den Gefühlen des Autors dar und erhält deswegen die Ironiestufe 3“ (ebd., S. 9).

(3.8) „Das ist doch schön! :-(“

„Die Nutzung von Superlativen mit gegensätzlicher Produktbeurteilung erhöht die Verwendung von Ironie. Für diese Analyse wird somit die Sternenvergabe [s. Kapitel 2, Abschnitt 2.1.1.1.3 (Seite 15)] der Kundenrezension für die Produktbewertung herangezogen. Die Beiträge mit positiven (/negativen) Superlativen und einer negativen (/positiven) Bewertung erhalten die Ironiestufe 4“ (ebd.). Aus 71 Kundenbewertungen galten diejenigen als ironieenthaltende Rezensionen, die Ironiestufen 3 – 5 erhielten (ebd., S. 10). Ironiestufen 1 – 2 zeigten dementsprechend, dass die Texte keine ironischen Aussagen enthielten (ebd.). Der Abgleich der durchgeführten automatischen Identifizierung der Ironie erfolgte mit den im Voraus festgelegten ironischen Bewertungen, denen ebenfalls im Voraus die entsprechenden Stufen zugeordnet wurden (ebd.). Zusätzlich wurden zu jeder Bewertung Aussagen getroffen, ob diese Ironie beinhalten oder nicht, „da die Wirkungen von Ironie bei

jedem Leser unterschiedlich sind“ (ebd.). „Bei der Evaluation des Systems ist eine korrekte Erkennung dann gegeben, sobald die Einstufung über das Vorhandensein von ironischen Aussagen mit der manuell vorher festgelegten Benotung übereinstimmt“ (ebd.). In der hier beschriebenen Arbeit wurden ironische Aussagen in Bewertungen in 69% der Fälle korrekt erkannt (ebd.). Die Ausführungen des aktuellen Abschnitts sind für die vorliegende Arbeit in mehrfacher Hinsicht interessant. Zum einen kann man Ironie als ein sozialpsychologisches Phänomen interpretieren, die, wie auch kognitive Effekte, persönliche Einstellungen ausdrückt (Wolfgruber, 2015, S. 24). Zum anderen hängt die automatische Ironie-Identifikation von den Möglichkeiten des gegebenen Online-Dienstes ab, was auf die Effekte ebenfalls zutrifft. Im Vergleich zu Tweeter haben die Bewertungsportale in diesem Sinn eher eingeschränkte Möglichkeiten, da die Bewertungen dort anonym sind und daher weniger Indikatoren zur automatischen Erkennung solcher Phänomene bieten.
Muster- und regelbasierte Lernverfahren eignen sich zur Erkennung sozialpsychologischer Phänomene wie kognitive Effekte besser, da diese nicht aus einzelnen Wörtern allein erkannt werden können, sondern erst innerhalb bestimmter Muster kontextbezogen einen Sinn ergeben. Als Indikatoren für Widersprüche, durch die Ironie, aber auch einige Effekte (s. Kapitel 2, Abschnitt 2.1.2.1.2 (Seite 30)) definiert sind, können Bewertungstexte und die Vergabe nummerischer Werte in Bezug auf die Übereinstimmung ihrer Polaritäten analysiert werden. Aufgrund der hier mehrmals betonten wenigen Möglichkeiten der Bewertungsdomäne stellen Bewertungstexte dabei einen besonderen Wert dar, aus denen solche explizite Indikatoren wie Anführungszeichen, Ausrufezeichen, Großbuchstaben, Emoticons etc. zur automatischen Analyse herangezogen werden können. Schließlich ist das von Schieber et al. (2012) eingeführte Ironie-Stufensystem insofern interessant, dass man in Bezug auf die Identifikation kognitiver Effekte ebenfalls ein nummerisches System entwickeln kann. Dieses würde dann die Wahrscheinlichkeiten aufzeigen, so dass im Ergebnis eine konkrete Aussage zum Ausmaß des Vorhandenseins eines Effekts pro Bewertung getroffen werden könnte.

3.2.1.2 Idiome und feste Wendungen

Die automatische Erkennung von Idiomen und festen Wendungen ist im Kontext der Stimmungsanalyse nicht wegzudenken, da sie als wichtige Indikatoren zum Ausdruck der Emotionen von Personen dienen können (Garg und Goyal, 2014, S. 13; Wolfgruber, 2015, S. 24). Idiome und feste Wendungen

werden hier aufgrund eines ähnlichen Vorgehens in der wissenschaftlichen Literatur synonym behandelt. Idiome sind Phrasen oder Ausdrücke mit einer nicht kompositorischen Bedeutung, sprich: die Bedeutungen solcher Phrasen sind verschieden von den Bedeutungen einzelner Wörter, aus denen sie bestehen, dementsprechend sind sie nicht voraussagbar (Muzny und Zettlemoyer, 2013, S. 2; Garg und Goyal, 2014, S. 12; Verma und Vuppuluri, 2015, S. 1). Wolfgruber (2015, S. 26) erstellte eine Sammlung fester Redewendungen, die dann von den lokalen Grammatiken verarbeitet wurde. Außer manueller Korpusanalyse bediente sie sich dabei externer Quellen wie dem Duden der Redewendungen und das Internet (ebd.). Mit 100%-iger Genauigkeit gelang ihr 81,8% fester Wendungen zu extrahieren. Auch Garg und Goyal (2014) erstellten eine große Datenbasis von Idiomen und deren Variationen und entwickelten ein Annotationssystem, das mit über 80% Genauigkeit in gegebenen Textdokumenten nach festen Wendungen sucht und diese annotiert. Mit Variationen fester Wendungen sind folgende Ausdrücke gemeint (ebd., S. 12):

- „Bad news travels fast“
- „Bad news has wings“

Ein überwachtes Verfahren zur Erhöhung der Idiomen-Anzahl in Wiktionary schlugen Muzny und Zettlemoyer (2013) vor. Zur Überwachung benutzten sie die Online-Ressource Wiktionary und entwickelten anhand der Phrasen, Definitionen und Beispielsätze von Wiktionary Dump (13.11.2012) fünf graphbasierte Features, die die Phrasen anhand ihrer Kompositionalität klassifizierten (ebd., S. 2f.). Auf den gleichen Daten getestet, erreichte das System 40,1% ohne Annotationen und 62,0% F-Score (s. Kapitel 6, Abschnitt 6.1.3 (Seite 187)) mit Annotationen von Idiomen. In der existierenden wissenschaftlichen Literatur zur automatischen Identifikation von Idiomen sehen Verma und Vuppuluri (2015, S. 1f.) das Problem in überwachten Verfahren, die manuell annotierte Daten benötigen, domänenspezifisch sind und mit dem Problem syntaktischer Einschränkungen wie z. B. Verbpartikel etc. konfrontiert werden usw. Aufgrund dieser Feststellungen entwickelten sie ein nicht überwachtes domänenunabhängiges Verfahren (IdiomExtractor), das keine annotierten Idiomen benötigt (ebd., S. 1). Das Arbeitsprinzip des in Python implementierten IdiomExtractor ist auf dem Vergleich der Differenz individueller Wortbedeutungen in einer Phrase mit der Bedeutung der Phrase als Ganzes aufgebaut (ebd., S. 3). Aus der Menge normalisierter Wörter, die

die Bedeutung einer Phrase ausmachen, wurden normalisierte Wörter ‚abgezogen', die die Bedeutungen phrasenbildender Wörter beschreiben (ebd., S. 3f.). Im Beispiel (3.9) ist der Prozess der beschriebenen Substraktion für die Phrase „forty winks" aufgeführt (ebd.).

(3.9) (a) Normalisierte Definitionen der Phrase und phrasenbildender Wörter
$RD_p = \{sleep\ period\ time\ bed\}$ (Definition der Phrase)
$RD_{W1} = \{number\ product\ ten\}$ (Definition des 1. Wortes)
$RD_{W2} = \{time\ time\ eye\ blink\ heart\ beat,\ eye\ signal,\ reflex\}$ (Definitionen des 2. Wortes)

(b) Substraktion
$S = \{sleep\ period\ time\ bed\} - \{number\ product\ ten\ time\ time\ eye\ blink\ heart\ beat,\ number\ product\ ten\ eye\ signal,\ number\ product\ ten\ reflex\} = \{sleep\ period\ bed,\ sleep\ period\ time\ bed,\ sleep\ period\ time\ bed\}$

Wenn man die nach der Substraktion gebliebenen Wörter und deren Anzahlen betrachtet, hat man folgende Zusammensetzung: $\{sleep: 3,\ period: 3,\ time: 2,\ bed: 3\}$ (ebd., S. 4). Eine der möglichen Interpretationen dieses Ergebnisses ist, dass die Phrase als ein Idiom zu klassifizieren ist, wenn die dargestellte Differenz keine leere Menge bildet (ebd.), was im aufgeführten Beispiel der Fall ist. Die Evaluation des IdiomExtractor erfolgte auf verschiedenen Datensets (ebd., S. 5). Das System erreichte z. T. über 90% F_1-Score (ebd.).
Aus den statistischen Angaben zu den in den beschriebenen Untersuchungen verwendeten Daten ist ersichtlich, dass Idiome z. B. knapp 12% aller Phrasen bilden können (Muzny und Zettlemoyer, 2013, S. 2) und somit, wie auch kognitive Effekte, seltene Phänomene sind. Die automatische Erkennung dieser Phänomene fand statt, ausgehend von deren Definitionen (ebd., S. 3; Verma und Vuppuluri, 2015, S. 3). Dementsprechend erfolgte auch die Suche nach Features oder Indikatoren der Idiome (ebd.; ebd.). Auf gleiche Weise kann man bei der Aufstellung der Kriterienkataloge pro Effekt vorgehen. Die aufgestellten Kriterien würden dann als Indikatoren fungieren, die bei der Einführung eines ‚Punktensystems' (vgl. Schieber et al., 2012) mit entsprechenden Wahrscheinlichkeiten ihres Auftretens in einer Bewertung angezeigt werden würden. Was den Ausdruck bestimmter Sachverhalte betrifft, so muss man bei Idiomen von festen Phrasen ausgehen, die als Einheiten

erkannt werden sollen (Wolfgruber, 2015, S. 26). Dies betrifft ebenfalls bestimmte phrasenbasierte Indikatoren bei kognitiven Effekten, die Meinungen ausdrücken und schwer zu generalisieren sind (s. Beipsiele (2.1), (2.4) und (2.5)). Gleichfalls ist klar, dass die Anfertigung einer Sammlung fester Wendungen sehr aufwendig ist (ebd., S. 113) und besonders bei den frei formulierten Phrasen keinen Anspruch auf eine 100%-ige Vollständigkeit erheben könnte. Ausgehend vom Erkenntnisinteresse der vorliegenden Arbeit jedoch, genügt es, manche wertende Phrasen als feste Ausdrücke zu behandeln (z. B. Bestätigungsfehler).

3.2.2 Weitere sozialpsychologische Phänomene

3.2.2.1 Bias und Spam

Bias und Spam kann man im Wesentlichen als Synonyme betrachten. Wie in einem Artikel von „Zeit Online“[59] berichtet, bedeutet Spam in Online-Bewertungen nichts Anderes als ‚Meinungsmüll‘, der von den Bewertenden wissentlich produziert wird, um bestimmte Ziele zu erreichen. Einige davon sind:

- Anpreisung konkreter Produkte und Leistungen zum Zweck eines finanziellen Gewinns
- Abhängen der Konkurrenz
- Werbung[60] (Liu, 2010, S. 28f.; Xie et al., 2012a, S. 823; Mukherjee et al., 2012, S. 191; Patil und Bagade, 2012, S. 33)

In Bezug auf E-Mails und Webseiten ist Spam sicher den meisten Menschen bekannt. Auch in diesem Kontext ist er als ‚Müll‘ zu betrachten. „Email spam refers to unsolicited commercial emails selling products and services, while Web spam refers to the use of „illegitimate means“ to boost the search rank positions of target Web pages“ (Liu, 2010, S. 28). Web Spam unterteilt man in zwei Typen: Content Spam und Link Spam (Mukherjee et al., 2012, S. 192). Link Spam bezieht sich auf die Hyperlinks einer Webseite (ebd.). Bei Content Spam werden irrelevante Wörter den Webseiten hinzugefügt,

[59] http://www.zeit.de/2013/31/gefaelschte-online-bewertungen (10.05.2017).

[60] https://www.heise.de/tr/artikel/Online-Faelschern-auf-der-Spur-1585323.html

um z. B. die Suchmaschinen irre zu führen (ebd.). Bei E-Mail Spam handelt es sich um kommerzielle Anzeigen (ebd.). In den Nachrichten werden Spam oder Bias als eine nicht objektive, nicht neutrale Präsentation beispielsweise politischer Ereignisse verstanden. „The bias in the news media is an inherent flaw of the news production process. The resulting bias often causes a sharp increase in political polarization and in the cost of conflict on social issues such as Iraq war“ (Park et al., 2009, S. 443).
Ähnlich den kognitiven Effekten kann man Bias und Spam als Fehler oder Anomalien betrachten, die jedoch bewusst produziert werden, um die Verzerrung des entsprechenden Gesamtbildes zu erreichen (vgl. Definition von kognitiven Effekten im Kapitel 2, Abschnitt 2.1.1.2.3, Seite 21). Für die genannten Phänomene ist eine Tatsache charakteristisch, dass diese in einem Normalfall nicht in die entsprechende Online-Textsorte gehören und somit auch als Abweichung von der Norm jeweiliger Online-Textsorten zu sehen sind[61] (vgl. ebd.).
In den Ansätzen zu Bias-Identifikation[62] werden individuelle Vorgehensweisen zum entsprechenden jeweiligen Thema gezeigt. Die Indikatoren zur Erkennung verschiedener Phänomene in diesen Ansätzen wurden definiert und logisch begründet, woraufhin die entsprechende Evaluation betrieben wurde. Diese Prinzipien kann man ebenfalls auf die automatische Identifikation sowie die problemgerechte Evaluation kognitiver Effekte übertragen.
Verschiedene Zugänge zum Spam in Bewertungen, kreative Lösungsansätze, begründete Auswahl und Kombination von Features, Relationen von Features zur automatischen Identifikation von Spam werden in der Literatur vorgestellt. In diesem Zusammenhang wird in der vorliegenden Arbeit ebenfalls – ausgehend von Definitionen kognitiver Effekte (s. Kapitel 2, Abschnitt 2.1.1) – nach einer geeigneten Kriterienauswahl gesucht, die praktisch getestet und entsprechend evaluiert wird.
Auf konkrete Kriterien, die als Indikatoren für kognitive Effekte in Arztbewertungen interessant sein könnten, wird im nächsten Abschnitt eingegangen. Ein solches Kriterium wären z. B. Ausreißer (s. Abschnitt 3.3.2.1).
Im Kontext der Identifikation von Spam und Spammern sind sich Wissenschaftler einig, dass diese Phänomene eine Abweichung von einer Norm dar-

[61] Bei News Bias sind damit die Meinungen der Autoren von Artikeln gemeint.

[62] Wissenschaftliche Ansätze zu Identifikatoin von News Bias und Spam werden hier aus platzsparenden Gründen ausgelassen. Zu E-Mail und Web Spam s. ausführlicher z. B. Linke (2003) und `http://www2006.org/programme/files/xhtml/3115/fp3115-wu/fp3115-wu-xhtml.html` (20.05.2017).

stellen bzw. ein anderes Verhalten als die meisten Bewertenden aufweisen (Liu, 2010, S. 30; Lim et al., 2010, S. 3). Dieses Kriterium macht jedoch nur in einer sinnvollen Kombination mit anderen definitionsstützenden Features einen Sinn. „[...] deviating from the norm is the necessary condition for harmful spam reviews, but not sufficient because many outlier reviews may be truthful“ (Liu, 2010, S. 30). Die konkrete praktische Ausarbeitung des Kriteriums „Ausreißer“ wird im Kapitel 5, Abschnitt 5.3.2.1 beschrieben. Auch Evaluationsmethoden für entwickelte Systeme zur Spam-Identifikation sind vielfältig und mit der Beteiligung von Annotatoren (s. Abschnitt 3.3.1) verbunden. Wie dieses Konzept in der vorliegenden Arbeit umgesetzt wird, wird im Kapitel 5, Abschnitt 5.3.1 erläutert.

3.2.2.2 Inkonsistenzen in Arztbewertungen

In der Einleitung der vorliegenden Arbeit auf der Abbildung 1.1 (Seite 3) wurden Struktur und Komponenten einer Bewertung dargestellt. Bei der Beschreibung von qualitativen und quantitativen Bereichen einer Bewertung fiel auf, dass der frei formulierte Text zur Dimension „Behandlung“ im Widerspruch zu ihrer nummerischen Bewertung im Sinne der Polarität steht. Solche Erscheinungen werden hier *individuelle Inkonsistenzen* genannt: „the disagreement in polarity of review text and its corresponding numerical ratings (individual inconsistency)“ (Geierhos et al., 2015b, S. 1). Solche Inkonsistenzen können innerhalb einer Bewertung vermehrt auftreten. In der Arbeit von Geierhos et al. (2015b) kann man auf Figure 1 (ebd., S. 5) mehrere solche Unregelmäßigkeiten feststellen:

- Polaritäten des Titels („Unzufrieden“) und der Gesamtnote („2.6“ auf der Schulnoten-Skala von 1 bis 6) stimmen nicht überein
- Polaritäten der Dimensionen „Freundlichkeit“, „Behandlung“ und „Vertrauen“ im Bewertungstext und laut nummerischen Bewertungen stimmen nicht überein:
  - „Freundlichkeit“
    Text: „sehr unhöflich“, Note: „2.0“
  - „Behandlung“
    Text: „Einen Innenmeniskusriss hat er klar verkannt“, Note: „3.0“

  - „Vertrauensverhältnis“
    Text: „würde ich ihn nicht unbedingt empfehlen“[63], Note: „2.0“

- Der erste Satz der Bewertung „Dr. Behle ist für Schulter- und Rückenprobleme ein guter Ansprechpartner.“ beschreibt positive frühere Erfahrungen mit dem Arzt, bezogen auf einen konkreten Bereich. Das lässt vermuten, dass die vergebenen Noten zu mehreren Dimensionen eine Mischung von mehreren Besuchen dieses Arztes darstellen, wodurch sich die o. g. Inkonsistenzen in konkreten Dimensionen möglicherweise erklären lassen (ebd., S. 9).

Außer Geierhos et al. (2015b) erkennen auch andere Autoren solche Unregelmäßigkeiten in Bewertungen (vgl. Hu et al., 2013; vgl. Tsang und Prendergast, 2009). Die genannten Arbeiten beschäftigten sich konkret mit Bewertungstexten und entsprechenden nummerischen Ratings und stellten fest, dass die o. g. Komponenten bereits von vorne herein unterschiedliche Informationen enthalten, da Bewertungstexte im Vergleich zu Ratings die Kauf- bzw. Konsumentscheidungen verschieden beeinflussen (vgl. ebd.; vgl. ebd.). Hu et al. (2013) untersuchten Relationen zwischen Ratings, Sentiments und Umsätzen mit dem Ziel, die gegenseitige Beeinflussung dieser drei Komponenten zu erfassen. Aus den Produktbewertungen von Amazon.com extrahierten sie die Stimmungen auf der Ebene der Gesamtbewertung und berechneten die Sentiment Scores für jede Bewertung (ebd., S. 5ff.). Im Ergebnis fanden sie heraus, dass Ratings meist indirekte, während Bewertungstexte direkte Auswirkungen auf die Umsätze haben (ebd., S. 19), was folgendermaßen erklärt wurde: „Due to the nature of the complex task of searching and purchasing in an online environment, consumers may use different strategies to lessen the burden of their cognitive effort. The way that they may do this is by using ratings as the way to screen potential items and use text reviews to evaluate the limited set of screened items to make the final choice“ (ebd., S. 20). Zu ähnlichen Ergebnissen kamen Tsang und Prendergast (2009), die untersuchten, wie inkonsistente Text-Rating-Bewertungen Konsumverhalten beeinflussen (ebd., S. 3). Sie stellten mehrere Hypothesen auf und studierten gezielt Verbraucherreaktionen, ausgehend von verschiedenen Bewertungsszenarios (ebd., S. 7ff.). Als Untersuchungsgegenstand wählten die Autoren Filmbewertungen mit der Begründung: „We chose movie reviews because

[63] In der vorliegenden Arbeit wurden sprachliche Muster zu einer Empfehlung nicht der Dimension „Vertrauensverhältnis“ zugeordnet.

movies possess substantial experience attributes and are difficult to assess before watching“ (ebd., S. 9). Die wichtigsten Erkenntnisse sind wie folgt zusammengefasst:

- In Bewertungen, die Texte und Ratings beinhalten, spielen Texte eine signifikantere Rolle in Beeinflussung von Konsumenten-Entscheidungen.
- Die Bewertungen, bei denen Texte positive Polarität und Ratings positive oder negative Polaritäten aufweisen, resultieren in einem größeren Kaufwunsch bei Konsumenten als Bewertungen mit negativen Texten und negativen oder positiven Ratings. Dabei sind bei beiden genannten Gruppen kaum Unterschiede festzustellen, was die oben erwähnte höhere Signifikanz der Bewertungstexte bestätigt.
- Bezüglich der Zuverlässigkeit der Bewertungen wurde herausgefunden, dass meistens konsistente Bewertungen als zuverlässig empfunden werden (ebd., S. 11f.).

Neben den individuellen Inkonsistenzen, die am Beginn des aktuellen Abschnitts definiert wurden, beschäftigten sich Geierhos et al. (2015b) mit kollektiven Inkonsistenzen: „the differences in patients' rating behavior for the same service category (e.g. treatment) expressed by varying grades on the entire data set (collective inconsistency)“ (ebd., S. 1). Einige Gründe für diese Art von Inkonsistenzen sind:

- verschiedene Wahrnehmung von Skalenwerten der Bewertenden: „For instance, somebody's view of a '3' may be considerable different from another's“ (Mudambi, 2014, zitiert in Geierhos et al. (2015b, S. 3)).
- Komplexität in „mapping opinions to a single number“ (Centeno et al., 2014, zitiert in Geierhos et al. (2015b, S. 3)).

Zur Identifikation beider Typen von Inkonsistenzen extrahierten die Autoren zunächst typische Phrasen zu fünf vordefinierten Dimensionen aus den Arztbewertungen[64], Polaritäten derer sie dann mit den vergebenen Noten zu korrespondierenden Dimensionen verglichen (ebd., S. 9f.). Die Phrasenextraktion erfolgte mit den im Abschnitt 3.1.2.4 (Seite 78f.) kurz erwähnten

[64]Die Daten wurden dem Portal Jameda entnommen. Projektbedingt wurden 17 bestehende Dimensionen dieses Portals (s. Tabelle 2.1, Seite 16) selbständig in 5 übergeordneten Dimensionen zusammengefasst.

lokalen Grammatiken (ebd., S. 8), deren Vor- und Nachteile im Kapitel 4, Abschnitt 4.2.1 näher erläutert werden. Die weitere Vorgehensweise unterschied sich aufgrund verschiedener Spezifik beider Typen von Inkonsistenzen voneinander (ebd., S. 9f.). Während individuelle Inkonsistenzen auf der Bewertungsebene (s. u.) analysiert wurden, berechnete man für die top-ten-Meinungsphrasen zu jeder Dimension die Anzahl vergebener Noten innerhalb der Gesamtdaten, um kollektive Inkonsistenzen zu untersuchen (ebd.). Im Durchschnitt wurde bei individuellen 31,4% und bei kollektiven, die anhand konsistenten Bewertungen analysiert wurden, 87% F-Score erreicht (ebd., S. 11f.). Die schlechten Ergebnisse bei individuellen Inkonsistenzen erklären sich durch niedrige Genauigkeit, die anhand der Unstimmigkeiten in Polarität von nummerischen und textuellen Bewertungen gemessen wurde, was auf unterschiedliche Schwächen regelbasierter Verfahren wie lokale Grammatiken hindeutet (ebd., S. 11; s. Kapitel 4, Seite 108).
Laut den im Kapitel 2, Abschnitt 2.1.3 (Seite 38ff.) aufgestellten Definitionen zu automatisch identifizierbaren kognitiven Effekten wird sichtbar, dass es mehrere Ebenen zur Aufstellung der Indikatoren automatischer Erkennung existieren:

- Bewertungsebene:
  - Muster zu Dimensionen
  - Phrasen zu Bestätigungsfehlern und Diskriminierungen
  - Nummerische Bewertungen zum Vergleich der Phrasen-Polaritäten
  - Bestimmung der Leitdimension
- Arztpraxisebene:
  - Ausreißer, die eine Abweichung von einer Norm darstellen, die durch die Mehrheit der Bewertungen definiert ist (s. Abschnitt 2.1.1.2.2, Seite 20)
- Gesamtdatenebene:
  - Korrelation / Zusammenhang der Dimensionen (für den Halo-Effekt (vgl. Pollock, 2012), s. auch Abschnitt 3.1.1.1, Seite 55)

Für die Analyse auf der Bewertungsebene wurde in diesem Kapitel bereits eine Reihe von wissenschaftlichen Arbeiten mit automatischen Verfahren vorgestellt, wodurch sich ein klares Bild in Bezug auf mögliche Vorgehensweisen in der aktuellen Arbeit abzeichnet. Was die Arztpraxisebene betrifft, so müssten Ausreißer innerhalb einer Praxis definiert und deren automatische Identifikation ausgearbeitet werden. Die Korrelation der Dimensionen muss auf den Gesamtdaten berechnet werden, um die Tendenzen und Abweichungen auf der Bewertungsebene feststellen zu können. Beide genannten Punkte kann man dem Preprocessing der Daten zuordnen, die Auseinandersetzung mit entsprechenden Konzepten aus dem statistischen Bereich erfolgt im nächsten Abschnitt dieses Kapitels.

3.3 Weitere Aspekte der Textverarbeitung

In diesem letzten Abschnitt des aktuellen Kapitels geht es nicht um die automatische Textverarbeitung, sondern um die sogenannten Begleitarbeiten dazu. In der vorliegenden Arbeit betreffen diese mehrere Einsätze von Helfern / Annotatoren sowie die Berechnungen einiger Indikatoren kognitiver Effekte, die manuell, halb- und z. T. vollautomatisch erledigt werden können. Die Auseinandersetzung mit aufgezählten Problematiken erfolgt in den nachfolgenden Abschnitten, deren theoretische Konzeption, praktische Realisierung und Auswertung innerhalb dieser Arbeit werden in Kapiteln 4, 5 und 6 verarbeitet (s. entsprechende Verweise jeweils am Ende der Abschnitte).

3.3.1 Annotatoren

Bei der automatischen Textverarbeitung wird man oft mit Problemen in mehreren Kontexten konfrontiert, bei denen eine manuelle Hilfe aufgrund der Subjektivität vorgenommener Aufgaben oder zum Zweck der Anfertigung von annotierten Daten (überwachte Verfahren) gebraucht werden. Meistens werden Annotatoren engagiert, die manuell solche Arbeiten wie Datenklassifikation, Systemevaluation etc. o. ä. durchführen. Zur Angabe zu Klassenzugehörigkeiten der Wörter, die für das überwachte Lernen gebraucht wurden, setzte Kaiser (2012, S. 21) zwei Annotatoren ein. Typisch ist dabei, dass eine Aufgabe von mehreren Helfern gelöst wird, denn „verschiedene Menschen können unter dem gleichen Begriff verschiedene Produkteigenschaften verstehen“ (ebd.). Die Übereinstimmung ihrer Arbeit wird dann entsprechend

gemessen, z. B.: „Zur Übereinstimmung der beiden Menschen, wird die Annotation eines Menschen als Goldstandard festgelegt“ (ebd.). Bei dem im Abschnitt 3.1.2.1.3, Seite 65 beschriebenen „Bootstrapping“-Verfahren zur Akquise der Adjektive von Vázquez et al. (2012) evaluierten die Autoren ihre Arbeit, indem sie auf 200 Dokumenten, die 12% des Korpus ausmachten, den Goldstandard selbst manuell annotierten (ebd., S. 1276). Alle polaren Adjektive in jedem Text, die sich im Lexikon befinden sollten, wurden mit den Werten ihrer Semantischen Orientierung (s. Abschnitt 3.1.2.1.2, Seite 63) versehen (positiv oder negativ), woraufhin ihre manuelle Markierung von einem Annotator erfolgte (ebd., S. 1276f.).
Im Weiteren wird lediglich zusammenfassend auf einige Arbeiten verwiesen, in denen Annotatoren eingesetzt und die im aktuellen Verfahren bereits angesprochen wurden.
Im Kontext der Identifikation von Produkteigenschaften (Aspekten) benutzten Archak et al. (2011) „Crowdsourcing-Based Technique“, die im Abschnitt 3.1.2.3 (Seite 74) beschrieben wurde. Die Aspekte wurden dabei von drei Annotatoren in freier Form beschrieben (ebd., S. 1489f.). Park et al. (2009) führten eine umfassende Evaluation des entwickelten Systems NewsCube zur Reduzierung von News Bias durch. Dabei wurden verschiedene Methoden wie Clickstream-Analyse, Fragebögen und Interviews verwendet (ebd., S. 450ff.).
Zur Evaluation von Kriterien, die durch die „verdächtigen“ Hotels im Sinne von Spam charakterisiert wurden, setzten Wu et al. (2010) 55 Annotatoren, die 46 Hotels auf ihre „Verdächtigkeit“ überprüften (ebd., S. 5). Dabei enthielten 41 davon bestimmte Sets an ausgearbeiteten Kriterien, 5 jedoch keine (ebd.). Die Ergebnisse zeigten, dass die o. g. fünf Hotels die niedrigsten „Verdächtigkeits-Scores“ aufwiesen (ebd., S. 5f.).
In der vorliegenden Arbeit sind ebenfalls mehrere Problematiken präsent, für die Annotatoren eingesetzt werden. Für welche Zwecke und wie genau dies geschieht, wird im Kapitel 5, Abschnitt 5.3.1 (Seite 167) erläutert. Wie die Übereinstimmung zwischen den Annotatoren (Inter-Annotator-Agreement) gemessen wird, wird im Kapitel 6, Abschnitt 6.2 (Seite 187ff.) erläutert.

3.3.2 Einige Fragestellungen aus der Statistik

3.3.2.1 Ausreißer

Zu einem der Kriterien fehlerfreier Bewertungen gehört die Feststellung, dass eine normative Bewertung zur Mehrheit aller Bewertungen zählt (s. Kapitel 2, Abschnitt 2.1.1.2.2, Seite 20f.). Im gleichen o. g. Abschnitt wurden Ausreißer angesprochen als Werte, die stark von der Masse der Daten abweichen, definiert und der Umgang mit ihnen angedeutet. Die Problematik der Ausreißer in den Datensätzen kann man folgendermaßen zusammenfassen: „Extreme Werte in den Datensätzen können die Ergebnisse substantiell beeinflussen. [...] Diese extremen Werte werden häufig als so genannte Ausreißer klassifiziert. Teilweise werden diese dann mit sehr einfachen Regeln identifiziert und eliminiert. [...] Bevor Ausreißer eliminiert werden, sollte versucht werden, festzuhalten, wie es zu diesen extremen Werten gekommen ist. Sobald man die vermeintlichen Ausreißer versteht, kann die Entscheidung, wie mit ihnen umzugehen ist, besser getroffen werden. Und die Elimination ist dann eine von mehreren Möglichkeiten.“ (Goerke, 2016, S. 23). Wenn Ausreißer in den Daten identifiziert wurden, müssen deren Gründe ermittelt werden (ebd., S. 38). Wenn notwendige Nacherhebungen wie die Ansprache der Befragten nicht möglich sind, da ihnen z. B. die Anonymität garantiert wurde, kann eine Elimination der Ausreißer einen Sinn ergeben (ebd.). Ein anderes Beispiel für Elimination ist ein Test mehrerer Internet-Dienste zur Datenauswertung von Stadler et al. (2007). Ein Gegenbeispiel, bei dem die Ausreißer nicht eliminiert werden sollten, wäre das Vorhandensein von verschiedenen wertvollen Segmenten in einem Datensatz, die fehlerhaft als Ausreißer eingestuft werden könnten (Goerke, 2016, S. 40). Diese Segmente wären z. B. ein großer Anteil von durchschnittlichen Kunden, die auf Werbemaßnahmen homogen reagieren und zur gleichen Zeit ein kleiner Anteil von Enthusiasten, die eine stärkere Reaktion darauf zeigen (ebd.). In diesem Fall sollten beide Segmente getrennt ermittelt und gezielt angesprochen werden (ebd.).
Bei der Auseinandersetzung mit Spam im Abschnitt 3.2.2.1 wurde auf der Seite 90 festgehalten, dass dieses sozialpsychologische Phänomen eindeutig zu Ausreißern gezählt werden kann. Gleichzeitig wurde festgestellt, dass Ausreißer als ein Indikator / ein Kriterium nur in Verbindung mit anderen charakteristischen Merkmalen zur Definition von Spam herangezogen werden kann. Ob als „Meinungsmüll“ (s. Seite 89) zur absichtlichen Datenverzerrung oder durch Zufall auftretende Inkonsistenzen (s. Abschnitt 3.2.2.2) in den Bewer-

tungen, ist die Elimination der beiden Phänomene sinnvoll. Das Kriterium „Ausreißer“ gehört auch zur Definition kognitiver Effekte, die in Kombination mit anderen Indikatoren unterscheidbar verschiedene Phänomene darstellen, die für die Weitererforschung wertvoll sein könnten. Die Elimination kognitiver Effekte würde dann eindeutig ein Verlust für die wissenschaftliche Auseinandersetzung mit ihnen bedeuten. Mit der Bestimmung der Ausreißer in Arztbewertungen befasst sich das Kapitel 4, Abschnitt 4.3.

3.3.2.2 Korrelationskoeffizient

Im Abschnitt 3.1.1.1 (Seite 55) wurde kurz der Nachweis des Halo-Effekts durch die Untersuchung von Pollock (2012) vorgestellt, wobei die Feststellung eines Zusammenhangs der Eigenschaften „Attraktivität“ und „Promiskuität“ angesprochen wurde. Dieser Zusammenhang wurde anhand des Korrelationskoeffizienten berechnet (ebd., S. 36). Bei der Aufzählung mehrerer Ebenen am Ende des Abschnitts 3.2.2.2 (Seite 95), auf denen nach Indikatoren zur Identifikation kognitiver Effekte gesucht wird, wurde festgehalten, dass der Zusammenhang / die Korrelation der Dimensionen auf den Gesamtdaten ausgerechnet werden muss. Zur Messung des linearen Zusammenhangs zwischen zwei Variablen existieren zwei Maße: Kovarianz und Korrelation (Leonhart, 2013, S. 261). „Ein positiver Zusammenhang liegt vor, wenn mit höherer Ausprägung in der Variablen X auch eine höhere Ausprägung in der Variablen Y gegeben ist und umgekehrt. [...] Ein negativer Zusammenhang liegt vor, wenn bei höherer Ausprägung in der Variablen X eine niedrigere Ausprägung in der Variablen Y vorliegt und umgekehrt. [...] Kein Zusammenhang liegt vor, wenn eine Ausprägung in der Variablen X keine Aussage über die Ausprägung in der Variablen Y erlaubt“ (ebd., S. 264). Im Vergleich zur Kovarianz ist die Korrelation ein standardisiertes Maß, das stets die Werte zwischen -1 und +1 annimmt, die dadurch besser vergleichbar und interpretierbar sind (ebd., S. 265). Es gibt eine Reihe von verschiedenen Formeln, mit denen man den Korrelationskoeffizienten berechnen kann (Clauß und Ebner, 1977, S. 117).

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \tag{3.9}$$

Eine dieser Formeln (Formel 3.9) beschreibt den PEARSON-BRAVEISschen Maßkorrelationskoeffizienten r, wobei

- $(x_i - \overline{x})$ Abweichung jedes X-Wertes von seinem Mittelwert,
- $(y_i - \overline{y})$ Abweichung jedes Y-Wertes von seinem Mittelwert,
- n Anzahl der Messwertpaare,
- s_x Standardabweichung der X-Werte,
- s_y Standardabweichung der Y-Werte

bedeuten (ebd.). Die Interpretation der Stärke des Zusammenhangs zwischen zwei Variablen hängt jedoch stark von dem untersuchten Problem ab (ebd., S. 122). „Sollte sich ein r = 0,40 für den Zusammenhang zwischen Leistungseigenschaften von Schülern und pädagogischen Eigenschaften ihres Lehrers ergeben, dann wäre dies Ausdruck einer überraschend „hohen“ Korrelation. Ergäbe sich dagegen r = 0,40 für den Zusammenhang zwischen Erst- und Zweitleistungen bei derselben Aufgabenart, dann müßte man dies als verwunderlich „niedrig“ betrachten“ (ebd.). Einer der Indikatoren zur Bestimmung des Halo-Effekts ist, wie bei Pollock (2012), der paarweise Zusammenhang zwischen den Bewertungsdimensionen. Durch die Erkenntnis, wie diese Dimensionen miteinander korrelieren, erhält man die Aussage über deren Abhängigkeit voneinander sowie die Abgrenzung der unscharfen Definitionen von Dimensionen (s. Kapitel 4, Abschnitt 4.1.2, Seite 103f.). Wie die Korrelation in der vorliegenden Arbeit ausgerechnet und in Bezug auf Effekte-Definitionen interpretiert wird, ist der Gegenstand der Abschnitte 4.3.1.2 auf der Seite 119f. (Kapitel 4) und 5.3.2.2 auf der Seite 175f. (Kapitel 5). Wie viele solche Berechnungen nötig sind, wird ebenfalls im Kapitel 5 im zuletzt genannten Abschnitt erläutert.

Kapitel 4

Methodische Vorgehensweise

Wie bereits mehrfach festgestellt, wird man bei der Identifikation kognitiver Effekte mit Meinungsäußerungen konfrontiert. Um diese Meinungsäußerungen automatisch zu erfassen, wurde sich mit den computerlinguistischen Disziplinen (aspektbasierte Stimmungsanalyse und domänenspezifische Informationsextraktion) auseinandergesetzt. Was die zuerst genannte Disziplin in der vorliegenden Arbeit betrifft, so kann man hier von einer eingeschränkten Stimmungsanalyse sprechen, die sich auf Meinungen zu Bewertungsdimensionen als Objekte konzentriert. Diese Stimmungsanalyse wird daher als Musterextraktion begriffen, wobei die irrelevanten Muster „überlesen“ werden (s. Seite 41). Die Muster beziehen sich hauptsächlich auf die explizit genannten Bewertungsobjekte (Dimensionen), die in typischen, diese Objekte bewertenden Phrasen vorkommen (Subsprachen, s. Seite 42).
Für die Musterextraktion eignen sich die im Kapitel 3, Abschnitt 3.1.2.4 beschriebenen regelbasierten Verfahren, zu denen auch die Methode lokaler Grammatiken zählt. Einige Gründe dafür sind:

- Notwendigkeit der Auseinandersetzung mit semantischen und syntaktischen Satzstrukturen
- Aspektbasierte relationale Beziehungen zwischen Bewertungsobjekten und dazu gehörige wertende Aussagen
- Korpusanalyse auf der Satz- und Phrasenebene
- Kleine Korpora der gewählten Domäne

- etc.[65]

Außer konkreten wertenden Mustern benötigt man auch andere Kriterien, anhand derer kognitive Effekte automatisch erkannt werden können. Aus diesem Grund werden alle Identifikationskriterien pro Effekt in so genannten Kriterienkatalogen systematisch zusammengefasst und erläutert (s. Abschnitt 4.3).

Nach der Ausführung der Grundlagen relevanter Fachdisziplinen, der Aufstellung der Definitionen und der Bestimmung relevanter automatisch identifizierbarer kognitiver Effekte im Kapitel 2 sowie nach der Auseinandersetzung mit aktuellen Arbeiten und Verfahren automatischer Sprachverarbeitung im Kapitel 3 kann die methodische Vorgehensweise für die vorliegende Arbeit entwickelt werden, was im aktuellen Kapitel geschieht.
Im Abschnitt 4.1 erfolgt eine Auseinandersetzung mit ausgewählten Problematiken der Domäne der Arztbewertungen, die im Sinne jeglicher automatischer Verarbeitung von Texten mehrere Herausforderungen darstellen. In den entsprechenden Abschnitten wird darauf eingegangen, wie mit den betreffenden Problemen umgegangen wird. In Abschnitten 4.2 und 4.3 werden Extraktions- und Identifikationsansätze vorgestellt, die als zwei Schritte des Systems CognIEffect bei der Identifikation kognitiver Effekte fungieren, wobei jeder von ihnen entsprechend evaluiert wird (s. Kapitel 6, Abschnitte 6.3 und 6.4). Einen besonderen Rang haben lokale Grammtiken als musterbasiertes Extraktionsverfahren (s. Abschnitt 4.2.1), das mit dem Korpusverarbeitungssystem *UNITEX* (s. Abschnitt 4.2.2) realisiert wird. Während im ersten Schritt die Extraktion wertender Muster erfolgt, wird im zweiten, nach der Aufstellung von Kriterienkatalogen, die anhand der im Kapitel 2, Abschnitte 2.1.2.1 erläuterten Selektionskriterien diskutiert und ausgearbeitet wurden, die Identifikation und Klassifikation automatisch identifizierbarer kognitiver Effekte (s. Kapitel 2, Abschnitt 2.1.3) durchgeführt. Das Identifikationssystem CognIEffect sowie die durchgeführten Begleitarbeiten dazu werden im Abschnitt 4.4 zusammengefasst und visualisiert.

[65] s. ausführlicher Abschnitt 4.2.1.1

4.1 Ausgewählte Domänenproblematiken

4.1.1 Sprachspezifik der Bewertungen

Jeder Patient, der eine Bewertung verfasst, hat (s)eine eigene Ausdrucksweise, verfügt über individuelle und z. T. einzigartige Formulierungen. Die Sprachstile der Bewertungen reichen von einer gesprochenen Sprache bis hin zu literarischen Ausdrücken. Syntaktisch gesehen, sind unvollständige, abgebrochene, aber auch stark verschachtelte Sätze vorhanden. Problematisch ist die Rechtschreibung. Das musterbasierte Verfahren „lokale Grammatiken“ erlaubt es, linguistische Muster auf der Satz- bzw. Phrasenebene zu extrahieren, wobei man sich auf die begrenzte Anzahl der Wörter und Wendungen konzentrieren kann, zu denen wertende Äußerungen von Patienten getätigt werden (hier: Bewertungsdimensionen und spezielle Ausdrücke zu Effekten). Was das medizinische Fachvokabular betrifft, so werden von Patienten teilweise Fachausdrücke gebraucht, um z. B. die gestellten Diagnosen zu beschreiben oder eine wertende Aussage zu machen, dass z. B. eine richtige Diagnose dieser oder jener Krankheit von einem Arzt gestellt wurde. Selbstverständlich kann hier, wie auch bei den Hotelbewertungen (Wolfgruber, 2015, S. 62), nicht von einer wissenschaftlichen Fachsprache die Rede sein, da die Bewertungen von Verbrauchern verfasst werden und „meist nur einen kleinen Teil der Branchenterminologie“ (ebd.) beinhalten, der „oft sehr nah am allgemeinen Sprachgebrauch“ (ebd.) liegt. Da jedoch Fachausdrücke in den wertenden Aussagen gebraucht werden, die sich auf eine der Hauptdimensionen (hier hauptsächlich: „Behandlung“, s. Seite 8) des Bewertungsportals Jameda beziehen, wird deren korpusbasierte Gewinnung (s. Kapitel 5, Abschnitt 5.1.2.2.1) notwendig sein.

4.1.2 Mehrdeutigkeit der Definitionen

Bei jeder zu bewertenden Dimension werden diese bei Jameda in Form von Info-Buttons definiert. Jede Dimension stellt sich für Bewertende in Form einer Frage dar, auf die man mit einer Note antworten muss. Zusätzlich kann man Informationen und Hilfen zu jeder Frage / Dimension erhalten, indem man auf den Info-Button mit dem Mauszeiger gehen würde. Z. B. Dimension „Behandlung“ wird einem Kunden, wie auf der Abbildung 4.1 gezeigt, dargestellt. Diese Definitionshilfen weisen jedoch Problematiken auf. So ist z. B. die Dimension „Betreuung“ als „Freundlichkeit des gesamten Praxisteams“ defi-

niert, wobei sich die Definition der Dimension „Freundlichkeit“ lediglich auf einen Arzt bezieht. In den Bewertungstexten wird jedoch die Freundlichkeit eines Arztes und eines Praxisteams angesprochen. In einigen solchen Fällen kann nur anhand der für die Dimension „Freundlichkeit“ vergebenen nummerischen Bewertung nachvollzogen werden, wessen Freundlichkeit tatsächlich gemeint wurde. Die individuelle Wahrnehmung der Menschen sorgt für individuelle Interpretationen zu dargebotenen Bewertungsdimensionen. Laut der beschriebenen Definition der Dimension „Behandlung“ von Jameda (Abbildung 4.1) sollten sich die Ausführungen von Bewertenden in freien Textfeldern auf die Therapie, ärztliche Kompetenz, die tatsächliche Besserung des gesundheitlichen Zustands, die Richtigkeit einer gestellten Diagnose u. ä. beziehen. In der Praxis wird diese Dimension jedoch in einigen Fällen als „Patientenumgang“ interpretiert (s. Beispiel (4.1)). Daraus lässt sich schließen, dass das Verstehen der zu bewertenden Dimensionen bei Patienten individuell verläuft, wie auch folglich die Bewertungen zu denselben.
In dieser Arbeit werden Definitionen der von Jameda vordefinierten Dimensionen nicht berücksichtigt. Der Grund dafür ist die Annahme, dass es selten der Fall ist, dass sich Patienten bei der Anfertigung ihrer Bewertungen an die Definitionen des jeweiligen Portals halten würden, sondern die Begriffe individuell interpretieren. Zur Extraktion wertender Muster zu genannten Dimensionen wird sich hauptsächlich nach expliziten Benennungen dieser Dimensionen von Patienten orientiert (s. z. B. Kapitel 2, Abschnitt 2.2.2.2.2) sowie Annotatoren zur Objektivierung dieser Muster herangezogen (s. Kapitel 3, Abschnitt 3.3.1; Abschnitt 4.2.3.1; Kapitel 5, Abschnitt 5.3.1.1).

Abbildung 4.1: Dimension „Behandlung“ bei Jameda

(4.1) „[...] man nur von diesen angeschnauzt und super schlecht behandelt wird, nein... sogar herr herlein hat mir ggü. heute erst so abwertende kommentare abgegeben [...]“.

4.1.3 Wie viele Dimensionen gibt es

Im Vergleich zu Noten / Sternen, die man für jede Dimension vergeben kann (und zu den Hauptdimensionen auch muss), lässt man den Patienten für die textuellen Bewertungen einen freien Raum, so dass diese ihre Erfahrungen mit den Ärzten nicht zu jeder Dimension ausdrücken müssen, sondern frei ausdrücken können. Dies bedeutet, dass die Äußerungen der Patienten mehr als bloße Bewertungsphrasen zu den Dimensionen darstellen und nicht von den skalierten Bewertungssystemen abgedeckt werden können (Geierhos et al., 2015b, S. 2). Es wird beispielsweise „allgemeine Zufriedenheit“ mit Pattern der Form „alles top“, „bin absolut zufrieden“ etc. ausgedrückt, obwohl es die genannte, jedoch denkbare Dimension gar nicht gibt. Im Beispiel (4.2) ist ebenfalls eine wertende Äußerung zu einer möglichen Dimension „Entfernung“ gezeigt, ohne dass diese Dimension von dem Bewertungsportal vordefiniert wurde. Die Aussage an sich macht semantisch durchaus einen Sinn, nur die angesprochene mögliche Dimension „Entfernung“ ist bei Jameda nicht definiert, so dass es für diese keine nummerische Bewertungsmöglichkeit gibt, was die Identifikation der Effekte aufgrund mehrerer fehlender Kriterien (s. Abschnitt 4.3) verhindern würde.

(4.2) „Ich ziehe jetzt zwar aus Düsseldorf weg aber für diese Ärztin nehme ich gerne 75 km Fahrtweg auf mich.“

Die hier beschriebene Problematik argumentiert die bereits mehrmals benannte Einschränkung wertender Muster, die sich nur auf die Äußerungen zu den vordefinierten Dimensionen beziehen.

4.1.4 Inkonsistenzen

Das Problem der Inkonsistenzen in Arztbewertungen wurde im Kapitel 3, Abschnitt 3.2.2.2 (Seite 91ff.) beschrieben. Aufgrund der von Geierhos et al. (2015b) getroffenen Unterscheidung der individuellen und kollektiven Inkonsistenzen wurden mehrere Ebenen zur Bestimmung der Inkonsistenzen in der vorliegenden Arbeit festgelegt. Individuelle Inkonsistenzen bilden eine Basis

für die vorliegende Arbeit, um die ersten Differenzen, die Kriterien der Norm einer Bewertung (s. Abschnitt 2.1.1.2.2, Seite 20) verletzen und somit auf kognitive Effekte hindeuten, aufzuspüren. Kollektive Inkonsistenzen, die anhand textueller und nummerischer Bewertungen derselben Arzt- oder Praxisleistungen verschiedener Patienten nachvollzogen werden können, erlauben Aussagen z. B. zu globalen Tendenzen und Trends, zu Zusammenhängen einzelner Dimensionen miteinander.

4.1.5 Skalentransformation

Wie die vergleichende Charakteristik dreier Bewertungsportale im Abschnitt 2.1.1.1.3 zeigt, sind die nummerischen Bewertungssysteme unterschiedlich. Das Problem, um dessen Lösung sich Islam (2014, zitiert in Geierhos et al. (2015b, S. 3)) bemüht und ein einheitliches Bewertungssystem vorschlägt. Dabei werden Stimmungen aus dem UGC abgeleitet, um nummerische Bewertungen, basierend auf der Polarität der gesamten Textbewertungen, zu generieren (ebd.).
In der vorliegenden Arbeit wird das beschriebene Problem insofern berücksichtigt, dass bei der Evaluation beider Schritte des Systems (s. Kapitel 6, Abschnitte 6.3 und 6.4) die Trainings- und Testkorpora unterschiedlich aufbereitet werden. Die Musterextraktion wird auf dem Jameda-Korpus trainiert und auf dem DocInsider-Korpus evaluiert, weil das Verfahren lediglich sprachliche Ausdrücke betrifft. Sobald es jedoch um die Identifikation kognitiver Effekte geht, wird das aktualisierte Jameda-Korpus in ein Trainings- und ein Testkorpus geteilt (s. Kapitel 5, Abschnitt 5.1.1.1.2, Seite 130 und Abschnitt 5.1.1.2.2, Seite 135), wodurch man die Vereinheitlichung des nummerischen Bewertungssystems umgehen kann, was den Rahmen dieser Arbeit andernfalls sprengen würde.

4.2 Extraktionsansatz

4.2.1 Lokale Grammatiken aspektbasierter Stimmungsanalyse und domänenspezifischer Informationsextraktion

In diesem Abschnitt werden lokale Grammatiken beschrieben. Das Augenmerk wird dabei auf die Punkte gelegt, die für die vorliegende Arbeit relevant sind.

4.2.1.1 Lokale Grammatiken als Extraktionsverfahren

Bei der Identifikation kognitiver Effekte geht es zunächst darum, bestimmte Muster, die Meinungen zu den vordefinierten Dimensionen ausdrücken, und deren Polarität zu erkennen. Diese Muster sind aus frei formulierten Texten der Nutzer zu extrahieren. Es gibt keine vorgeschriebene Vorgehensweise zur Lösung des angesprochenen Problems, gleichzeitig kann man sich nicht auf eine „umfassende Analyse des gesamten Inhalts“ (Neumann, 2004, S. 502) der Korpora einlassen. „IE- Systeme [...] sollen nur die Textpassagen analysieren bzw. „verstehen“, die relevante Informationen beinhalten. Was als relevant gilt, wird dabei durch vordefinierte domänenspezifische Lexikoneinträge oder Regeln dem System fest vorgegeben“ (ebd.). Die noch relativ geringen, jedoch für Textanalysen representativen Datenmengen im Bereich der Bewertungen von Arztpraxen veranlassen dazu, lokale Grammatiken als Verfahren zur Extraktion solcher relevanten Informationen zu verwenden. Selbsverständlich stellt die aus den Daten resultierende Korpusgröße nicht den einzigen Grund dar, lokale Grammatiken zur Musterextraktion einzusetzen. Bereits in der Einleitung zu dieser Arbeit auf der Seite 9 wurden Vorteile diesen Lernverfahrens angedeutet. Im Kapitel 3 (Abschnitt 3.1.2.4) wurden sie als muster- und regelbasierte Verfahren in mehreren aktuellen wissenschaftlichen Arbeiten hervorgehoben und deren Einsatz in diesen beschrieben. Gerade für das Erreichen des Teilziels der Extraktion von sprachlichen Mustern, die zur Identifikation kognitiver Effekte unabdingbar sind, scheint diese Methode durch die Effizienz der Graphen in der Beschreibung der Sprache (Nagel, 2008, S. 7) am besten geeignet zu sein. Aufgrund der im Abschnitt 4.1.1 beschriebenen syntaktisch bedingten Problematik eignen sich lokale Grammatiken zur automatischen Textverarbeitung durch die Möglich-

keit partieller syntaktischer Analysen der Teilsätze oder Phrasen (ebd.; Geierhos, 2010, S. 30; Wolfgruber, 2015, S. 123). Schließlich lassen sich lokale Grammatiken einfach modifizieren und wiederverwenden. Zu diesem Punkt wird im Abschnitt 4.2.1.2 modularer Aufbau und Universalität lokaler Grammatiken beschrieben. Im Rahmen der domänenspezifischen Informationsextraktion (s. Kapitel 2, Abschnitt 2.2.1.2) erfolgt in der vorliegenden Arbeit eine lokale Textanalyse, die auf Bewertungsdimensionen, deren Polaritäten und speziellen Phrasen zu kognitiven Effekten eingegrenzt ist. Solche semantisch bedingte Einschränkungen können im Rahmen statistischer Verfahren nicht gewährleistet werden. Schließlich kann man mit lokalen Grammatiken kontextbezogen arbeiten sowie korpusspezifische Akquise sprachlicher Ressourcen realisieren, wobei der Vorteil einer Interaktion von Syntax und Semantik (Stotz, 2018, S. 48) deutlich wird.

„Lokale Grammatiken sind nicht dazu bestimmt, die gesamte Grammatik einer Sprache zu bestimmen, sondern ihre syntaktischen und lexikalischen Phänomene in Bezug auf die jeweilige Domäne abzubilden“ (Geierhos, 2010, S. 79). Es scheint vernünftig, diese „Phänomene“, die im gebotenen Kontext nicht unbedingt immer nach den Grammatikregeln der deutschen Sprache gebildet wurden, aus dem Korpus selbst zu gewinnen, wie im Kapitel 5, Abschnitt 5.1.2 erläutert wird. „In der Regel werden lokale Grammatiken in Form von Graphen visualisiert. [...]. Graphen sind sehr benutzerfreundliche und intuitive Repräsentationen für lokale Grammatiken, welche äquivalenten Formalismen wie z. B. regulären Ausdrücken weit überlegen sind. Mithilfe diverser Grafikprogramme können lokale Grammatiken geradezu „kinderleicht“ erstellt, erweitert oder modifiziert werden“ (Geierhos, 2010, S. 79f.). Auf das in dieser Arbeit verwendete Programm wird ebenfalls im Abschnitt 4.2.2 eingegangen.

Auch die Nachteile des regelbasierten Verfahrens lokale Grammatiken sind nicht wegzudenken. Zum einen ist der Aufwand deren Erstellung immens groß, was zu einer hohen Genauigkeit der zu erzielten Ergebnisse zwar führt, gleichzeitig jedoch dadurch die Abdeckung der gesuchten Muster in den Daten auf einem geringen Stand hält. Zum anderen können zahlreiche Übergeneralisierungen der Muster erfolgen, was einen umgekehrten Effekt provozieren kann (höhere Abdeckung, geringe Genauigkeit). In jeder Hinsicht ist man mit der Problematik konfrontiert, dass „neither these dictionaries nor their extensions can cover all variants of relevant opinion phrases“ (Geierhos et al., 2015b, S. 7).

Die Musterextraktion kann hier daher als eine Aufgabe begriffen werden, bei

der nach einem optimalen Score gesucht wird, der als bestmögliches Ergebnis entsprechend der Ziele der jeweiligen Arbeit angepasst wird. Für die vorliegende Arbeit wird dieser Score auf 80% gesetzt, da, wie bereits an einigen Stellen erwähnt (s. Seiten 84, 88), kognitive Effekte wie auch andere sozialpsychologische Phänomene in der Minderheit auftreten. Durch einen hoch gesetzten Score hat ihre Identifizierbarkeit größere Chancen.

4.2.1.2 Modularer Aufbau und Universalität

Der Aufbau lokaler Grammatiken erfolgt modular (Gross, 1997, S. 329ff.), was bedeutet, dass zur Erkennung sprachlicher Phänomene eigene, möglichst universelle Grammatiken entwickelt werden. Eine linguistische Auseinandersetzung mit dem vorliegenden Korpus erlaubt es, für kleinere semantische Einheiten Regeln zu formulieren, diese dann miteinander zu kombinieren, um auf diese Weise zu immer größeren Aussagen, Phrasen oder ganzen Sätzen zu gelangen. Eine solche Zusammenfassung sprachlicher Phänomene in den sogenannten Modulen ermöglicht es, lokale Grammatiken in verschiedenen Kontexten bei der Verwendung derselben Phänomene universell einzusetzen. Gross (1997) stellt in seiner Arbeit „The Construction of Local Grammars“ anhand mehrerer Beispiele ein Modell zur Entwicklung lokaler Grammatiken aus kleinen semantischen Einheiten und deren Einbindung in die komplexeren syntaktischen Strukturen vor. Phänomene gleicher oder ähnlicher Natur aus der Sicht der Morphologie, Semantik und Syntax werden entsprechend in Gruppen eingeteilt und symmetrisch in den entsprechenden Graphen untergebracht (ebd., S. 349). Kleinere Module oder semantisch definierte Subgraphen werden kompakt und platzsparend in weitere Graphen eingebunden (ebd., S. 347), was diesen Übersichtlichkeit und Eleganz verleiht. Bei der linguistischen Sprachanalyse werden präzise die Regeln formuliert, um die Phänomene, wo möglich, zu generalisieren, wodurch man mehr Treffer erzielen kann, gleichzeitig werden Ausnahmen und verbotene Konstruktionen definiert, um Übergeneralisierungen zu vermeiden. Es wird außerdem darauf geachtet, dass bei der Generierung der Sätze alle Pfade sinnvoll definiert sind, was eine Benutzerfreundlichkeit sichert (ebd., S. 337). Obwohl das Modell lokaler Natur ist, kann man dieselben Graphen auch für die Beschreibung anderer Korpora verwenden (ebd., S. 331).
Die Entwicklung und das Aufbausystem der Graphen zur Extraktion sprachlicher wertender Muster in der vorliegenden Arbeit wird im Kapitel 5, Abschnitt 5.2 beschrieben.

4.2.1.3 Bootstrapping

In mehreren Abschnitten des Kapitels 3 (s. z. B. Abschnitte 3.1.2.1.2, 3.1.2.1.3 und 3.1.2.2.2) wurde bereits die Funktionsweise der „Bootstrapping“-Methode, die zu den hybriden Verfahren zählt, angedeutet. Im aktuellen Abschnitt wird „Bootstrapping“ in Verbindung zu lokalen Grammatiken erläutert sowie in der vorliegenden Arbeit u. a. zur Adjektiv-Akquise verwendet (s. Kapitel 5, Abschnitt 5.1.2.2.1, Seite 137ff.).
„Bootstrapping“ kann man nach Gross (1999, S. 229) als eine Methode zur Entwicklung lokaler Grammatiken, die um ein Schlüsselwort oder um eine semantische Einheit gebildet werden, beschreiben. In seiner Arbeit demonstriert er diese Methode am Beispiel des Worts „health“, indem er für dessen unterschiedliche Kontexte korpusbasiert lokale Grammatiken entwickelt, die über 33.000 relevante Strings erkennen (ebd., S. 238). Wichtig ist bei dieser Methode festzustellen, dass die Konstruktion lokaler Grammatiken einen empirischen Charakter beibehält (ebd., S. 237) und stark von einem konkreten Korpus abhängig ist. Auch in dieser Arbeit wird „Bootstrapping“-Methode zur Findung relevanter rechter und linker Kontexte um die Bewertungsobjekte eingesetzt (ebd., S. 236). Dieses Vorgehen dient dem Aufbau wertender Phrasen (s. Abbildung 5.13). Geierhos (2010) benutzt den „Bootstrapping“-Ansatz, um Lexika um Instanzen einer Objektklasse der Entitäten (z. B. Orte) zu erweitern. Durch eine detaillierte Beschreibung eines Kontextes einer Prädikat-Argument-Struktur spürt sie neue Instanzen dieser Klasse iterativ auf und reichert damit das Toponymlexikon um weitere Einträge an. Durch die Automatisierung dieses iterativen Prozesses läuft man dabei jedoch die Gefahr, bei zu vielen Iterationen ebenfalls zu viele Fehler zu produzieren, die das Gesamtergebnis unbrauchbar machen könnten. Daher scheint bei diesem Vorgehen eine semiautomatische überwachte Akquise lexikalischer korpusbasierter Ressourcen sinnvoll (ebd.). Ähnlich dem Vorgehen von Vázquez et al. (2012) wird „Bootsrapping“ in dieser Arbeit als Methode zur Akquise der Polaritätswörter mittels lokaler Grammatiken verwendet. Die Vorgehensweise basiert dabei auf den im Kapitel 3 (Aschnitt 3.1.2.1.3, Seite 64) beschriebenen Arbeiten von Hatzivassiloglou und McKeown (1997) (Seite 65) und Vázquez et al. (2012) (Seite 65). In Anlehnung auf die zuletzt genannten Autoren (Seite 1275) wird der „Bootstrapping“-Prozess in der vorliegenden Arbeit im Kapitel 5 (Abschnitt 5.1.2.2.1, Seite 137ff., Abbildung 5.7, Seite 143) dargestellt.

4.2.2 Korpusverarbeitungssystem *UNITEX*

UNITEX ist eine freie Software[66], ein Korpusverarbeitungssystem, mit dem lokale Grammatiken in Form von Graphen (s. 4.2.1.1) erstellt werden können. Derzeit unterstützt *UNITEX* 16 Sprachen [67]. Die Vorteile, die *UNITEX* im Vergleich zu anderen Softwares wie z. B. *GATE* bietet, bestehen u. a. in einer graphischen Oberfläche, durch die eine visuelle Übersichtlichkeit zu den entwickelten Grammatiken sowie deren Verschachtelungen gewährleistet werden kann (Stotz, 2018, S. 109).
In den folgenden Abschnitten werden die wichtigsten Funktionen von *UNITEX* angesprochen, die einen Überblick zur beschriebenen Software und deren Möglichkeiten geben. Die Ausführungen in Bezug auf die Anwendung dieser Funktionen in der vorliegenden Arbeit sind im Kapitel 5, Abschnitt 5.2 zu finden.

4.2.2.1 Vorverarbeitung

Bevor das Korpus mit lokalen Grammatiken verarbeitet wird, werden von *UNITEX* folgende Vorverarbeitungsschritte durchgeführt:

- Normalisierung: Mit dem Programm *Normalize* werden Folgen von Leerzeichen durch ein Leerzeichen ersetzt
- Tokenisierung: Mit dem Programm *Tokenize* wird eine Tokenliste mit der Häufigkeit der Token jeweiliger Sprache erstellt
- Satzenderkennung: Die Satzenden werden erkannt und mit {S} markiert

4.2.2.2 Graphen, Subgraphen und Transduktoren

Wie im Abschnitt 4.2.1.1 bereits angedeutet, stellen Graphen eine Visualisierung lokaler Grammatiken dar. Wie z. B. auf der Abbildung 4.2 zu sehen ist, haben sie einen Start- (Rechtspfeil) und einen Finalzustand (doppelt umrandet). Die Pfade, die vom Start- zum Finalzustand führen, werden von links nach rechts gelesen, sofern sie auch in die besagte Richtung verlaufen

[66] `http://www-igm.univ-mlv.fr/~unitex/` (12.09.2014).
[67] `http://www-igm.univ-mlv.fr/~unitex/index.php?page=7#lex-gram` (23.08.2016).

wie die betreffende Sprache in dieselbe Richtung gelesen wird, und symbolisieren Muster, die gefunden werden sollen (Geierhos, 2010, S. 80). Die Muster, die sich in jedem Zustand befinden, können als Lexikonkategorien (z. B. <ADJPOS>), einfache sprachliche Muster (z. B. „zeitnahme“), Subgraphen (z. B. „nicht_wörter_positiv“) usw. auftreten. Lokale Grammatiken haben in diesem Sinn die gleiche Bedeutung wie endliche Automaten mit der Möglichkeit, als Transduktoren eine Ausgabe (z. B. { ... ,. gzpos+GZ_EXP_zeitnahme}) zu produzieren (ebd., S. 86). Diese Ausgabe benötigt man zur Weiterverarbeitung bzw. zu Korrekturen beim Trainingsprozess erkannter Pattern (s. ausführlicher pragmatische Verwendung der Graphen in dieser Arbeit im Kapitel 5, Abschnitt 5.2.3). Die Tags des auf der o. g. Abbildung dargestellten Graphen enthalten die Informationen zur Dimension und Polarität der Aussage (gzpos = „Genommene Zeit“, positive Polarität; gzneg = „Genommene Zeit“, negative Polarität) sowie die Bezeichnung des Graphen („GZ_EXP_zeitnahme“). Wie im Abschnitt 4.2.1.2 ausgeführt, ist es sinnvoll, lokale Grammatiken modular aufzubauen, um von der Extraktion kleinerer Aussagen zu immer größeren zu gelangen. *UNITEX* bietet in diesem Zusammenhang eine ‚Verschachtelung‘ von Graphen, die Konstruktion sogenannter Subgraphen (s. o.). In diesem Fall fungiert der Graph „nicht_woerter_positiv“ auf der Abbildung 4.2 als Subgraph, in dem Negationswörter („nicht“, „nie“ etc.) zusammengefasst sind.

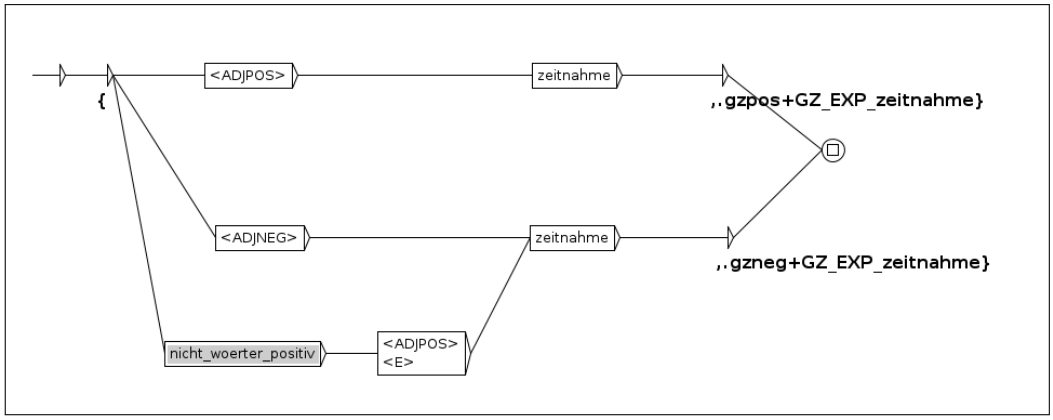

Abbildung 4.2: Ein Beispielgraph zur Erkennung positiver und negativer Aussagen zur Dimension „Genommene Zeit“ (Zustände, Subgraph, Output) (vgl. Paumier, 2015, S. 45)

4.2.2.3 Einbindung der Lexika

UNITEX stellt eigene Lexika zur Verfügung, die man auf das zu verarbeitende Korpus anwenden kann. Die Lexika enthalten Beschreibungen zu einfachen und zusammengesetzten Wörtern (Paumier, 2015, S. 45). Die Beschreibungen implizieren Kodierungen zu grammatischen Kategorien dieser Wörter, deren Flexionskodierungen (optional) und verschiedene semantische Informationen (ebd.). Einen Überblick zu einigen Kodierungen für englische Sprache bietet Paumier (2015, S. 45). Für andere Sprachen werden dieselben Kodierungen verwendet, es sei denn, es gibt für einige Sprachen eigene Spezifika (ebd). Da in dieser Arbeit mit der deutschen Sprache und auch mit einem deutschen Lexikon CISLEX (s. Abschnitt 3.1.2.1.2, Seite 62) gearbeitet wird, sind häufige grammatische Kategorien mit deren Kodierungen für das Deutsche in der Tabelle 4.1 dargestellt.

Code	**Beschreibung**	**Beispiele**
ADJ	Adjektive	gut, schön, laut
ADV	Adverb	einerlei, einerseits
KONJ	Konjunktion	aber, weil
DET	Artikel	alle, der
PREP	Präposition	durch, entlang, für
PDET	Präposition + Artikel	im, am, vorm
INTJ	Interjektion	ach, zack
N	Nomen	Achse, Gewusel, Iodierung
PRON	Pronomen	du, sie
V	Verb	entwässern, mögen

Tabelle 4.1: Häufige grammatische Kodierungen des *UNITEX*-Lexikons CISLEX

4.2.2.4 Verwendung von Variablen

Die Verwendung der Variablen erfüllt den Zweck, bestimmte Textpassagen im Korpus zu verändern. Dafür kann man diese Textpassagen in Variablen setzen und die entsprechende lokale Grammatik in einem Replace-Modus anwenden (Paumier, 2015, S. 138). Man könnte z. B. die Folgen ‚Adjektiv Nomen‘ im Text umkehren (ebd., S. 140) oder die Schreibweise des Datums automatisch verändern (vorher: 20.02.2015, nachher: 02.20.2015).

In der vorliegenden Arbeit wird diese Methode zur automatischen Erzeugung der Kodierungen für die erstellten Lexika verwendet. Zur Annotation des CISLEX werden lokale Grammatiken entwickelt, wobei man das Lexikon selbst als Korpus verwendet, auf das man die Liste mit den aus dem SentiWS (Abschnitt 3.1.2.1.2, Seite 63) extrahierten und kodierten (<APOS> für positive, <ANEG> für negative Adjektive) Adjektiven wiederum als Lexikon anwendet. Ein Beispiel für eine solche lokale Grammatik ist auf der Abbildung 4.3 zu sehen. Die Einführung zweier Variablen („adj_pos“ und „satz_end“) und die Anwendung des Graphen im „Replace-Modus“ auf das Korpus erlauben es, die bei der Vorverarbeitung (Abschnitt 4.2.2.1) markierten Satzenden zu löschen und dem gefundenen Muster eine neue Kategorie (ADJPOS) hinzuzufügen. Bei einem CISLEX-Eintrag mit Satzendmarkierung wie „schön,.ADJ:up{S}“ würde diesem somit die semantische Kategorie „ADJPOS“ hinzugefügt („schön,.ADJ:up+ADJPOS“) und das Satzendzeichen „{S}“ gelöscht werden. Daraufhin erfolgt später eine Sortierung der Reihenfolge von Kategorien („schön,.ADJ+ADJPOS:up“, vgl. Abbildung 5.4, Seite 139).

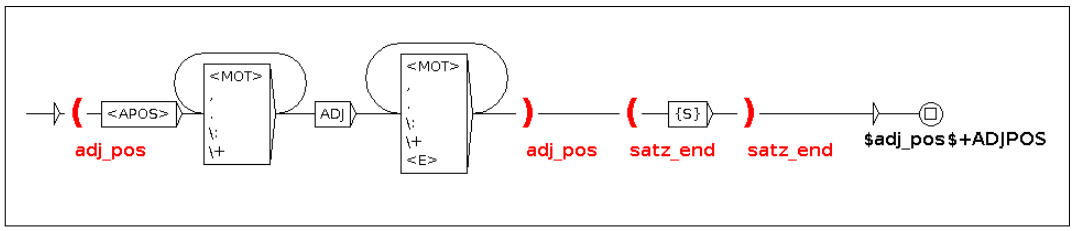

Abbildung 4.3: Abgleich polarer Adjektive aus SentiWS im CISLEX

4.2.2.5 Morphologischer Modus

Die Möglichkeit, auf einer Wortebene zu arbeiten, eröffnet sich durch die Erstellung lokaler Grammatiken im *UNITEX* und deren Anwendung in einem morphologischen Modus (Paumier, 2015, S. 130f.). Dies kann sehr nützlich sein, wenn man nach Wortarten mit bestimmter morphologischer Bildung sucht. Dabei kann man auf Lexika zugreifen, die im morphologischen Modus durchsucht werden (ebd., S. 131). In der vorliegenden Arbeit wird diese Methode zur Neugewinnung von Adjektiven zum Zweck der Lexikon-Erweiterung angewendet. Dabei werden morphologische Filter eingesetzt, mit deren Hilfe auf der Wortebene gearbeitet werden kann. Auf der Abbildung 4.4 wird nach Adjektiven gesucht, die mit in der ersten Box aufgezählten Präfi-

xen beginnen. Zwischen den Boxen, die in eckigen Klammern untergebracht sind, werden keine Leerzeichen zugelassen. Das in den Graphen eingebundene CISLEX wird dabei im morphologischen Modus verwendet (s. ausführlicher Kapitel 5, Abschnitt 5.1.2.2.1, Seite 142f.).

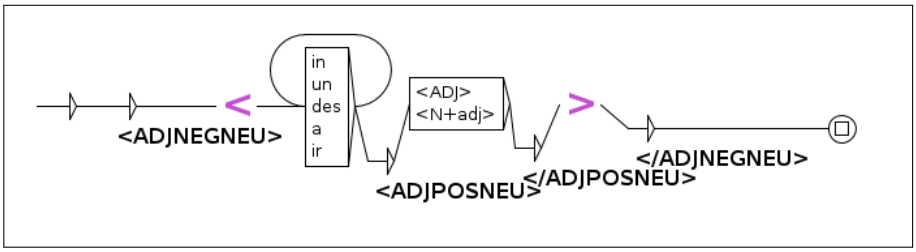

Abbildung 4.4: Graph zur Suche nach Adjektiven, die durch verneinende Präfixe gebildet werden

4.2.2.6 Kaskade

Mit der Kaskade „CasSys“ bietet *UNITEX* eine Möglichkeit, die entwickelten lokalen Grammatiken auf ein Korpus nacheinander anzuwenden (Paumier, 2015, S. 249). Die Modifikationen am Text, die mit einem Graphen

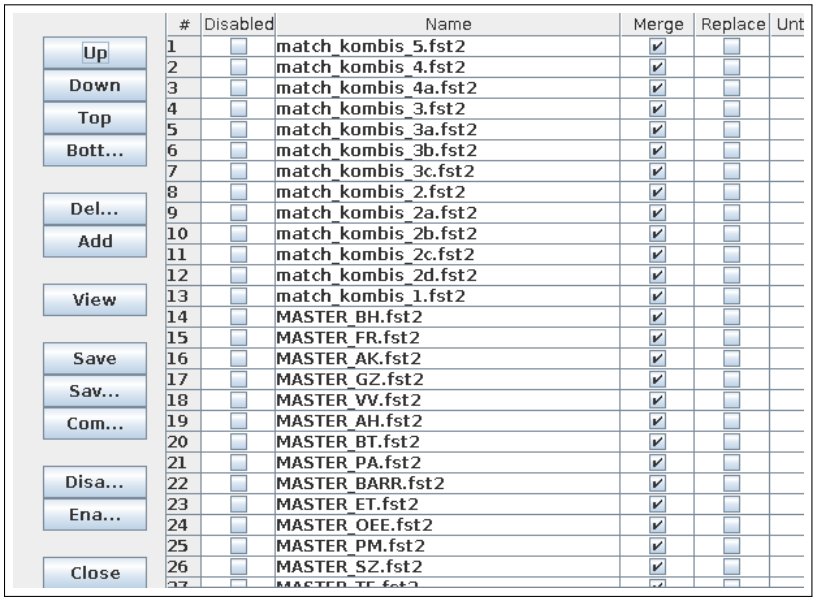

Tabelle 4.2: Kaskade der Musterextraktion

durchgeführt wurden, können für weitere Bearbeitungen mit den nächsten Graphen nützlich sein (ebd.). Wie aus der Tabelle 4.2 ersichtlich, werden mehrere Graphen, die durchnummeriert sind, im Merge-Modus nacheinander angewandt. In dieser Arbeit wird mit der in der o. g. Tabelle dargestellten Kaskade das Korpus mit den wertenden Mustern annotiert. Die Graphen zu wertenden Ausdrücken zu Dimensionen in der Kaskade werden dabei nach den Längen der Pattern sortiert. Bei Mastergraphen (s. Kapitel 5, Abschnitt 5.2.2.4, Seite 160) wird ihre Disjunktion in meisten Fällen angenommen, was bedeutet, dass ihre Reihenfolge in der Kaskade willkürlich ist. Weitere Ausführungen zur Kaskade sind im Kapitel 5, Abschnitt 5.2.3.1, Seite 162 zu finden.

4.2.3 Objektivierung der Entscheidungen

Bei der Arbeit mit UGC wird man häufig mit einer Reihe von subjektiven Entscheidungen konfrontiert. Als Lösung dafür kann man Annotatoren bzw. Beurteiler engagieren (s. Kapitel 5, Abschnitt 5.3.1.1, Seite 167), die unabhängig voneinander subjektive Entscheidungen treffen, wonach die Qualität deren Übereinstimmung nach einem bestimmten Maß ausgerechnet wird (s. Kapitel 6, Abschnitt 6.2, Seite 187). In folgenden Abschnitten werden zwei Problematiken der Subjektivität, die diese Arbeit betreffen, thematisiert.

4.2.3.1 Pattern-Definitionen

Die Entscheidung, zu welcher Dimension dieser oder jener in der natürlichen Sprache formulierte Ausdruck gehört, ist subjektiv. Wenn ein Patient z. B. eine Phrase in seiner Bewertung auf folgende Weise formuliert „Der Doktor hat mir die Angst vor der OP genommen“, könnte man dieses Pattern der Dimension „Vertrauen“ zuordnen mit der Begründung: Wenn man keine Angst vor einer Operation mehr hat, so hat man Vertrauen zum Arzt, der einem eben diese Angst nehmen konnte. Mit diesem Gedanken stellt sich jedoch die Frage, ob solch eine Vertrauensbasis nicht automatisch die fachliche und die soziale Kompetenz eines Arztes impliziert, so dass diese Aussage dann eben zwei oder sogar noch mehr Dimensionen zugeordnet werden kann.
Die Widersprüchlichkeit der von Jameda vorgeschlagenen Definitionen der Bewertungsdimensionen sowie die individuelle Wahrnehmung der Bewertenden diesbezüglich wurden im Abschnitt 4.1.2 thematisiert. Tatsächlich, wenn man sich in eine mögliche Situation eines Patienten versetzen würde, der ge-

rade eine Arztbewertung schreiben möchte, so würde er sich auch sicher nicht viel Zeit nehmen, um jede der Dimensionen zu definieren. Eine Bewertung wird oft schnell abgegeben, so dass man nicht einmal davon ausgehen kann, dass Patienten auch nur einen Blick auf die Erläuterungen der zur Verfügung stehenden Dimensionen werfen. Jeder betrachtet und versteht die Dimensionen individuell. Das bedeutet, dass es hier darum geht, eine Objektivierung der Dimensionszuordnung von subjektiv geäußerten Meinungen zu ermöglichen. Die Erläuterungen zur Arbeit der Annotatoren, um das beschriebene Problem zu lösen, sowie die erzielten Ergebnisse ihrer Arbeit sind jeweils im Kapitel 5, Abschnitt 5.3.1.1.1, Seite 167 und im Kapitel 6, Abschnitte 6.2.2.1, 6.5.1, und 6.5.2.1 zu finden.
Ein weiteres Problem der Subjektivität von sprachlichen Mustern bilden wertende Ausdrücke in natürlicher Sprache (explizit oder implizit) zu kognitiven Effekten wie Bestätigungsfehler und Diskriminierungen.

4.2.3.2 Bestimmung der Leitdimension

Wie im Abschnitt 2.1.3.1, Seite 38 definiert, neigen Patienten bei einem Halo-Effekt dazu, eine oder mehrere Dimensionen zu über- bzw. zu unterbewerten. Das passiert, weil sie sich von einer Ankerdimension leiten lassen. Da der Eindruck von der Anker- oder besser Leitdimension bei Bewertenden vermutlich groß ist, wird hier angenommen, dass die Patienten gerade diese Dimension in ihren Bewertungstexten zum größten Teil ansprechen und oder beschreiben. Das bedeutet, dass es möglich ist, automatisch herauszufinden, welche Dimension in einer Bewertung eine Leitdimension sein soll. Das Vorhandensein einer solchen Dimension pro Bewertung sollte – bei der Auffindbarkeit weiterer entsprechender Kriterien (s. Abschnitt 4.3) – als einer der Hinweise auf einen Halo-Effekt interpretiert werden. Das Programm, das diese Entscheidung automatisch treffen soll, wird ausgearbeitet und im Kapitel 5, Abschnitt 5.3.1.1.2 beschrieben. Jedoch ist die maschinelle Auswertung der qualitativen Daten nicht immer einfach, da jegliche Zuordnung dieser Art auch per Hand nicht immer eindeutig ist. Zur Objektivierung der Leitdimensionsbestimmung benötigt man auch an dieser Stelle Annotatoren, die Ergebnisse des o. g. Programms bewerten sollten (s. Kapitel 5, Abschnitt 5.3.1.1.2, Seite 168 und Kapitel 6, Abschnitte 6.2.2.2, 6.5.1 und 6.5.2.1).

4.3 Identifikationsansatz – Kriterienkataloge

Das Ergebnis des im Abschnitt 4.2 ausgeführten Extraktionsansatzes soll das mit den Mustern annotierte Korpus sein, wobei sich der hier zu beschreibende Identifikationsansatz u. a. auf die Weiterverarbeitung und Interpretation annotierter Muster bezieht. Die Tags der extrahierten und annotierten Muster sollten die Informationen zu den von Jameda vordefinierten Dimensionen enthalten bzw. die entsprechenden Effekte benennen und differenzieren. Gleichzeitig soll die Polarität wertender Aussagen aus den Annotationstags ersichtlich sein. Die eben beschriebene Extraktion gehört dabei zu einem der Kriterien / Indikatoren von Effekten. Ausgehend von den beschriebenen Meta-Informationen im Korpus wird nach anderen für jeden Effekt typischen Kriterien / Indikatoren gesucht, die dann in ihrer Kombination mit den annotierten Mustern auf einen oder einen anderen Effekt hindeuten werden. Die Klassifikation identifizierbarer kognitiver Effekte ergibt sich anhand der für jedes Kriterium vergebenen Wahrscheinlichkeiten (s. Kapitel 5, Abschnitt 5.3.2.3).

4.3.1 Halo-Effekt

4.3.1.1 Anker- und Anpassungsdimensionen

In der domänenbezogenen Diskussion zum Halo-Effekt (s. Abschnitt 2.1.2.2.1, Seite 31) wurde die Urteilsheuristik „Anker und Anpassung“ angesprochen, durch die dieser Effekt zustande kommen könnte. Ausgehend von der Domänenspezifik wurde diese Heuristik insofern interpretiert, dass es eine Anker- oder ‚Leitdimension‘ geben sollte, durch die eine oder mehrere andere Dimensionen (Anpassungsdimensionen) über- bzw. unterbewertet werden (s. Abschnitt 2.1.3.1). Wichtig ist bei diesen Ausführungen festzuhalten, dass die Anker- und Anpassungsdimensionen gleiche Polarität aufweisen müssen. Wie im Abschnitt 4.2.3.2 angedeutet, wird ein Programm entwickelt, das anhand sprachlicher Äußerungen der Bewertenden innerhalb einer Bewertung entscheiden soll, welche Dimension eine Anker- bzw. Leitdimension darstellt. Wie entscheidet man jedoch, welche Dimension(en) von dieser Leitdimension abhängig ist / sind, sprich: (Wie) Kann man Anpassungsdimension(en) bestimmen? Selbst wenn man alle anderen Dimensionen, die gleiche Polarität wie Leitdimension aufweisen (s. o.), bestimmt, müssen sie nicht unbedingt als abhängig von der Leitdimension interpretiert werden. Rückblickend auf

die Kriterien fehlerfreier Bewertungen im Abschnitt 2.1.1.2.2 kann man eine gewisse Norm z. B. innerhalb einer Arztpraxis definieren (s. Kapitel 3, Abschnitt 3.2.2.2, Seite 94), für die eine Mehrheit der Bewertungen sprechen würde. Mit anderen Worten: um die Wahrscheinlichkeit des Vorhandenseins von Anpassungsdimension(en) zu bestimmen, benötigt man eine Analyse, die über eine Bewertung hinaus gehen würde (weitere Kriterien). Auf diese Weise würde man ein umfassenderes Bild bzw. mehr Informationen zu einem oder mehreren Aspekte erhalten, der / die nur innerhalb einer einzigen Bewertung zu wenig Aussagekraft hätte(n). Um eine Norm, die durch die Mehrheit der Bewertungen auf der Arztpraxisebene deutlich wird, zu bestimmen, wird für jede Dimension ausgerechnet, ob sie mehr positive oder mehr negative Bewertungen aufweist. Da diese Berechnungen anhand der Textbewertungen zu aufwendig wären und man dabei kaum von einer quantifizierbaren Größe sprechen kann (Hu et al., 2013, S. 14), werden dazu nur nummerische Bewertungen herangezogen.

4.3.1.2 Korrelation der Dimensionen

Im Kapitel 3, Abschnitt 3.1.1.1 (Seite 55) wurde in einem durchgeführten Experiment eine hohe Korrelation zwischen bestimmten menschlichen Eigenschaften untersucht und nachgewiesen. Die beobachtbare Tendenz dabei war, dass eine hohe Korrelation zwischen zwei Eigenschaften als Halo-Effekt zu interpretieren war, was impliziert, dass diese Eigenschaften im Normalfall nicht stark miteinander korrelieren sollten. Pollock (2012, S. 36) hat die Korrelation mittels des Pearson's Korrelationskoeffizienten gemessen. Dieses Maß scheint auch für die paarweisen Messungen des linearen Zusammenhangs der Bewertungsdimensionen gut geeignet zu sein (s. Kapitel 3, Abschnitt 3.3.2.2). Das Problem der Fehlinterpretationen zeichnet sich hier jedoch durch die im Abschnitt 4.1.2 angesprochenen ‚unscharfen' Definitionen von Dimensionen sowie die individuellen Wahrnehmung und Interpretationen derselben von Patienten selbst. Trotzdem lässt sich aufgrund der im Abschnitt 2.1.1.2.2 definierten Aspekte einer normativen Bewertung empirisch ermitteln, wie stark oder schwach die Mehrheit der Dimensionen (paarweise) miteinander korrelieren (s. Seite 99 im Kapitel 3). Auf diese Weise könnte man z. B. die mit einer Leitdimension zwar schwach korrelierten, jedoch überbewerteten Anpassungsdimensionen (gleiche Polarität) bestimmen. Wie im vorigen Abschnitt begründet, werden die Berechnungen auch bezüglich der Korrelation der Dimensionen anhand nummerischer Werte vorgenommen. Allerdings ist

dies auf der Ebene der Gesamtdaten sinnvoll (s. Kapitel 3, Seite 94).

4.3.1.3 Ausreißer

Im Abschnitt 2.1.1.2.2, Seite 20 wurden Ausreißer definiert und im Sinne der Minderheit der Bewertungen beschrieben, die fehlerhaft sind. Ebenso kann man in einem kleineren Rahmen die Ausreißer auf einzelne Dimensionen innerhalb der Bewertungen zu einem Arzt / einer Arztpraxis einschränken, wie dies im Abschnitt 2.1.2.1.3, Seite 30 angedeutet ist. Bei der Auffindbarkeit solcher Dimensionen und bei anderen Auffälligkeiten, d. h. durch die beiden oben benannten Kriterien oder zumindest durch eines davon könnte man mit entsprechender Wahrscheinlichkeit (je nachdem, wie viele von beschriebenen Kriterien gleichzeitig vertreten sind) einen Halo-Effekt identifizieren. Mit der praktischen Realisierung des ersten genannten Kriteriums und gleichzeitig im Fall der Zugehörigkeit der Anpassungsdimension(en) zu den Ausreißern würde die im Abschnitt 2.1.3.1, Seite 38 aufgestellte Definition der Halo-Effekte erfüllt. Diese Besonderheit sollte den Halo-Effekt von anderen Effekten deutlich abgrenzen. Weitere für den Halo-Effekt charakteristische Kriterien können zusätzlich die Wahrscheinlichkeit des Auftretens diesen Effekts steigern.

4.3.1.4 Gleichheit nummerischer Bewertungen

Ein zusätzliches Kriterium, das auf den Halo-Effekt hindeuten würde, ist die Tatsache, dass alle Dimensionen gleich gut bzw. gleich schlecht bewertet werden. Wenn die nummerischen Bewertungen aller Dimensionen gleich sind, kann man daraus schließen, dass keine Differenzierung der Leistungen getroffen wurde. Der Grund dafür kann die Beeinflussung eines Bewertenden durch die Anker- bzw. Leitdimension (s. o.) sein.

4.3.2 Überbewertung

4.3.2.1 Anker- und Anpassungsdimensionen

Ähnlich den im obigen Abschnitt 4.3.1.1 vorgenommenen Ausführungen verhält es sich mit Anker- und Anpassungsdimension(en) bei Überbewertungen. Die Hauptunterschiede liegen in der Abhängigkeit zwischen den Dimensionen und deren eingeschränkter Kombination. Laut der Definition im Abschnitt 2.1.3.3, Seite 39 wird die Anpassungsdimension „durch das Vorhan-

densein der Ankerdimension(en) auf irgendeine Weise „gerechtfertigt“ und trotz des negativen Empfinden von derselben positiv benotet bzw. überbewertet, wodurch eine Verzerrung entsteht“. Diese Rechtfertigung kann nur in den Fällen erfolgen, wenn die Leitdimension eine der Hauptdimensionen mit positiver Polarität ist. Es ist gut vorstellbar, dass schlechte Parkplätze oder lange Wartezeiten durch eine ausgezeichnete Behandlung ‚entschuldigt‘ werden, jedoch nicht umgekehrt.

4.3.2.2 Inkonsistenzen in der Polarität

Wenn man erneut auf die Definition im Abschnitt 2.1.3.3, Seite 39 zurückblickt, so geht es offensichtlich bei den Anpassungsdimensionen um die Inkonsistenzen individuellen Charakters, wie diese bei Geierhos et al. (2015b) beschrieben wurden. Beim Vorhandensein dieser Inkonsistenzen und in der Kombination mit der entsprechenden Leitdimension, die auch zu Hauptdimensionen zählen soll (s. o.), bildet dieses Kriterium einen wesentlichen Grund zur Annahme, dass eine Patientenrezension mit einer Überbewertung versehen ist.

4.3.2.3 Ausreißer

Zählt eine Anpassungsdimension zusätzlich zu den Ausreißern innerhalb der Bewertungen zu einem Arzt / einer Arztpraxis (s. Abschnitt 4.3.1.3), so erhöht sich automatisch die Wahrscheinlichkeit, dass man eine Überbewertung identifizierte.

4.3.3 Bestätigungsfehler

4.3.3.1 Linguistische Muster

Bestätigungsfehler kommen zustande, wenn man einseitig nach Bestätigung eigener Hypothesen sucht (vgl. Definition im Abschnitt 2.1.3.2, Seite 39). In Arztbewertungen kann man sie anhand einiger typischer linguistischer Muster identifizieren (s. Abschnitt 2.1.2.2.3, Seite 32). Die Beispiele solcher Muster sind im eben genannten Abschnitt zu finden (s. Bsp. (2.1), Seite 33). Die im Beispiel aufgeführten Phrasen werden mittels lokaler Grammatiken aufgebaut[68]. Selbstverständlich sind die aufgeführten Muster nicht alle möglichen

[68] Weitere Erläuterungen dazu s. Kapitel 5, Abschnitte 5.2.2.3 und 5.2.2.4.

Phrasen, mit denen man die Bestätigungsfehler sprachlich realisieren kann. Diese Arbeit ist jedoch auf gewisse Einschränkungen angewiesen, was hier die Extraktion lediglich einiger ausgewählter Phrasen zu Bestätigungsfehlern impliziert.

4.3.3.2 Ausreißer

Ausreißer wurden bereits in mehreren Abschnitten (s. o.) beschrieben. Es bleibt hier zu bemerken, dass die Dimensionen, die in einer Bewertung als Ausreißer definiert werden können, in Verbindung mit oben erläuterten linguistischen Mustern einen sicheren Hinweis auf Bestätigungsfehler liefern sollen, wobei eine klare Abgrenzung zu anderen Effekten stattfinden sollte, unabhängig davon, dass das Identifikationskriterium „Ausreißer“ bei allen Effekten vertreten ist.

4.3.4 Diskriminierung

4.3.4.1 Linguistische Muster

Aus den Beispielen im Abschnitt 2.1.2.3.2, Seite 36ff. ist ersichtlich, dass eine Diskriminierung nicht isoliert, sondern aufgrund bestimmter Merkmale stattfindet. In Beispielen (2.4) finden die Abwertungen der Ärzte oder der Patienten aufgrund der Nationalität, des Alters und des Aussehens statt. Außer der im o. g. Abschnitt bereits gemachten Einschränkung auf die Extraktion explizit geäußerten Diskriminierungen, werden ebenfalls deren Merkmale auf das Alter und die Nationalität eingeschränkt. Linguistische Muster werden – wie auch bei Bestätigungsfehlern (s. Abschnitt 4.3.3.1, Seite 121) – mit lokalen Grammatiken extrahiert.
Zu erwähnen ist hier noch, dass die Extraktion der Diskriminierungen aufgrund derselben Domänenspezifik eine Herausforderung darstellt. Mit der Einverständniserklärung der Patienten mit den AGBs der Bewertungsportale nehmen diese in Kauf, dass es keine abwertende bzw. diskriminierende Äußerungen erlaubt sind[69].

[69] `https://www.jameda.de/qualitaetssicherung/nutzungsrichtlinien/`(24.01.2017), `http://www.docinsider.de/agb#/` (24.01.2017)

4.3.4.2 Ausreißer

Wie bereits beschrieben, würden auch hier die Ausreißer-Dimensionen in Verbindung mit linguistischen Mustern zu Diskriminierungen diesen Effekt eindeutig identifizieren.

4.4 Ablauf des Verfahrens CognIEffect

Aus den Ausführungen vorangegangener Kapitel und der Aufstellung der Identifikationskriterien im aktuellen Kapitel wird die automatische Identifikation kognitiver Effekte konzipiert und dabei die methodische Vorgehensweise mit Angaben betreffender Kapitel oder Abschnitte visualisiert. Der Ablauf der gesamten Arbeit zur Identifikation kognitiver Effekte, die das Identifikationsverfahren CognIEffect einschließt, ist auf der Abbildung 4.5 schematisch dargestellt. Das automatische Verfahren CognIEffect selbst ist gesondert abgebildet (s. Abbildung 4.6).
Wie anfangs des Kapitels angedeutet, besteht das CognIEffect aus zwei Schritten. Der erste Schritt betrifft die Extraktion linguistischer Muster aus dem UGC (s. Abschnitt 2.2.2.1, Seite 45) mit lokalen Grammatiken, dem eine besondere Rolle aufgrund des Aufwands und der Komplexität zuzuschreiben ist. Das Extraktionsverfahren wird daher als ein besonderer Bestandteil des gesamten Identifikationsverfahrens CognIEffect begriffen. Der zweite Schritt beinhaltet die entsprechende Kombination der Kriterien, wodurch Effekte identifiziert werden.

4.4.1 Vorarbeiten von CognIEffect

Bevor das automatische Verfahren in Gang gesetzt wird, werden die Vorarbeiten (Abbildung 4.5) wie folgt beschrieben. Ausgehend davon, dass kognitive Effekte aufgrund der Meinungsbildung auch in der Domäne der Arztbewertungen auftreten, benötigt man eine domänenbezogene Definition dieser Phänomene sowie eine empirische Analyse des UGC. Da die automatische Identifikation nicht für alle Effekte der getroffenen Auswahl möglich ist, findet die Überprüfung deren Identifizierbarkeit statt: ist der Effekt automatisch identifizierbar, wird für diesen ein Kriterienkatalog aufgestellt, in dem Indikatoren aufgezählt werden, die durch CognIEffect automatisch erkannt werden. In anderem Fall wird der Effekt verworfen, wodurch dessen Identifikation

beendet wird. Die oben genannte Abbildung zeigt die oben beschriebenen Verarbeitungsschritte mit den Angaben entsprechender Abschnitte, in denen sie ausführlicher erläutert wurden und werden.

4.4.2 Automatisches Verfahren CognIEffect

Für identifizierbare Effekte wird die automatische Erkennung aufgestellter Kriterien erfolgen (Abbildung 4.6 mit Angaben der Abschnitte). Ein Kriterium bei allen Effekten ist die Extraktion wertender Muster mit lokalen Grammatiken mit *UNITEX*. Einerseits werden Graphen zur Extraktion von Mustern zu vordefinierten Dimensionen, die die Leistungen der Arztpraxen betreffen, entwickelt. Andererseits werden damit wertende Aussagen zu zwei konkreten kognitiven Effekten (Bestätigungsfehler und Diskriminierungen) extrahiert. Der Einsatz lokaler Grammatiken betrifft in der vorliegenden Arbeit außerdem z. B. die vorherige Akquise lexikalischer Ressourcen, wodurch die Anfertigung und Anwendung eigener korpusbasierter Wörterbücher möglich wird. Zum Preprocessing gehören außerdem der

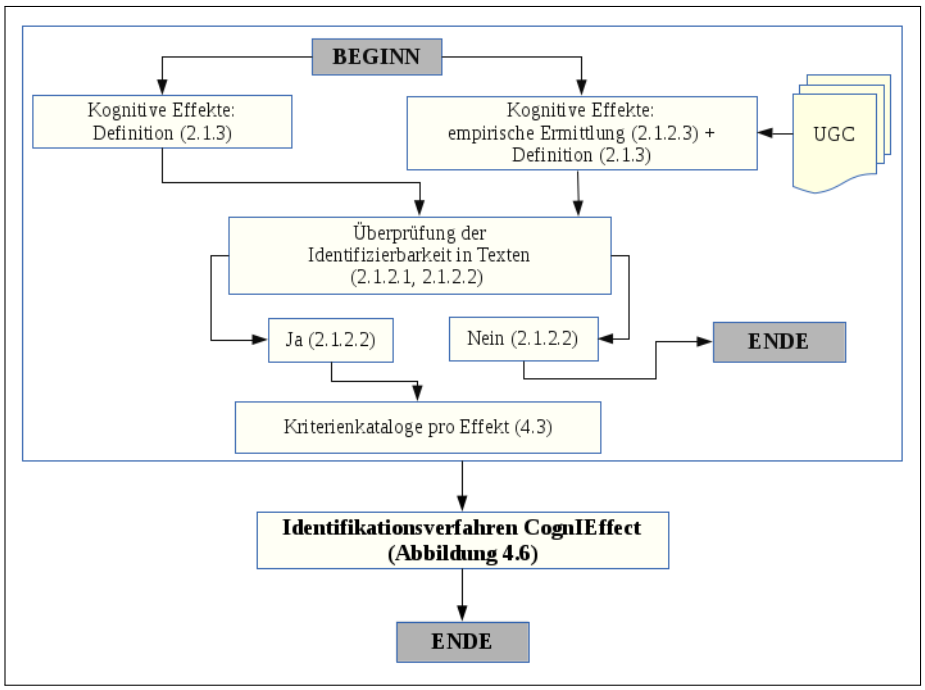

Abbildung 4.5: Vorarbeiten zum Identifikationsprozess CognIEffect

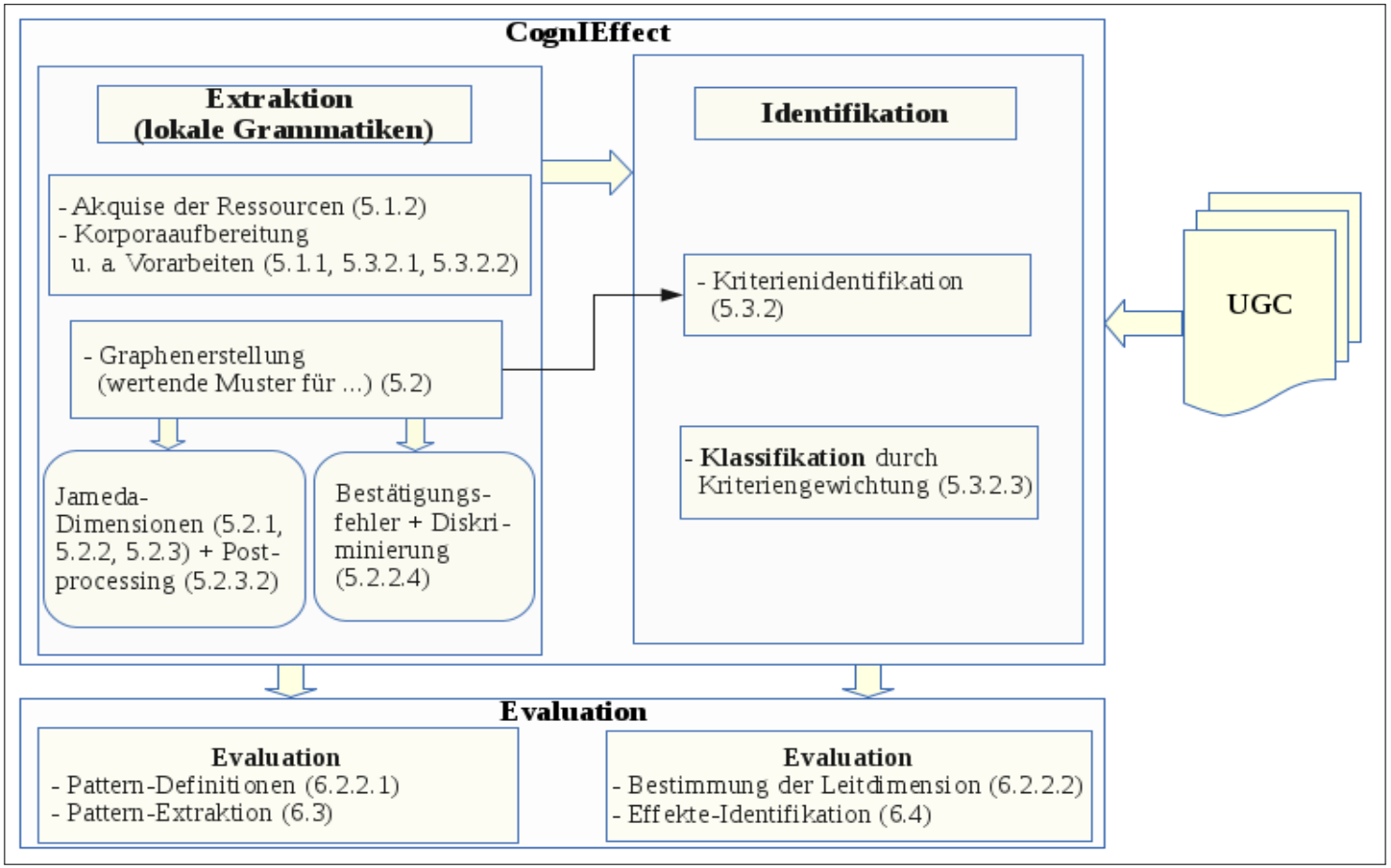

Abbildung 4.6: Ablaufdiagramm von CognIEffect und Evaluation

Einsatz der Annotatoren und andere Analysen, die auf Gesamtdaten erfolgen. Die beschriebene Extraktion ist, wie oben gesagt, eines der Identifikationskriterien von CognIEffect, von deren Ergebnissen die Ergebnisse des Gesamtsystems abhängig sind. Daher wird für die Extraktion ein bestimmtes Qualitätsziel gesetzt, das durch die Evaluation der Musterextraktion überprüft wird. Die eigentliche Identifikation kognitiver Effekte erfolgt im zweiten Schritt des Verfahrens. Neben den anderen Kriterien werden extrahierte und annotierte Muster automatisch identifiziert und mit einem entsprechenden Score versehen, wodurch die Klassifikation der Effekte für jede Bewertung entsteht. Die Evaluation des beschriebenen zweiten Schrittes erfolgt mit Berücksichtigung richtig extrahierter wertender Muster. Die Evaluation des Gesamtverfahrens CognIEffect ist dann als arithmetisches Mittel aus den Evaluationen beider Schritte zu verstehen.

Kapitel 5

Umsetzung des Identifikationsverfahrens CognIEffect

Die Identifikation kognitiver Effekte setzt sich, wie aus dem vorigen Kapitel bekannt (s. Abbildungen 4.5 und 4.6), aus zwei Schritten zusammen:

- Extraktion wertender Aussagen von Patienten zu vordefinierten Dimensionen (Patternextraktion) und zu zwei kognitiven Effekten
- Identifikation und Klassifikation ausgewählter kognitiver Effekte anhand aufgestellter Kriterienkataloge

In diesem Kapitel geht es um die Umsetzung bzw. praktische Realisierung der o. g. Schritte, deren ausgewählte Aspekte teilweise in einigen auf den o. g. Abbildungen angegebenen Abschnitten des vorigen Kapitels angesprochen wurden. Es werden weitere Aspekte beleuchtet, so dass ein umfassenderes systematisches Bild zum Gesamtverfahren entsteht. Zunächst werden Ressoursen beschrieben, die mittels lokaler Grammatiken korpusbasiert akquiriert und dann in weiteren zur Extraktion wertender Muster entwickelten Graphen eingebunden wurden. Auf die wichtigen semantischen, syntaktischen und pragmatischen Aspekte der Konstruktion lokaler Grammatiken wird im Abschnitt 5.2 eingegangen, was mit einigen Beispielen gestützt wird. Die Identifikationskriterien, die auf die Phrasenextraktion abzielen, werden in den Abschnitten erläutert, die sich mit den lokalen Grammatiken beschäftigen. Auf andere Kriterien, die nach der durchgeführten Musterextraktion

automatisch zu identifizieren sind, wird in den letzten Abschnitten dieses Kapitels eingegangen. Auf diese Weise entsteht eine klare, durch die eingesetzten Verfahren und zeitliche Ablaufprozesse abgrenzbare Vorstellung des durchgeführten Identifikationsmechanismus – CognIEffect.

5.1 Ressourcen

Unter Ressourcen werden in dieser Arbeit

- die Korpora, auf denen die Extraktions- und Identifikationsverfahren entwickelt und evaluiert werden,
  und

- externe und korpusbasierte Ressourcen, die zum Aufbau von für die Musterextraktion notwendigen Lexika dienen,

verstanden.

Im Abschnitt 5.1.1 werden die zuerst und im Abschnitt 5.1.2 die zuletzt genannten beschrieben.

5.1.1 Korpora

In den nachfolgenden Abschnitten werden Korpora aufgezeigt, die zum einen als Trainings- und Testdaten für die Musterextraktion und Effekte-Identifikation eingesetzt werden. Dass die Aufteilung der Korpora bei beiden Schritten des Identifikationsverfahrens unterschiedlich ist, liegt an den unterschiedlichen nummerischen Bewertungssystemen verschiedener Portale, was bedeutet, dass die Musterextraktion auf zwei Portalen trainiert und getestet wird. Die Identifikation von Effekten kann wegen der Einheitlichkeit nummerischer Werte nur auf einem Portal trainiert und getestet werden. Zum anderen werden thematische Subkorpora aus den Jameda- und z. T. DocInsider-Portalen gebildet, die einerseits zur korpusbasierten lexikalischen Ressourcen-Akquise verwendet und andererseits zur Identifikation der Kriterien entsprechend zusammengestellt werden.

5.1.1.1 Aufteilung in Trainings- und Testkorpora

5.1.1.1.1 Musterextraktion

Das größte Webportal im Bereich der deutschsprachigen Bewertungen von Arztpraxen ist Jameda. Die Rezensionen dieses Portals werden als Trainingskorpus zur Extraktion der wertenden Aussagen zu entsprechenden Dimensionen eingesetzt (s. Tabelle 5.2[70].). Für die Musterextraktion sind lediglich Bewertungen (Spalte 4) und deren Titel (Spalte 3) relevant, was bedeutet, dass die nummerischen Bewertungen und andere Angaben außer Acht gelassen werden. Die Bewertungstexte mit ihren Überschriften / Titeln werden daher in einer Datei als ein Korpus zusammengefasst. Wie aus der Tabelle 2.1 ersichtlich, sind die nummerischen Bewertungssysteme unterschiedlicher Portale verschieden. Da jedoch die Bewertenden ihre Bewertungstexte frei formulieren und somit auch jede Dimension, unabhängig von der Dimensionsklassifikation des jeweiligen Portals, ansprechen können, werden bei der Patternextraktion auf dem beschriebenen Jameda-Korpus das Training und auf dem auf die gleiche Weise zusammengestellten DocInsider-Korpus der Test durchgeführt[71].

	Jameda	**DocInsider**
Anzahl der Bewertungen	199 442	5 227
Ärzte	51 561	535
Sätze	766 087	14 309
Sätze pro Bewertung ∅	3,84	2,74
Wörter	11 225 609	172 453
Wörter pro Bewertung ∅	56,29	32,99

Tabelle 5.1: Statistische Angaben der Korpora zur Musterextraktion

[70]Hier wird lediglich ein Auszug des Korpus dargestellt, bei dem zwei Bewertungen zu einem Arzt (ArztID) zu sehen sind.

[71]Die Musterextraktion zu Diskriminierungen wurde auf den Testdaten von Jameda (1400 Bewertungen) durchgeführt (s. Tabelle 5.3 (Jameda (Test)) und die beiliegende DVD/Evaluationsdaten/DISKR_EVALUATION.txt).

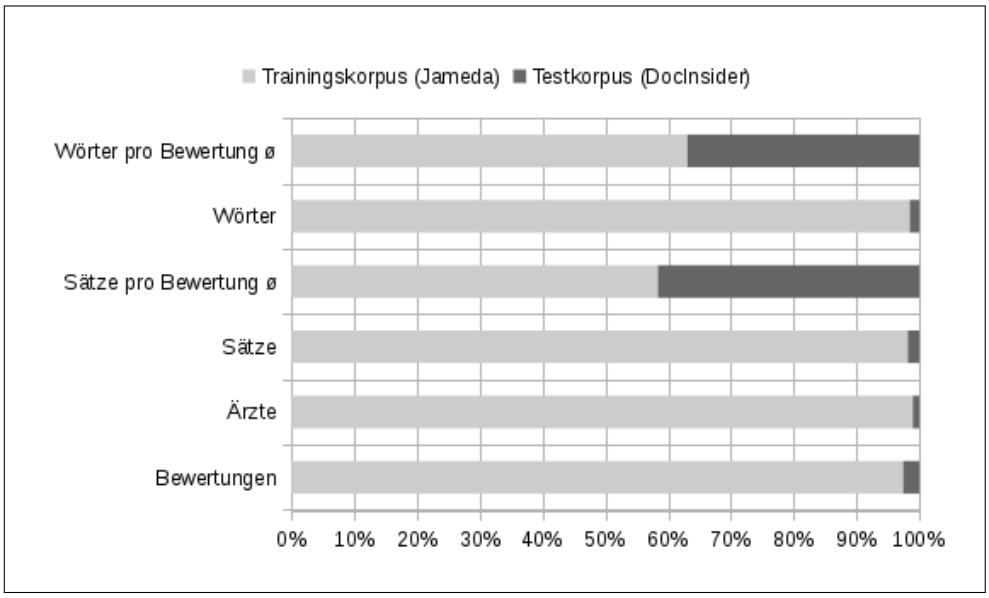

Abbildung 5.1: Prozentuelle Anteile von Trainings- und Testdaten der Musterextraktion

Tabelle 5.1 zeigt einige statistische Angaben der Trainings- und Testkorpora, was auf der Abbildung 5.1 in prozentuellen Anteilen visualisiert wird.

5.1.1.1.2 Effekte-Identifikation

Bei der Identifikation kognitiver Effekte sind nicht nur die frei formulierten und nach der Musterextraktion entsprechend annotierten Arztbewertungen, sondern auch andere Angaben, die in der Tabelle 5.2 gezeigt wurden, notwendig. Vor allem sind nummerische Bewertungen vordefinierter Dimensionen (hier: ‚Noten zu Jameda-Dimensionen') für den Identifikationsprozess relevant. Aufgrund der erwähnten Unterschiede in den Klassifikationssystemen von Bewertungsdimensionen verschiedener Portale (s. Tabelle 2.1) und wie bereits am Anfang des Abschnitts 5.1.1 angedeutet, ist es sinnvoll, die Klassifikation eines Bewertungsportals (in diesem Fall Jameda) komplett zu übernehmen. Eine Anpassung auf ein gemeinsames Klassifikationssystem mehrerer Bewertungsportale würde den Rahmen der vorliegenden Arbeit übersteigen und scheint daher nicht sinnvoll zu sein. Nach Ausarbeitung des Extraktionsverfahrens werden die Muster-Annotationen (s. Abschnitt 5.3.1.2.1) auf

BewertungID	**ArztID**	**Titel**	**Bewertung**	**Datum**	**Kassenart**	**Gesamtnote**	**Noten zu Jameda-Dimensionen**	**Zeitstempel**	**Alter**
1 022 776	80 069 935	Tolle Behandlung, super Ärztin	Ich bin zur Frau Sonntag, da ein anderer Arzt mein Leid nicht ernst genommen hat und alles was ich ihm gesagt habe nur wiederholt hat und alles auf meine Körpergröße geschoben hat.	12.08.2013	Kassenpatient	1.4	Vertrauensverhältnis = 2.0; Aufklärung = 1.0; ...	2013-10-11 21:46:57	unter 30
872 007	80 069 935	sehr zu empfehlen	sehr zu empfehlen.Nimmt sich zeit.erklärt bei Rückfragen des Patienten die Behandlung oder den op eingriff.	09.04.2013	Kassenpatient	1.0	Vertrauensverhältnis = 1.0; Aufklärung = 1.0; ...	...	...

Tabelle 5.2: Auszug aus dem Korpus zur Identifikation kognitiver Effekte

das Jameda-Korpus übertragen, das nicht nur Bewertungstexte und -titel, sondern auch vollständige Angaben, wie in der Tabelle 5.2 gezeigt, enthält. Die Anzahl der Arztbewertungen in diesem Korpus wird für die Effekte-Identifikation lediglich auf diejenigen Arztpraxen reduziert, die mindestens über drei Patientenbewertungen verfügen. Dies ist nötig, um den im Kapitel 2, Abschnitt 2.1.1.2.2, Seite 20 erläuterten Aspekt der Mehrheit von Bewertungen zu einem Arzt / einer Arztpraxis zu gewährleisten. Einige der aussortierten Bewertungen werden für die Annotatorenarbeit bei der Ermittlung der Leitdimension eingesetzt, was einer sinnvollen Ressoursennutzung dient (s. Abschnitt 5.3.1.1.2, Seite 168). Vor der Durchführung des Identifikationsverfahrens wird das Jameda-Korpus[72] in eine Trainings- und Testmenge geteilt, deren prozentuelle Anteile in der Tabelle 5.3 aufgeführt und auf der Abbildung 5.2 gezeigt sind.

5.1.1.2 Bildung thematischer Subkorpora

Sogenannte Subkorpora benötigt man, um korpusspezifische Teilaufgaben wie die Ressourcenakquise, Annotatorenarbeiten etc. zu erledigen sowie zielspezifische Anpassungen wie z. B. bestimmte Anordnungen der Bewertungen im Korpus vorzunehmen, um die Identifikationsvoraussetzungen zu erfüllen.

	Jameda (Training)	**Jameda (Test)**
Anzahl der Bewertungen	158 085	17 514
Ärzte	24 674	2 742
Sätze	428 336	47 385
Sätze pro Bewertung ∅	2,71	2,7
Wörter	6 631 748	728 328
Wörter pro Bewertung ∅	41,95	41,59

Tabelle 5.3: Statistische Angaben der Korpora zur Effekte-Identifikation

[72]Weitere Informationen zum beschriebenen Korpus sind dem Abschnitt 5.1.1.2.2 (Seite 135) zu entnehmen.

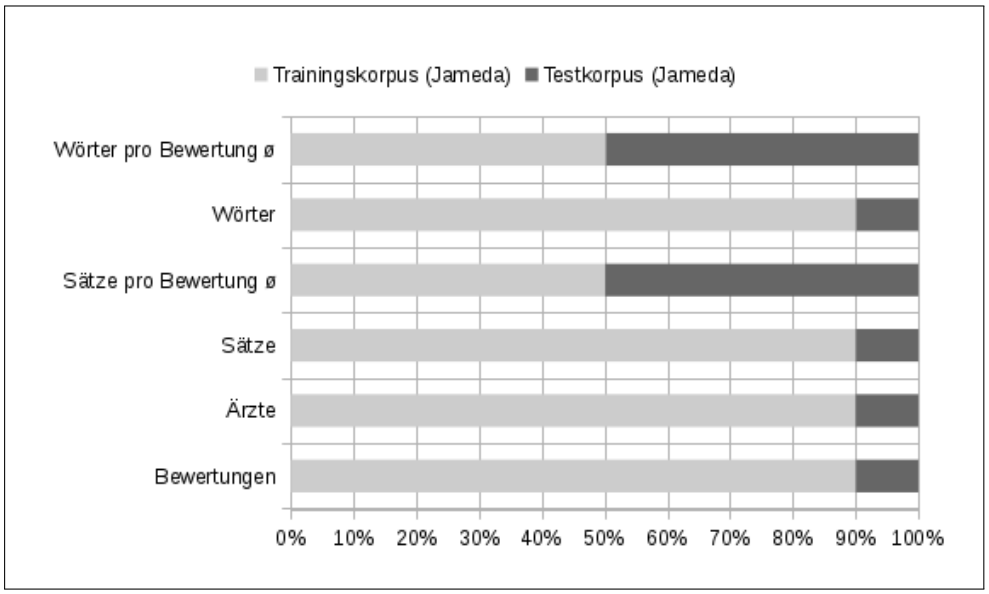

Abbildung 5.2: Prozentuelle Anteile von Trainings- und Testdaten der Effekte-Identifikation

Was die Ressourcen betrifft, so können z. B. die Namen von Ärzten sowie Fachwörter, die Patienten in ihren Bewertungen verwenden, von Interesse sein. Als Identifikationsvoraussetzungen sind z. B. einzelne Bewertungsangaben, wie diese in einer ‚Tabellenreihe' (Tabelle 5.2) dargestellt wurden, relevant, um z. B. die ‚Leitdimension' pro Bewertung zur Identifikation des Halo-Effekts zu bestimmen. Außerdem ist die Mindestanzahl und die Anordnung der Bewertungen zu einem Arzt / einer Arztpraxis als weitere Identifikationsvoraussetzungen bei allen Effekten zu verstehen, da diese Kriterien eine Basis zur Bestimmung der Ausreißer bilden.
In den nachfolgenden Abschnitten werden Korpora, die Bewertungen pro Facharztgruppe und pro Arzt(praxis) beinhalten, vorgestellt.

5.1.1.2.1 Bewertungen pro Facharztgruppe

Bei korpusbasierten Ressourcen (Abschnitt 5.1.2.2) geht es darum, Entitäten, Polaritätswörter, Fachausdrücke u. ä. aus den Korpora selbst zu gewinnen, um diese dann in die Lexika einzutragen, welche in die lokalen Grammatiken eingebunden werden sollen. Eine der Ressourcen bildeten dabei fachspezifische Ausdrücke, die Patienten in Bewertungen oft verwenden und somit z. B.

eine der wichtigsten Bewertungsdimensionen von Jameda „Behandlung“ charakterisieren und beurteilen. Um ein Lexikon mit Fachausdrücken zu erstellen, ist es vernünftig, die Bewertungen zu einigen Fachärzten zusammenzufassen. Rezensionen von vier Fachärzten (HNO, Zahnarzt, Schönheitschirurg und Augenarzt), die ihrer Popularität entsprechend auffällig waren, wurden aus den Jameda- und DocInsider-Portalen zu einem Korpus zusammengefasst.

Facharzt	**Lexikon-Tag**	**annotiert**	**neu gewonnen**
Augenärzte	AUGNOM	51	92
Behandlung allgemein	*BHNOM*	*351*	*239*
HNO-Ärzte	HNONOM	46	53
Schönheitschirurgen	SCHNOM	220	110
Zahnärzte	ZNOM	189	376

Tabelle 5.4: Allgemeine und fachbezogene Äußerungen der Patienten

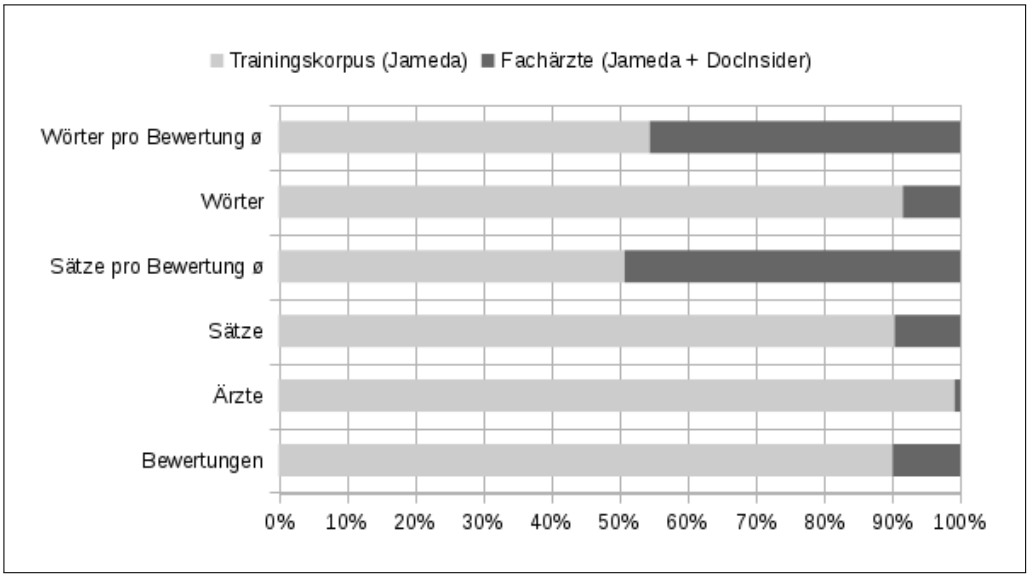

Abbildung 5.3: Prozentuelle Anteile von Trainingsdaten und Fachärzte-Korpus

Die Größe des erstellten Korpus in Bezug auf die Größe des gesamten Jameda-Korpus ist auf der Abbildung 5.3 zu sehen. Aus dem erstellten Korpus wurden Nomen extrahiert, manuell von allgemeinen Begriffen, die in jeder Arztpraxis verwendet werden können, aussortiert und dem Lexikon CISLEX_SENTIWS (s. Abschnitt 5.1.2.2.1, Seite 137) hinzugefügt (falls nicht vorhanden) bzw. ins Lexikon neu eingetragen. Als Beispiel für einen allgemeinen Begriff wäre das Wort „Besprechung“ zu nennen. „Transplantation“ wäre allerdings ein Begriff, den man eher in einer Zahnarztpraxis oder in der Schönheitschirurgie verwendet. Tabelle 5.4 zeigt, wie viele Fachwörter pro Facharztgruppe annotiert und wie viele davon neu gewonnen wurden. Den Auszug mit einigen Fachwörtern aus dem Lexikon CISLEX_SENTIWS und weitere Erläuterungen kann man im Abschnitt 5.1.2.2.1, Seite 144 finden.

5.1.1.2.2 Bewertungen pro Arzt(praxis)

Das Kriterium, das bei allen Effekten vertreten ist, sind Ausreißer. Dieses kann nur durch die entsprechende Anordnung der Bewertungen identifiziert werden, wobei es sich in allen Fällen um Ausreißer-Dimensionen immer nur innerhalb einer Praxis handelt und nicht im Rahmen des gesamten Bewertungsportals. Dies geschieht aus dem Grund, dass es nicht das Ziel der vorliegenden Arbeit ist, allgemein die Ausreißer in der Arztpraxendomäne zu untersuchen, um z. B. bestimmte Trends bzgl. konkreter Bewertungsdimensionen festzustellen, sondern mögliche individuelle Unstimmigkeiten in einzelnen Bewertungen einzelner Patienten herauszufinden. Die Einschränkung der Anzahl der Bewertungen auf drei innerhalb einer Praxis begründet sich dadurch, dass zur Objektivierung der Patientenurteile davon ausgegangen wird, dass zwei der Bewertungen in ihrer Mehrheit eine objektive Einschätzung zu Arzt- bzw. Arztpraxisleistungen darstellen. Zu diesem Zweck müssen zunächst Bewertungen nach ihrer Zugehörigkeit zu einer Praxis geordnet und dann diejenigen Praxen aussortiert werden, die weniger als drei Bewertungen enthalten. Somit reduziert sich die Anzahl unterschiedlicher Ärzte / Arztpraxen von 57 663 auf 27 416 (vgl. Tabellen 5.1 und 5.3). Bei der Anzahl der Ärzte handelt es sich hier um ein am 22.10.2013 aktualisiertes Jameda-Korpus, das zu Beginn 217 841 Bewertungen mit allen Meta-Daten in xml-Format enthielt, und nach beschriebenen Modifikationen über insgesamt 175 599 Bewertungen verfügt (s. Tabelle 5.3). Das Korpus stellt dasselbe Korpus dar, wie dieses im Abschnitt 5.1.1.1.2, Seite 130 beschrieben wurde, und hat die in der Tabelle 5.2 dargestellte Struktur. Jede Bewertung mit den aufgezählten

Angaben (BewertungID, ArztID, Titel usw. sind Metainformationen in den die entsprechenden Daten umschließenden Tags) steht in einer Zeile. Hier erfolgte lediglich die Beschreibung und die Begründung der Anordnung von Bewertungen im Korpus zur Effekte-Identifikation.

5.1.2 Lexikalische Ressourcen

In diesem Abschnitt wird die korpusbasierte Akquise lexikalischer Ressourcen beschrieben. Sie wurde zum größten Teil auf dem Trainingskorpus Jameda durchgeführt. Sicher kann man auch Daten anderer Bewertungsportale zum beschriebenen Zweck einbeziehen, jedoch wird hier nicht – aufgrund einer relativ kleinen Menge von Bewertungsportalen der Arztpraxendomäne – von einem großen lexikalischen Gewinn ausgegangen, sondern eher von einem größeren Zeitaufwand. Außer der von *UNITEX* (s. Kapitel 4, Abschnitt 4.2.2) zur Verfügung gestellten Lexika, besteht die Möglichkeit, eigene Lexika in das Programm einzubinden. Bei der korpusbasierten Arbeit ist das besonders interessant, da man domänenspezifische Wörter bzw. Wendungen entsprechend kodieren oder bereits bestehende Kodierungen erweitern kann. Diese Wörter und Wendungen können in den lokalen Grammatiken über die vergebenen Kategorien angesprochen werden.

5.1.2.1 Korpusunabhängige Ressourcen

Im Kapitel 3, Abschnitt 3.1.2.1.2 (Seite 62) wurden das Lexikon CISLEX vorgestellt und einige Arbeiten aufgeführt, die dieses Lexikon in verschiedenen Kontexten verwendeten. Da das Lexikon umfangreich ist und viele lexikalische Entitäten (u. a. auch die Stimmungswörter) enthält (s. ebd.), findet es seine Anwendung in der vorliegenden Arbeit. Ein anderes Lexikon SentiWS, das in demselben Kapitel, Abschnitt 3.1.2.1.2 (Seite 63) beschrieben wurde, bildet für diese Arbeit eine Grundlage zur Akquise polarer Wörter. Das CISLEX wird mit den Stimmungswörtern bzw. den Kodierungen von denselben aus dem SentiWS erweitert und als eine neu aufgearbeitete Ressource zur Extraktion wertender Muster eingesetzt. Außerdem wird mit Hilfe von einem Stimmungslexikon die korpusbasierte Akquise betrieben, was ebenfalls zur Erweiterung von CISLEX mit neuen Einträgen / Kodierungen führt und im Abschnitt 5.1.2.2 ausführlich beschrieben wird. Aufgrund der oben beschriebenen Art der Erstellung des Ergebnislexikons wird diesem die Bezeichnung *CISLEX_SENTIWS* vergeben. Aus dem Auszug aus CISLEX_SENTIWS auf

der Abbildung 5.4 im Abschnitt 5.1.2.2.1 sind die bestehenden („ADJ“, „N“) und neu eingeführten Kodierungen („ADJPOS“, „WZ(NEG)“) ersichtlich.

5.1.2.2 Korpusbasierte Ressourcen

Wie genau die Akquise erfolgte, welche Wörter auf welche Weise gewonnen wurden und wie gut die Ergebnisse des Vorgehens waren, wird in diesem Abschnitt erläutert. Außer CISLEX_SENTIWS wird auf die anderen selbst entwickelten Lexika eingegangen.

5.1.2.2.1 Erweitertes CISLEX-Lexikon (CISLEX_SENTIWS)

Wie angesprochen, ist CISLEX_SENTIWS ein CISLEX-Lexikon, das mit Stimmungswörtern aus SentiWS erweitert wurde. Nach der durchgeführten Akquise (Gewinn: 1671 Adjektive) sowie durch die systematische manuelle Erweiterung des Lexikons um Adjektive im Prozess des Patternmatching wurden 3181 neue Adjektive gewonnen.

a) Akquise wertender Adjektive

Wertende Adjektive als „Polaritätsanker“ spielen eine entscheidende Rolle bei der Extraktion der Sentiments (Shailesh, 2015, S. 116; Vázquez und Bel, 2013, S. 3557; Wolfgruber, 2015, S. 24f.). Dieser Feststellung kann man sich anschließen, besonders im Kontext der Bewertungsportale, auf denen die Meinungsausdrücke oft auf einzelne Adjektive beschränkt sind. In diesem Sinne wird sich in der vorliegenden Arbeit auf die korpusbasierte Gewinnung der Adjektive zum Aufbau von Lexika konzentriert.
Wie im Kapitel 3, Abschnitt 3.1.2.1.3 beschrieben, wurden Adjektive im Ansatz von Wolfgruber (2015) „nach dem Muster <ADJ> (= alle Adjektive im Text)“ (ebd., S. 66) extrahiert, manuell analysiert und in „separate Graphen“ (ebd.) (positiv und negativ) aufgeteilt (ebd.). Im eigenen Ansatz wurden Adjektive mit dem beschriebenen Wörterbuch SentiWS von Robert Remus und Khurshid Ahmad (vgl. Remus und Ahmad, 2010) im CISLEX abgeglichen. Insgesamt wurden 11 923 positive und 10 841 negative Grund- und Flektionsformen der Adjektive im SentiWS gefunden. Von 708 Grundformen der Adjektive mit negativer Polarität waren 537 dem CISLEX bekannt. Von 790 Grundformen der Adjektive mit positiver Polarität waren es 662.

Diesen Grundformen sowie deren Flektionsformen wurden im CISLEX lediglich die Kategorien für positive (<ADJPOS>) und negative (<ADJNEG>) Adjektive hinzugefügt. Restliche unbekannte Adjektive (Grund- und Flektionsformen sowie polaritätsspezifische Kategorien, s. o.) wurden automatisch bzw. halbautomatisch hinzugefügt (s. Tabelle 5.5 für Grundformen).

		SentiWS	**CISLEX**	
Grundformen	**pos**	790	aus SentiWS bekannt	hinzugefügt
			662	128
	neg	708	537	171

Tabelle 5.5: Erweiterung des CISLEX mit Grundformen aus SentiWS

Eine korpusspezifische Erweiterung der im CISLEX kodierten Adjektive erfolgte auf zwei Ebenen – einer syntaktisch-semantischen und einer morphosyntaktischen Ebene, auf die in weiteren Abschnitten eingegangen wird. Neue Adjektive, die nicht im CISLEX vorhanden waren, wurden semiautomatisch mit allen flektierten Formen eingetragen. Ein Beispiel für einen neuen Lexikon-Eintrag ist auf der Abbildung 5.4 zu sehen (die hinzugefügten Kodierungen sind markiert).

Syntaktisch-semantische Ebene

Nach dem Ansatz von Hatzivassiloglou und McKeown (1997) (s. Kapitel 3, Abschnitt 3.1.2.1.3, Seite 65) wurde das Korpus nach neuen Adjektiven auf folgende Weise durchsucht:

- Nach bekannten, im CISLEX bereits vorhandenen Adjektiven wurden neue Adjektive mit gleicher Polarität gesucht, indem diese mit „und“, „sowie“, „als auch“ u. ä. verbunden wurden (Abbildung 5.5, Subgraph „KONJ_Dasselbe“).

- Nach bekannten, im CISLEX bereits vorhandenen Adjektiven wurden neue Adjektive mit gegensätzlicher Polarität gesucht, indem diese mit „aber“, „sondern“, „trotzdem“ u. ä. verbunden wurden (Abbildung 5.5, Subgraph „KONJ_Gegenteil“).

```
enorm,.ADJ+ADJPOS+WZ(NEG):up
enorme,enorm.ADJ+ADJPOS+WZ(NEG):aeFxp:aeFyp:aeFzp:aeNyp:amUxp:neFxp:neFyp:neFzp:neMyp:neNyp:nmUxp
enormem,enorm.ADJ+ADJPOS+WZ(NEG):deMxp:deNxp
Enormem,Enorme.N+adj
Enormem,Enorme.N+adj:deMxp
Enormem,Enorme.N+adj:deNxp
Enorme,.N+adj:aeFxp:aeFyp:aeFzp:amFxp:neFxp:neFyp:neFzp:nmFxp
Enorme,.N+adj:aeNyp:amNxp:neNyp:nmNxp
Enorme,.N+adj:amMxp:neMyp:nmMxp
enormen,enorm.ADJ+ADJPOS+WZ(NEG):aeMxp:aeMyp:aeMzp:amUyp:amUzp:deFyp:deFzp:deMyp:deMzp:deNyp:deNzp
eFzp:geMxp:geMyp:geMzp:geNxp:geNyp:geNzp:gmUyp:gmUzp:nmUyp:nmUzp
Enormen,Enorme.N+adj:aeMxp:aeMyp:aeMzp:amMyp:amMzp:deMyp:deMzp:dmMxp:dmMyp:dmMzp:geMxp:geMyp:geMzp
Enormen,Enorme.N+adj:amFyp:amFzp:deFyp:deFzp:dmFxp:dmFyp:dmFzp:geFyp:geFzp:gmFyp:gmFzp:nmFyp:nmFzp
Enormen,Enorme.N+adj:amNyp:amNzp:deNyp:deNzp:dmNxp:dmNyp:dmNzp:geNxp:geNyp:geNzp:gmNyp:gmNzp:nmNyp
enormere,enorm.ADJ+ADJPOS+WZ(NEG):aeFxk:aeFyk:aeFzk:aeNyk:amUxk:neFxk:neFyk:neFzk:neMyk:neNyk:nmUx
enormerem,enorm.ADJ+ADJPOS+WZ(NEG):deMxk:deNxk
Enormerem,Enormere.N+adj
Enormerem,Enormere.N+adj:deMxk
Enormerem,Enormere.N+adj:deNxk
```

Abbildung 5.4: Auszug aus dem CISLEX_SENTIWS mit Polaritätskodierungen der Adjektive

Nach der Auseinandersetzung mit den vorliegenden Korpora war auffällig, dass es zahlreiche Aufzählungen der Adjektive gibt, die als wertende Beschreibungen der Objekte / Dimensionen zu interpretieren sind (5.1):

(5.1) (a) „Sehr netter und freundlicher, aufgeschlossener Arzt“

(b) „Er ist ein sehr sympathischer, netter, kompetenter und rücksichtsvoller Arzt“

(c) „Freundlich, gründlich, kompetent und sehr gut organisierte Praxis“

Diese Auffälligkeit bot den Anlass, nach solchen Aufzählungen im Korpus zu suchen und die aus dem Lexikon bereits bekannten um neue Adjektive zu erweitern. Die Erweiterung der Adjektive erfolgte in beiden Fällen mittels lokaler Grammatiken nach Bootstrapping-Methode. In Anlehnung auf Vázquez et al. (2012, S. 1275) kann man die hier beschriebenen Vorgehensweisen wie auf der Abbildung 5.7 darstellen.
Beide beschriebenen Methoden auf der syntaktisch-semantischen Ebene haben lediglich zur Erweiterung der Adjektive nur um eine sehr geringe Menge, deren Zahl in der Summe im zweistelligen Bereich liegt[73], geführt. Die Gründe

[73]So konnten bei der ersten Methode lediglich fünf positive und drei negative und bei der zweiten zwölf positive und drei negative Adjektive gewonnen werden.

dafür kann man wie folgt zusammenfassen:

- Die meisten richtig gefundenen Adjektive waren bereits aus dem SentiWS bekannt und dementsprechend im CISLEX vorhanden.

- Die gefundenen Adjektive waren unspezifisch bzw. nicht eindeutig, was ihre Polarität betraf („breit“, „kurz“, „offen“, „zügig“ etc), oder gehörten zu anderen Wortarten („weil“, „rede“, „das“, „sehr“ etc.). Dies geschah durch Übergeneralisierungen der angewandten Methode der sogenannten „guessing strategies“ (Erläuterung s. u.).

Wie im Kapitel 3, Abschnitt 3.1.2.1.3, Seite 66 erläutert, hat Mikheev (1996) die Technik von „guessing strategies“ angewandt, um neue Wörter für Part-of-Speech-Tagging zu gewinnen. Dabei wurden „guessing rules“ für Präfixe, Suffixe und Endungen der Wörter entwickelt, wobei von den bekannten Wörtern im Lexikon ausgegangen wurde. So impliziert z. B. die Regel

$$A^p : [un\ (VBD \quad VBN) \quad (JJ)], \tag{5.1}$$

dass ein unbekanntes Wort, das Präfix „un“ in einem aus dem Lexikon als „a past verb and participle“ bekannten Verb aufweist, als Adjektiv interpretiert wird (ebd., S. 770). Auf der Abbildung 5.6 wird eine ähnliche Suchstrategie bei der Akquise der Adjektive auf der Satzebene angewandt, wobei man nach und vor bekannten Adjektiven mit bestimmter Polarität, die von „und“ oder „,“ umgeben sind, nach beliebigen klein geschriebenen Wörtern sucht (die Box mit <MIN>-Kodierung)[74] und davon ausgeht, dass diese Wörter ebenfalls Adjektive mit derselben Polarität sind.
Nach den aufgeführten Erläuterungen zur Qualität der durchgeführten Adjektiverweiterung wurden die Ergebnisse für Adjektivakquise durch Konjunktionen und Aufzählungen in den Tabellen 5.6 und 5.7 zusammengefasst:

- richtig: Richtig gefundene Adjektive (auch die aus dem SentiWS bekannten Adjektive)

- falsch: Adjektive mit gegensätzlich erwarteten Polarität, andere Wortarten durch Übergeneralisierungen (durch <MIN>, s. o.) oder Adjektive mit neutraler Polarität

[74] In *UNITEX* bedeutet die Kodierung <MIN> ein klein geschriebenes Wort.

	Positive Adjektive	Negative Adjektive
richtig	83,7%	83,3%
falsch	16,3%	16,7%

Tabelle 5.6: Extraktionsergebnisse der Adjektivakquise durch Konjunktionen

	Positive Adjektive	Negative Adjektive
richtig	85,0%	47,0%
falsch	15,0%	53,0%

Tabelle 5.7: Extraktionsergebnisse der Adjektivakquise durch Aufzählungen

Die Qualität und insbesondere die geringe Menge der gewonnenen Adjektive führen zu folgenden Erkenntnissen:

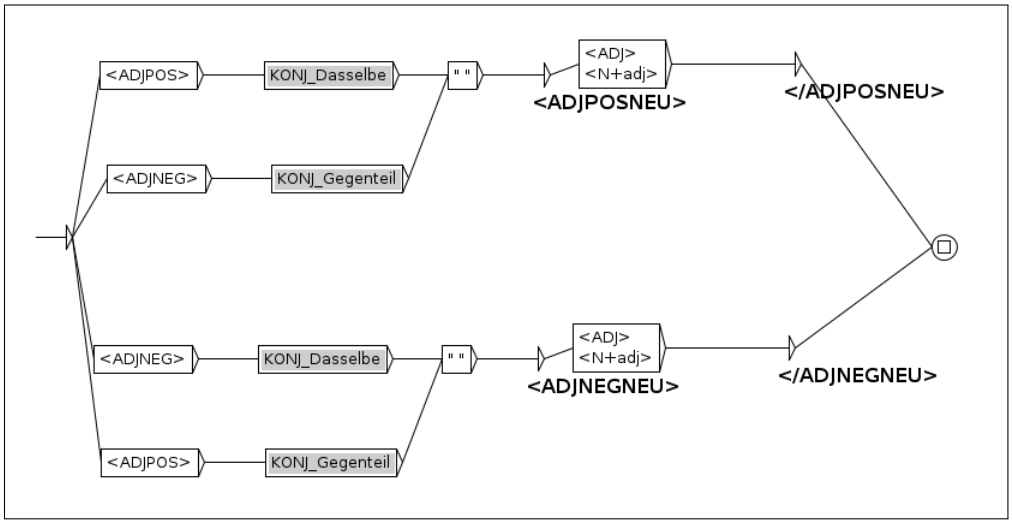

Abbildung 5.5: Graph zur Suche nach Adjektiven, die mit Konjunktionen verbunden sind

- Die Kombination der Adjektiv-Akquise mit gleichzeitigem Lexikonabgleich führt nicht zu brauchbaren Ergebnissen und sichert keinen vernünftigen Gewinn an neuen Wörtern. Einzeln genommen, ist jedes Vorgehen produktiver. Eine korpusbasierte Akquise (ohne SentiWS) sollte man daher mit einer handverlesenen Menge an polaren Adjektiven durchführen.

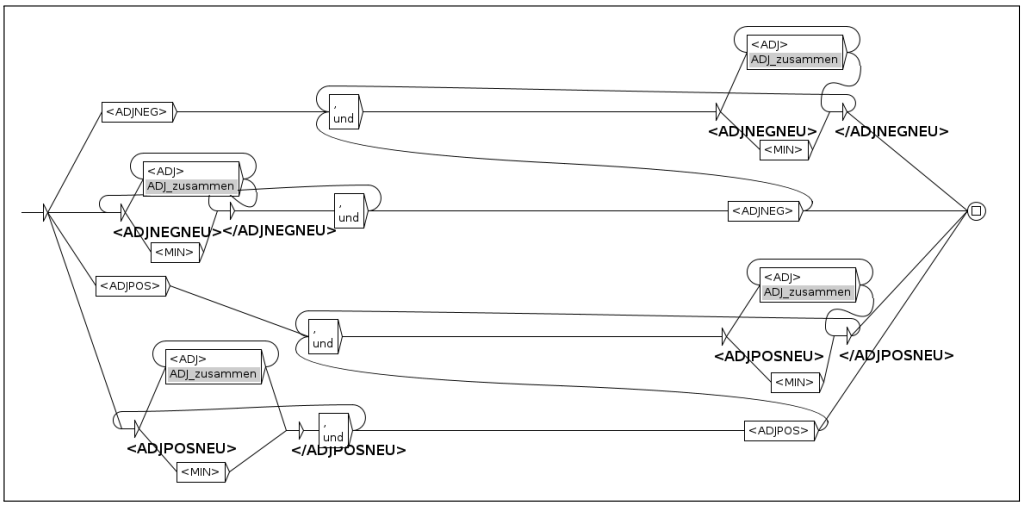

Abbildung 5.6: Graph zur Suche nach Aufzählungen von Adjektiven

- Bei der Anwendung von „guessing strategies“ sollten die Übergeneralisierungen vermieden werden, wodurch präzisere Ergebnisse erreicht werden können. In diesem konkreten Fall sollte in der Grammatik auf der Abbildung 5.6 nicht nach beliebigen klein geschriebenen Wörtern, sondern nach Adjektiven gesucht werden.

Morpho-syntaktische Ebene

Neue Adjektive wurden ebenfalls mittels wortbildender Elemente gewonnen. Auf der Basis einiger Präfixe mit verneinenden Aussagen[75] wurde ein Graph (siehe Abbildung 4.4, Seite 115) entwickelt, mit dem Adjektive mit entsprechender Polarität aus dem Korpus extrahiert wurden. Mit dieser Grammatik, die im morphologischen Modus (s. Abschnitt 4.2.2.5, Seite 114) eingesetzt wird und auf Wortebene arbeitet, erfolgte eine deutlich bessere Erweiterung des Lexikons als mit beiden im vorigen Abschnitt beschriebenen Methoden (Gesamtgewinn: 1671 Adjektive). Wenn man sich die „guessing strategies“ und die von Mikheev (1996, S. 770) aufgestellte Regel (s. Seite 140) in Erinnerung ruft, so könnte man eine ähnliche Regel, die durch die lokale Grammatik

[75] `http://www.mein-deutschbuch.de/lernen.php?menu_id=17#praefixe` (12.07.2015). Die Auswahl der Präfixe und Suffixe ist der genannten Quelle entnommen und erhebt keinen Anspruch auf Vollständigkeit.

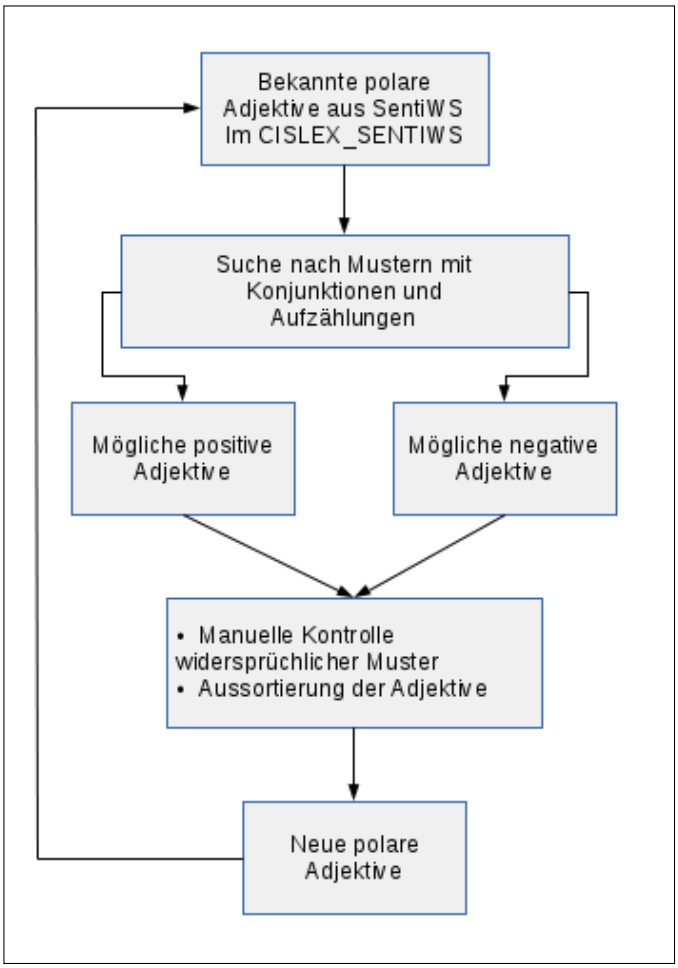

Abbildung 5.7: Akquise der Adjektive nach „Bootstrapping“-Methode (nach Vázquez et al. (2012, S. 1275))

auf der o. g. Abbildung realisiert wird, wie folgt, formulieren:

$$A^{pneg} : [in\ (ADJPOS) \quad (ADJNEG)] \tag{5.2}$$

Diese Regel impliziert, dass wenn ein Adjektiv mit positiver Polarität gefunden wird, wird seine Polarität durch das Anhängen entsprechender Präfixe (hier: „in“) umgekehrt. In der Tabelle 5.8 wird die Qualität der auf morphologischer Ebene erzielten Ergebnisse zusammengefasst.
Die Problematiken der Adjektivakquise auf der Wortebene erklären sich teilweise am korpusspezifischen hohen Anteil an unterschiedlichen Adjektiven

und an der Vielfältigkeit der Sprache. Das Letztere impliziert, dass nicht alle Adjektive, die mit aufgeführten verneinenden Präfixen beginnen, automatisch eine negative Polarität im entsprechenden Kontext aufweisen: „informativ“, „desinfiziert“, „unverkrampft“, „unverändert“, „unweit“ etc. Hinzu kommt, dass bei der beschriebenen Vorgehensweise mit einigen Lexikon-Fehlern zu rechnen ist. Wenn mit der lokalen Grammatik auf der Abbildung 4.4 (Seite 115) solche Adjektive wie „unglaublich“, „involviert“, „unfassbar“ gefunden wurden, so impliziert dieses, dass im CISLEX die Wörter „glaublich“, „volviert“, „fassbar“ als eigenständige sinnvolle Adjektive aufgeführt wurden, was tatsächlich der Fall ist.

	Positive Adjektive	**Negative Adjektive**
richtig	53,0%	63,5%
falsch	47,0%	36,5%

Tabelle 5.8: Extraktionsergebnisse der auf der Wortebene extrahierten Adjektive

b) Fachvokabular

Im Abschnitt „Bewertungen pro Facharztgruppe“ auf der Seite 133f. wurde die Erstellung der fachspezifischen Wörter aus einem Subkorpus beschrieben. In diesem Sinn stellt das Fachvokabular kein eigenes Lexikon dar. Die ermittelten Wörter werden im CISLEX_SENTIWS mit entsprechenden Kodierungen versehen bzw. dem Lexikon – falls noch nicht vorhanden – hinzugefügt und dann kodiert. Einige fachspezifische Wörter sind im Beispiel (5.2) zu sehen. Die hinzugefügten Elemente – sei es Tags (z. B. „ZNOM“ etc., s. ausführlicher Tabelle 5.4 im o. g. Abschnitt) oder die neu gewonnenen Wörter – sind im Beispiel kursiv markiert.

(5.2) (a) Überkronungen,Überkronung.N+*ZNOM*:amF:dmF:gmF:nmF
Überkronung,.N+*ZNOM*:aeF:deF:geF:neF
überkronungs,überkronung.N+FF+*ZNOM*
überkronung,.N+FF+ZNOM
überkronungen,.N+FF+ZNOM

(b) Trigemini,Trigeminus.N+*HNONOM*:amM:dmM:gmM:nmM
Trigeminus,.N+*HNONOM*:aeM:deM:geM:neM
trigeminus,.N+FF+*HNONOM*

(c) *unterspritzung,.N+FF+SCHNOM*
unterlidstraffung,.N+FF+SCHNOM

(d) Tropfen,.N+*BHNOM*:aeM:amM:deM:dmM:gmM:neM:nmM
Tropfen,.N+*BHNOM*:aeN:deN:neN
tropfen,.N+FF+*BHNOM*

c) Wörter zur Nationalität

Für die Extraktion wertender Muster zu Diskriminierungen wegen der Nationalität wurde ebenfalls kein eigenes Lexikon angefertigt, sondern eine Liste entsprechender Wörter im CISLEX_SENTIWS untergebracht. Diese Liste wurde aus dem „Verzeichnis der Staatennamen für den amtlichen Gebrauch in der Bundesrepublik Deutschland“[76] extrahiert. Das genannte Verzeichnis verfügt über 199 Länderbezeichnungen, die jeweils in Kurz- und Vollformen aufgeführt sind. Weiterhin sind zu jedem Staat entsprechende Adjektive, Staatsangehörige in männlicher und weiblicher Formen und die Ländercodes nach der Norm DIN EN ISO 3166-1 angegeben. Das entsprechende Ländervokabular wird ins CISLEX_SENTIWS eingetragen und, wie im Beispiel (5.3) gezeigt, kodiert. Bei Musterextraktion werden die Phrasen mit diesem Vokabular mit dem Graphen „DISKR_Nationalitaet.grf“ angesprochen (s. beiliegende DVD/Lokale_Grammatiken/Diskriminierungen). Im Anhang A.1 befindet sich die Übersicht zu den Einträgen.

(5.3) (a) Kurzform: „Kasachstan“, Kodierung: „LAND(KF)“

(b) Vollform: „Republik Kasachstan“, Kodierung: „LAND(VF)“

(c) Adjektiv: „kasachisch“, Kodierung: „ADJNAT“

(d) Staatsangehörige (m): „Kasache“, Kodierung: „NAT“

(e) Staatsangehörige (f): „Kasachin“, Kodierung: „NAT“

[76] http://www.auswaertiges-amt.de/DE/Infoservice/Terminologie/Staatennamen-DE/Uebersicht.html?nn=373568 (27.08.2017)

5.1.2.2.2 Phrasenlexikon (PHRASE_LEX)

Die mit lokalen Grammatiken extrahierten Pattern beinhalten teilweise Ausdrücke, die auf zwei und mehr Dimensionen zutreffen. Im Beispiel (5.4)) werden gleichzeitig drei Dimensionen angesprochen:

- „Behandlung“ („sehr kompetente“)
- „Freundlichkeit“ („nette“)
- „Betreuung“ („Betreuung“)

(5.4) „sehr kompetente und nette Betreuung“

Eine automatische Erkennung aller drei Dimensionen in einem Schritt wäre bei der gewählten Vorgehensweise beim Phrasenaufbau zu umständlich realisierbar (s. Abschnitt 5.2.2.2, Seite 157), da man zum einen die möglichen Kombinationen der ‚Adjektiv+Nomen‘-Folgen auf ihre Kongruenz überprüfen müsste, wobei bei dimensions- und polaritätsbestimmenden Adjektiven (s. Abschnitt 5.2.1.2.2) jede denkbare Kombination einzeln aufzuführen wäre. Eine lokale Grammatik mit solchen Mustern wäre zum anderen unübersichtlich. Viel einfacher wäre es, zunächst eine Dimension als Objekt (in diesem Fall „Betreuung“) und ihre positive Polarität anhand der voranstehenden Adjektive zu erkennen und erst in einem zweiten Schritt eine weitere mögliche Unterscheidung dieser Adjektive zu treffen. Aus diesem Grund wird hier ein Phrasenlexikon (PHRASE_LEX) entwickelt, das fertige Phrasen, wie aus dem oben aufgeführten Beispiel, enthält. Zu jeder Dimension werden spezielle ‚Module‘ (s. Abbildung 5.18 (linke Seite)) zusammengestellt, die dimensionsbeschreibende typische Wortarten wie Adjektive, Nomen etc. enthalten. Dann werden in den zu jeder Dimension mit Mastergraphen (s. Abschnitt 5.2.2.4) extrahierten Phrasen mittels der genannten Module nach Hinweisen auf Benennungen anderer Dimensionen gesucht und annotiert. Die annotierten Phrasen für solche kombinierte Dimensionenbeschreibungen werden semiautomatisch mit den Tags, die entsprechende Dimensionen und Polaritäten beinhalten (s. Abbildung 5.8), versehen. Die entstandenen Phrasen wurden im Lexikon PHRASE_LEX zusammengefasst und mit den Dimensions- und Polaritätskategorien kodiert. Ein Auszug dessen ist auf der eben genannten Abbildung zu sehen.

```
sehr kompetenter\, erfahrener\, freundlicher und sehr ehrlicher arzt,.bhpos+frpos+vvpos
sehr kompetenter erfahrener arzt\, sehr freundliches personal,.bhpos+frpos
sehr kompetenter\, entgegenkommender arzt,.bhpos+frpos
sehr kompetenter\, engagierter\, vertrauenswürdiger arzt,.bhpos+vvpos
sehr kompetenter\, engagierter und netter arzt,.bhpos+frpos
sehr kompetenter\,engagierter und freundlicher arzt,.bhpos+frpos
sehr kompetenter\, einfühlsamer\, zuvorkommender arzt,.bhpos+frpos
sehr kompetenter\, einfühlsamer\, verständnisvoller und erfolgreicher arzt,.bhpos+vvpos
sehr kompetenter\,einfühlsamer und vorallem menschlicher kinderchirurg,.bhpos+frpos
```

Abbildung 5.8: Auszug aus dem Phrasenlexikon PHRASE_LEX

Die semiautomatische Bearbeitung des Phrasenlexikons führte dazu, dass nicht nur kombinierte Tags zu Dimensionen, sondern – aufgrund manueller Nachkorrekturen einiger ‚Falschannotationen' – auch einzelne davon in diesem Lexikon vorhanden sind. Von 18 377 Lexikoneinträgen sind 8 241 diejenigen mit nur einem Tag. Außer dem Aufwand manueller Eintragungen besteht der Nachteil dieses Lexikons darin, dass einige Phrasen zu lang und / oder unvollständig sind, z. B. „arzt der ein sehr gutes fachwissen hat auf die bedürfnisse der patienten eingeht ein arzt dem man sein vertrauen". Solche Phrasen betreffen individuelle Formulierungen der Bewertenden und sind bei anderen Korpora derselben Domäne unbrauchbar. 763 Phrasen aus dem Lexikon wurden mit 10 und mehr Wörtern als zu lang und / oder unvollständig eingeschätzt.
Trotz der aufgezählten Nachteile ist das Phrasenlexikon als ein strategisch gelungenes Werkzeug zur Annotation wertender Patientenausdrücke einzustufen. Durch dessen Anwendung bei dem Annotationsverfahren werden die Übergeneralisierungen und folglich die Nichterkennungen der Pattern durch Mastergraphen (s. Abschnitt 5.2.2.4) ausgeglichen.

5.1.2.2.3 Ärztenamen (ARZTNAME_LEX)

Um zu unterscheiden, ob die Patienten andere, von Personen verschiedene Objekte (z. B. Praxis, Wartezeit) bewerten, sowie um die Vollständigkeit wertender Phrasen zu gewährleisten, wurden im Rahmen des in der Einleitung (Abschnitt 1.1) erwähnten Projekts „More Than Words" (Seite 3) die Ärztenamen akquiriert. Zum einen wurden alle Namensangaben zu den Ärzten aus den Portalen Jameda und DocInsider extrahiert (Trefferquote = 100%). Zum anderen erfolgte eine weitere Akquise aus den Bewertungstexten

mit regulären Ausdrücken[77]. Insgesamt wurden 48 046 Ärztenamen gewonnen (Geierhos et al., 2015b, S. 7), die in einem Lexikon ARZTNAME_LEX gespeichert und als ‚ANAME' kodiert wurden.

5.2 Aufbau lokaler Grammatiken

Bereits im Kapitel 3, Abschnitt 3.1.2.4 wurden einige aktuelle Arbeiten mit lokalen Grammatiken erwähnt, in denen diese zur Extraktion bestimmter Muster aus strukturierten und unstrukturierten Daten eingesetzt wurden. Im Kapitel 4, Abschnitt 4.2.1, Seite 107ff. fand die Beschreibung lokaler Grammatiken als musterbasiertes Extraktionsverfahren statt, wobei einige Vorteile und Möglichkeiten für die vorliegende Arbeit aufgezeigt wurden. Mit dem Korpusverarbeitungssystem *UNITEX* wurden bereits im o.g. Kapitel, Abschnitt 4.2.2 außerdem einige Funktionen und praktische Realisierungsmöglichkeiten lokaler Grammatiken in Form von Graphen gezeigt.
In diesem Abschnitt wird der systematische Aufbau der Graphen beschrieben. Aufgrund der großen Menge von lokalen Grammatiken (rund 400 Graphen) kann nicht auf alle davon eingegangen werden, was auch überflüssig wäre. Es werden lediglich wichtige Aspekte der Vorgehensweise bei der Musterextraktion aufgegriffen, die in Hinsicht auf ihre semantische, syntaktische und pragmatische Konzeptionen an einigen Beispielen erläutert werden.

5.2.1 Semantisch

Die Semantik als Teilgebiet der Linguistik beschreibt Bedeutung und Inhalt sprachlicher Zeichen und Zeichenfolgen[78]. Im Sinne von lokalen Grammatiken kann dies als Beschreibung der inhaltlichen Struktur von Graphen, die zur Erfassung von semantischen Einheiten wie z. B. Schlüsselwörter, von denen im Abschnitt 4.2.1.3, Seite 110 die Rede war, interpretiert werden. In erster Linie sind in Bezug auf diese Arbeit die Bewertungsobjekte (Wolfgruber, 2015, S. 80ff.), die Dimensionen oder Schlüsselwörter als semantische Einheiten zu verstehen, da diese Ausgangspunkt des Aufbaus lokaler Grammatiken sind. Weitere semantische Einheiten betreffen Gruppierungen bestimm-

[77] Die akquirierten Ärztenamen wurden mir von Kollegen des genannten Projekts freundlicherweise zur Verfügung gestellt. Auf weitere Details der Namensakquise wird hier daher nicht eingegangen.

[78] `http://www.duden.de/rechtschreibung/Semantik(31.01.2017)`

ter Wortarten, die kontextbezogen für jede der Dimensionen entwickelt und an entsprechenden Stellen in die Graphen eingebaut werden. Auch im Sinne des Phrasenaufbaus zu Bestätigungsfehlern und Diskriminierungen kann von einer Menge an charakteristischen Schlüsselwörtern ausgegangen werden.

5.2.1.1 Bewertungsobjekte

5.2.1.1.1 Benennungen der Dimensionen

Um sich mit dem Kontext der expliziten (s. z. B. Seite 37) Dimensionsbenennungen auseinander zu setzen und um Bewertungsobjekte zu konstruieren, eignet sich die im Abschnitt 4.2.1.3, Seite 110 beschriebene Methode „Bootstrapping“. Ähnlich dem im genannten Abschnitt angesprochenen Beispiel der Entwicklung lokaler Grammatiken für unterschiedliche Kontexte des Wortes „health“ werden hier die Objekte, die Nomen, mit denen viele Dimensionen bezeichnet werden, als Schlüsselwörter eingesetzt. Für die Dimension „Wartezeit“ kann man beispielsweise als eine solche Benennung die Nomen „Wartezeit“, „Wartezeiten“ und „Warten“ sowie weitere zusammengesetzte Nomen wie „Wartezimmerzeit“ oder solche Ausdrücke wie „Zeit des Wartens“ verstehen. Diese können mit dem auf der Abbildung 5.9 dargestellten Graphen extrahiert werden, wobei man für die Nomen reguläre Ausdrücke verwendet, die durch den Anfang des Wortes „warte...“ sicherstellen, dass die Zeit, um die es geht, wirklich „Wartezeit“ betrifft. Wie aus der Tabelle 2.1 (Seite 16) im Kapitel 2 bekannt, gibt es bei Jameda zwei Dimensionen, die Wartezeiten betreffen: „Wartezeit (Praxis)“ und „Wartezeit (Termin)“. Um diese Bewertungsobjekte voneinander zu unterscheiden, kann nach deren weiteren Konkretisierungen gesucht werden. Wie auf der Abbildung 5.10 zu sehen ist, wurde die „Wartezeit“ um ein Dativobjekt erweitert. In der Kombination mit dem Graphen auf der Abbildung 5.9 bildet die beschriebene lokale Grammatik (s. Abbildung 5.11) die explizite Benennung der Dimension „Wartezeit (Praxis)“. Auf derselben Abbildung ist auch die Konkordanz des eben beschriebenen Graphen dargestellt. Genauso wird mit der Dimension „Wartezeit (Termin)“ verfahren, indem der Graph auf der Abbildung 5.9 mit den Ausdrücken wie „auf den Termin“ erweitert wird, wodurch gleichzeitig die Disambiguierung der Benennungen beider Dimensionen erfolgt. Bei den Wartezeiten, die ohne aufgeführte Differenzierungen angesprochen werden, wird angenommen, dass es sich um das Warten in der Praxis handelt. Die Situationen, in denen dies jedoch nicht der Fall ist und in den Phrasen explizite

Benennungen von Wörtern wie z. B. „Termin“ vorzufinden sind, werden mit der Ausarbeitung des Phrasenlexikons (PHRASE_LEX, Abschnitt 5.1.2.2.2, Seite 146ff.) neu annotiert bzw. die bestehenden Annotationen korrigiert oder erweitert.

5.2.1.1.2 Fachvokabular

Die Dimension „Behandlung“ ist im Kontext der ärztlichen Leistungen sicher das wesentlichste Bewertungsobjekt. Die Behandlung wird in den Bewertungen am häufigsten angesprochen, da die Patienten offensichtlich große Erwartungen in deren Erfolg setzen, was in der Domäne der Arztpraxen nur logisch erscheint. Dementsprechend zahlreich sind die expliziten Benennungen der betreffenden Dimension. In Abschnitten 5.1.1.2.1 (Seite 133) und 5.1.2.2.1 (Seite 144) wurde erläutert, wie die fachspezifischen Ausdrücke zur Charakterisierung der Behandlung verarbeitet und im CISLEX_SENTIWS zusammengefasst und kodiert werden (s. auch Beispiel (5.2)). Mit diesen Kategorien werden die Fachausdrücke in entsprechenden lokalen Grammatiken angesprochen. Diese lokale Grammatiken werden in Form von Subgraphen (s. auch Kapitel 4, Abschnitt 4.2.2.2, Seite 111) in die Grammatiken zu wertenden Phrasen der Dimension „Behandlung“ eingebaut.

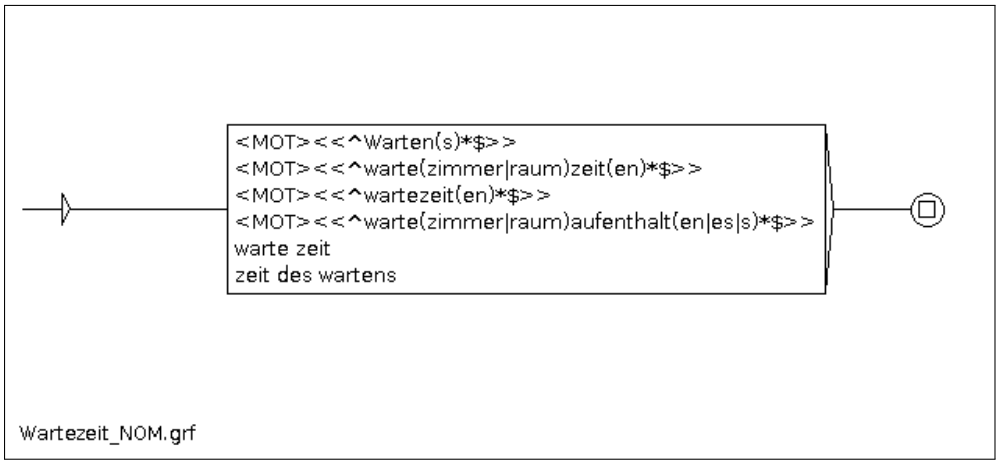

Abbildung 5.9: Graph zur Erkennung der Nomen zu „Wartezeit“

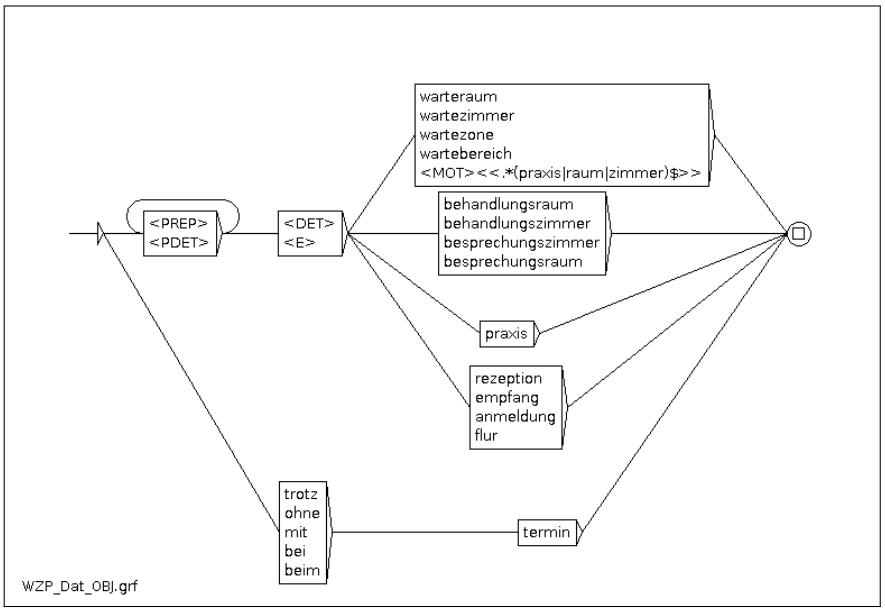

Abbildung 5.10: Graph zur Erweiterung der Nomen zu „Wartezeit“

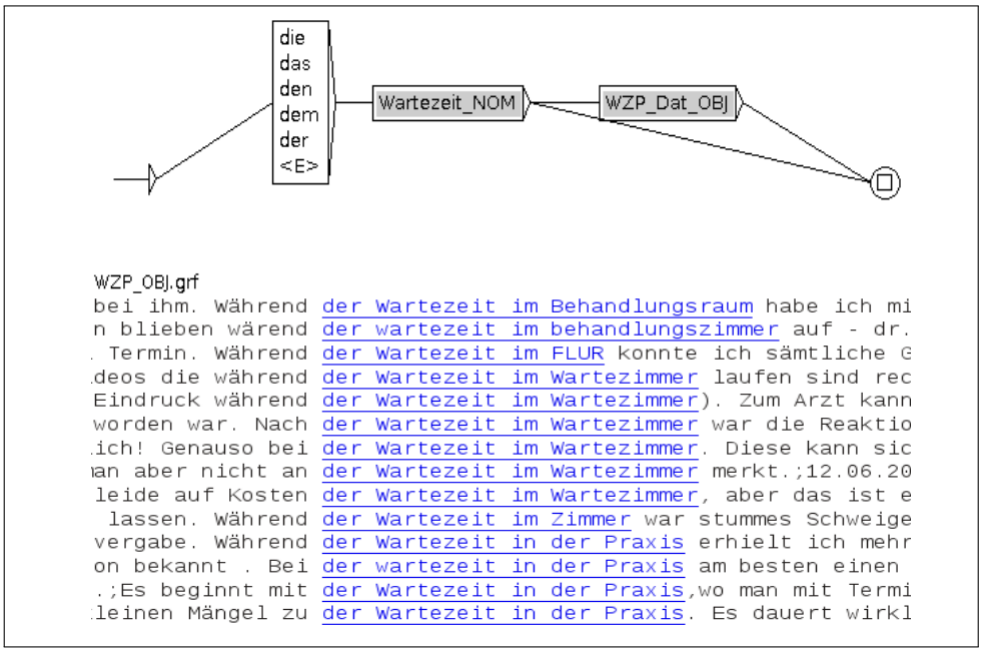

Abbildung 5.11: Graph zur Erweiterung der nominalen Objekte zur Dimension „Wartezeit (Praxis)“ und dessen Konkordanz

5.2.1.1.3 Wertende Phrasen

Bei der Ermittlung wertender Muster zu Bewertungsdimensionen fällt auf, dass viele dieser Phrasen die Muster ‚Adjektiv + Nomen' aufweisen (s. z. B. Beispiel (5.8)(a) und (5.8)(b)). Außer den Nomen, die Dimensionen explizit benennen, werden Bewertungsobjekte auch mit anderen Wortarten wie Verben, Adjektiven, Partizipien etc. oder mit kompletten Wortgruppen (s. Abbildung 5.12) charakterisiert. Wenn man die Bewertungsdimension „Behandlung" u. a. mit den Folgen ‚Adjektiv + Nomen' beurteilen kann („gute Behandlung", „mangelhafte Behandlung"), so ist es bei der „Freundlichkeit" nicht unbedingt der Fall. Hier werden meist Adjektive verwendet („sehr freundlich", „ein netter Arzt"). Das bedeutet, dass bei dieser Dimension nicht die Nomen, sondern die Adjektive als Schlüsselwörter, als Ausgangspunkte zum Aufbau wertender Phrasen fungieren (s. Abbildung 5.12c)). In solchen Fällen, speziell im Falle von Adjektiven (s. auch Abschnitt 5.2.1.2), werden diese im CISLEX_SENTIWS entsprechend kodiert und in den Graphen mittels dieser vergebenen Kategorien (hier: ‚FR<POS>') angesprochen. Für andere Wortarten, die lediglich innerhalb einer Bewertungsdimension verwendet werden (z. B. „gut ausgestattete Praxis", Abbildung 5.12b)), werden keine gesonderten Kodierungen im Lexikon eingetragen. Diese Wortarten werden in

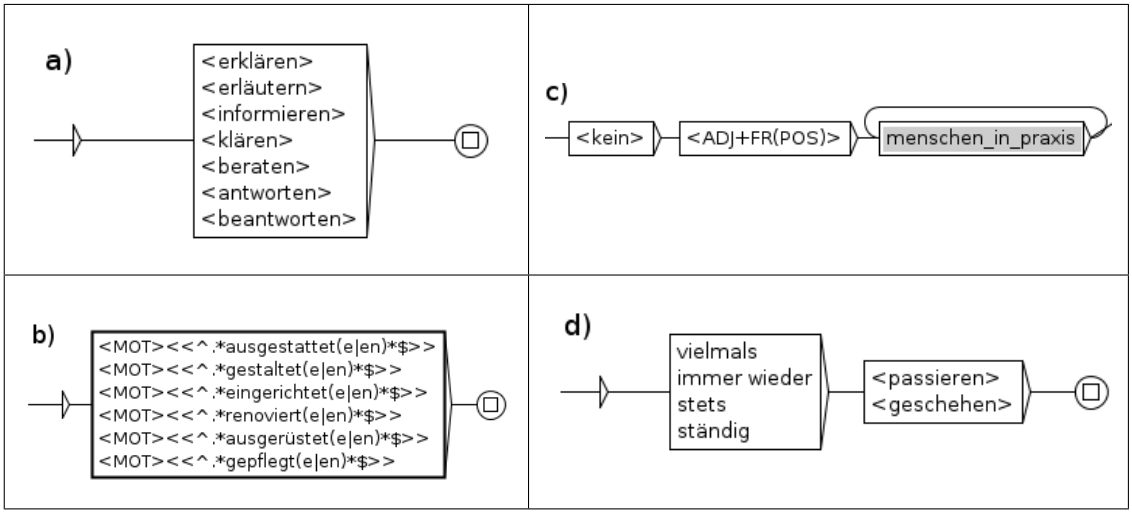

Abbildung 5.12: Beispiele zu einigen Modulen mit dimensionsbeschreibenden Wortarten

einzelnen Modulen / Graphen zusammengefasst. Für die expliziten Benennungen der Dimensionen sind oft nicht nur einzelne Wörter, sondern Wortgruppen, Wendungen oder Ausdrücke zuständig. Wenn man die positive Meinung zur Bewertungsdimension „Freundlichkeit“ mit einem Adjektiv „freundlich“ ausdrücken kann, so funktioniert dieses mit der Dimension „Praxisausstattung“ nicht immer. Wenn z. B. die Partizipien auf der Abbildung 5.12b) genauer betrachtet werden, die in ihrer Funktion als Adjektive fungieren, so kann man sich die Ausdrücke „ausgestattete Praxis“ oder „eingerichtete Praxis“ zwar vorstellen, jedoch kann man den ganzen Ausdruck nicht als positive oder negative Meinung auswerten (vgl. „gut ausgestattete Praxis“, „schlecht ausgestattete Praxis“). Genauso verhält es sich im Falle der Dimension „Aufklärung“ (Abbildung 5.12a)). Sowohl bei einigen Dimensionen als auch bei Mustern zu kognitiven Effekten kann der Gegenstand, den man beurteilt, meist nur mit ganzen Phrasen angesprochen werden. Mit dem Ausdruck „Er hat sich viel Zeit genommen“ wird eine positive Meinung zur Dimension „Genommene Zeit“ mit einem kompletten Satz ausgedrückt, deren Polarität durch die subjektiv empfundene Menge dieser Zeit und schließlich durch das Wort „viel“ definiert ist. Bei den kognitiven Effekten wie Diskriminierungen und Bestätigungsfehler, zu denen textuelle Muster gefunden werden sollen und die keine Dimensionen beschreiben, wird keine Polarität markiert. In den meisten Fällen wird die negative Polarität der Ausdrücke angenommen (z. B. Diskriminierungen sind an sich selten positiv). Genau wie bei Dimensionen, konzentrieren sich auch hier die gesuchten Muster auf möglichst explizite Benennungen der definierten Problematiken.

Wie kann man jedoch einen kompletten Set an Ausdrucksmöglichkeiten z. B. zu den „Tendenzen, nach den Bestätigungen eigener Hypothesen einseitig zu suchen“ (s. Kapitel 2, Abschnitt 2.1.3.2) zusammenstellen und in lokalen Grammatiken verarbeiten?

Dieses Problem kann durch den Einsatz von Annotatoren auf eine ähnliche Weise wie dies im Kapitel 4, Abschnitt 4.2.3.1 (Seite 116f.) beschrieben wurde, gelöst werden, was in dieser Arbeit aufgrund des zeitlichen Aufwands unterlassen wird. Ein Beispiel für eine mögliche Variante des Ausdrucks für Bestätigungsfehler ist auf der Abbildung 5.12d) zu sehen.

5.2.1.2 Wertende Adjektive

5.2.1.2.1 Domänenabhängigkeit der Adjektive

Bei der Auseinandersetzung mit Adjektiven aus drei verschiedenen Korpora stellten Vázquez und Bel (2013) fest, dass über die Hälfte aller Adjektive eine domänenabhängige Polarität haben (s. ausführlicher Kapitel 3, Abschnitt 3.1.2.1.3 , Seite 65). In der Domäne der Arztpraxen kristallisiert sich dasselbe Phänomen heraus. Auf der Abbildung 5.4 ist das Adjektiv „enorm“ folgenderweise kodiert:

ADJ+ADJPOS+WZ(NEG)

Daraus kann geschlossen werden, dass dieses Wort ein Adjektiv ist, das generell eine positive Polarität aufweist (z. B. „enorme Kompetenz“), im Kontext der Wartezeit jedoch als negativ zu werten ist, z. B. im Ausdruck „enorme Wartezeiten“. Um das genannte Adjektiv bei den Dimensionen, die Wartezeiten betreffen, nicht als positiv zu annotieren, wird hier eine zusätzliche Kodierung vorgenommen, die in einem entsprechenden Graphen in einer Box z. B. vor den Nomen zu Wartezeiten untergebracht wird, um negative Phrasen zu Dimensionen „Wartezeit(Praxis)“ und „Wartezeit(Termin)“ zu extrahieren. Ein anderes Beispiel dafür wäre das Adjektiv „bescheiden“, das im Sinne einer menschlichen Eigenschaft die positive Polarität aufweist, jedoch bei der Bewertungsdimension „Parkmöglichkeiten“ eindeutig als negativ zu werten ist. Die entsprechenden Kategorien (s. o.) dieser Adjektive werden zum Zeitpunkt des Aufbaus von lokalen Grammatiken während der Auseinandersetzung mit dem Kontext der Bewertungsobjekte (s. Abbildung 5.13) getätigt.

5.2.1.2.2 Dimensions- und Polaritätsspezifik der Adjektive

In der vorliegenden Arbeit beschreiben Adjektive nicht nur die Polarität, wie z. B. „tolle Behandlung“ (positiv), sondern auch die Dimension einer Aussage – „freundlicher Arzt“ („Freundlichkeit“). Dass Adjektive sowohl bei der Polaritäts-, als auch Dimensionsbestimmung entscheidend sind, wird anhand des Beispiels (5.5) deutlich. In (5.5)(a) ist das Adjektiv „tolle“ lediglich polaritätsspezifisch, weil durch seine Kombination mit anderen Bewertungsobjekten (z. B. „Aufklärung“) die Polarität der Aussage positiv bleibt. In (5.5)(b)

dagegen ist das Adjektiv „freundlicher“ gleichzeitig polaritäts- und dimensionsspezifisch.

(5.5) (a) „tolle Behandlung“ (Dimension = „Behandlung“; Wortart = „Nomen“; Polarität = positiv; Wortart = „Adjektiv“)

(b) „freundlicher Arzt“ (Dimension = „Freundlichkeit“; Wortart = „Adjektiv“; Polarität = positiv; Wortart = „Adjektiv“)

(c) „sehr gut behandelt“ (Dimension = „Behandlung“, Wortart = „Verb“; Polarität = positiv; Wortart = „Adverb“)

Diese Beobachtung impliziert, dass man Adjektive in der Arztpraxendomäne nach ihrer Polarität und Dimension klassifizieren muss. Für dimensionsspezifische Adjektive werden, wie im vorigen Abschnitt beschrieben, Kategorien im CISLEX_SENTIWS eingeführt. Die Polaritäten vieler solcher Adjektive stimmen teilweise mit denen aus dem SentiWS bekannter Adjektive überein. Problematisch erscheint hier die Klassifikation der Adjektive bei deren zahlreichen Aufzählungen auf der Phrasenebene. Dieses Problem wird im Abschnitt 5.2.3.2.1 behandelt.

5.2.1.3 Zeit als Bewertungsobjekt

Bei den Bewertungsdimensionen wie „Wartezeit (Praxis)“, „Wartezeit (Termin)“ und „Genommene Zeit“ gilt die Zeit als Bewertungsobjekt. Die ersten zwei o. g. Dimensionen betreffen die Wartezeit, die man in Minuten, Stunden oder auch Tagen, Wochen und Monaten konkretisieren kann (vgl. auch Tabelle 2.1, Seite 16). Auch die „Genommene Zeit“ kann man in Minuten angeben. Die Wertung der Zeiten – seien diese relativ mit Angaben wie „viel“, „wenig“, „keine“ etc. beschrieben oder in o. g. Maßeinheiten aufgeführt – sind subjektiv. Während bei den Wartezeiten eher von niedrigeren Zeiten im positiven Kontext gesprochen wird, verhält es sich bei der Dimension „Genommene Zeit“ anders. Diese bewerten Patienten positiv, je mehr Zeit sich der Arzt für sie nimmt. Betrachtet man konkrete Zeitangaben der Patienten genauer, so lässt sich empirisch feststellen, dass Wartezeit in der Praxis über 30 Minuten im Durchschnitt als negativ empfunden wird (Geierhos et al., 2015a, S. 314). Die Zeit jedoch, während derer der Arzt die Patienten behandelt, gerade ab 20 Minuten und mehr als positiv gewertet wird. Solche Gesetzmäßigkeiten werden in kleineren Modulen implementiert und in die Phrasen zu genannten Dimensionen eingebaut, die in der Folge mit entsprechenden Polaritäten versehen werden.

5.2.2 Syntaktisch

Die Syntax wird als „übliche Verbindung von Wörtern zu Wortgruppen und Sätzen“, „korrekte Verknüpfung sprachlicher Einheiten im Satz“[79] etc. definiert. Für die vorliegende Arbeit kann man Syntax als Phrasenkonstruktionen aus semantischen Einheiten mittels lokaler Grammatiken interpretieren. Wie bereits betont (s. Abschnitt 5.1.2.2.1), spielen Adjektive eine wesentliche Rolle bei der Polaritätsbestimmung der Phrasen, was ebenfalls bei der Auseinandersetzung mit den Kontexten der Bewertungsobjekte auffällt. In den nachfolgenden Abschnitten werden einige Aspekte des Phrasenaufbaus mit wertenden Adjektiven beschrieben und einige Beispiele aufgeführt. Auf die Konstruktionsmöglichkeiten weiterer wertender Phrasen wird ebenfalls eingegangen, woraufhin die Erläuterungen zum organisatorischen System wertender Phrasen folgen.

5.2.2.1 Negation

Bevor das Aufbausystem und die Organisation wertender Phrasen beschrieben werden, ist es wichtig, auf einen polaritätsentscheidenden Modifikator, die Negation, einzugehen. Negation ist ein linguistisches Phänomen, das die Polarität eines Wortes oder einer Phrase umkehrt (Wolfgruber, 2015, S. 32). Negationsträger können auf der lexikalischen („nein“, „nicht“, „niemand“, „keiner“, „ohne“ etc.) und auf der morphologischen (Präfixe wie „un-“, „nicht-“, „ent-“ etc.; Suffixe wie „-los“, „-frei“ etc.) Ebene unterschieden werden (ebd., S. 31f.). Die Negation auf morpho-syntaktischer Ebene wurde in der vorliegenden Arbeit bei der Akquise wertender Adjektive im Abschnitt 5.1.2.2.1 (Seite 142ff.) behandelt, wobei deren Polarität durch verneinende Präfixe umgekehrt wurde. Auch bei dem Aufbau wertender Phrasen spielt Negation eine entscheidende Rolle, da es sich hier um die Polaritäten ganzer Aussagen handelt. An dem unten aufgeführten Beispiel (5.6) sind einige wertende Phrasen zu sehen, bei denen die Umkehrung der Polarität durch die Negation erfolgt (s. auch den nächsten Abschnitt).

(5.6) (a) {wirklich gut beraten,.akpos+AK_EXP_Aufkl_1}
{nicht wirklich gut beraten,.akneg+AK_EXP_Aufkl_1a}

(b) {eigene Parkplätze,.pmpos+Park_moeglichkeiten_1}
{Keine eigenen Parkplätze,.pmneg+Park_moeglichkeiten_1}

[79] `http://www.duden.de/rechtschreibung/Syntax` (04.02.2017).

(c) {Hat sich wirklich Zeit genommen,.gzpos+GZ_EXP_ZEIT_NEHMEN}
{Hat sich keine Zeit genommen,.gzneg+GZ_EXP_ZEIT_NEHMEN}

5.2.2.2 Adjektiv + Bewertungsobjekt

Betrachtet man die Konkordanz erkannter Objekte, so kann man dimensionsspezifische Phrasen entdecken. Die kürzesten oft vorkommenden Phrasen sind Kombinationen aus polaritätsaufweisenden Adjektiven und Bewertungsobjekten, z. B. „langen Wartezeiten“. Eine weitere Auseinandersetzung mit beschriebenen Phrasen zeigt, dass sich deren zunächst scheinbar eindeutig identifizierbare Polarität anhand der Adjektive durch die Negation umgekehrt werden kann (Wolfgruber, 2015, S. 32): „keine langen Wartezeiten in der Praxis“. Auf der Abbildung 5.13 ist der Prozess der Entwicklung von lokalen Grammatiken zu wertenden Phrasen mit Adjektiven und Negation schematisch dargestellt. Als Ausgangspunkt ist hier das Bewertungsobjekt zu verstehen, um welches die wertenden Phrasen aufgebaut werden, indem man deren unterschiedliche Kontexte analysiert und klassifiziert. Im ersten Schritt werden mit einer lokalen Grammatik die Bewertungsobjekte extrahiert und deren Kontexte betrachtet. Diese Kontexte werden ihren syntaktischen Strukturen nach klassifiziert. Die o. g. Abbildung beschreibt schematisch den Aufbau von ‚Adjektiv+Bewertungsobjekt‘-Konstruktionen, die um eine Negation erweitert werden, was die Polarität der Aussagen umkehrt und durch die entsprechende Annotation markiert wird. Im zweiten Pfad der Abbildung werden ebenfalls die Konstruktionen mit Adjektiven und Negation aufgebaut, es geht jedoch um den rechten Kontext des Bewertungsobjektes, wobei andere Wortarten wie z. B. Verben zwischen den Objekten und Adjektiven zugelassen werden können.
Wie im Abschnitt 5.2.1.1.3, Seite 152 gezeigt, können nicht nur Nomen in der Rolle der Bewertungsobjekte auftreten. Auf der Abbildung 5.14 ist ein Auszug aus lokaler Grammatik zu sehen, mit der man Phrasen zur Dimension „Betreuung“ extrahieren kann. Aus der auf derselben Abbildung dargestellten Konkordanz ist ersichtlich, dass als Bewertungsobjekte alle möglichen Formen der Verben „begleiten“ und „betreuen“ auftreten.

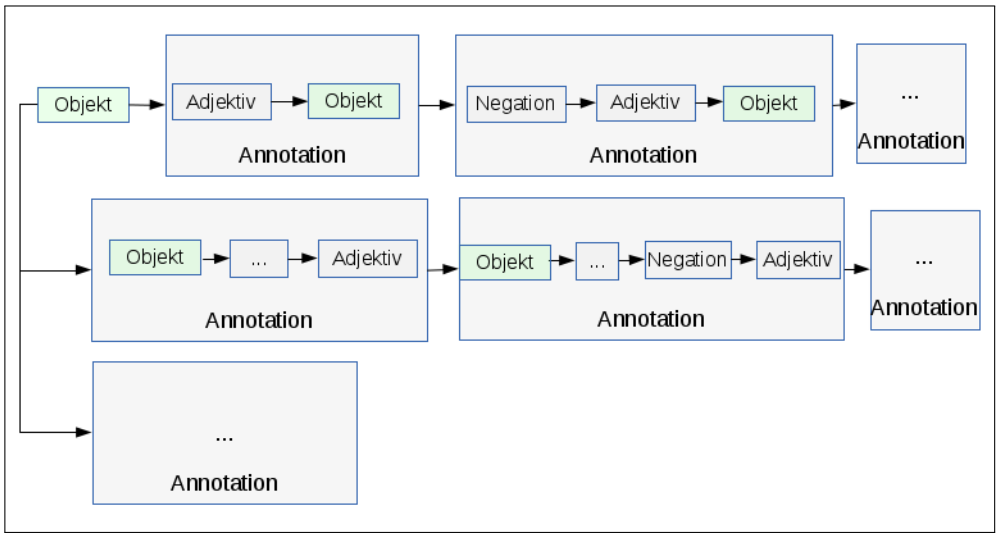

Abbildung 5.13: Aufbau wertender Phrasen

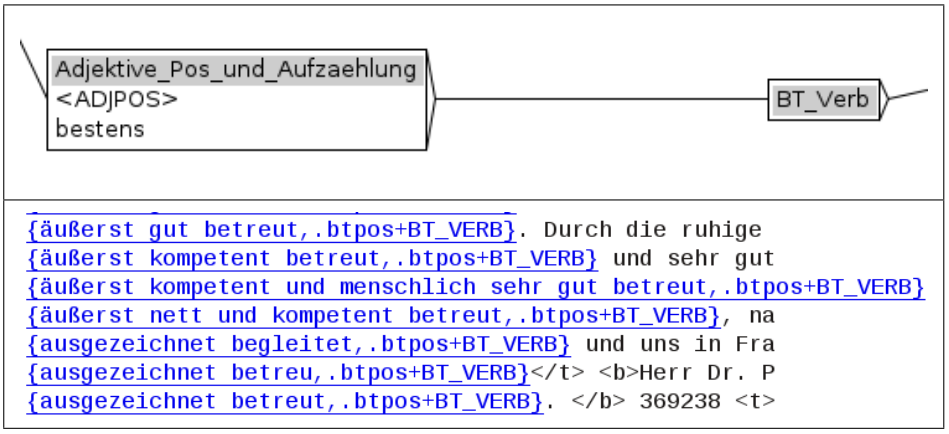

Abbildung 5.14: Wertende Phrasen zur Dimension „Betreuung“ und Konkordanz

5.2.2.3 Konstruktion anderer wertender Ausdrücke

Andere wertende Ausdrücke sind Phrasen, die die Meinungen mit anderen sprachlichen Mitteln als Wortgruppen mit wertenden Adjektiven bilden. Im Ausdruck „Er hat sich viel Zeit genommen“ muss man gleichzeitig den rechten und den linken Kontext des Bewertungsobjekts „Zeit“ berücksichtigen,

zumal das Verb „genommen“ eine entscheidende Rolle spielt. Das bedeutet, dass man beim Aufbau eines Graphen von dem ganzen Ausdruck „Zeit genommen“ ausgeht und die Klassifikation der Zeitmengen und nicht der Adjektive vornehmen muss (s. auch Abschnitt 5.2.1.3, Seite 155). Ähnlich wie bei den Ausdrücken zur „Genommenen Zeit“ wird mit den wertenden Phrasen zu Diskriminierungen und Bestätigungsfehlern verfahren. Entsprechend den im Kapitel 2, Abschnitt 2.1.3 aufgestellten Definitionen der Effekte (hier: Bestätigungsfehler (Seite 39) und Diskriminierungen (Seite 40)), werden Sets an möglichen sprachlichen Ausdrucksmöglichkeiten dieser Phänomene zusammengestellt und entsprechende lokale Grammatiken entwickelt. Auf der Abbildung 5.15 ist eine lokale Grammatik zu sehen, mit der man Phrasen extrahieren kann, mit denen explizit die Bestätigung eigener Vermutungen, Erfahrungen u. ä. (s. Subgraph auf derselben Abbildung) ausgedrückt wird.

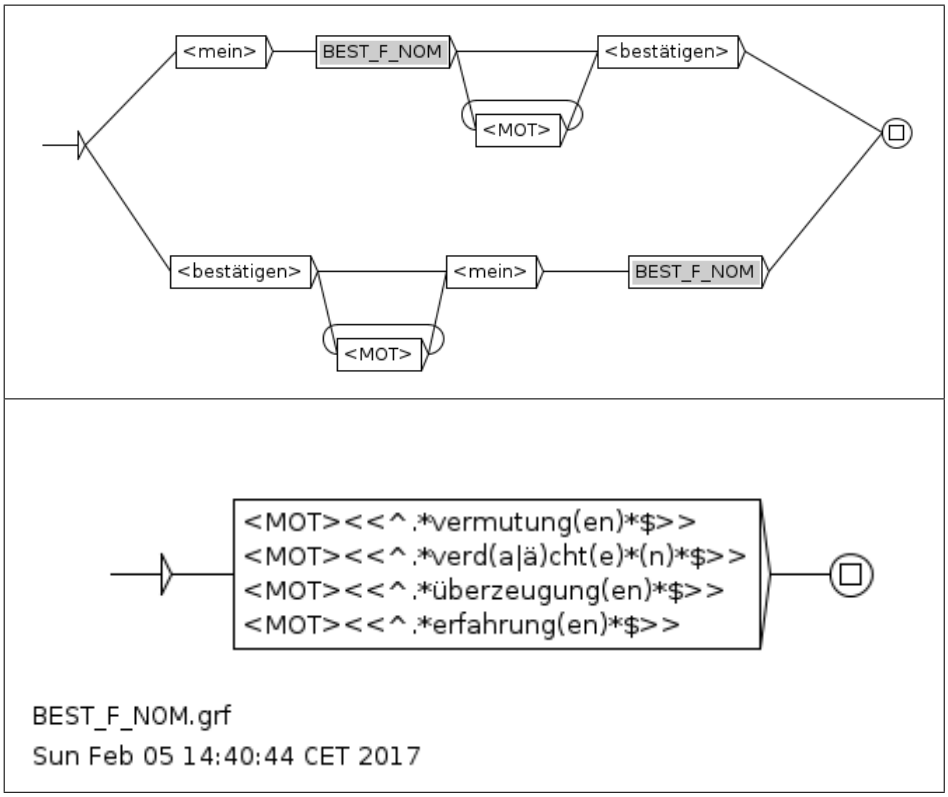

Abbildung 5.15: Graph und Subgraph zur Extraktion wertender Phrasen zu Bestätigungsfehlern

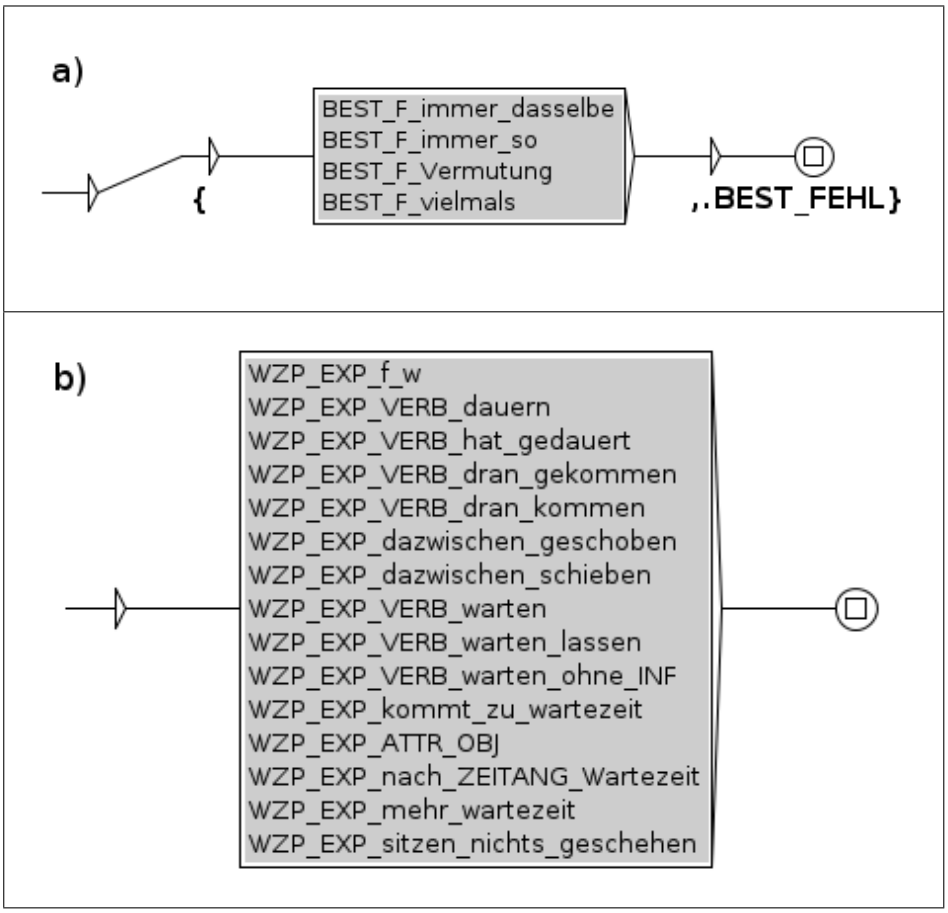

Abbildung 5.16: Mastergraphen zu Bestätigungsfehlern und zur Dimension „Wartezeit (Praxis)“

5.2.2.4 Master- und Phrasengraphen

Die iterativ konstruierten Grammatiken, deren Ausgangspunkte Bewertungsobjekte sind und jene die wertenden Phrasen zu diesen Objekten bilden, werden sowohl für jede Dimension als auch für jeden der bekannten zwei Effekte in den sogenannten Mastergraphen zusammengefasst. Das bedeutet, dass die Mastergraphen nur aus Subgraphen bestehen, die wiederum Phrasen zu Dimensionen oder zu Effekten zusammenfassen (Phrasengraphen). Die Annotationen bei den Effekten erfolgen in den Mastergraphen selbst, da in allen Fällen von einer negativen Polarität der Aussagen ausgegangen wird (s. auch Seite 153, s. Abbildung 5.16a)). Bei Phrasen zu Dimensionen

verhält es sich anders: hier werden die Annotationen innerhalb der Phrasengraphen vorgenommen (s. Abbildung 5.16b)), wobei positive und negative Polaritäten der Ausdrücke entweder in einem oder in verschiedenen Graphen untergebracht werden, was von der Größe und der Übersichtlichkeit jeweiliger Grammatiken abhängt. Auf der Abbildung 5.17 ist ein Phrasengraph zur Dimension „Praxisausstattung“ gezeigt, dessen zu extrahierenden Phrasen in einigen Ausdrücken auch die Dimension „Entertainment“ betreffen. Im Beispiel (5.7) sind Phrasen zu jedem der zehn Pfade ((5.7)(a) bis (5.7)(j)) der o. g. Grammatik zusammengestellt (jeweils ein Beispiel zu einem Pfad). Die vier fehlenden Beispiele erklären sich dadurch, dass es eine Teilung der Bewertungsobjekte in ‚Entertainment-Objekte‘ und ‚Ausstattungsobjekte‘ gibt, jedoch nicht alle sprachlichen Ausdrücke mit allen Objekten in den Bewertungen vertreten sind.

(5.7) (a) {Praxis-TV ist sehr informativ,.etpos+papos+Ausstattung_2}
(b) –
(c) {immer genügend Zeitschriften,.etpos+papos+Ausstattung_2}
(d) {mit Spielzeug ausgestattet,.etpos+papos+Ausstattung_2}
(e) –
(f) –
(g) {Geräteausstattung ist sehr gut,.papos+Ausstattung_2}
(h) {Geräte sind auf dem neuesten Stand,.papos+Ausstattung_2}
(i) {Sehr gute medizinische Geräteausstattung,.papos+Ausstattung_2}
(j) –

5.2.3 Pragmatisch

Als linguistische Disziplin beschäftigt sich die Pragmatik mit der „Beziehung zwischen sprachlichen Zeichen und den Benutzern sprachlicher Zeichen“. Im außerlinguistischen Kontext bedeutet die Pragmatik „Orientierung auf das Nützliche“, „Sachbezogenheit“ u. ä.[80]. Im Sinne der Pattern-Extraktion mit lokalen Grammatiken wird die Pragmatik als der sinnvolle, durch die begründete Anordnung der Graphen durchzuführende Extraktionsprozess von

[80] `http://www.duden.de/rechtschreibung/Pragmatik` (08.02.2017).

wertenden Phrasen aus dem Korpus mit Arztbewertungen verstanden. Die aufgegriffenen beschriebenen Aspekte der Musterextraktion im semantischen und syntaktischen Sinne sind lediglich Bestandteile des oben angedeuteten Prozesses. Für eine gute Musterextraktion ist es nicht genug, sich lediglich auf eine begrenzte Anzahl der Phrasen zu den vordefinierten Bewertungsdimensionen zu konzentrieren und diese der Reihe nach abzuarbeiten. Zwischen diesen Dimensionen bestehen Zusammenhänge, die auch sprachlich in einem Satz, einem Teilsatz, einer Phrase vorkommen können, wodurch eine Kombination mehrerer Dimensionen erfolgt (s. Beispiel (5.4)). Ein Problem stellt die Berücksichtigung weiterer Kontexte dar, die oft als Aufzählungen nicht nur durch Adjektive realisiert werden. Dies erfordert einer Auseinandersetzung mit den Phrasenkonstruktionen, die sich innerhalb und außerhalb der extrahierten Muster zu Dimensionen befinden. In den nächsten Abschnitten erfolgt eine Beschreibung der Musterextraktion und des Postprocessing, was zusammen das Extraktionsverfahren darstellt und automatisch mit einem Shell-Skript (s. beiliegende DVD/Tools/CognIEffect...) realisiert wird.

5.2.3.1 Musterextraktion

Die Extraktion wertender Muster erfolgt in einer Kaskade (s. Kapitel 4 Abschnitt 4.2.2.6). Zunächst sollen die längeren und oder kombinierten sprachlichen Muster extrahiert werden, die Phrasen, die das Phrasenlexikon PHRASE_LEX enthält. Dafür werden mehrere lokale Grammatiken geschrieben, die die Lexikoneinträge mit ihren Kodierungen ansprechen, wodurch Phrasen mit Aussagen zu zwei und mehr Dimensionen erkannt und entsprechend annotiert werden. Nach diesen ‚kombinierten' Phrasen erfolgt die Extraktion der Phrasen zu Bewertungsdimensionen und schließlich zu den „Bestätigungsfehlern" und „Diskriminierungen" mit den im Abschnitt 5.2.2.4 beschriebenen Mastergraphen.

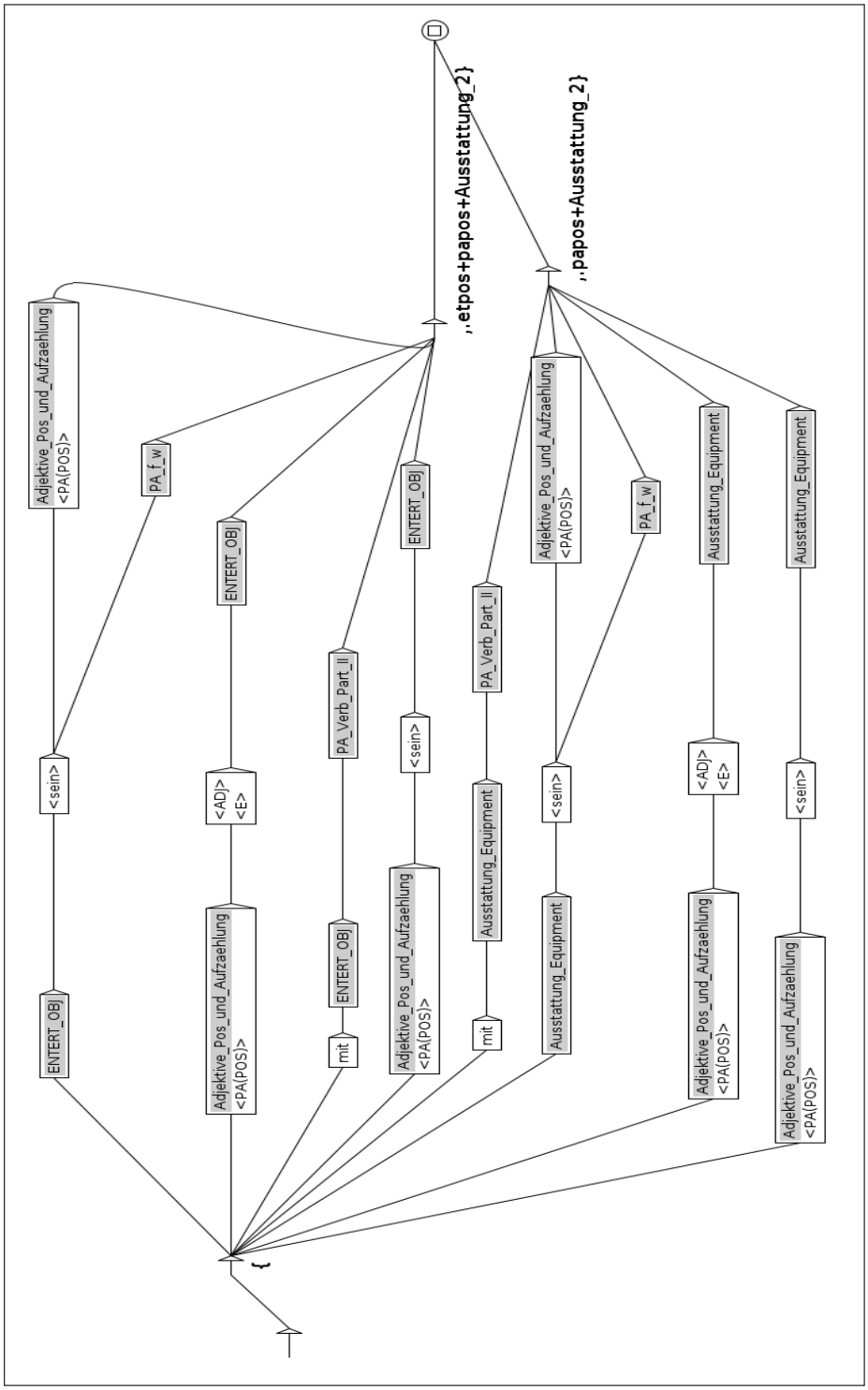

Abbildung 5.17: Phrasengraph der Dimension „Praxisausstattung“

5.2.3.2 Postprocessing

5.2.3.2.1 Explizite Kontexte

Das Korpus, das nach dem Durchlauf der ersten Kaskade mit wertenden Phrasen annotiert wurde, wird zur Weiterverarbeitung, d. h. zur Erweiterung um einige explizite Kontexte eingesetzt. Dabei geht es um die rechten und die linken Kontexte der bereits extrahierten Muster. Auf der Abbildung 5.18 ist eine lokale Grammatik gezeigt, die den linken Kontext extrahierter Pattern auf mögliche Erweiterungen überprüft und annotiert. Mit der rechten Box des Graphen erkennt man durch die Aufzählung aller Kombinationen von verwendeten Tags sämtliche extrahierten Muster zu Bewertungsdimensionen. Ein Beispiel von einer der 16 lokalen Grammatiken ist auf derselben Abbildung zu sehen: Der Graph „module_BT.grf“ fasst Benennungen von Bewertungsobjekten der Dimension „Betreuung“ zusammen, beim Vorhandensein derer eine zusätzliche Annotation (hier: ‚(bt)‘) stattfindet, die später bei einer Weiterverarbeitung mit der Polarität der Aussage ergänzt wird.
Aufgrund zahlreicher Möglichkeiten von sprachlichen Formulierungen werden wertende Äußerungen der Patienten nicht nur bei den Aufzählungen der ärztlichen Leistungen übersehen. Besonders wichtig sind dabei einige Hauptdimensionen (s. Seite 8), deren explizite Benennungen von Bewertungsobjekten in Mastergraphen nicht berücksichtigt werden konnten. Von der Polarität der bereits getätigten wertenden Phrasen abhängig, kann man daher in der Nähe (rechter und linker Kontext) nach expliziten Benennungen anderer Dimensionen suchen, wie dies ebenfalls beschrieben wurde. Hier geht es um eine differenziertere Vorgehensweise, die in einem Graphen auf der Abbildung 5.19 gezeigt wird. Im Vergleich zur Grammatik auf der Abbildung 5.18, die Kombinationen von Bewertungsobjekten der Dimensionen miteinander beliebig erlaubt, werden hier einerseits die Bewertungsobjekte der Dimension „Aufklärung“ im Kontext der Dimensionen „Vertrauen“ und „Behandlung“ und andererseits die der Dimension „Behandlung“ im Kontext der „Freundlichkeit“ extrahiert und annotiert. Im Beispiel (5.8) sind einige extrahierte Sätze nach dem Durchlauf beschriebener Kaskaden aufgeführt.

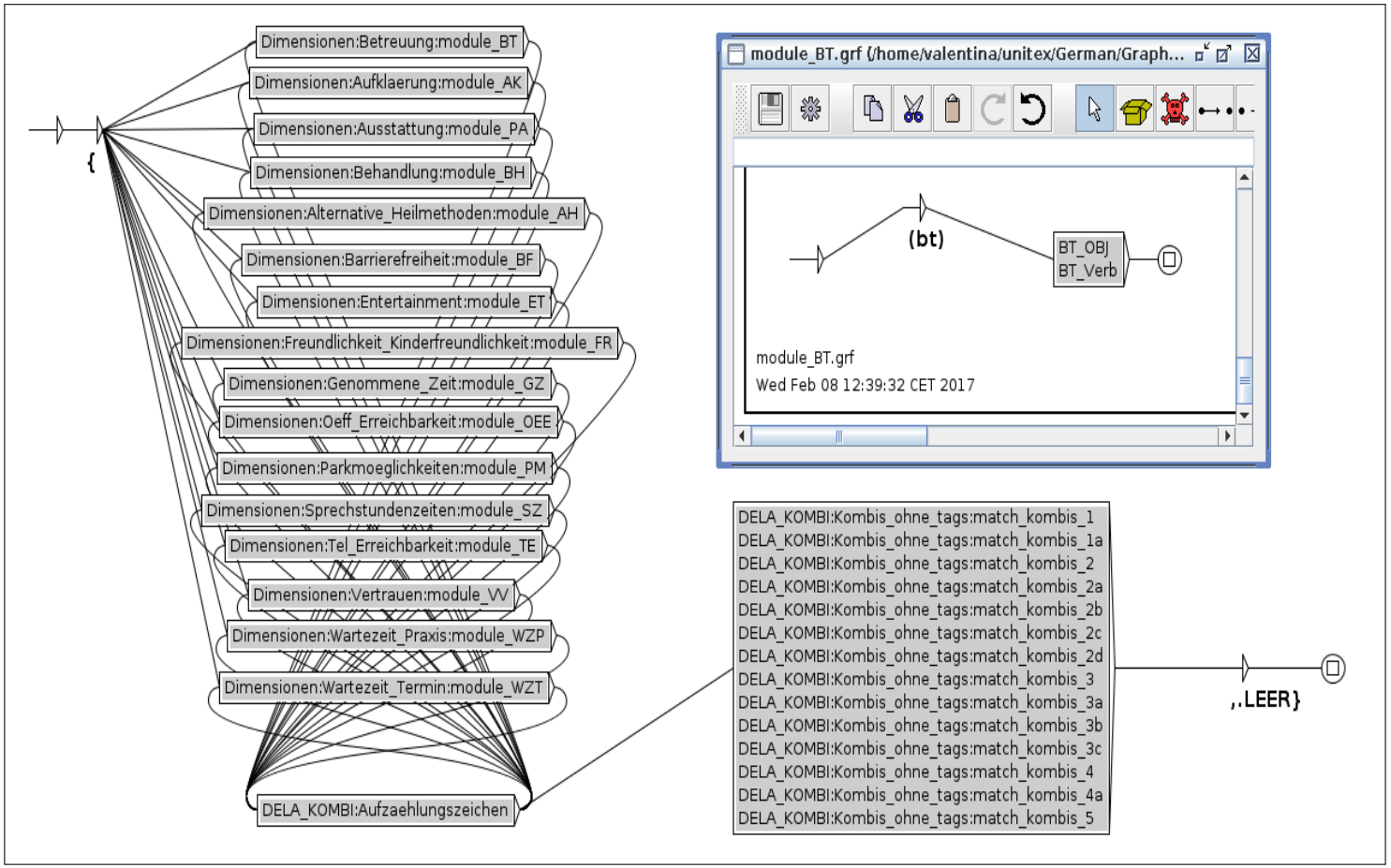

Abbildung 5.18: Erweiterung der Extraktionsmuster um explizite Kontexte und Bewertungsobjekte zur Dimension „Betreuung“

(5.8) (a) {{Sehr gründliche Untersuchung,.bhpos+Behandlung_2} und verständliche Aufklärung,.akpos}

(b) {{sehr kompetenter Arzt,.bhpos} mit umfassender Beratung,.akpos}

(c) {{Ich fühle mich hier gut aufgehoben,.vvpos+Vertrauen_9} und habe das Gefühl soweit auch auf alle meine Fragen oder Anliegen eine Antwort,.akpos}

5.2.3.2.2 Annotationskorrekturen

Mit dem selbst entwickelten Perl-Skript werden Korrekturen des Extraktionsprozesses mit regulären Ausdrücken durchgeführt. Bei einer annotierten Phrase im Beispiel (5.9)(a) sind einige Tags überflüssig, die mit dem regulären Ausdruck gelöscht werden, so dass das Ergebnis die Phrase im Beispiel (5.9)(b) darstellt.

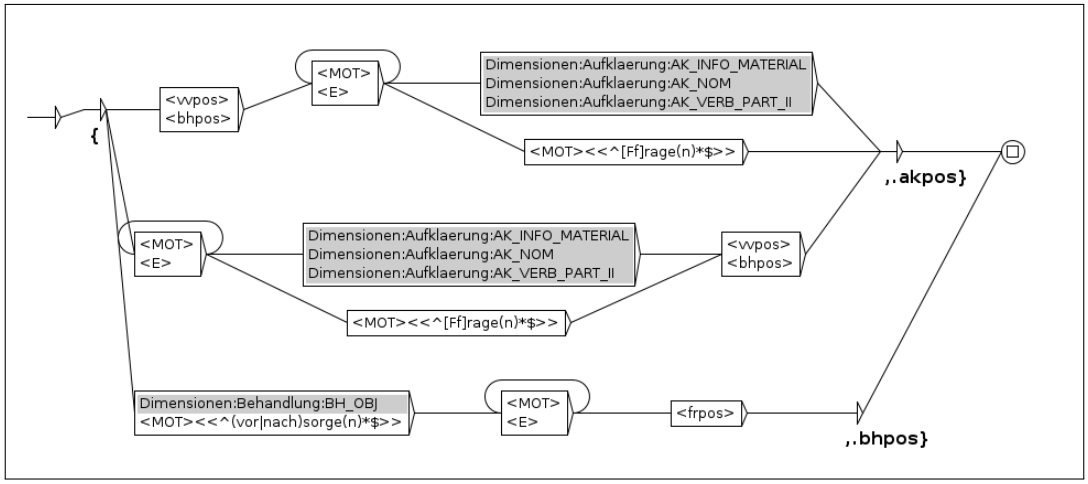

Abbildung 5.19: Erweiterung der Extraktionsmuster um explizite Kontexte einiger Hauptdimensionen

(5.9) (a) Phrase: {(fr){unfreundlich,.frneg+Freundlichkeit_13}, {lange Wartezeit,.wzpneg+WZP_EXPR_ATTR_OBJ},.LEER}

(b) Ergebnis: {unfreundlich,.frneg+Freundlichkeit_13}, {lange Wartezeit,.wzpneg+WZP_EXPR_ATTR_OBJ}

Außer ähnlichen Korrekturen werden mit dem o. g. Programm die Polaritäten einiger Phrasen bzw. Objekte ergänzt, deren Annotationen z. B. lediglich die Angaben zur Dimension und nicht zur Polarität enthalten (s. Abbildung 5.18, Tag: ‚(bt)'). Nach dem Durchlauf des hier kurz beschriebenen Postprocessing-Programms gilt der Extraktionsprozess als beendet. Dessen Ergebnisse werden im Kapitel 6, Abschnitt 6.3 evaluiert.

5.3 Annotationen und Kriterienidentifikation

In diesem Abschnitt werden Annotationen beschrieben sowie die automatische Identifikation der Kriterien erläutert.

5.3.1 Annotationen

Unter Annotationen wird zum einen der Einsatz der Annotatoren verstanden, die manuelle Arbeiten in verschiedenen Kontexten erledigen (s. z. B. Kapitel 3, Abschnitt 3.3.1, Seite 95f.). Im Abschnitt 4.2.3 wurden Probleme thematisiert, die den Einsatz von Annotatoren erfordern. Die Vorgehensweise bei der Annotatorenarbeit wird im Abschnitt 5.3.1.1 beschrieben. Zum anderen werden Annotationskonventionen erläutert, die in der vorliegenden Arbeit zur Annotation wertender Muster etc. vorgenommen wurden (Abschnitt 5.3.1.2).

5.3.1.1 Einsatz der Annotatoren

In der vorliegenden Arbeit werden Annotatoren bei den Zuordnungen sprachlicher Muster den vorgegebenen Jameda-Dimensionen (Pattern-Definitionen) und bei der Qualitätsauswertung des Programms zur Bestimmung der Leit- bzw. Ankerdimension pro Bewertung eingesetzt.

5.3.1.1.1 Pattern-Definitionen

Im Kapitel 3, Abschnitt 3.3.1 wurden die Einsätze der Annotatoren in aktuellen Arbeiten zusammengefasst. So werden für diese Arbeit ebenfalls zwei Annotatoren[81] engagiert, die unabhängig voneinander die Zuordnung der

[81] Zwei Annotatoren und eigenhändig definierter ‚Goldstandard'.

automatisch extrahierten Patientenausdrücke den Dimensionen vornehmen. Dabei wird wie folgt vorgegangen:

- Von den für jede Dimension mit Mastergraphen annotierten Pattern werden zwanzig häufigsten extrahiert.
- Aus dem Phrasenlexikon (PHRASE_LEX) werden zwanzig Pattern extrahiert, die vorher nach der Anzahl der Tags sortiert werden. Die Pattern mit den meisten Tags werden dabei priorisiert, wodurch die Auseinandersetzung mit möglichst vielen Dimensionen erfolgt.
- Die Annotationstags mit Informationen zu Bewertungsdimensionen werden bei extrahierten Pattern (insgesamt 355) gelöscht, wodurch eine neue Zuordnung von diesen Dimensionen von Annotatoren vorgenommen werden kann. Bei der Auswertung werden ihre Annotationen mit den anfänglich gelöschten Tags (‚Goldstandard') verglichen (s. Kapitel 6 Abschnitt 6.2.2.1).
- Alle für Annotatorenarbeit extrahierten Pattern werden nach dem Zufallsprinzip sortiert und in einer Tabelle gespeichert. In die Tabelle werden außerdem alle Dimensionsbezeichnungen eingetragen, so dass die Annotatoren die Zuordnung der Pattern den Dimensionen vornehmen können (s. Tabelle 5.9).
- Eine Datei mit Arbeitsanweisungen wird erstellt (s. DVD/Evaluationsdaten/Patternmatching).
- Die zuletzt genannte und die Zuordnungsdatei werden an die Annotatoren verschickt.

5.3.1.1.2 Bestimmung der Leitdimension

Zur Bestimmung der Leitdimension werden die Bewertungstexte zunächst mit den Transduktoren bearbeitet (s. Abschnitt 5.2), die die Benennungen zu Bewertungsobjekten der Dimensionen zusammenfassen und auf der Abbildung 5.18 zu sehen sind. Dadurch werden die Bewertungsobjekte in den Bewertungstexten annotiert. Danach wird mit einem Perl-Skript (s. DVD/Tools/ CognIEffect...) pro Bewertung ausgerechnet, welche Wörter von welcher Dimension am meisten vertreten sind, indem die gesetzten Tags pro Bewertung

	1. Behandlung (bh)	2. Aufklärung (ah)	3. Vertrauensverhältnis (w)	4
die praxis ist sehr sauber				
den Arzt sehr gut, freundlich und kompetent, ausreichende Aufklärung, es blieb ein Gefühl des Vertrauens				
Sehr kompetente und freundliche Ärztin				
sehr kompetent und freundlich betreut				
unfreundliche praxis				
eigene parkplätze				
sich zeit für den patienten nimmt				
optimal betreut				
habe mich sehr wohl gefühlt				
termin sofort				
hat sich viel zeit genommen				
kurze wartezeiten				
die praxis ist schön eingerichtet				
gut versorgt				
sehr gute beratung				
alten praxis				
sehr gute erreichbarkeit				
telefonisch immer zu erreichen				
verständnisvoller arzt				
nettes praxisteam				
telefonische erreichbarkeit sehr gut				

Tabelle 5.9: Auszug aus der Zuordnungsdatei der Pattern-Objektivierung

ausgezählt werden. Die Ergebnisdimension wird dann als Leitdimension definiert. Sollten zwei oder mehr Dimensionen in einer Bewertung über eine gleiche Anzahl von dimensionsspezifischen Wörtern verfügen, so wird keine Leitdimension für die betreffende Bewertung definiert und als Ergebnis ‚undef' ausgegeben. Für die Annotatorenarbeit bzw. zur Kontrolle der Qualität von dem o. g. Programm wurde zunächst ein Korpus aus den Bewertungen zusammengestellt, die aufgrund ihrer geringen Anzahl (2 Bewertungen für eine Praxis) aus dem Jameda-Trainingskorpus aussortiert wurden (s. Abschnitt 5.1.1.1.2, Seite 130; s. DVD/Evaluationsdaten/Leitdimension/Korpus _Leit_Dim_annotiert.txt). 500 Bewertungen des erstellten Korpus wurden mit dem Programm zur Bestimmung der Leitdimension bearbeitet. Für alle Bewertungen wurden die Ergebnisse in einer Datei (s. DVD/Evaluationsdaten/Leitdimension/LEITDIM_AUSGABE.txt) gespeichert. Diese Datei, die Datei mit Anweisungen sowie 500 extrahierter Bewertungen wurden an die Annotatoren verschickt. Ihre Aufgabe bestand darin, jede der 500 Bewertungen zu lesen, die Programmergebnisse pro Bewertung zu betrachten und danach zu entscheiden, ob die für die jeweilige Bewertung definierte Leitdimension richtig (‚+') oder falsch (‚–') ist.
Bei der beschriebenen Vorgehensweise lässt sich die Voreingenommenheit der Annotatoren nicht vermeiden. Jedoch wäre – der Erfahrung nach – die Aufgabenstellung, bei der die individuelle Entscheidung für oder gegen bestimmte Leitdimension erfolgen würde, ohne dass der Vergleich mit Programmergebnissen stattfindet, für Annotatoren zu komplex und in der dafür vorgesehenen Zeit nicht zu bewältigen.

5.3.1.2 Annotationskonventionen

Bei den Annotationskonventionen werden Outputs beschrieben, die einerseits von lokalen Grammatiken und andererseits von dem Identifikationsprogramm produziert werden.

5.3.1.2.1 Dimensionen und Polarität

Für die Annotationen extrahierter Muster verwendet man bestimmte Tags, die mit verschiedenen Attributen versehen werden (Wolfgruber, 2015, S. 79). Im Fall der Effekte-Identifikation in den von Patienten verfassten Bewertungen sind Polaritäten der Aussagen und deren Zugehörigkeit zu den Bewertungsdimensionen interessant. In der Tabelle 5.10 sind die innerhalb der

Tags verwendeten Abkürzungen der Dimensionen und Polaritäten der Aussagen aufgeführt. In den mit geschweiften Klammern eingeführten Tags (s. z. B. Beispiel (5.9)(b), Seite 166) sind außer den extrahierten Pattern Metainformationen (s. Tabelle 5.10) aufgeführt, die die Zuordnung der Aussage zur betreffenden Dimension und deren Polarität beinhalten. Dabei wird jede Dimension (auf dem o. g. Beispiel: ‚wzp') mit der entsprechenden Polarität (‚neg') kombiniert, so dass sich im Ergebnis das zusammengesetzte Attribut (‚wzpneg') ergibt[82]. Üblich sind die Annotationen an das XML-Format angelehnt, um eine weitere maschinelle Verarbeitung extrahierter Pattern zu ermöglichen (vgl. Wolfgruber, 2015; vgl. Geierhos, 2010). Im Rahmen der vorliegenden Arbeit werden einheitlich lexikalische Tags benutzt, was bei Bedarf selbstverständlich leicht auf XML-Format zu ändern wäre.

Bezeichnung	**Abkürzung**	**Erläuterung**
Dimension	bh	„Behandlung“
	ak	„Aufklärung“
	gz	„Genommene Zeit“
	fr	„Freundlichkeit“
	vv	„Vertrauensverhältnis“
	wzt	„Wartezeit (Termin)“
	wzp	„Wartezeit (Praxis)“
	sz	„Sprechstundenzeiten“
	et	„Entertainment“
	bt	„Betreuung“
	kfr	„Kinderfreundlichkeit“
	bf	„Barrierefreiheit“
	pa	„Praxisausstattung“
	te	„Telefonische Erreichbarkeit“
	oee	„Öffentliche Erreichbarkeit“
	pm	„Parkmöglichkeiten“
	ah	„Alternative Heilmethoden“
Polarität	pos	„Positiv“
	neg	„Negativ“

Tabelle 5.10: Verwendete Attribute innerhalb der Annotationstags

[82] „WZP_EXPR_ATTR_OBJ“ aus dem genannten Beispiel ist die Bezeichnung des Graphen, mit dem entsprechende Ausdrücke extrahiert wurden. Seine Benennung erfüllt einen praktischen Zweck und wird nicht weiter erläutert.

5.3.1.2.2 Programmannotationen

Während im vorigen Abschnitt das Annotationssystem der mit lokalen Grammatiken extrahierten Muster beschrieben wurde, wird hier auf den Output, der mit verschiedenen Programmen (s. DVD/Tools/...) produziert wurde, eingegangen.

a) Allgemeine Angaben

Allgemeine Angaben betreffen die Programmausgaben, die hauptsächlich zur Weiterverarbeitung bei der Effekte-Identifikation oder zur schlichten Übersichtlichkeit produziert werden. Bei der im Abschnitt 5.3.1.1.2 beschriebenen Bestimmung der Leitdimension werden den Bewertungen die Tags mit <LEITDIM>...</LEITDIM> hinzugefügt, in die die Ergebnisse des Programms eingeschlossen werden (z. B. ‚(fr)' für Dimension „Freundlichkeit"). Die Ausreißer-Dimensionen werden in <AUSR>...</AUSR> aufgeführt, wobei ebenfalls die Polarität angegeben wird (z. B. ‚Behandlung -> neg', s. auch Abbildung 5.21). Weitere Angaben, die den Bewertungen hinzugefügt werden, sind z. B.:

- „Anzahl von Ausreissern = 5"
- „Korrelationskoeffizient Punkte = 0" (Halo-Effekt)
- „Count für gleiche Noten = 2" (Halo-Effekt)
- etc.

b) Effekte und Scores

Die mit lokalen Grammatiken extrahierten Phrasen zu Bestätigungsfehlern und Diskriminierungen werden in die Tags {... ‚.BEST_FEHL} (s. Abbildung 5.16a)) und {... ‚.DISKR} eingeschlossen. Durch das Identifikationsprogramm (s. Abschnitt 5.3.2) werden pro Effekt die Ausgaben zu vergebenen Punkten und deren prozentuellen Entsprechungen (s. Abschnitt 5.3.2.3) produziert, z. B. „Halo Effekte: 3 von 30; in Prozenten 10".

5.3.2 Kriterienidentifikation

Wie im Kapitel 3, Abschnitt 3.2.2.2 (Seite 94) festgestellt, gibt es mehrere Ebenen zur Aufstellung der Kriterien automatischer Effekte-Identifikation. Auf der Bewertungsebene wurden mit lokalen Grammatiken Muster zu Bewertungsdimensionen sowie Phrasen zu Bestätigungsfehlern und Diskriminierungen extrahiert. Es bleibt hier lediglich Vergleiche zwischen nummerischen und textuellen Bewertungen durchzuführen (Konsistenzüberprüfung). Die Bestimmung der Leitdimension wurde im Abschnitt 5.3.1.1.2 erläutert. Auf der Arztpraxisebene wird im Abschnitt 5.3.2.1 die Berechnung der Ausreißer in 3 Schritten beschrieben, wobei auf der Abbildung 5.21 die Ergebnis-Annotation des Korpus gezeigt wird. Die Berechnung des Korrelationskoeffizienten auf der Gesamtdatenebene wird im Abschnitt 5.3.2.2 beschrieben, die Ergebnisse werden in einer Datenstruktur im entsprechenden Programm gespeichert (s. DVD/Tools/../effect.pl), worauf dann bei der Identifikation der Halo-Effekte zugegriffen wird.
Bevor man zu der Beschreibung der erwähnten Berechnungen der Kriterienzusammensetzung sowie der Vergabe der Scores pro Effekt übergeht, bleibt im aktuellen Abschnitt auf einige allgemeine Programmdaten und Funktionen einzugehen. Zunächst werden Datenstrukturen angelegt, in denen bekannte bzw. im Voraus ausgerechnete Sachverhalte wie Korrelationskoeffizient gespeichert werden. Ein weiteres Beispiel für solche Daten wären die Bezeichnungen der Haupt- und Nebendimensionen mit deren Abkürzungen, die in den Annotationstags zu entsprechenden Pattern (s. Tabelle 5.10, Seite 171) zu finden sind. Für jeden Effekt werden Perl-Subroutinen angelegt, aus denen man auf die genannten Datenstrukturen und weitere Funktionen zugreifen kann. Solche Funktionen (Subroutinen) differenzieren z. B. nummerische Werte für positive (Noten: 1.0, 2.0) und negative (Noten: 3.0 bis 6.0) Polaritäten, benennen die Haupt- und Leitdimensionen sowie positive und negative Ausreißer, geben die Bewertungsnummer zurück oder berechnen prozentuelle Anteile der pro Effekt vergebenen Scores.

5.3.2.1 Berechnung der Ausreißer

Die im Abschnitt 4.3.1.3, Seite 120 beschriebenen Ausreißer werden in drei Schritten berechnet:

5.3.2.1.1 Schritt 1

Zur automatischen Identifikation der Ausreißer wird das Korpus, das im Abschnitt 5.1.1.2.2, Seite 135 beschrieben wurde, herangezogen. Für jede Arztpraxis mit mehr als drei Bewertungen (zulässige Mindestanzahl, s. Abschnitt 5.1.1.2.2) wird zunächst ausgerechnet, wie viele Patientenbewertungen eine Mehrheit bilden. Zwei von drei Bewertungen bilden ca. 67%, was für alle Arztpraxen als Mindestmaß für die gesuchte Mehrheitszahl gelten soll. Danach wird eine Statistik zu positiven („1.0“ oder „2.0“) und negativen („3.0“ bis „6.0“) Noten pro Dimension in jeder Arztpraxis angefertigt, wodurch ausgerechnet wird, für welche Note(n) bei einer konkreten Dimension sich die Mehrheit der Patienten entschieden haben. Alle anderen Noten der Patienten zu dieser Dimension sind logischerweise als Ausreißer zu interpretieren, da diese die Minderheit bilden. Die nicht vergebenen Noten („n/a“) werden nicht interpretiert und dementsprechend nicht berücksichtigt. Eine mögliche Begründung dafür wurde in Bezug auf MUM-Effekt im Kapitel 2, Seite 34 angegeben. Die Ergebnisse werden in einer Datei gespeichert, deren Auszug auf der Abbildung 5.20 zu sehen ist. Zu jeder Arztpraxis werden die Dimensionen aufgeführt, die durch ihre positiven (‚pos‘) oder negativen (‚neg‘) nummerischen Bewertungen die Ausreißer bilden. Insgesamt wurden beispielsweise 16432 Bewertungen in 9709 Arztpraxen (Trainingskorpus zur Effekte-Identifikation) gefunden, die mit Ausreißern versehen sind.

5.3.2.1.2 Schritt 2

Im zweiten Schritt wird im Trainingskorpus zu kognitiven Effekten (s. Abschnitt 5.1.1.1.2) nach entsprechenden Bewertungen mit Ausreißern gesucht, indem die Identifikationsnummer der Praxen mit Ausreißern (ArztID, s. Abbildung 5.20) im Trainingskorpus identifiziert und bei den Bewertungen dieser Praxen die Ergebnisse aus der Datei der o. g. Abbildung vermerkt werden.

5.3.2.1.3 Schritt 3

Im letzten Schritt werden einige Korrekturen vorgenommen, so dass die mit den Ausreißern versehenen Bewertungen zum Ende zusätzliche Tags sowie die Auszählung der Ausreißer erhalten, wie dies auf der Abbildung 5.21 gezeigt wird.

```
ArztID = 80408022 => Freundlichkeit = pos Behandlung = pos Zeit = pos Betreuung = pos
Vertrauensverhaeltnis = pos Aufklaerung = pos
ArztID = 80026742 => Zeit = neg Vertrauensverhaeltnis = neg
ArztID = 81031251 => WartezeitPraxis = neg WartezeitTermin = neg Sprechstundenzeiten =
neg
ArztID = 81039649 => Freundlichkeit = neg Behandlung = neg Zeit = neg
Vertrauensverhaeltnis = neg Aufklaerung = neg
ArztID = 80080613 => Freundlichkeit = neg Behandlung = neg Vertrauensverhaeltnis = neg
ArztID = 81094536 => Behandlung = pos Zeit = pos Vertrauensverhaeltnis = pos Aufklaerung
= pos
ArztID = 81044033 => Behandlung = neg Betreuung = neg Vertrauensverhaeltnis = neg
ArztID = 81181164 => Aufklaerung = neg
ArztID = 81110962 => Freundlichkeit = neg Behandlung = neg Zeit = neg
Vertrauensverhaeltnis = neg Aufklaerung = neg
ArztID = 80332523 => Behandlung = neg Zeit = neg Vertrauensverhaeltnis = neg Aufklaerung
= neg
```

Abbildung 5.20: Dateiauszug zur Identifikation der Ausreißer (Schritt 1)

```
<column name="BewertungID">974209</column> <column name="ArztID">80280382</column>
<LEITDIM>notdef</LEITDIM> <column name="Titel">Nicht empfehlenswert !</column> <column
name="Bewertung">Diesen Arzt kann ich leider nicht empfehlen - Einzelheiten erspare ich mir
!</column> <column name="Datum">03.07.2013</column> <column
name="Kassenart">Kassenpatient</column> <column name="Gesamtnote">5.2</column> <column
name="b_Vertrauensverhaeltnis">5.0</column> <column name="b_Aufklaerung">5.0</column> <column
name="b_Behandlung">5.0</column> <column name="b_Zeit">5.0</column> <column
name="b_Freundlichkeit">6.0</column> <column name="b_WartezeitTermin">4.0</column> <column
name="b_WartezeitPraxis">3.0</column> <column name="b_Sprechstundenzeiten">4.0</column>
<column name="b_Betreuung">4.0</column> <column name="b_Praxisausstattung">4.0</column>
<column name="b_ErreichbarkeitTEL">4.0</column> <column
name="b_ErreichbarkeitOEFF">3.0</column> <column name="b_Heilmethoden">n/a</column> <column
name="b_Parkmöglichkeiten">5.0</column> <column name="b_Entertainment">n/a</column> <column
name="b_Barrierefreiheit">4.0</column> <column name="b_Kinderfreundlichkeit">n/a</column>
<column name="timestamp">2013-10-11 23:14:52</column> <column name="Age"> über 50</column>
<AUSR>Behandlung -> neg</AUSR> <AUSR>Freundlichkeit -> neg</AUSR> <AUSR>Zeit -> neg</AUSR>
<AUSR>Aufklaerung -> neg</AUSR> <AUSR>Praxisausstattung -> neg</AUSR>
<AUSR>Vertrauensverhaeltnis -> neg</AUSR> <AUSR>ErreichbarkeitOEFF -> neg</AUSR> Anzahl von
Ausreissern = 7
```

Abbildung 5.21: Dateiauszug zur Identifikation der Ausreißer (Schritte 2 und 3)

5.3.2.2 Korrelation der Dimensionen

Bei der methodischen Vorgehensweise in Bezug auf den Halo-Effekt wurde im Abschnitt 4.3.1.2 auf die Tendenz einer hohen Korrelation der Eigenschaften aufmerksam gemacht sowie auf die Möglichkeit deren Ausrechnung mit dem Pearson's Korrelationskoeffizienten hingewiesen. Gleichzeitig wurde festgestellt, dass mit möglichen Fehlinterpretationen der Definitionen von Bewertungsdimensionen etc. zu rechnen ist. Allerdings existiert ein für alle Bewertenden einheitliches nummerisches Bewertungssystem, was die Möglichkeit der einheitlichen Vergabe nummerischer Werte impliziert. Ohne weitere Differenzierung der individuellen Interpretationen von Bewertungsdimensionen

werden die nummerischen Werte als Basis zur Ausrechnung des Korrelationskoeffizienten innerhalb des gesamten Jameda-Korpus herangezogen. Wie dieser berechnet wird und wie viele solche Berechnungen nötig sind, wird wie folgt erläutert.
Die Korrelation der Dimensionen wird nach der im Kapitel 3, Abschnitt 3.3.2.2 (Seite 98) erläuterten Formel paarweise ausgerechnet. Wie werden alle Dimensionen paarweise miteinander kombiniert? Als Teilgebiet der Wahrscheinlichkeitstheorie (Leonhart, 2013, S. 133) beschäftigt sich die Kombinatorik mit den Möglichkeiten der Zusammenstellung von Elementen einer Menge (Clauß und Ebner, 1977, S. 136). Dabei werden Permutationen und Kombinationen unterschieden, je nach den Regeln der Anordnung von Objekten (ebd., S. 136f.). Wie bekannt (s. Kapitel 2, Tabelle 2.1, Seite 16), hat Jameda 17 Bewertungsdimensionen, die man nummerisch bewerten kann, definiert. Um die paarweise Korrelation dieser Dimensionen auszurechnen, müssen aus ihnen Paare konstruiert werden, wobei aus den nummerischen Werten jeder Dimension des Paars die Tupel gebildet und in Form einer zweispaltigen Tabelle im Programm „LibreOffice“ gespeichert werden. Wie in der Einleitung zu dieser Arbeit (Kapitel 1, Seite 8) erläutert, sind für die Abgabe einer Praxis-Bewertung die nummerischen Werte von Hauptdimensionen verpflichtend, nicht jedoch für die Nebendimensionen. Aus diesem Grund orientiert sich die Menge der paarweise zusammengestellten Werte nach der kleineren Anzahl der von Patienten vergebenen nummerischen Werte zu einer Dimension. Ob die kleinere oder die größere Anzahl nummerischer Werte bzw. die Werte von welcher Dimension in der ersten Spalte der gebildeten Tabellen stehen, ist irrelevant, da der Korrelationskoeffizient unabhängig von der Reihenfolge der Werte ist (s. Formel 3.9, Seite 98).
Aus den eben erfolgten Ausführungen lässt sich Folgendes zusammenfassen:

- Die Reihenfolge der Dimensionspaare ist irrelevant
- Die Wiederholungen der Dimensionen sind überflüssig und für die formulierte Zielsetzung ebenfalls irrelevant

„Jede Zusammenstellung von k aus n Elementen, bei der die Anordnung der Elemente unberücksichtigt bleibt, heißt eine Kombination k-ter Klasse. Man unterscheidet Kombinationen ohne Wiederholung und Kombinationen mit Wiederholung, je nachdem, ob die k Elemente voneinander verschieden sind oder nicht“ (ebd., S. 137). Die Anzahl der Kombinationen ohne Wiederholung und ohne Reihenfolge lässt sich nach folgender Formel ausrechnen (ebd.):

$$K_{n;k} = \frac{n!}{(n-k)!\ k!} = \binom{n}{k} \tag{5.3}$$

Im Fall der Kombinationen der Bewertungsdimensionen ist k = 2 (Paare) und n = 17 (Dimensionen). Die Anzahl der Paare zur Ausrechnung des Korrelationskoeffizienten ist somit 136. Für jedes der 136 Dimensionspaare wird der Korrelationskoeffizient mit Hilfe von der Funktion „PEARSON(Daten1;Daten2)“ von dem Programm „LibreOffice“ ausgerechnet. Die Werte werden in einem Programm in der entsprechenden Datenstruktur gespeichert (s. DVD/Tools/../effect.pl).

5.3.2.3 Kombinationen der aufgestellten Kriterien

In den folgenden Abschnitten werden die Kombinationen der im Kapitel 4, Abschnitt 4.3 aufgestellten Kriterien pro Effekt beschrieben. Gleichzeitig werden die festgesetzten Scores pro Kriterium erläutert. Für jeden Effekt wird der Mindestscore bestimmt, der zur dessen Identifikation in einer Bewertung zu erreichen wäre.

5.3.2.3.1 Halo-Effekt

Laut den im Kapitel 4, Abschnitt 4.3.1 für Halo-Effekte aufgestellten Kriterien ist dieser Effekt nach vier Indikatoren automatisch zu erkennen. Das erste Kriterium ‚Anker- und Anpassungsdimensionen‘ wird insofern bestimmt, dass zunächst die mit dem im Abschnitt 5.3.1.1.2 beschriebenen Perl-Skript die annotierte Leitdimension festgehalten wird. Nach der Feststellung bezüglich der Leitdimension für Überbewertungen im Kapitel 4, Abschnitt 4.3.2.1 (Seite 121), dass diese zu Hauptdimensionen zählen sollte, ist nicht schwer zu begreifen, dass dies auch der Fall bei Halo-Effekten ist. Allein aus der Definition geht hervor, dass der Ankerdimension eine größere Bedeutung zugeschrieben werden sollte. Mit anderen Worten: Gehört die Leitdimension zu den Hauptdimensionen, wird der Score für dieses Kriterium höher (2 Punkte für Leitdimension als Hauptdimension, 1 Punkt für Leitdimension als Nebendimension). Die Anpassungsdimensionen, die in diesem Sinn zu Nebendimensionen gehören, werden nach ihrer Polarität in zwei Gruppen (Hashes) sortiert. Sowohl für Leitdimension als auch für Nebendimensionen werden die Polaritäten überprüft. Zum einen sollen Anker- und potentielle Anpassungsdimensionen gleiche Polaritäten aufweisen. Zum anderen werden die

Polaritäten bei allen Dimensionen auf ihre Konsistenz im Sinne von textuellen und nummerischen Werten kontrolliert. Schließlich wird überprüft, ob die Nebendimensionen zu den Ausreißern gehören. Diejenigen, die gleichzeitig gleiche Polarität wie die Leitdimension haben, konsistent sind und zu Ausreißern mit derselben Polarität gehören, werden entsprechend hohe Scores erhalten. Werden wertende Muster im Text und gleichzeitig nummerische Bewertung zu Leitdimension gefunden, so werden 2 Punkte vergeben. Jede der 12 Neben- bzw. Anpassungsdimensionen erhält auf gleiche Weise 2 Punkte sowie für ihre Zugehörigkeit zu den entsprechenden Ausreißern 1 Punkt. Bei Nebendimensionen wird ihre Zugehörigkeit zu Ausreißern vor der Konsistenzprüfung kontrolliert, wodurch der zufälligen Polaritätsübereinstimmungen der Anker- und Anpassungsdimensionen vorgebeugt wird. Nach den beschriebenen Berechnungen wird nun die paarweise Korrelation der Anker- und Anpassungsdimensionen aus dem im Voraus angelegten Hash (s. Seite 173) entnommen und die Stärke der Korrelation überprüft. Wie im Kapitel 4, Abschnitt 4.3.1.2 erläutert, sollte man eher nach einer schwächeren Korrelation der Dimensionen suchen, weil diese dann definitionsbedingt weniger miteinander verbunden sind. Sollten diese wenig miteinander korrelierten Dimensionen eine Abhängigkeit aufgrund der oben beschriebenen Kriterien aufweisen, so scheinen sie ‚verdächtiger' zu sein als diejenigen Dimensionen, die stark miteinander korrelieren. Als Orientierungsscore wurde 0,7 für die hohe Korrelation empirisch ermittelt, d. h. die niedrig korrelierten Dimensionspaare weisen den Score auf, der niedriger als 0,7 ist ($< 0{,}7$). Diesem Kriterium können maximal 12 Punkte vergeben werden (entsprechend der Anzahl von Nebendimensionen (s. o.)). Beim letzten Kriterium ‚Gleichheit nummerischer Bewertungen' wird kontrolliert, ob alle Dimensionen einer Arztpraxisbewertung gleiche nummerische Werte erhalten. Dabei wird angenommen, dass ein Bewertender keine Differenzierung von Leistungsaspekten vorgenommen hat und – möglicherweise von Leitdimension beeinflusst – alles entweder gut oder schlecht bewertete. Diesem Kriterium werden 2 Punkte vergeben. Die für jedes Kriterium vergebenen Scores werden aufsummiert, was ein Endscore für Halo-Effekt (für Positivität oder Negativität) ergibt. Dieser maximale Score setzt sich wie folgt zusammen: 1 (Leitdimension) + 1 (Leitdimension ist Hauptdimension) + 2 (Konsistenz der Leitdimension in Polarität) + 12 (Nebendimension gehört zu Ausreißern) + 24 (Konsistenz der Nebendimensionen in Polarität und dieselbe Polarität wie bei Leitdimension) + 12 (Korrelationskoeffizient) + 2 (gleiche nummerische Werte) = 54. Selbstverständlich wäre es unmöglich, den maximalen Score von 54 Punkten

zu erreichen, denn dies würde heißen, dass z. B. alle Nebendimensionen in einer Bewertung gleichzeitig zu Ausreißern zählen, für alle von ihnen wertende Muster gefunden wurden, die auch mit nummerischen Bewertungen konsistent sind, und alle Dimensionen auch eine niedrige Korrelation mit der Leitdimension aufweisen. Das Auftreten einer solchen Situation wäre ziemlich unwahrscheinlich, da allein die Anzahl von Ausreißern im Trainingskorpus in 99,67% Fällen < 12 ist. Daher werden die Scores für Kriterien, bei denen man 12 Punkte für Nebendimensionen erreichen kann, halbiert, was impliziert, dass nun der angenommene maximal zu erreichende Score für Halo-Effekte 54 - 6 x 4 = 54 - 24 = 30 wäre. Bei der Annahme, dass mindestens eine Bewertungsdimension eine Anpassungsdimension sein muss, muss der Mindestscore für Halo-Effekte mehr als 4 Punkte ($> 13{,}333...\%$) sein. Empirisch wurde der Score von mindestens 6 Punkten gesetzt (20%).

5.3.2.3.2 Überbewertung

Ähnlich dem Halo-Effekt, wird bei Überbewertungen das Vorhandensein der Anker- und Anpassungsdimensionen festgestellt. Die Vorgehensweise ist dabei allerdings etwas anders: Zunächst werden die Hauptdimensionen einer Bewertung auf die positive Polarität nummerischer und textueller Bewertungen (2 Punkte) überprüft. Sollte solch eine Hauptdimension eine Leitdimension sein, so wird zusätzlich noch 1 Punkt vergeben. Während die Hauptdimensionen auf ihre Konsistenz kontrolliert werden, werden Nebendimensionen auf ihre Inkonsistenz überprüft: Bei den nummerischen Werten muss laut der Definition im Kapitel 2, Abschnitt 2.1.3.3 eine positive und bei den Äußerungen eine negative Polarität vorhanden sein (3 Punkte). Wenn eine solche Nebendimension außerdem zu den Ausreißern mit positiver Polarität gezählt werden kann, so werden dafür 2 Punkte vergeben. Somit kann man bei Überbewertungen maximal 8 Punkte erreichen, wobei mindestens 4 Punkte (≥ 4) nötig sind, um diesen Effekt als einen solchen identifizieren zu können. Bei dem Erreichen von genau 4 Punkten könnte z. B. eine Hauptdimension mit positiver und konsistenter Polarität sowie ein positiver Ausreißer unter den Nebendimensionen gefunden werden.

5.3.2.3.3 Bestätigungsfehler und Diskriminierung

Die automatische Identifikation von empirisch ermittelten Effekten wie Bestätigungsfehlern und Diskriminierungen erfolgt nach dem gleichen Muster. Wie

im Kapitel 4, Abschnitten 4.3.3 und 4.3.4 beschrieben, sind bei diesen Effekten jeweils zwei Kriterien zu identifizieren: ‚linguistische Muster' und ‚Ausreißer'. Auf die Extraktion linguistischer Muster zu diesen Effekten wurde im Abschnitt 5.2.2.3 eingegangen. Die Aufgabe des Programms in diesem Sinne ist es, die annotierten Muster mit regulären Ausdrücken zu identifizieren. Bei der Auffindbarkeit solcher Muster pro Bewertung werden für dieses Kriterium jeweils 40 Punkte vergeben. Allerdings sind linguistische Muster nur in Kombination mit Ausreißern für beide Effekte interessant. Durch das quantifizierbare Kriterium ‚Ausreißer' (s. Seite 119) wird überprüft, inwiefern die diskriminierenden Äußerungen bzw. einseitige Suche nach Bestätigungen eigener Hypothesen die Bewertungen beeinflussen bzw. verzerren. Für jeden identifizierten Ausreißer werden 3,5 Punkte vergeben, so dass man insgesamt maximal 99,5 Punkte[83] erreichen kann, was dann in Prozente umgerechnet wird. Logisch ist, dass man von einem Bestätigungsfehler oder einer Diskriminierung spricht, wenn man neben identifizierten linguistischen Mustern mindestens einen Ausreißer in einer Bewertung nachweisen kann, was prozentuell einem Score von ca. 43,7% entspricht ($\geq$ 43,7%).

5.4 Zwischenfazit

Für die Übersichtlichkeit des beschriebenen Verfahrens werden in diesem Abschnitt die wichtigen Komponenten des CognIEffect zusammengefasst.

5.4.1 Allgemein

Es wurden zwei Schritte der Identifikation kognitiver Effekte ausführlich erläutert: **Extraktion wertender Aussagen** und **Identifikation und Klassifikation der Effekte**. Dabei stellt der zuerst genannte Schritt ein Teil des zweiten Schrittes dar. Durch den Aufwand des Extraktionsprozesses wurde dieser Schritt jedoch gesondert aufgeführt, wobei er in zwei Abschnitten (5.1 und 5.2) des aktuellen Kapitels beschrieben wurde. Die Komponenten des Identifikationsprozesses wurden im Abschnitt 5.3 erläutert.

[83] 40 (linguistische Muster) + 17 (Dimensionen) x 3,5 (pro Ausreißer) = 99,5

5.4.2 Lexikoneinträge

Im Abschnitt 5.1 wurden Ressourcen beschrieben, die sich aus den Korpora und aus den Lexikoneinträgen zusammensetzen. Die Korpora wurden in Trainings- und Testkorpora sowie zu Extraktions- und Identifikationszwecken thematisch aufgeteilt. Was lexikalische Ressourcen betrifft, die zur Erstellung von Lexikoneinträgen notwendig waren, so wurden diese teils aus den o. g. Korpora gewonnen, teils aus zwei externen Quellen (CISLEX und SentiWS) übernommen.

5.4.2.1 CISLEX_SENTIWS

Aus den beiden o. g. Wörterbüchern wurde eine Quelle angefertigt (CISLEX_SENTIWS), die als Ausgangspunkt zur korpusbasierten Adjektiv-Akquise mittels Bootstrapping-Methode verwendet wurde. Anstatt einer üblicherweise kleinen Seed-Liste (s. Kapitel 3, Abschnitt 3.1.2.1.3) wurde von einer großen Menge der aus beiden Lexika bekannten Adjektiven ausgegangen in der Hoffnung, bessere Ergebnisse zu erzielen. Allerdings fand man dadurch lediglich eine geringe Menge der vorher unbekannten Adjektive (syntaktisch-semantische Ebene). Auf der genannten Ebene wurden Adjektive auf der Basis unterschiedlicher Relationen wie bestimmte Konjunktionen und Aufzählungen akquiriert. Die Verwendung der „guessing rules“ bei den Aufzählungen führte zu zahlreichen Übergeneralisierungen. Produktiver im Sinne von der Anzahl gewonnener Wörter (1671 Adjektive) war die Akquise auf der morpho-syntaktischer Ebene mit verneinenden Präfixen.
Außer den auf die beschriebene Weise gewonnenen Adjektiven wurden fachspezifische Wörter sowie Angaben zu Nationalitäten (für Diskriminierungen) in das CISLEX_SENTIWS eingetragen. Die Fachwörter wurden aus den entsprechend erstellten Subkorpora zu Fachärzten (s. Abschnitt 5.1.1.2.1, Seite 133f.) semiautomatisch akquiriert. Die nationalitätsspezifischen Wörter wurden dem im Abschnitt 5.1.2.2.1 (Seite 145) beschriebenen Verzeichnis entnommen.

5.4.2.2 Eigene Lexika

Außer der Erweiterung des externen Wörterbuchs wurden zwei eigene Lexika mit Ressourcen angefertigt, die aus den Jameda- und DocInsider-Korpora (s. Abschnitt 5.1.1.1.1) gewonnen wurden. Eines davon besteht aus Phrasen, die

wertende Aussagen zu gleichzeitig mehreren Dimensionen beinhalten. Das andere Wörterbuch enthält die Namen der Ärzte.

5.4.3 Graphen

Die entwickelten lokalen Grammatiken wurden in dieser Arbeit zu mehreren Zwecken aufgebaut. Die Mastergraphen fassen wertende Phrasen zusammen, während funktionale Graphen weitere Vor- und Nacharbeiten erledigen. Um wertende Muster zu extrahieren, werden beide Graphentypen in entsprechenden Kaskaden zusammengefasst und auf die Bewertungstexte angewandt.

5.4.3.1 Bewertungsobjekte und -phrasen

Um die Bewertungen zu Objekten mit entsprechender Polarität zu erhalten, wurden lokale Grammatiken mit *UNITEX* entwickelt. Zu Bewertungsobjekten zählen in erster Linie die Dimensionen. Weiterhin gehören zu Bewertungsobjekten die Fachausdrücke, durch die die Dimension „Behandlung“ charakterisiert wird. Die Graphen wurden nach dem Prinzip ‚von Bewertungsobjekten zu kompletten Phrasen‘ aufgebaut. Bei wertenden Adjektiven in den Phrasen wurde festgestellt, dass diese domänenabhängig sind und eigene Dimensions- und Polaritätsspezifika aufweisen, was teilweise durch entsprechende Kodierungen im Lexikon gelöst wurde. Die Phrasen wurden auf verschiedene Weise aufgebaut, je nach Spezifik der Ausrücke zu Dimensionen und weiteren Bewertungsobjekten.

5.4.3.2 Funktionale Graphen

Lokale Grammatiken, die Vor- und Nacharbeiten zur eigentlichen Musterextraktion leisten, sind funktionale Graphen. Diese wurden bereits bei der Akquise lexikalischer Ressourcen eingesetzt. Weitere solche Grammatiken wurden für folgende Funktionen entwickelt:

- Aufbau und Extraktion wertender Phrasen zu gleichzeitig mehreren Dimensionen aus dem Phrasenlexikon PHRASE_LEX
- Erweiterung der Kontexte bereits extrahierter Pattern
- Extraktion der Bewertungsobjekte pro Bewertung zur Bestimmung der Leitdimension

5.4.4 Aufstellung der Identifikationskriterien

Wie bereits im Abschnitt 5.4.1 beschrieben, gehören die Akquise der Ressourcen sowie die Entwicklung des größten Teils von Grammatiken zum ersten Schritt des Identifikationsverfahrens CognIEffect. Gleichzeitig bilden die extrahierten wertenden Muster diesen Schrittes eines der Identifikationskriterien für kognitive Effekte. Weitere Kriterien wurden mit unterschiedlichen Methoden und Mitteln aufgestellt:

- Ausreißer: automatisch mit Perl-Programm
- Korrelation der Dimensionen: semiautomatisch mit Funktionen des „LibreOffice“-Programms
- Bestimmung der Leitdimension: automatisch mittels der Kombination von lokalen Grammatiken und dem Perl-Programm
- (In)Konsistenzen in textuellen und nummerischen Bewertungen: automatisch mit Perl-Programm

Durch entsprechende Kombinationen aller aufgestellten Kriterien sowie die Vergabe der Scores pro Effekt soll die Identifikation und eindeutige Klassifikation kognitiver Effekte gewährleistet werden.

5.4.5 Innovation

Das Innovative bei der beschriebenen Vorgehensweise ist:

- Die Findung und Ausarbeitung der Definition von Effekten innerhalb der Domäne der Arztbewertungen anhand domänenspezifischer Merkmale
- Aufstellung und entsprechende Kombination der Kriterien pro Effekt (laut der Definition), so dass durch die Scores-Vergabe die eindeutige Klassifikation ermöglicht wird
- Die Auswahl der Methoden zur Identifikation aufgestellter Kriterien, wobei die multifunktionale Anwendung von lokalen Grammatiken zum Erreichen unterschiedlicher Ziele besonders hervorzuheben ist.

Kapitel 6

Evaluation

Die Evaluation bezieht sich auf die in der vorliegenden Arbeit durchgeführten Analysen. Da diese recht unterschiedlich sind, wird wie folgt vorgegangen: Es werden die Qualitätsmaße erläutert und im Rahmen der Auswertung der Textanalysen interpretiert. Die erzielten Ergebnisse werden betrachtet.

6.1 Qualitätsmaße

Für die Qualitätsanalyse beider vorgestellten Schritte des Verfahrens CognIEffect werden die Precision- und Recall-Werte verwendet, die seit den im Kapitel 2, Abschnitt 2.2.1.2.2 (Seite 43) erwähnten MUC-Konferenzen als Standardmaße zur Qualitätsbeurteilung von IE-Systemen dienen (Grishman und Sundheim, 1996, in: Geierhos (2010, S. 210)). Zusätzlich werden die F_1 und $F_{1,5}$-Scores aus beiden o. g. Maßen ausgerechnet, um balancierte Evaluationsergebnisse (Manning et al., 2009, S. 156) darzustellen.

6.1.1 Precision

Die Genauigkeit der Ergebnisse wird entsprechend dem Precision-Wert (P) beurteilt (Geierhos, 2010, S. 211). Im Information Retrieval-Kontext bedeutet das konkret die Menge relevanter im Verhältnis zu allen gefundenen Dokumenten (Manning et al., 2009, S. 155). Aus diesem Gedanken geht logisch hervor, dass die Menge aller Dokumente in vier Teilmengen klassifiziert werden kann (Manning et al., 2009, S. 155; Geierhos, 2010, S. 211) (s. Kontingenztabelle unten), wobei

- true positives (TP) richtig gefundene und relevante Dokumente
- false positives (FP) falsch gefundene und nicht relevante Dokumente
- false negatives (FN) nicht gefundene und relevante Dokumente
- true negatives (TN) nicht gefundene und nicht relevante Dokumente

	Relevant	**Nicht relevant**
Gefunden	true positives (TP)	false positives (FP)
Nicht gefunden	false negatives (FN)	true negatives (TN)

Tabelle 6.1: Qualitätsmaß Precision (nach Geierhos (2010, S. 212))

Die Bedeutung von P kann man in eine Frage umformulieren: „Wie hoch ist der Anteil gefundener Dokumente, die relevant sind?“ und wie in der Formel 6.1 ausrechnen.

$$P = \frac{TP}{TP + FP} \tag{6.1}$$

6.1.2 Recall

Durch den Recall-Wert (R) wird beschrieben, inwiefern erzielte Ergebnisse vollständig sind (Geierhos, 2010, S.212), was im Kontext der Dokumentensuche in der Frage „Wie hoch ist der Anteil relevanter Dokumente, die gefunden sind?“ formuliert werden kann.

	Relevant	**Nicht relevant**
Gefunden	true positives (TP)	false positives (FP)
Nicht gefunden	false negatives (FN)	true negatives (TN)

Tabelle 6.2: Qualitätsmaß Recall (nach Geierhos (2010, S. 212))

Hier ist die Menge von FN-Dokumenten interessant, die zusammen mit TP-Dokumenten die Anzahl aller relevanter Treffer bilden. Daraus ergibt sich die Formel 6.2 für R:

$$R = \frac{TP}{TP + FN} \tag{6.2}$$

6.1.3 F-Score

Das Evaluationsmaß, das Precision und Recall kombiniert, wobei eine gezielte Gewichtung eines der beiden Maße ermöglicht wird, ist F-Score (Geierhos, 2010, S. 213). Durch die Wahl des Parameters α (s. Formel 6.3) kann nach individueller Zielsetzung die Gewichtungsentscheidung getroffen werden (ebd.).

$$F_\alpha = \frac{(1+\alpha) \times (P \times R)}{(\alpha \times P + R)} \tag{6.3}$$

Wird weder der Precision- noch der Recall-Wert gewichtet, wird der Parameter α auf 1 gesetzt (Geierhos, 2010, S. 213). Somit wird F_1-Score oder F_1-Maß (ebd.) nach der Formel 6.4 ausgerechnet.

$$F_1 = \frac{2 \times (P \times R)}{(P + R)} \tag{6.4}$$

Wolfgruber (2015, S. 118) und Geierhos (2010, S. 213) verzichten in ihren Arbeiten auf die Ausrechnung des F_1-Score mit der Begründung eines besseren Überblicks zur Abdeckungsrate.
In der vorliegenden Arbeit werden zwei Evaluationen durchgeführt. Von der Qualität der Pattern-Extraktion (eines der Identifikationskriterien) hängen die Ergebnisse der Identifikation von kognitiven Effekten ab. Bei der Evaluierung des ersten Schrittes spielt der Recall eine wesentlichere Rolle als Precision und wird daher ‚aufgewertet'.

6.2 Inter-Annotator-Agreement

In diesem Abschnitt werden Evaluationsmaße im Kontext der Übereinstimmung zwischen Annotatoren (Inter-Annotator-Agreement) erläutert, neu formuliert und interpretiert. Daraufhin folgen die entsprechenden Berechnungen der Ergebnisse.

6.2.1 Qualitätsmaße

Um qualitative Daten auszuwerten, verwendet man am häufigsten den Cohens-Kappa-Koeffizienten, da dieser die zufälligen Übereinstimmungen zwischen Annotatoren berücksichtigt (Hammann et al., 2014, S. 1).

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{6.5}$$

Die Bestimmung des Cohens-Kappa-Koeffizienten erfolgt nach der Formel 6.5, wobei

- p_0 = Anteil tatsächlich beobachteter Übereinstimmungen (der prozentuale Anteil der Übereinstimmungen an der Gesamtanzahl der Kodierungen (Annotationen))
- p_e = Anteil zufälliger Übereinstimmungen

sind.
Die Qualität der Ergebnisse kann dabei folgendermaßen interpretiert werden: „[...] 0.0 to 0.2 indicating slight agreement, 0.21 to 0.40 indicating fair agreement, 0.41 to 0.60 indicating moderate agreement, 0.61 to 0.80 indicating substantial agreement, and 0.81 to 1.0 indicating almost perfect or perfect agreement“ (Hallgren, 2012, S. 6).
Oft wird allerdings der Grad der Übereinstimmung zwischen den Annotatoren anhand den aus dem Abschnitt 6.1 bereits bekannten Maßen Precision, Recall und F-Score gemessen (Kaiser, 2012, S. 21). „Zur Ermittlung der Übereinstimmung der beiden Menschen, wird die Annotation eines Menschen als Goldstandard festgelegt. Precision gibt an, wie viele der anderen Annotationen korrekt sind. Recall zeigt auf, wie viele der Goldstandardannotationen gefunden werden“ (ebd.).

6.2.2 Interpretation der Qualitätsmaße

Die Auswertung der stattgefundenen Objektivierung der Entscheidungen (s. Kapitel 4, Abschnitt 4.2.3, Seite 116ff.) mittels des oben erläuterten Cohens-Kappa-Koeffizienten scheitert leider in beiden in dieser Arbeit beschriebenen Fällen (Pattern-Definitionen und Bestimmung der Leitdimension).
Im Fall der Pattern-Zuordnung geht es um polyhierarchische Klassenzuordnung, was impliziert, dass die Annotatoren die Patientenausdrücke nicht nur einer, sondern mehreren Dimensionen zuordnen durften. Obwohl zur Ausrechnung des Cohens-Kappa-Koeffizienten die Arbeit von mehr als zwei Annotatoren zulässig ist, gibt es leider keine Möglichkeit, die Muster mehreren Klassen zuzuordnen, was dieses Maß im genannten Kontext unbrauchbar macht.

Im zweiten Fall – bei der Bestimmung der Leitdimension – scheitert der Cohens-Kappa-Koeffizient an der Auswertung der Datenverteilung. „Für eine belastbare Einschätzung der Beurteiler-Übereinstimmung sind daher Daten günstiger, bei denen die Ratings sich eher gleichmäßig auf die verschiedenen Kategorien verteilen (geringe zufällige Übereinstimmung p_e)“ (Hammann et al., 2014, S. 4).

6.2.2.1 Pattern-Definitionen

Die im Abschnitt 5.3.1.1 (Seite 167) beschriebene Zuordnung der Pattern von zwei Annotatoren wird hier entsprechend den bereits bekannten Evaluationsmaßen interpretiert. Die vor dem Annotatoreneinsatz den Pattern vergebenen Tags mit Angaben zu entsprechenden Dimension und Polarität wurden als ‚Goldstandard‘ (‚G.Std.‘) verwendet. Die Zuordnungen beider Annotatoren wurden mit dem ‚G.Std.‘ verglichen, wodurch sich folgende Werte bestimmen ließen:

6.2.2.1.1 Precision

$$\mathbf{P} = \frac{x}{x+z} \begin{cases} \mathbf{x} = \textit{Übereinstimmung der Annotatoren und ‚G.Std.‘} \\ \mathbf{z} = \textit{keine Übereinstimmung der Annotatoren und ‚G.Std.‘} \\ \mathbf{x} + \mathbf{z} = \textit{Summe aller zugeordneten Pattern} \end{cases} \tag{6.6}$$

Als ‚x‘ sind hier diejenigen Pattern-Zuordnungen der Annotatoren definiert, die den Zuordnungen des ‚G.Std.‘ entsprechen (true positives). Die als ‚z‘ markierten Pattern sind dementsprechend diejenigen, die Annotatoren anders definiert haben als dies ursprünglich (vor dem Löschen der Tags) der Fall war (false positives).

6.2.2.1.2 Recall

$$\mathbf{P} = \frac{x}{x+y} \begin{cases} \mathbf{x} = \textit{Übereinstimmung der Annotatoren und ‚G.Std.‘} \\ \mathbf{y} = \textit{keine Zuordnung der Annotatoren im Vgl. z. ‚G.Std.‘} \\ \mathbf{x + y} = \textit{Summe aller zuzuordnenden Pattern} \end{cases} \tag{6.7}$$

Als ‚y‘ sind hier diejenigen Pattern-Zuordnungen der Annotatoren zu verstehen, die laut dem ‚G.Std.‘ richtig wären, von Annotatoren jedoch nicht getätigt wurden (false negatives).
Einige weiter aufgeführte Erläuterungen mit Beispielen werden den Berechnungsprozess zugeordneter Pattern verdeutlichen:

- Wenn zwei Annotatoren einen Pattern wie im ‚G.Std.‘ zugeordnet haben, so werden die Annotationen als dreimal true positives gerechnet (TP = 3)
- Wenn einer der Annotatoren einen Pattern wie im ‚G.Std.‘ zugeordnet hat und der andere demselben Pattern eine andere Bewertungsdimension zugewiesen hat, so werden die Annotationen als zweimal true positives, einmal false positives und einmal false negatives gerechnet (TP = 2; FP = 1; FN = 1)
- Wenn keiner der Annotatoren einen Pattern wie im ‚G.Std.‘ zugeordnet hat, dafür jedoch eine eigene Zuordnung vorgenommen hat, so werden die Annotationen als einmal true positives, zweimal false positives und zweimal false negatives gerechnet (TP = 1; FP = 2; FN = 2)

6.2.2.2 Bestimmung der Leitdimension

		Bewerter B		**Randsumme**
		+	−	
Bewerter A	+	305	90	395
	−	70	35	105
Randsumme		375	125	500

Tabelle 6.3: Übereinstimmung zwischen Annotatoren (Leitdimension)

In der Tabelle 6.3 sind Datenverteilungen zweier Annotatoren (Bewerter A und Bewerter B) dargestellt, wobei die Werte bei

- ‚++' – Übereinstimmung der Annotatoren, dass das Programm die Leitdimension richtig bestimmt hat (305)
- ‚––' – Übereinstimmung der Annotatoren, dass das Programm die Leitdimension falsch bestimmt hat (35)
- ‚+–' – Fälle, bei denen Bewerter A für eine richtige und Bewerter B für eine falsche Programmentscheidung ist (90)
- ‚–+' – Fälle, bei denen Bewerter B für eine richtige und Bewerter A für eine falsche Programmentscheidung ist (70)

bedeuten. Positiv auffällig an der in der genannten Tabelle dargestellten Verteilung ist, dass das entwickelte Programm in 305 von 500 Fällen die Leitdimension in den Bewertungen richtig bestimmt. Dabei sind es wiederum lediglich diejenigen Fälle, bei denen sich beide Annotatoren einig sind. Ohne diese Übereinstimmung zu berücksichtigen, würde die Qualität des Programms von beiden Annotatoren (unabhängig voneinander) noch höher eingeschätzt:

- Bewerter A = 395 von 500
- Bewerter B = 375 von 500

Was hier jedoch als positiv für die Qualität des Programms zu werten ist, wird bei der Ausrechnung der Ergebnisse mit dem Cohens-Kappa-Koeffizienten als ungleichmäßige Verteilung interpretiert, wodurch die Wahrscheinlichkeit zufälliger Übereinstimmungen (p_e) zu hoch ausfällt und das Ergebnis verzerrt und im Verhältnis zu den in der Tabelle 6.3 ausgerechneten und erläuterten Daten unbrauchbar wird (0,099).
Die in der obigen Tabelle dargestellten Ergebnisse lassen sich auf die Precision- und Recall-Werte folgendermaßen übertragen:

6.2.2.2.1 Precision

$$\mathbf{P} = \frac{x}{x+z} \begin{cases} \mathbf{x} = \textit{Beide Annotatoren} - \textit{richtiges Programmergebnis} \ (++) \\ \mathbf{z} = \textit{Beide Annotatoren} - \textit{falsches Programmergebnis} \ (--) \\ \mathbf{x} + \mathbf{z} = \textit{Summe aller bestimmten Programmergebnisse} \end{cases} \tag{6.8}$$

Als ‚x' ist hier die Ergebnisspalte mit ‚++' aus der Tabelle 6.3 (Anzahl = 305) definiert, was bedeutet, dass sich beide Annotatoren in 305 Fällen für ein richtiges Programmergebnis entschieden haben. Dementsprechend impliziert ‚z' die Ergebnisspalte mit ‚– –', wobei die Annotatoren in 35 gleichen Fällen das Programmergebnis als falsch befunden haben.

6.2.2.2.2 Recall

$$\mathbf{R} = \frac{x}{x+y} \begin{cases} \mathbf{x} = \textit{Beide Annotatoren} - \textit{richtiges Programmergebnis} \ (++) \\ \mathbf{y} = \textit{Beide Annotatoren} - \textit{keine Übereinstimmung} \ (+- \textit{und} \ -+) \\ \mathbf{x} + \mathbf{y} = \textit{Summe aller zu bestimmenden Programmergebnisse} \end{cases} \tag{6.9}$$

Als false negatives (‚y') wurden diejenigen Entscheidungen der Annotatoren definiert, die voneinander verschieden waren, wodurch das Programmergebnis als subjektiv und nicht bzw. nicht eindeutig zuordenbar zu werten ist.

6.3 Pattern-Extraktion

Im Sinne der Informationsextraktion mit einem konkreten Ziel der Patternextraktion und Stimmungsanalyse müssen die Evaluationsmaße neu formuliert werden. Die nachfolgenden Abschnitte beschreiben die Evaluationsmaße für das IE-System.

6.3.1 Precision

Geierhos (2010, S.212) spricht von Textsequenzen oder extrahierten Elementen und adaptiert die Formel 6.1, wie in 6.10 gezeigt:

$$\mathbf{P} = \frac{x}{x+z} \begin{cases} \mathbf{x} = \textit{Anzahl der korrekt extrahierten Elemente} \\ \mathbf{z} = \textit{Anzahl der falsch extrahierten Elemente} \\ \mathbf{x} + \mathbf{z} = \textit{Summe aller extrahierten Elemente} \end{cases} \tag{6.10}$$

In der vorliegenden Arbeit sind Elemente als die extrahierten Pattern zu interpretieren, deren Korrektheit (‚x') man durch drei wesentliche Eigenschaften charakterisieren kann:

- **syntaktische Vollständigkeit** eines Pattern
- **Polarität** eines Pattern
- **Dimensionszuordnung** eines Pattern

Im Wesentlichen ist ein sprachlicher Ausdruck dann korrekt, wenn er

- **semantisch** und **pragmatisch** eindeutig ist
- **syntaktisch** alle dazu gehörigen Komponenten (falls im Text vorhanden) enthält

Im Kapitel 5, Abschnitt 5.2 wurden die Begriffe der Semantik, der Syntax und der Pragmatik definiert und im Rahmen des Aufbaus lokaler Grammatiken interpretiert. Im Sinne extrahierter Pattern sind mit Semantik einer Bewertung die Bedeutung und der Inhalt einzelner Bestandteile (Wörter) eines Pattern gemeint und mit Pragmatik die Nachvollziehbarkeit der Gesamtaussage verstanden, selbst wenn der Ausdruck – syntaktisch gesehen – unvollständig ist. Im Beispiel (6.1) konnte die in die Tags gesetzte Aussage vom Extraktionssystem eindeutig nachvollzogen werden, obwohl das Satzsubjekt fehlt.

(6.1) dr. meyer {hat sich viel zeit genommen,.gzpos}

Syntaktische Vollständigkeit impliziert eine korrekte Verknüpfung sprachlicher Einheiten im Satz (s. Definition im Abschnitt 5.2.2), hier: das Vorhandensein aller bedeutungstragenden zu extrahierenden Satzkomponenten wie Subjekt, Prädikat, Objekt usw., die zur Formulierung einer Meinung herangezogen werden. Funktionswörter wie Artikel, Hilfsverben, Pronomen, Präpositionen u. ä. werden nicht dazu gezählt bzw. nicht berücksichtigt.
Wenn man die oben beschriebenen Bestandteile des Precision-Wertes prozentuell aufteilen würde, so ist zu betonen, dass für die extrahierten Pattern, die Meinungen bilden, welche für die Identifikation kognitiver Effekte notwendig sind, die Polarität und Dimensionszuordnung der Pattern wesentlich mehr Gewicht haben als die syntaktische Vollständigkeit eines Ausdrucks. Aus diesem Grund werden prozentuelle Anteile der drei oben aufgezählten Eigenschaften für die Precision-Werte wie folgt verteilt:

- syntaktische Vollständigkeit eines Pattern = 4% (0,04)

- Polarität eines Pattern = 48% (0,48)
- Dimensionszuordnung eines Pattern = 48% (0,48)

Im Beispiel (6.2) sind einige unvollständig annotierte Pattern und deren Zuordnungen den im Abschnitt 6.1 erläuterten Elementen zur Ausrechnung der Qualitätsmaße aufgeführt. Obwohl die Aussage im (a) durch den Modifikator „sehr“ eine höhere Intensität (s. Seite 23) erhält, wird sie zu 100% zu richtig gefundenen Pattern gezählt, da die Intensitätsgrade für diese Arbeit nicht differenziert werden. Das Beispiel im (b) erhält 96% als „true positive“, da hier lediglich die syntaktische Unvollständigkeit vorliegt. Die Ausdrücke (c) und (d) wurden nur ca. zur Hälfte extrahiert. Im (c) ist die bestimmte Bewertungsdimension falsch („Freundlichkeit“, nicht „Behandlung“): syntaktische Vollständigkeit (4%) + Polarität (48%) = 52%. Im (d) ist bei dem Ausdruck lediglich die Dimension richtig bestimmt: 48%. In weiteren Beispielen (e) bis (g) erfolgt keine Wertung ärztlicher Leistungen: Im (e) geht es um eine Überschrift, im (f) eine Anrede und im (g) einen Verweis auf die Leistungen anderer Personen.

(6.2) (a) sehr {geduldig,.gzpos} = 100% als true positive

(b) dr. meyer {hat sich viel zeit genommen,.gzpos} = 96% als true positive

(c) von {ganz netter Schwester,.bhpos} = 52% als true positive (gleichzeitig „Freundlichkeit“ als false negative)

(d) super {Vorgespräch,.ak} = 48% als true positive

(e) normale {Vorsorge,.bh} = nicht berücksichtigt

(f) {liebe Frau,.frpos} Jansen = false positive

(g) in Zusammenarbeit mit {hervorragenden Zahntechnikern,.bhpos +Behandlung_10} = false positive

6.3.2 Recall

Ähnlich dem Precision-Wert wird die Formel 6.2 für Recall, wie im 6.10 gezeigt, adaptiert (s. 6.11).

$$\mathbf{P} = \frac{x}{x+z} \begin{cases} \mathbf{x} = \textit{Anzahl der korrekt extrahierten Elemente} \\ \mathbf{y} = \textit{Anzahl der } fehlenden \textit{ Elemente} \\ \mathbf{x} + \mathbf{y} = \textit{Summe aller zu extrahierenden Elemente} \end{cases} \tag{6.11}$$

In diesem Abschnitt bleibt lediglich zu präzisieren, was unter den fehlenden Elementen verstanden wird. Mit Problematiken der Pattern-Definition wurde sich im Kapitel 4, Abschnitt 4.2.3.1 auseinandergesetzt. Ausgehend davon, wurde die Interpretation der individuell formulierten und am öftesten vorkommenden Meinungsausdrücke durch den Einsatz zweier Annotatoren objektiviert (s. ebd., s. Kapitel 5, Abschnitt 5.3.1.1.1). Zusätzlich wurden in mehreren Kontexten dieser Arbeit die Begriffe der Explizitheit und der Implizitheit eingeführt (s. z. B. Kapitel 2, Abschnitt 2.2.2.2.2; s. auch Seiten 37 und 44).

Die Vollständigkeit extrahierter Pattern wird in der vorliegenden Arbeit durch das automatische Erkennen der mit Hilfe von Annotatoren objektivierten Pattern bzw. jeglicher expliziter Benennungen der Bewertungsdimensionen definiert. Folgende Beipiele ((6.3)) geben Überblick und Wertungen zu nicht erkannten bzw. fehlenden Elementen. Bei den Pattern von (a) bis (d) handelt es sich aufgrund der Fehler, die in den lokalen Grammatiken bzw. beim Preprocessing hätten berücksichtigt werden können, um nicht erkannte Ausdrücke. Bei (a) betrifft die Problematik die bei der Adjektivaquise nicht berücksichtigten zusammengesetzten Adjektive, bei (b) die Zusammensetzungen von Bewertungsobjekten, die außerdem bei anderen Dimensionen behandelt wurden (s. z. B. den Graphen auf der Abbildung 5.9). Die Ausdrücke in (c) und (d) wurden aufgrund des mangelnden Preprocessing nicht erkannt, während die Phrase zur Freundlichkeit der Schwestern im (g) durch den Schreibfehler nicht zu den nicht erkannten Pattern gehört. Schreibfehler betreffen eine umfangreiche Problematik, mit der sich die vorliegende Arbeit nicht befasst. Im (e) hat man mit dem Adjektiv „modern“ zu tun, das im Kontext von „Praxis“ nicht berücksichtigt wurde. Im (f) erfolgt eine implizite Beschreibung der Dimension „Vertrauen“, da, wie im (g) gezeigt, laut der Definition der Muster, mit dem Wort (hier: Adjektiv) „ruhig“ die genannte Bewertungsdimension assoziiert wird. Dadurch hätte semantisch konsequent

das Konzept eines Wortes mit allen zugehörigen Wortarten ausgearbeitet werden können. Obwohl der sprachliche Ausdruck aus dem (f) mit dem Nomen „Ruhe“ den Annotatoren nicht vorlag, wird dieser zu ‚false negatives‘ gezählt. Schließlich und rückblickend auf das Beispiel (c), werden Ausdrücke in ähnlichen Fällen konzeptuell zu false negatives gezählt.

(6.3) (a) supernettes Team = false negatives

(b) tolle Prophylaxebehandlung = false negatives

(c) und kompetentHat sich = false negatives

(d) Sehr aufmerksamNett = false negatives

(e) die Praxis ist insgesamt modern und schön = false negatives

(f) und stets Ruhe und Optimismus vermittelt = false negatives

(g) durch die {ruhigen,.vvpos}, freudlichen Schwestern und Ärzte = nicht berücksichtigt

6.4 Effekte-Identifikation

Die Evaluation der im Kapitel 4, Abschnitt 4.3 ausgearbeiteten Kriterien mit Berücksichtigung der im Kapitel 5, Abschnitt 5.3.2.3 angenommenen Wahrscheinlichkeiten pro Effekt werden nach Precision- und Recall-Maßen wie folgt ausgerechnet:

6.4.1 Precision

Precision-Werte setzen sich aus den Kombinationen der Kriterien pro Effekt zusammen, deren Menge in Prozenten berechnet wird (z. B. vier Kriterien bei einem Effekt = 25% pro Kriterium). Wenn man z. B. die Ausführungen in 6.10 (Seite 192) nochmal betrachtet, so würde die Anzahl der korrekt extrahierten Elemente 100% (oder 1) betragen, wenn alle nötigen Kriterien

für einen jeweiligen Effekt gefunden wurden. Sollte nur ein Teil der Kriterien richtig identifiziert worden sein, so erhält der Precision-Wert die entsprechenden prozentuellen Anteile. Ein Beispiel dafür könnten nicht gefundene und dementsprechend nicht annotierte wertende sprachliche Muster oder falsch ausgerechnete Ausreißer etc. sein. Weiterhin wird auf die Kriterien der Bewertungs-, Arztpraxis- und Gesamtdatenebene eingegangen (s. Kapitel 3, Abschnitt 3.2.2.2, Seite 94).

6.4.1.1 Kriterien auf der Bewertungsebene

Was wertende Muster zu Dimensionen betrifft, so wurde dafür im Abschnitt 6.3 eine gesonderte Evaluation durchgeführt, so dass mit deren Ergebnissen die Gesamtergebnisse der Effekte-Identifikation relativiert werden, indem der Durchschnittswert aus den F-Scores[84] beider Evaluationen berechnet wird. Somit werden bei der Effekte-Evaluation nur die Bewertungen berücksichtigt, die richtig erkannte Pattern aufweisen. Genauso wird mit der Leitdimension verfahren. Was andere Kriterien auf der Bewertungsebene betrifft, so wird sich deren falsche Identifikation in den Precision-Werten widerspiegeln. Bekommt ein Effekt dadurch einen zu hohen Wahrscheinlichkeitswert und wird als solcher falsch identifiziert, so gehört dieser zu falsch extrahierten Elementen (false positives). Im Beispiel auf der Abbildung 6.1 ist der Halo-Effekt falsch identifiziert: Mit der Leitdimension „Behandlung“ und der Anpassungsdimension „Heilmethoden“, wobei die zuletzt genannte zu positiven Ausreißern zählt, wären die nötigen Kriterien für einen Halo-Effekt scheinbar erfüllt. Das Problem ist jedoch, dass die alternativen Heilmethoden in der Textbewertung bewusst positiv hervorgehoben wurden. Somit erfolgte die Bewertung der Heilmethoden nicht automatisch und auch nicht von der Ankerdimension beeinflusst: Die Situation hier ist eine andere als diejenige mit den Parkplätzen, wie sie im Kapitel 2 auf der Seite 31 beschrieben wurde.

[84] Die Differenzierung entsprechender F-Scores (s. Formeln 6.3 und 6.4, Seite 187) erfolgt im Abschnitt 6.5.2.

Abbildung 6.1: Beispiel eines falsch identifizierten Halo-Effekts (false positive) (a)

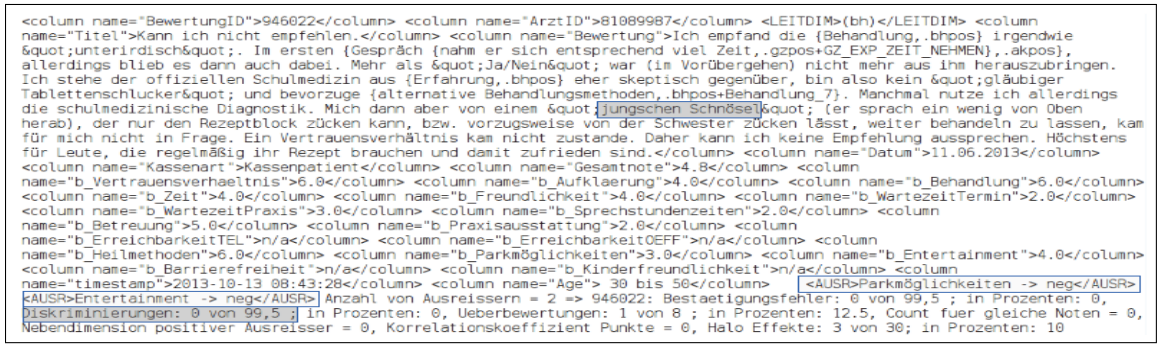

Abbildung 6.2: Beispiel einer nicht identifizierten Diskriminierung (false negative)

6.4.1.2 Kriterien auf der Arztpraxisebene

Zu den Kriterien auf der Arztpraxisebene gehören Ausreißer. Die Überprüfung deren Richtigkeit und Vollständigkeit sollte auf einem entsprechend zusammengestellten Korpus erfolgen (s. Kapitel 5, Abschnitt 5.1.1.2.2). In dieser Arbeit wurden dafür 100 zufällige Arzt-IDs mit insgesamt 540 Bewertungen (s. DVD/Evaluationsdaten) zusammengestellt und manuell kontrolliert. Für alle 100 Arztpraxen waren Ausreißer richtig berechnet, so dass dieses Ergebnis für alle Ausreißer angenommen wird.

6.4.1.3 Kriterien auf der Gesamtdatenebene

Die im Abschnitt 5, Abschnitt 5.3.2.2 beschriebene Berechnung des Korrelationskoeffizienten setzt voraus, dass dessen 100%-tige Korrektheit für die Daten in der vorliegenden Arbeit angenommen werden kann.

6.4.2 Recall

Die Vorgehensweise bei dem Recall-Wert ist weniger kompliziert. Sollte ein Effekt durch nicht identifizierte Kriterien einen zu niedrigen Wahrscheinlichkeitswert erhalten, so ist eine solche Bewertung zur Anzahl der fehlenden Elemente (false negatives) zu zählen. Auf der Abbildung 6.2 wurde eine im Text geäußerte Diskriminierung eines Arztes nicht erkannt, wodurch dieser Effekt die Wahrscheinlichkeit 0 erhielt und somit als false negative gewertet wurde. Durch zwei negative Ausreißer hätte er jedoch erkannt werden müssen.

6.5 Ergebnisse

6.5.1 Zusammenfassung

Die Ergebnisse werden in zwei übersichtlichen Tabellen zusammengefasst. In der Tabelle 6.4 sind Ergebnisse aller durchgeführten Analysen, deren Qualitätsmaße in obigen Abschnitten beschrieben wurden, aufgeführt. In der Tabelle 6.5 werden Ergebnisse für kognitive Effekte einzeln zusammengefasst. In den weiter folgenden Abschnitten wird auf die Erläuterungen der o. g. Ergebnistabellen eingegangen.

	Precision	**Recall**	F_1**-Score**	$F_{1,5}$**-Score**
Pattern-Def.	0,68	0,89	**0,75**	
Best. der Leitdim.	0,90	0,66	**0,76**	
Pattern-Extrakt.	0,79	0,81		**0,8**
Effekte-Identif.			**0,78**	

Tabelle 6.4: Ergebnisse von durchgeführten Korpusanalysen

Effekt	**Precision**	**Recall**	F_1**-Score**	F**-Score mit** $\varnothing$
Halo-Effekt	0,97	0,81	**0,88**	**0,81**
Überbewertung				
Diskriminierung	0,85	0,65	**0,74**	
Bestät.-fehler				

Tabelle 6.5: Ergebnisse von einzelnen Effekten

6.5.2 Erläuterungen

6.5.2.1 Inter-Annotator-Agreement

Die für die Annotationen definierten Ergebnisse wurden sowohl für Pattern-Definitionen als auch für die Bestimmung der Leitdimension nach der Formel 6.4 (Seite 187) für F-Score ausgerechnet. Der Parameter α wurde auf 1 gesetzt, so dass keine Gewichtung der Precision- und Recall-Werte stattfand. Der F_1-Score hat in beiden durchgeführten Analysen einen Wert von 75% und mehr der Übereinstimmungen geliefert, was insgesamt als eine gute Qualität zu werten ist (s. Seite 188).

6.5.2.2 Pattern-Extraktion

Der F-Score für die Pattern-Extraktion wurde aufgrund der Recall-Aufwertung nach der Formel 6.3 (Seite 187) ausgerechnet. Durch die Seltenheit des Auftretens von kognitiven Effekten in Meinungsäußerungen (Ausgangspunkt: die Mehrheit der Bewertungen ist fehlerfrei) und die relativ kleinen Korpora in der Domäne der Arztbewertungen wird der Vollständigkeit von Extraktionsergebnissen mehr Wert als deren Genauigkeit zugeschrieben. Durch unvollständig erkannte Pattern können immer noch die aufgestellten Kriterien für kognitive Effekte und somit auch die Effekte selbst identifiziert werden. Anders ist es bei nicht erkannten Pattern, was die Nicht-Identifikation der Effekte zur Folge hat. Der α-Parameter wurde daher auf 1,5 gesetzt, was nach einigen empirischen Experimenten als ein angemessener Wert festgelegt wurde.
Einige Fehlertypen der nicht bzw. falsch erkannten Pattern wurden im Rahmen der Beschreibungen von Qualitätsmaßen Precision (Abschnitt 6.3.1, Seite 195ff.) und Recall (Abschnitt 6.3.2, Seite 195ff.) erläutert. Im aktuellen Abschnitt werden weitere ausgewählte Fehlertypen mit Beispielen in zwei Tabellen (Tabelle 6.6 und 6.7) zusammengefasst. Bei Beispielen 1a) bis 1c) aus der zuerst genannten Tabelle handelt es sich um eine Anrede (1a)), einen Namen (1b)) und eine Danksagung (1c)), die im Sinne der Stimmungsanalyse falsch extrahiert wurden. Die Beispiele 2a) bis 2f) zeigen Adjektive, die im Kontext einiger Bewertungsobjekte nicht berücksichtigt und daher entweder nicht oder falsch erkannt wurden. Im Beispiel 2c) wurde das Adjektiv „kurz“ im Kontext von „Wartezeiten“ und in 2d) im Kontext von „Operation“ im Lexikon nicht berücksichtigt. Das Adjektiv „enorm“ wurde wiederum im Phrasenlexikon PHRASE_LEX (s. Kapitel 5, Abschnitt 5.1.2.2.2) in der in 2f) aufgeführten Phrase nicht berücksichtigt, woraufhin eine falsche Dimensionszuordnung („Wartezeit (Praxis)“) aufgrund eines übergeneralisierenden Pfads im angegebenen Graphen erfolgte. In diesem Beispiel ist dies kein Lexikonfehler, da „enorm“ eine „WZ(NEG)“-Kodierung aufweist (s. Abbildung 5.4). Die Beispiele 3a) und 3b) wurden bereits im (6.3) ((c) und (d)) aufgeführt und erläutert. Bei 4a) bis 4c) handelt es sich um negative Kontexte, wobei diese in meisten Fällen ‚zu weit‘ von dem Muster selbst entfernt sind (4b) und 4c)). Aus dem richtig extrahierten Gegenbeispiel in 4a) ist außerdem ersichtlich, dass das Kompositum „Prophylaxebehandlung“, das bereits in 2b) vorgekommen ist, als Bewertungsobjekt der Dimension „Behandlung“ richtig erkannt

Ambige Wörter und Phrasen	**Unberücksichtigte (zusammengesetzte) Adjektive**	**Kein Preprocessing**	**Negative Kontexte nicht berücksichtigt**
1a) {liebe Frau,.frpos} Jansen 1b) {kluge,.bhpos} 1c) {Herzlichen,. frpos} Dank für die ...	2a) supernettes Team 2b) tolle Prophylaxebehandlung 2c) kurze wartezeiten (Gegenbeispiel: {lange wartezeiten,.wzpneg+WZP_EXP_ATTR_OBJ}) 2d) Die {Operation war kurz,.bhneg+Behandlung_8} 2e) sehr modernwirkende Praxis 2f) {die Wartezeiten auf einen Termin sind enorm,.wzpneg+WZP_EXP_f_w}	3a) Sehr aufmerksamNett 3b) und kompetentHat sich	4a) gar nicht {so schmerzhaft,.bhneg} (Gegenbeispiel: {keine gute Prophylaxebehandlung,.bhneg +Behandlung_7}) 4b) Stellvertretende Ärzte, Schwester etc arbeiten nicht sorgfältig und {gewissenhaft,. bhpos} genug 4c) sind nicht hektisch und {unfreundlich,.frneg +Freundlichkeit_13}

Tabelle 6.6: a) Beispiele falsch / nicht erkannter Pattern der Musterextraktion

Unberücksichtigte Einschübe	**Falsch / nicht berücksichtigte Kontexte (Postprocessing)**	**Verweise auf Andere**	**Neutrale Aussagen**
5a) Dr. Schuhmacher nimmt sich für seinen Patienten Zeit 5b) Er nimmt sich jedoch jedes Mal Zeit für meine Probleme	6a) Auch nach der {OP {hab ich bei Bedarf sehr schnell einen Termin bekommen,.wztpos+WZT_EXP_termin_bekommen},.bhpos} 6b) Die {{Operation war kurz,.bhneg+Behandlung_8} und schmerzlos,.bhneg} 6c) Auf einen Termin muss man {nicht sehr lang warten,.wzppos+WZP_EXP_ATTR_OBJ}	7a) Nach einer {sehr schlechten Erfahrung,.bhneg+Beh_26} in einer anderen Praxis 7b) weil alle anderen Arzte {{ keine Erklärung,.akneg+AK_EXP_Aufkl_1a},.akneg}	8a) Nach abschluss des Gespräches {bekam ich den Termin,.wztpos+WZT_EXP_termin_bekommen} zur OP 8b) {starke Rueckenschmerzen,.bhneg+Beh_26} 8c) {Professionelle Zahnreinigung,.bhpos+Behandlung_7} 8d) Hatte eine {schwere HüftOP,.bhneg+Behandlung_2}

Tabelle 6.7: b) Beispiele falsch / nicht erkannter Pattern der Musterextraktion

wurde. Die nächste Tabelle (6.7) beginnt mit Beispielen (5a) und 5b)), bei denen Einschübe nicht berücksichtigt und dadurch die sämtlichen Muster nicht erkannt wurden. Bei der Entwicklung lokaler Grammatiken ist man bei diesem Problem mit einem Dilemma konfrontiert, bei dem eine Wahl zwischen den nicht erkannten und übergeneralisierten (und somit falsch erkannten) Mustern getroffen werden muss. Die nächsten Beispiele (6a) bis 6c)) zeigen falsch oder nicht erkannte Kontexte beim Postprocessing (s. auch Kapitel 5, Abschnitt 5.2.3.2.1). Die Gründe dafür sind unterschiedlich: beginnend mit einer neutralen Beschreibung des Zeitpunktes, an dem ein Termin vergeben wurde, in 6a) und endend mit 6c), bei dem sich das Bewertungsobjekt „Termin“ zwei Wörter entfernt vom erkannten Muster befindet. Wenn man die Beispiele 2d) und 6b) (Tabellen 6.6 und 6.7) betrachtet, so kann man verfolgen, wie falsch zugeordnete Polarität (2d)) im nächsten Schritt das Postprocessing beeinflusst hat (6b)). Die übrigen Beispiele beziehen sich auf die Muster, die in bewertenden Situationen richtig erkannt sein könnten, jedoch bezüglich der Verweise auf andere Praxen oder in allgemein beschreibenden Vorgängen als falsch erkannt gelten müssen.

6.5.2.3 Effekte-Identifikation

Die zu identifizierenden Effekte kann man in zwei Gruppen einteilen:

- Halo-Effekte und Überbewertungen
- Diskriminierungen und Bestätigungsfehler

6.5.2.3.1 Halo-Effekte und Überbewertungen

Halo-Effekte und Überbewertungen sind in den Arztbewertungen vertreten und auffindbar. Beide Effekte zeigen jedoch unterschiedliche Ergebnisse. Während man im Testkorpus (Kapitel 5, Abschnitt 5.1.1.1.2) 64 richtig identifizierte Halo-Effekte findet, sind in demselben keine Überbewertungen nach den vordefinierten Kriterien auffindbar. Unter den nicht gefundenen Überbewertungen wäre die durch die falsch zugeordnete Dimension („Wartezeit (Praxis)“ anstatt „Wartezeit (Termin)“) auf der Abbildung 6.3 zu erwähnen. Die Leit- und Ankerdimension „Behandlung“ beeinflusst in dieser Bewertung die Anpassungsdimension „Wartezeit (Termin)“, die im Text negativ beschrieben („Wartezeit für einen normalen Kontrolltermin wie bereits vorher beschrieben etwas lang“), jedoch nummerisch positiv bewertet wird. Im

gleichen Text wird die aufgeführte Aussage mit positiven Gegenbeispielen relativiert („Wenn aber ein akutes Problem besteht bekommt man auch kurzfristig einen Termin“), so dass die Dimension „Wartezeit (Termin)“ innerhalb einer Bewertung sowohl negativ als auch positiv beschrieben wird, was wiederum die automatische Identifikation einer Überbewertung erschwert. Sucht man nach Überbewertungen im Trainingskorpus, so findet man eine einzige Bewertung mit diesem Effekt, wobei die Anpassungsdimension „Öffentliche Erreichbarkeit“ mit dem vordefinierten Pattern „weiten Weg“ ausgedrückt wird. Die niedrige Präsenz von Überbewertungen macht deren Evaluation überflüssig, dementsprechend wird in der vorliegenden Arbeit darauf verzichtet. Anders verhält es sich bei Halo-Effekten. Einer der richtig identifizierten Halo-Effekte ist auf der Abbildung 6.4 zu sehen. Die Leit- und Ankerdimension „Freundlichkeit“ beeinflusst hier die Anpassungsdimension „Wartezeit (Praxis)“, die auch zu positiven Ausreißern gehört. Die Ergebnisse der Halo-Effekte-Identifikation sind in der Tabelle 6.5 zu sehen. Die Spalte ‚F-Score mit ∅‘ impliziert dabei den F-Score, der durch arithmetisches Mittel aus den Ergebnissen der Scores der Halo-Effekte (0,88), der Pattern-Extraktion (0,8) und der Bestimmung der Leitdimension (0,76) berechnet wurde (s. auch Tabelle 6.4), wie dies im Abschnitt 6.4.1.1 erläutert wurde.

6.5.2.3.2 Diskriminierungen und Bestätigungsfehler

Bei diesen zwei Effekten musste man gleichfalls feststellen, dass für einen Effekt die Evaluation nicht viel Sinn macht. Die Bestätigungsfehler sind es, bei denen vordefinierte sprachliche Muster in meisten Fällen nicht eindeutig sind, um daraus auf diesen Effekt schließen zu können. Die im Beispiel (6.4) aufgeführten Ausdrücke beinhalten zweifelhaft „Tendenzen, nach den Bestätigungen eigener Hypothesen einseitig zu suchen“ (s. Seite 39), sondern beziehen sich eher auf die Einzelfälle und Einzelereignisse, die Patienten mit konkreten Arztpraxen in Zusammenhang bringen. Dieses entspricht jedoch nicht der auf der o. g. Seite aufgestellten Definition diesen Effekts.

Abbildung 6.3: Beispiel einer nicht identifizierten Überbewertung

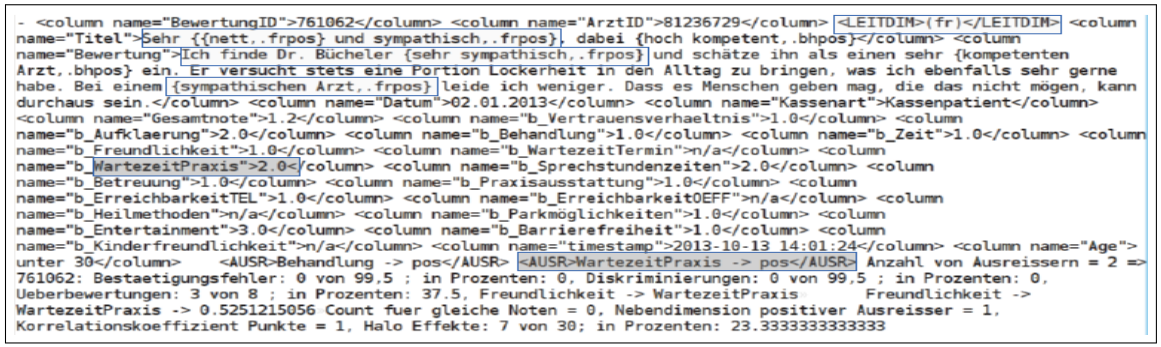

Abbildung 6.4: Beispiel eines identifizierten Halo-Effekts

(6.4) (a) Leider hat die Ärztin mir nicht richtig zugehört, war {immer schon,.BEST_FEHL} mit einem Auge auf dem Terminkalender (nächste Patientin?)

(b) aber Fr. Dr. Brender schafft es immer wieder, das ich für ein paar Tage / Wochen komplett schmerzfrei bin

Wenn man an dieser Stelle auf das Beispiel (2.1) (Kapitel 2, Seite 33) zurückblickt, so fällt auf, dass Phrasen (2.1)(b) und (2.1)(c) einen allgemeinen generalisierenden Charakter haben, wobei vermutlich nach den Bestätigungen eigener Hypothesen (aus der Erfahrung) in konkreten Praxen gesucht wird. Bei den Phrasen (2.1)(a), (2.1)(d) und (2.1)(e) sind eher Vermutungen und Erwartungshaltungen im engeren Rahmen, nur auf konkrete Praxen bezogen, beschrieben.
Was die Diskriminierungen betrifft, so sind mehrere explizite sprachliche Muster trotz der eingeführten Einschränkung auf das Alter und die Herkunft auffindbar (s. Beispiel (6.5)). Diese Muster sind mit den Ausdrücken in Beispielen (3.1) und (3.2), Seite 57 vergleichbar.

(6.5) (a) er ist selbst schon relativ alt

(b) und erst recht nicht für Kinder

(c) {junger {dynamischer Arzt,.bhpos+Behandlung_10},.DISKR}

(d) für {ältere leute,.DISKR} geeignet nicht für kinder

Die Ergebnisse der Identifikation von Diskriminierungen sind in der Tabelle 6.5 aufgeführt. Bei diesem Effekt wird kein arithmetisches Mittel aus den Ergebnissen der Pattern-Extraktion (zu Bewertungsdimensionen) und den der Bestimmung der Leitdimension berechnet, da die zwei genannten Kriterien für diesen Effekt nicht relevant sind.

Kapitel 7

Fazit und Ausblick

In diesem Kapitel wird nach einer kurzen Zusammenfassung der Dissertation betrachtet, ob und wie gestellte Ziele erreicht und die Forschungsfragen beantwortet bzw. gelöst wurden. Es werden die Grenzen des entwickelten Identifikationssystems aufgezeigt und die Perspektiven weiterer Forschung beleuchtet.

7.1 Zusammenfassung der Arbeit

Online-Rezensionen bilden einen riesigen Datenbestand und stellen eine wertvolle Informationsquelle dar. Dabei gewinnen u. a. die Bewertungen von Arztpraxen immer mehr an Bedeutung, wobei gleichzeitig diese Domäne wenig erforscht ist im Sinne computerlinguistischer Arbeiten. Dies bot den Anlass zur Aufbereitung der Korpora für die vorliegende Arbeit, die aus Bewertungen zu Arztpraxen bestehen. Als Untersuchungsgegenstand waren sozialpsychologische Phänomene – kognitive Effekte – interessant, da diese als wahrnehmungsbedingte, unbewusst erfolgte Fehler zu verstehen sind, durch die Bewertungen verzerrt werden. Eine automatische Erkennung solcher Fehler der Meinungsbildung stellt seitens Computerlinguistik eine Herausforderung dar und erlaubt z. B. das Aussortieren fehlerhafter Reviews (Ausreißer) für statistische Analysen. Computerlinguistisch interessant war zudem eine Unterscheidung bzw. Klassifikation kognitiver Effekte anhand entsprechender domänenbedingter Merkmale (Kriterien). Um das beschriebene Phänomen zu begreifen, wurde ein Exkurs in die Kognitive Psychologie durchgeführt, wobei eine Auseinandersetzung mit kognitiven Prozessen menschlicher Infor-

mationsverarbeitung erfolgte. In Arztbewertungen allerdings treten lediglich Ergebnisse dieser Prozesse auf, die in Form sprachlich ausgedrückter Meinungen sichtbar und anhand der o. g. Kriterien erkennbar werden. Solch ein Vorhaben – automatische Identifikation kognitiver Effekte in Arztbewertungen – wurde noch nie realisiert. In diesem Sinne, um Identifikationskriterien pro erkennbarer Effekt zu bestimmen, wurden zunächst Definitionen von diesen für die Domäne der Arztbewertungen adaptiert, nachdem die Kriterienaufstellung erfolgte. Um eine automatische Erkennbarkeit der Effekte zu realisieren, wurde sich mit computerlinguistischen maschinellen Lernverfahren der Informationsextraktion und Stimmungsanalyse auseinandergesetzt und eine eigene Vorgehensweise bestimmt. Informationsextraktion, wegen der Extraktion linguistischer Muster, und Stimmungsanalyse, da kognitive Effekte aufgrund der fehlerhaften Meinungsbildung auftreten, was die Meinungsextraktion impliziert. Eines der Identifikationskriterien ist gerade die Meinungsextraktion zu Bewertungsdimensionen wie „Behandlung“, „Aufklärung“, „Freundlichkeit“ etc. (17 Dimensionen) sowie die Extraktion der diskriminierenden Phrasen, die mit der Methode der lokalen Grammatiken gelöst wurden. Die Extraktion weiterer Kriterien wie Ausreißer, Korrelation der Dimensionen, Bestimmung der Leitdimension wurden teils mit Perl-Skripten, teils mit lokalen Grammatiken in Kombination mit Perl-Skripten realisiert. Die Musterextraktion mit lokalen Grammatiken, die eines der Identifikationskriterien kognitiver Effekte ausmacht, wurde durch den Aufwand als ein eigener Schritt des Identifikationsverfahrens CognIEffect angesehen und auch einzeln evaluiert. Bei der Ausarbeitung der Musterextraktion wurde versucht, eine entsprechende Qualität zu erreichen, da durch die Seltenheit des hier untersuchten Phänomens nur so dessen Identifikation möglich wurde.

7.2 Ziele und Forschungsfragen

Ziel der vorliegenden Arbeit war es, die seitens Patienten stattgefundenen kognitiven Effekte in Arztbewertungstexten zu definieren, automatisch zu identifizieren und zu klassifizieren. Dementsprechend waren drei Forschungsfragen (Existenz-, Identifikations- und Klassifikationsfrage, s. Seite 6f.) aufgestellt. Im Kapitel 2, Abschnitt 2.1.1.2.3 (Seite 21) wurde die allgemeine Definition kognitiver Effekte ausgearbeitet, nachdem die Auseinandersetzung mit der Domäne der Arztbewertungen erfolgte. Als zentraler Punkt dieser Definition kann man die Norm einer fehlerfreien Bewertung verstehen, aus-

gehend von welcher man die Kriterien für jeden identifizierbaren Effekt der fehlerbehafteten Bewertungen aufbauen konnte. Nach der Diskussion zu einer Auswahl von Effekten in Bezug auf ihre automatische Identifizierbarkeit in der gewählten Domäne wurden zwei aus der wissenschaftlichen Literatur und zwei empirisch ermittelte Effekte domänenspezifisch definiert. Ausgehend von diesen Definitionen und nach der Betrachtung maschineller Lernverfahren der Computerlinguistik vor dem Hintergrund sozialpsychologischer Experimente, wurden Identifikationskriterien pro Effekt aufgestellt. In der Praxis ließen sich die gestellten Ziele der Identifikation und der Klassifikation nicht allen gestellten Erwartungen entsprechend erreichen, was allerdings zu interessanten Erkenntnissen führte. Diese Erkenntnisse kann man mit der allgemeinen Beantwortung der aufgestellten Forschungsfragen, wie folgt, zusammenfassen:

- Zur Existenzfrage: Wie vermutet, existieren kognitive Effekte in der Domäne der Arztbewertungen, da diese immer in Bezug auf die Meinungsbildung auftreten. Als Nachweise dafür dienen zahlreiche Beispiele, begleitet mit ausführlichen Erläuterungen und Diskussionen. Da sich diese Beispiele allerdings auf die identifizierten Effekte beziehen, kann man hier nur behaupten, dass man lediglich die in Bewertungen ‚sichtbaren' Effekte nachweisen kann. Die Praxis der vorliegenden Arbeit hat außerdem gezeigt, dass
  - Effekte eine seltene Erscheinung sind und
  - sie in unterschiedlicher Menge (je nach Effekt) vertreten sind.
- Zur Identifikationsfrage: Bereits im Kapitel 2, Abschnitt 2.1.2.2 wurde festgestellt, dass nicht alle kognitiven Effekte in Arztpraxenbewertungen (maschinell) identifizierbar sind. Automatisch identifiziert werden können kognitive Effekte anhand der aufgestellten, für die gewählte Domäne charakteristischen Identifikationsmerkmale. Eine wesentliche Rolle spielen dabei die aufgestellte Definition eines jeweiligen Effekts und die Umsetzung dieser Definition in die Praxis, d. h. die Adaption / Interpretation der Definitionsbedeutung für die entsprechende Domäne, die sich in den Identifikationskriterien widerspiegelt. Die Diskussionsfrage, die hier offen bleibt, ist, ob diese Interpretation gelungen ist, mit anderen Worten: Kann man die aufgestellten Kriterien als eine genaue Entsprechung von Definitionsaspekten eines jeweiligen Effekts

betrachten? Die Ausarbeitung und das Testen der entwickelten Konzeption diesbezüglich zeigte, dass die Identifikation kognitiver Effekte anhand der vordefinierten Indikatoren möglich ist, die Qualität dieser Identifikation setzt allerdings

- eine eindeutige Interpretation der Identifikationskriterien, ausgehend von der aufgestellten Definition eines Effekts,
- eine tiefgründige praktische Umsetzung dieser Identifikationskriterien und
- eine hohe Qualität der Musterextraktion (aufgrund der Seltenheit des Phänomens und des Vorhandenseins von Kriterium „linguistische Muster“ bei allen ausgewählten Effekten)

voraus.

- Zur Klassifikationsfrage: Bei dieser Frage ist interessant – außer der Identifikationskriterien, nach welchen Effekte unterschieden / klassifiziert werden –, ob man verschiedene Effekte generell auseinander halten / differenzieren kann. Dabei ist nicht außer acht zu lassen, dass innerhalb einer Bewertung auch mehrere Effekte unabhängig voneinander auftreten können. Als Beispiel dafür könnte eine zusätzliche Eigenschaft einem Arzt zugeschrieben werden, wodurch eine / mehrere Bewertungsdimension(en) über- oder unterbewertet werden: z. B. „ein junger Arzt“, was in der vorliegenden Arbeit als eine ‚positive‘ Diskriminierung gelten muss und somit parallel zum Halo-Effekt auftreten kann. Da die Kriterien bei den zwei genannten Effekten sehr unterschiedlich sind, insbesondere bezüglich der wertenden Muster, lassen sich solche Effekte problemlos auseinander halten. Anders ist es bei den Überbewertungen in Kombination mit Halo-Effekt: man kann die Überbewertungen als ein Spezialfall der Halo-Effekte betrachten. Dabei wäre die Inkonsistenz in textuellen und nummerischen Bewertungen einer / der Anpassungsdimension / -en der Unterschied zwischen diesen zwei Effekten, der diesen Spezialfall ausmacht.
  Allgemein lässt sich aus den Ausführungen zur Klassifikationsfrage schließen, dass

  - je verschiedener die Identifikationskriterien sind, desto unterscheidbarer sind die Effekte,

– je feinkörniger die Kriterienprogrammierung erfolgt, desto bessere Klassifikationsergebnisse können erreicht werden.

7.3 Grenzen des CognIEffect

7.3.1 Allgemein

Nach der Beantwortung der gestellten Forschungsfragen und der Auseinandersetzung mit den Ergebnissen kann man verallgemeinernd zusammenfassen, dass kognitive Effekte

- ein sozialpsychologisches Phänomen sind, das bei der Meinungsbildung unbewusst, aber nicht zufällig auftritt,
- domänenabhängig durch Aufstellung der Erkennungsmerkmale (Identifikationskriterien) neu definiert werden müssen,
- je nach Möglichkeiten der Domäne identifizierbar sind,
- durch die Kombination unterschiedlicher Methoden identifiziert und klassifiziert werden können,
- zu ‚wertvollen' Ausreißern gezählt werden können (s. Kapitel 2, Abschnitt 2.1.1.2.2, Seite 20; Kapitel 3, Abschnitt 3.3.2.1, Seite 97), zumal dieses Kriterium bei allen Effekten vertreten ist.

Während sich die Existenzfrage bezüglich der Effekte allein aus den logischen Ausführungen mit „ja" beantworten ließ, gestaltete sich die Beantwortung und praktische Umsetzung der Identifikations- und Klassifikationsfragen viel umständlicher. Da es keine wissenschaftlichen Arbeiten zur automatischen Identifikation kognitiver Effekte gibt, musste sich mit anderen ähnlichen sozialpsychologischen Phänomenen und vor allem deren domänenspezifischen Kriterienaufstellung auseinandergesetzt werden. Festgestellt wurde außerdem, dass

- es keine allgemeine Klassifikation der Effekte selbst gibt,
- die gewählte Domäne Einschränkungen bezüglich der Auswahl an Kriterien aufweist. Gleichzeitig und gerade aus diesem Grund sollten mehr

Bewertungskomponenten (z. B. Gesamtnote) zur Aufstellung der Identifikationskriterien herangezogen werden (s. Kapitel 2, Abschnitt 2.1.1.1.2, Seite 15),

- automatische Textinterpretationen aufwendige Vorarbeiten implizieren, zumal die Sprachspezifik der Bewertungen die Auseinandersetzung mit weiteren zahlreichen Problematiken verlangt, z. B.
  - laienhafter Gebrauch vom medizinischen Fachvokabular
  - Rechtschreib- und Tippfehler
  - unvollständige Sätze,
- automatische Textinterpretationen anhand linguistischer Muster erfolgen, deren Zugehörigkeit zu vordefinierten Bewertungsdimensionen einen interpretativen Charakter beibehält. Die Möglichkeiten wertender Ausdrücke können nie zu 100% abgedeckt werden, da diese individuell sind. Genauso individuell ist das Empfinden verschiedener Personen in Bezug auf die Polarität der Sachverhalte: für eine Person kann die Wartezeit von 30 Minuten in einer Praxis akzeptabel sein, während eine andere Person nicht bereit ist, dieses zu tolerieren.
- etc.

Nach der durchgeführten Identifikation ausgewählter Effekte wurde klar, dass

- durch einige zu abstrakt gehaltene Annahmen die automatische Erkennung nicht aller Effekte gelungen ist. Dies impliziert nicht unbedingt, dass die Annahmen falsch, sondern z. B. dass nicht alle davon umsetzbar waren wie beispielsweise die Muster zu Bestätigungsfehlern. Die Feststellung, dass die erkannten Muster nicht bzw. nur teilweise die aufgestellte Definition der Bestätigungsfehler abbilden, impliziert eine weitere tiefere Auseinandersetzung mit den Fragen der möglichst objektiven, sprachlichen Musteraufstellung wie dies z. B. durch Annotatoreneinsatz bei den Pattern-Definitionen zu Bewertungsdimensionen gezeigt wurde. Ein weiteres Beispiel für eine solche Annahme ist die empirische Ermittlung der Überbewertungen, die eine Vertiefung in die domänenspezifische Klassifikationsproblematik kognitiver Effekte impliziert.

- die Kriterienumsetzung bei Effekten konzeptuell erweiterbar ist. Bei dem Kriterium ‚linguistische Muster', dessen automatische Identifizierung mit lokalen Grammatiken umgesetzt wurde, war man auf zahlreiche Einschränkungen aufgrund der teilweise thematisierten Problematiken angewiesen. Gleichzeitig musste aufgrund der Seltenheit des Vorkommens von Effekten eine entsprechend gute Qualität gewährleistet werden. Daher wurde im Voraus ein Score der Musterextraktion gesetzt, der erreicht werden konnte. Die relationalen Zusammenhänge zwischen den Identifikationskriterien müssen jedoch differenzierter ausgearbeitet werden, wodurch bessere Klassifikationsergebnisse erreicht werden können (z. B. der Frage nachgehen, ob Leitdimension ein Ausreißer sein darf).

- kognitive Effekte seltene Phänomene sind. Diese Tatsache ist auch bei anderen sozialpsychologischen Phänomenen wie bei rhetorischen Stilmitteln auffällig (z. B. Idiome, Kapitel 3, Abschnitt 3.2.1.2, Seite 86). Je nach der gewählten Domäne, in der man Effekte automatisch identifiziert, ist der Grad deren Vorkommens unterschiedlich hoch. Auch von der Art eines Effekts selbst ist die Häufigkeit seines Auftretens abhängig.

7.3.2 Ausgewählte Identifikationseinschränkungen

Rückblickend auf die Klassifikationsfrage ist positiv hervorzuheben, dass die Klassifikation von Effekten insofern gelungen ist, dass bei richtig identifizierten Kriterien durch die angenommenen Wahrscheinlichkeiten deren eindeutige Unterscheidung stattfindet, wobei pro Effekt gezeigt wird, zu wie viel Prozent dieser in einer Bewertung vertreten ist. Wenn man die Abbildung 6.3 im Kapitel 6 (Seite 206) nochmal betrachtet, so fällt auf, dass dadurch, dass die Dimension „Wartezeit (Termin)" sowohl positiv als auch negativ in einer und derselben Bewertung auftritt, gleichzeitig zwei Effekte (laut Kriterien) in dieser Bewertung vorhanden sind . Aus diesem Grund liegt der Gedanke nahe, dass mehrere Effekte in einer Bewertung auftreten können und dass sie sich nicht unbedingt ausschließen müssen. Im konkreten Fall der Überbewertung und Halo-Effekts ist es die Frage der Klassifikation, ob die Überbewertung als ein Sonderfall des Halo-Effekts zu betrachten wäre.
Das Vorkommen des in dieser Arbeit definierten Phänomens „Überbewertung" bleibt trotz der erzielten Ergebnisse unumstritten. Auch in den vor-

liegenden Daten und Datenmenge könnte das mehrfache Vorkommen dieses Effekts nachgewiesen werden, wenn man mehr von folgenden linguistischen Mustern wie „Wartezeit nehme ich gern in Kauf“ beim Patternmatching gezielt berücksichtigt.
Was den Halo-Effekt betrifft, so war bei diesem auffällig, dass die häufigsten Leit- oder Ankerdimensionen die Hauptdimensionen wie „Behandlung“, „Freundlichkeit“, „Vertrauen“ etc. sind und die häufigsten Anpassungsdimensionen zu den Nebendimensionen wie „Wartezeit (Praxis)“, „Wartezeit (Termin)“, „Entertainment“, „Telefonische Erreichbarkeit“ usw. gehören, was die anfänglichen konzeptuellen Überlegungen bestätigte (s. z. B. Seite 121 im Kapitel 4). Fehler bei der Identifikation des Effekts werden von dem entwickelten Programm z. B. durch die Nichtberücksichtigung der Polarität von Leitdimensionen verursacht, was heißt, dass der Zusammenhang zwischen der Anker- und Anpassungsdimensionen bezüglich der Polarität weder in Texten noch in nummerischen Werten für das Programm existiert. Das kann dazu führen, dass, wie auf der Abbildung 7.1 zu sehen ist, die ganze Bewertung negativ ist, wobei eine schlechte Behandlung beschrieben wird und die Dimension „Behandlung“ als Leitdimension fungiert, während zwei positive Ausreißer, die mit „Behandlung“ schwach korrelieren und dadurch als Anpassungsdimensionen interpretiert werden, existieren. Durch verschiedene Polaritäten zwischen Anker- und Anpassungsdimensionen stehen sie nicht in erwartetem Zusammenhang zueinander im Sinne von Halo-Effekt[85].
Wenn man Erkenntnisse zu Diskriminierungen und Bestätigungsfehlern zusammenfasst, so war die Annahme zur ausschließlich negativen Polarität bei beiden Effekten falsch (s. Seite 153). Die Tatsache z. B., dass ein Arzt jung ist, kann einen Bewertenden positiv beeinflussen. Die Suche nach Bestätigungen eigener Hypothesen kann ebenfalls positiv sein, z. B. im Fall, wenn jemand alle Ärzte prinzipiell für freundlich hält. Auch die Definitionen sprachlicher Muster sind mit gewissen Einschränkungen verbunden. Bei den Bestätigungsfehlern ist generell anzumerken, dass die Äußerungen für eine Bestätigung eigener Hypothesen zu erfassen, schwer ist, da man dies vermutlich selten in Sprache fassen würde. Diese Bestätigungen würden dann lediglich als kognitiv-emotionale Prozesse ‚im Kopf‘ existieren, was in den Texten unerkennbar bleibt. Bei den Diskriminierungen stellte sich der umstrittene

[85]Eine andere Problematik eines falsch erkannten Halo-Effekts wurde bereits im Kapitel 6 erläutert (Abbildung 6.1)

- <column name="BewertungID">729867</column> <column name="ArztID">80113655</column> <LEITDIM>(bh)</LEITDIM> <column name="Titel">Arrogant und{ Unkompetent,,bhneg}</column> <column name="Bewertung">Bei Schwangerschaft nicht zu emfehlen. Nimmt Probleme nicht ernst und gibt nicht einmal ein Ultraschallbild mit.</column> <column name="Datum">22.11.2012</column> <column name="Kassenart">n/a</column> <column name="Gesamtnote">5.4</column> <column name="b_Vertrauensverhaeltnis">6.0</column> <column name="b_Aufklaerung">6.0</column> <column name="b_Behandlung">6.0</column> <column name="b_Zeit">4.0</column> <column name="b_Freundlichkeit">5.0</column> <column name="b_WartezeitTermin">4.0</column> <column name="b_WartezeitPraxis">1.0</column> <column name="b_Sprechstundenzeiten">2.0</column> <column name="b_Betreuung">6.0</column> <column name="b_Praxisausstattung">3.0</column> <column name="b_ErreichbarkeitTEL">3.0</column> <column name="b_ErreichbarkeitOEFF">1.0</column> <column name="b_Heilmethoden">3.0</column> <column name="b_Parkmöglichkeiten">2.0</column> <column name="b_Entertainment">2.0</column> <column name="b_Barrierefreiheit">3.0</column> <column name="b_Kinderfreundlichkeit">2.0</column> <column name="timestamp">2013-10-11 18:56:03</column> <column name="Age">n/a</column> <AUSR>Entertainment -> pos</AUSR> <AUSR>Kinderfreundlichkeit -> pos</AUSR> Anzahl von Ausreissern = 2 => 729867: Bestaetigungsfehler: 0 von 99,5 ; in Prozenten: 0, Diskriminierungen: 0 von 99,5 ; in Prozenten: 0, Ueberbewertungen: 1 von 8 ; in Prozenten: 12.5, Behandlung -> Entertainment Behandlung -> Entertainment -> 0.5911976953 Behandlung -> Kinderfreundlichkeit Behandlung -> Kinderfreundlichkeit -> 0.6932851895 Count fuer gleiche Noten = 0, Nebendimension positiver Ausreisser = 2, Korrelationskoeffizient Punkte = 2, Halo Effekte: 8 von 30; in Prozenten: 26.6666666666667

Abbildung 7.1: Beispiel eines falsch identifizierten Halo-Effekts (false positive) (b)

Zusammenhang beider zu identifizierenden Kriterien durch die o. g. falsche Annahme der Polarität heraus. Z. B. es wurden ein Muster „junger kompetenter Arzt“ und gleichzeitig ein Ausreißer „Parkmöglichkeiten“ mit einer negativen Polarität in einer Bewertung gefunden. Dass diese beiden Kriterien in irgendeiner Relation zueinander stehen, ist nicht nachzuvollziehen. Auffällig war außerdem ein typisches domänenspezifisches Diskriminierungsmerkmal ‚Kassenzugehörigkeit‘, wobei sich in den meisten Fällen die gesetzlich versicherten Patienten diskriminiert vorkommen. Tendenziell ist zu bemerken, dass sich Kassenpatienten eine Gleichbehandlung wünschen, während die privat Versicherten eine Sonderbehandlung erwarten.

7.4 Perspektiven weiterer Forschung

Im Grunde genommen, ist jeder Effekt ein Ausreißer, jedoch nicht jeder Ausreißer ein Effekt! Effekte fungieren als extreme Werte, die stark von der Masse der Daten abweichen. Solche Werte werden für die statistischen Erhebungen teilweise eliminiert (s. Kapitel 3, Abschnitt 3.3.2.1). An der eben genannten Stelle lag der Gedanke jedoch nahe, die Ausreißer zu verstehen, um den richtigen Umgang mit ihnen zu bestimmen. „Eine vorschnelle Elimination kann dazu führen, dass neue Erkenntnisse nicht aufgedeckt werden, wenn die extremen Werte beispielsweise aufgrund eines bislang nicht beobachteten Verhaltens entstanden sind. Dies würde einen Verlust für die Forschung darstellen“(Goerke, 2016, S. 23). Diese Tatsache bietet den Anlass für interdisziplinäre computerlinguistisch-sozialpsychologische Forschungsprojekte, in denen z. B. die Definitionsproblematiken wertender Muster gelöst, Interpretationen domänenspezifischer Kriterien ausgearbeitet werden können etc.

Anhang A

Übersicht zu Einträgen in Lexika

A.1 CISLEX_SENTIWS

CISLEX_SENTIWS				
Einträge	**Grund-formen**	**alle Formen**	**bekannt**	**hinzugefügt**
SentiWS	x		1199	299
Fachvokabular		x	857	870
Wörter zur Nationalität[a]		x	711	214

Tabelle A.1: Einträge im CISLEX_SENTIWS

[a]Hinzugefügt wurden hauptsächlich Vollformen der Staaten und Adjektive. Einige der im Kapitel 5, Abschnitt 5.1.2.2.1 (Seite 145) beschriebenen Formen fehlen im „Verzeichnis der Staatennamen für den amtlichen Gebrauch in der Bundesrepublik Deutschland“. Außerdem wurden 47 Formen (Staaten und Einwohnerbezeichnungen) in der Ausführung nicht eingetragen, z. B. „Syrien, Arabische Republik“; „SierraLeoner“; „Laos, Demokratische Volksrepublik “ u. ä.

A.2 PHRASE_LEX

PHRASE_LEX	
Einträge	18377

Tabelle A.2: Einträge im PHRASE_LEX

A.3 ARZTNAME_LEX

ARZTNAME_LEX	
Einträge	48046

Tabelle A.3: Einträge im ARZTNAME_LEX

Anhang B

Auszüge aus den Ressourcen

B.1 Grundformen dimensionsspezifischer wertender Adjektive

1. anspruchsvoll,.ADJ+ADJPOS+PA(POS):up
2. anständig,.ADJ+ADJPOS+VV(POS):up
3. aufrichtig,.ADJ+ADJPOS+VV(POS):up
4. barrierefrei,.ADJ+ADJPOS+BF(POS):up
5. dämmlich,.ADJ+ADJNEG+BH(NEG):up
6. fahrlässig,.ADJ+ADJNEG+BH(NEG):up
7. funktionsgerecht,.ADJ+ADJPOS+PA(POS):up
8. gemütlich,.ADJ+ADJPOS+PA(POS):up
9. geschmackvoll,.ADJ+ADJPOS+PA(POS):up
10. glaubwürdig,.ADJ+ADJPOS+VV(POS):up
11. großräumig,.ADJ+ADJPOS+PA(POS):up
12. großzügig,.ADJ+ADJPOS+PA(POS)+FR(POS)+PM(POS):up
13. gutherzig,.ADJ+ADJPOS+FR(POS):up

14. hochmodern,.ADJ+ADJPOS+PA(POS):up
15. hochwertig,.ADJ+ADJPOS+PA(POS):up
16. humorvoll,.ADJ+ADJPOS+FR(POS):up
17. kinderfreundlich,.ADJ+ADJPOS+KFR(POS):up
18. kindgerecht,.ADJ+ADJPOS+PA(POS):up
19. kostenfrei,.ADJ+ADJPOS+PM(POS):up
20. kunstvoll,.ADJ+ADJPOS+PA(POS):up
21. liebenswert,.ADJ+ADJPOS+FR(POS):up
22. liebenswürdig,.ADJ+ADJPOS+VV(POS):up
23. liebevoll,.ADJ+ADJPOS+FR(POS):up
24. nachlässig,.ADJ+ADJNEG+BH(NEG):up
25. nachsichtig,.ADJ+ADJPOS+FR(POS):up
26. sanftmütig,.ADJ+ADJPOS+FR(POS):up
27. stilvoll,.ADJ+ADJPOS+PA(POS):up
28. übersichtlich,.ADJ+ADJPOS+PA(POS):up
29. unschlagbar,.ADJ+ADJPOS+WZ(NEG):up
30. unverhüllt,.ADJ+ADJPOS+VV(POS):up
31. unwissend,.ADJ+ADJNEG+BH(NEG):up
32. unzureichend,.ADJ+ADJNEG+PM(NEG):up
33. unzuverlässig,.ADJ+ADJNEG+BH(NEG):up
34. vernünftig,.ADJ+ADJPOS+BH(POS):up
35. versiert,.ADJ+ADJPOS+BH(POS):up
36. verständlich,.ADJ+ADJPOS+WZ(NEG):up

37. verständnisvoll,.ADJ+ADJPOS+VV(POS):up
38. vertrauenserweckend,.ADJ+ADJPOS+VV(POS):up
39. vertrauensvoll,.ADJ+ADJPOS+VV(POS):up
40. vertrauenswürdig,.ADJ+ADJPOS+VV(POS):up
41. wahrhaftig,.ADJ+ADJPOS+VV(POS):up
42. warm,.ADJ+ADJPOS+PA(POS):up
43. weichherzig,.ADJ+ADJPOS+FR(POS):up
44. wohnlich,.ADJ+ADJPOS+PA(POS):up
45. zeitgemäß,.ADJ+ADJPOS+PA(POS):up
46. zielstrebig,.ADJ+ADJPOS+BH(POS):up
47. zugänglich,.ADJ+ADJPOS+PA(POS)+PM(POS):up
48. zuverlässig,.ADJ+ADJPOS+BH(POS):up
49. zuvorkommend,.ADJ+ADJPOS+FR(POS):up
50. zweckmäßig,.ADJ+ADJPOS+PA(POS):up

B.2 Grundformen der Nomen zu „Behandlung“

1. anastesie,.N+FF+BHNOM
2. anschlußbehandlung,.N+FF+BHNOM
3. antibiotics,.N+FF+BHNOM
4. arztbesuch,.N+FF+BHNOM
5. behandlungsabläufe,.N+FF+BHNOM
6. behandlungsablaufes,.N+FF+BHNOM

7. behandlungsablauf,.N+FF+BHNOM
8. behandlungsablaufs,.N+FF+BHNOM
9. behandlungsalternative,.N+FF+BHNOM+AHNOM
10. behandlungsalternativen,.N+FF+BHNOM+AHNOM
11. behandlungsansatz,.N+FF+BHNOM
12. behandlungsaufwand,.N+FF+BHNOM
13. behandlungsempfehlungen,.N+FF+BHNOM
14. behandlungserfolg,.N+FF+BHNOM
15. behandlungsergebnis,.N+FF+BHNOM
16. fachkompetenz,.N+FF+BHNOM
17. präparat,.N+FF+BHNOM
18. prognose,.N+FF+BHNOM
19. provisorium,.N+FF+ZNOM
20. rezept,.N+FF+BHNOM
21. risiko,.N+FF+BHNOM
22. röntgen,.N+FF+BHNOM
23. rosacea,.N+FF+SCHNOM
24. routine,.N+FF+BHNOM
25. säure,.N+FF+SCHNOM
26. schnitt,.N+FF+SCHNOM+AUGNOM
27. schnupfen,.N+FF+BHNOM
28. screening,.N+FF+BHNOM
29. sekret,.N+FF+BHNOM

30. septum,.N+FF+HNONOM
31. silikon,.N+FF+SCHNOM
32. sinusitis,.N+FF+HNONOM
33. sport,.N+FF+BHNOM
34. spray,.N+FF+BHNOM
35. stirn,.N+FF+SCHNOM
36. symptom,.N+FF+BHNOM
37. techniker,.N+FF+ZNOM
38. teleskop,.N+FF+ZNOM
39. test,.N+FF+BHNOM
40. therapie,.N+FF+BHNOM
41. tinnitus,.N+FF+HNONOM
42. tonsillaris,.N+FF+HNONOM
43. trigeminus,.N+FF+HNONOM
44. tropfen,.N+FF+BHNOM
45. vorsorge,.N+FF+BHNOM
46. wurzel,.N+FF+ZNOM
47. zahn,.N+FF+ZNOM
48. zahnprothetiker,.N+FF+ZNOM
49. zentrum,.N+FF+BHNOM
50. zygoma,.N+FF+ZNOM

B.3 Wörter zur Nationalität

1. Afghane,.N+NAT:neM
2. Afghanin,.N+NAT:aeF:deF:geF:neF
3. afghanisch,.ADJ+ADJNAT:up
4. Afghanistan,.N+LAND(KF):aeN:deN:neN
5. Ägypten,.N+LAND(KF):aeN:deN:neN
6. ägyptisch,.ADJ+ADJNAT:up
7. Albanerin,.N+NAT:aeF:deF:geF:neF
8. Albaner,.N+NAT:aeM:amM:deM:gmM:neM:nmM
9. Albanien,.N+LAND(KF):aeN:deN:neN
10. albanisch,.ADJ+ADJNAT:up
11. Algerien,.N+LAND(KF):aeN:deN:neN
12. Algerierin,.N+NAT:aeF:deF:geF:neF
13. Algerier,.N+NAT:aeM:amM:deM:gmM:neM:nmM
14. algerisch,.ADJ+ADJNAT:up
15. Amerikanerin,.N+NAT:aeF:deF:geF:neF
16. Amerikaner,.N+NAT:aeM:amM:deM:gmM:neM:nmM
17. amerikanisch,.ADJ+ADJNAT:up
18. Andorra,.N+LAND(KF):aeN:deN:neN
19. Andorranerin,.N+NAT:aeF:deF:geF:neF
20. Andorraner,.N+NAT:aeM:amM:deM:gmM:neM:nmM
21. andorranisch,.ADJ+ADJNAT:up

22. Angola,.N+LAND(KF):aeN:deN:neN
23. Angolanerin,.N+NAT:aeF:deF:geF:neF
24. Angolaner,.N+NAT:aeM:amM:deM:gmM:neM:nmM
25. angolanisch,.ADJ+ADJNAT:up
26. Antiguanerin,.N+NAT:aeF:deF:geF:neF
27. Antiguaner,.N+NAT:aeM:amM:deM:gmM:neM:nmM
28. antiguanisch,.ADJ+ADJNAT:up
29. Äquatorialguinea,.N+LAND(KF):aeN:deN:neN
30. Argentinien,.N+LAND(KF):aeN:deN:neN
31. Argentinierin,.N+NAT:aeF:deF:geF:neF
32. Argentinier,.N+NAT:aeM:amM:deM:gmM:neM:nmM
33. argentinisch,.ADJ+ADJNAT:up
34. Armenien,.N+LAND(KF):aeN:deN:neN
35. Armenierin,.N+NAT:aeF:deF:geF:neF
36. Armenier,.N+NAT:aeM:amM:deM:gmM:neM:nmM
37. armenisch,.ADJ+ADJNAT:up
38. Aserbaidschanerin,.N+NAT:aeF:deF:geF:neF
39. Aserbaidschaner,.N+NAT:aeM:amM:deM:gmM:neM:nmM
40. aserbaidschanisch,.ADJ+ADJNAT:up
41. Aserbaidschan,.N+LAND(KF):aeN:deN:neN
42. Äthiopien,.N+LAND(KF):aeN:deN:neN
43. äthiopisch,.ADJ+ADJNAT:up
44. Australien,.N+LAND(KF)+LAND(VF):aeN:deN:neN

45. Australierin,.N+NAT:aeF:deF:geF:neF
46. Australier,.N+NAT:aeM:amM:deM:gmM:neM:nmM
47. australisch,.ADJ+ADJNAT:up
48. Bahamaerin,.N+NAT:aeF:deF:geF:neF
49. Bahamaer,.N+NAT:aeM:amM:deM:gmM:neM:nmM
50. bahamaisch,.ADJ+ADJNAT:up

B.4 Phrasen aus dem PHRASE_LEX

1. arrogant und unfreundlich behandelt,.bhneg+frneg
2. arrogant und herablassend sehr unfreundlich unsaubere warteräume und,.frneg+paneg
3. arroganten und unfreundlichen eindruck gemacht,.bhneg+frneg
4. antwortet und beraet uns immer geduldig und zufriedenstellend,.akpos+gzpos
5. antwortete er sehr geduldig,.akpos+gzpos
6. antwortet auf fragen immer kompetent und geduldig,.akpos+gzpos
7. antworten und fühl mich dort gut aufgehoben,.akpos+vvpos
8. antworten nur kurz und unfreundlich,.akneg+frneg
9. antworten fiel es mir sehr leicht vertrauen,.akpos+vvpos
10. antibiotika oder etwas sanfter mit eigenbluttherapie behandelt,.bhpos+frpos
11. anständig untersucht,.bhpos+vvpos
12. ansprache auch stets freundlich weiterhilft,.bhpos+frpos
13. ansonsten netter kompetenter arzt,.bhpos+frpos
14. ansonsten nette kompetente ärztin,.bhpos+frpos

15. anscheinend nette und kompetente ärztin,.bhpos+frpos
16. ängsten sehr gut und beruhigend helfen,.bhpos+vvpos
17. angenehm untersucht,.bhpos+vvpos
18. angenehm freundlich behandelt,.bhpos+frpos+vvpos
19. angenehmes personal freundliche ärztin,.frpos+vvpos
20. angenehme schwangerschaftsbetreuung,.btpos+vvpos
21. angenehmer und vertrauensvoller arzt,.frpos+vvpos
22. angenehmer und ruhiger arzt,.frpos+vvpos
23. angenehme ruhige ärztin,.frpos+vvpos
24. angenehmer ruhiger arzt,.frpos+vvpos
25. angenehmer offener mensch,.frpos+vvpos
26. angenehmer neurochirurg,.frpos+vvpos
27. angenehmer mensch bei dem man sich sehr gut aufgehoben fühlt,.frpos+vvpos
28. angenehmer mensch & arzt mitarbeiter,.frpos+vvpos
29. angenehmer facharzt,.frpos+vvpos
30. angenehmer arzt und auch angenehmes klima,.frpos+vvpos
31. angenehmer arzt/team,.frpos+vvpos
32. angenehmer arzt & personal,.frpos+vvpos
33. angenehme praxisräume und nettes personal,.frpos+papos
34. angenehme praxisangestellte,.frpos+vvpos
35. angenehme personal,.frpos+vvpos
36. angenehme nachbetreuung,.btpos+vvpos
37. angenehme betreuung,.btpos+vvpos

38. angenehme ärztin nettes personal,.frpos+vvpos
39. angenehm betreut,.btpos+vvpos
40. angenehm behandelt,.bhpos+vvpos
41. anfangs noch motivierten freundlichen und kompetenten arzt,.bhpos+frpos
42. anamnese verliefen sehr entspannt,.bhpos+vvpos
43. alternativmedizin sehr aufgeschlossen,.ahpos+vvpos
44. alternativmedizin gegenüber aufgeschlossen,.ahpos+vvpos
45. alternativmedizin aufgeschlossen,.ahpos+vvpos
46. alle gleich freundlich und kompetent behandelt,.bhpos+frpos
47. alle fragen ausführlich und ist auch gegenüber alternativen heilverfahren,.ahpos+akpos+bhpos
48. alle betreuen mich wirklich ganz kompetent und lieb,.bhpos+btpos+frpos
49. alle belange ausreichend zeit genommen und entsprechend aufgeklärt über alternative heilmethoden,.ahpos+akpos+bhpos+gzpos
50. alle arbeiten kompetent freundlich,.bhpos+frpos

Anhang C

Ausgewählte lokale Grammatiken

C.1 Einzelne Module: Extraktion verschiedener Wortarten zu Bewertungsdimensionen

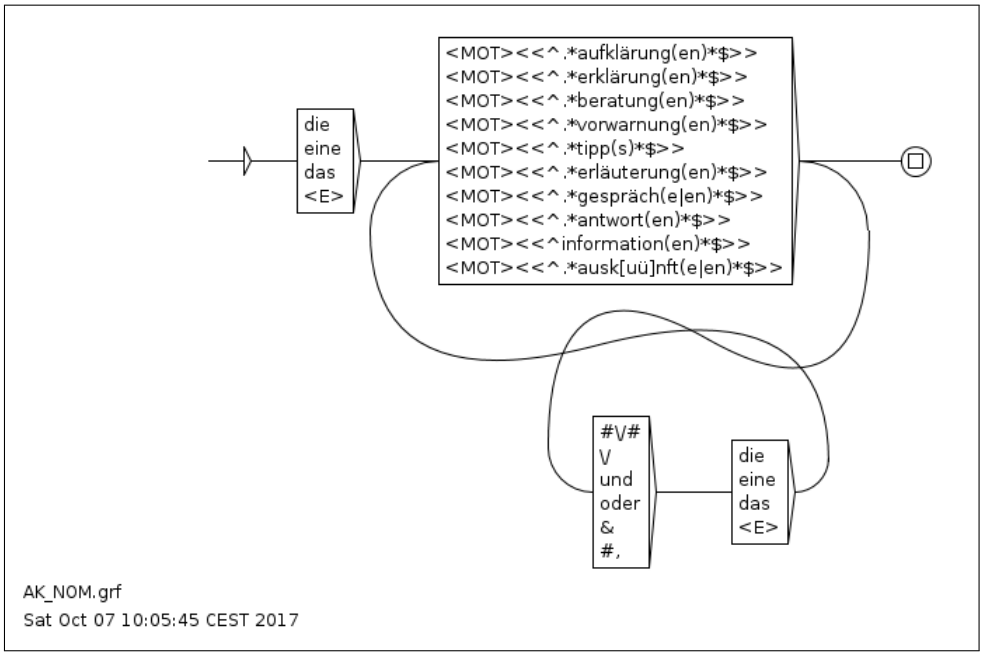

Abbildung C.1: Graph zur Erkennung der Nomen zu „Aufklärung“

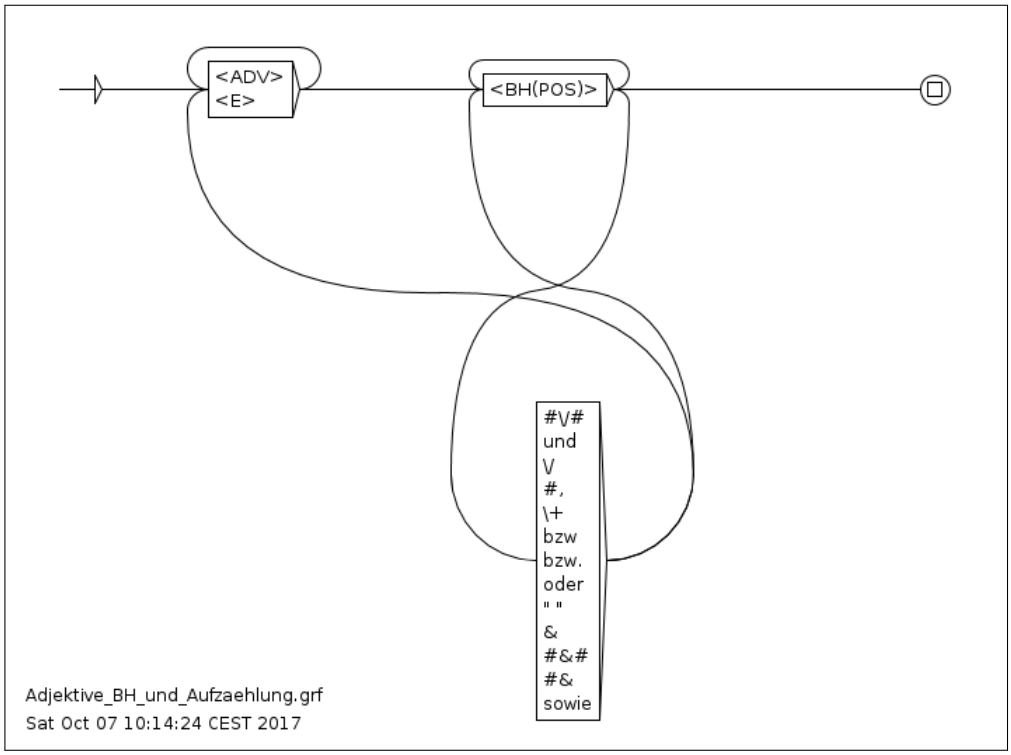

Abbildung C.2: Graph zur Erkennung der Adjektive zu „Behandlung“ mit positiver Polarität

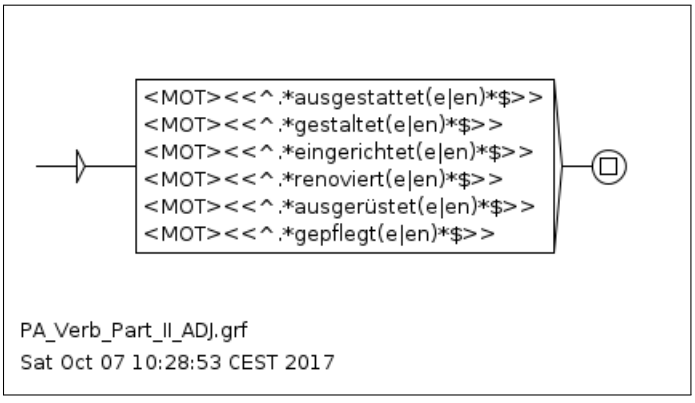

Abbildung C.3: Graph zur Erkennung der Partizipialadjektive zu „Praxisausstattung“

C.2 Phrasengraphen: Erkennung wertender Phrasen zu Bewertungsdimensionen

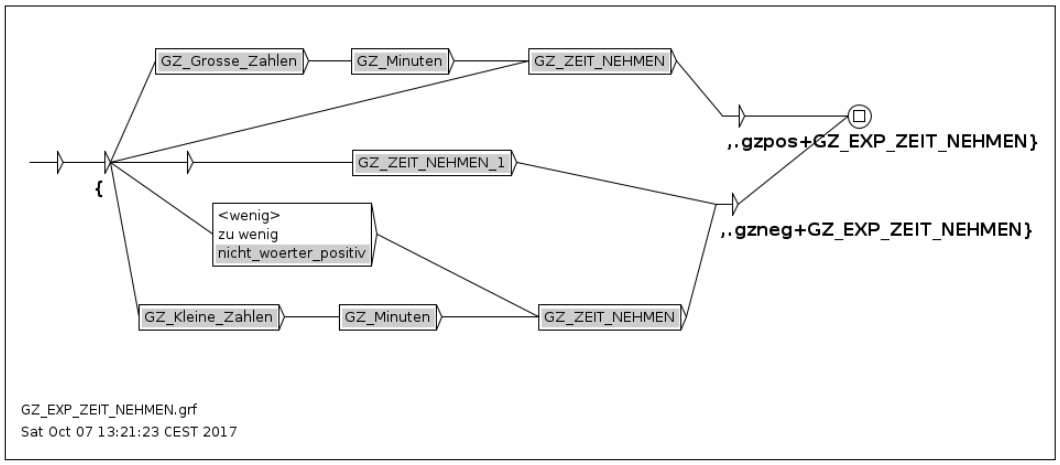

Abbildung C.4: Graph zur Erkennung der Phrasen zu „Genommene Zeit“

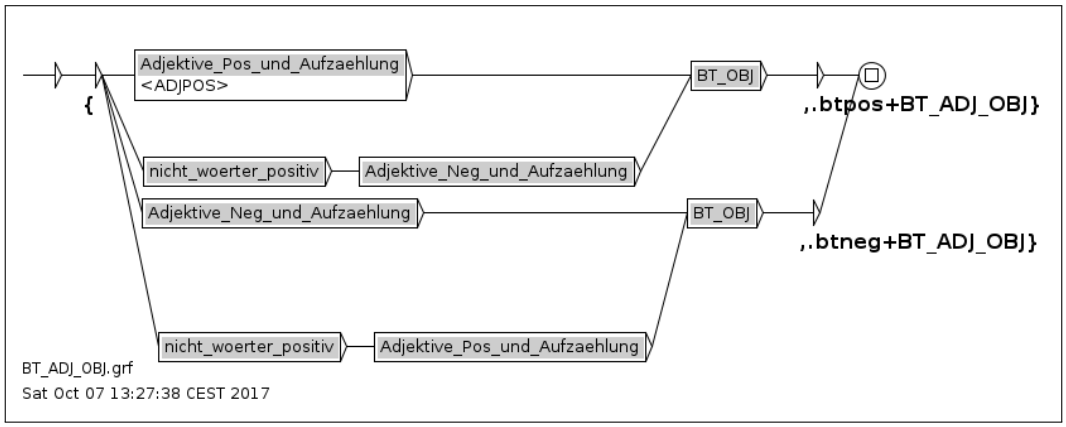

Abbildung C.5: Graph zur Erkennung der Phrasen zu „Betreuung“

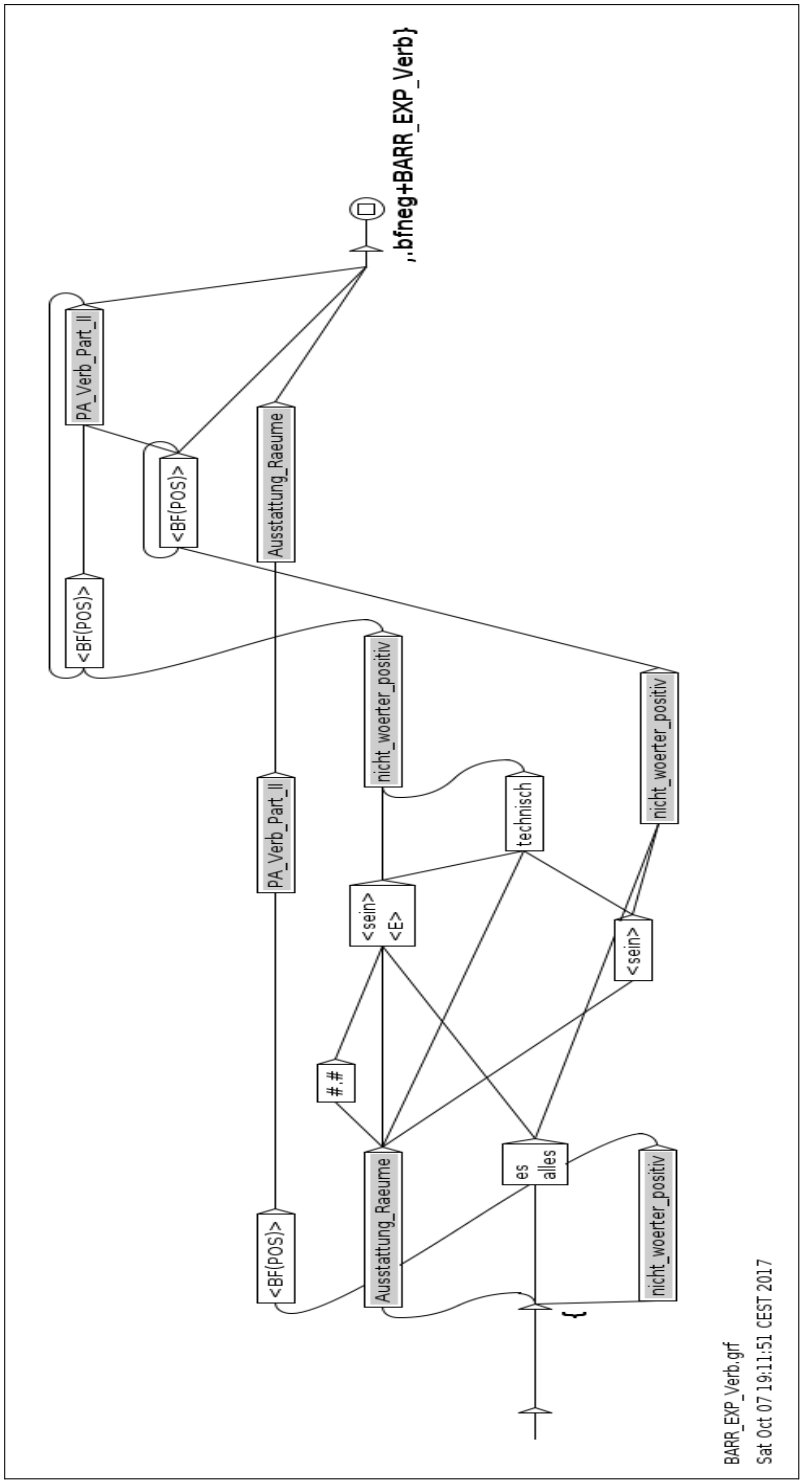

Abbildung C.6: Graph zur Erkennung der Phrasen zu „Barrierefreiheit“

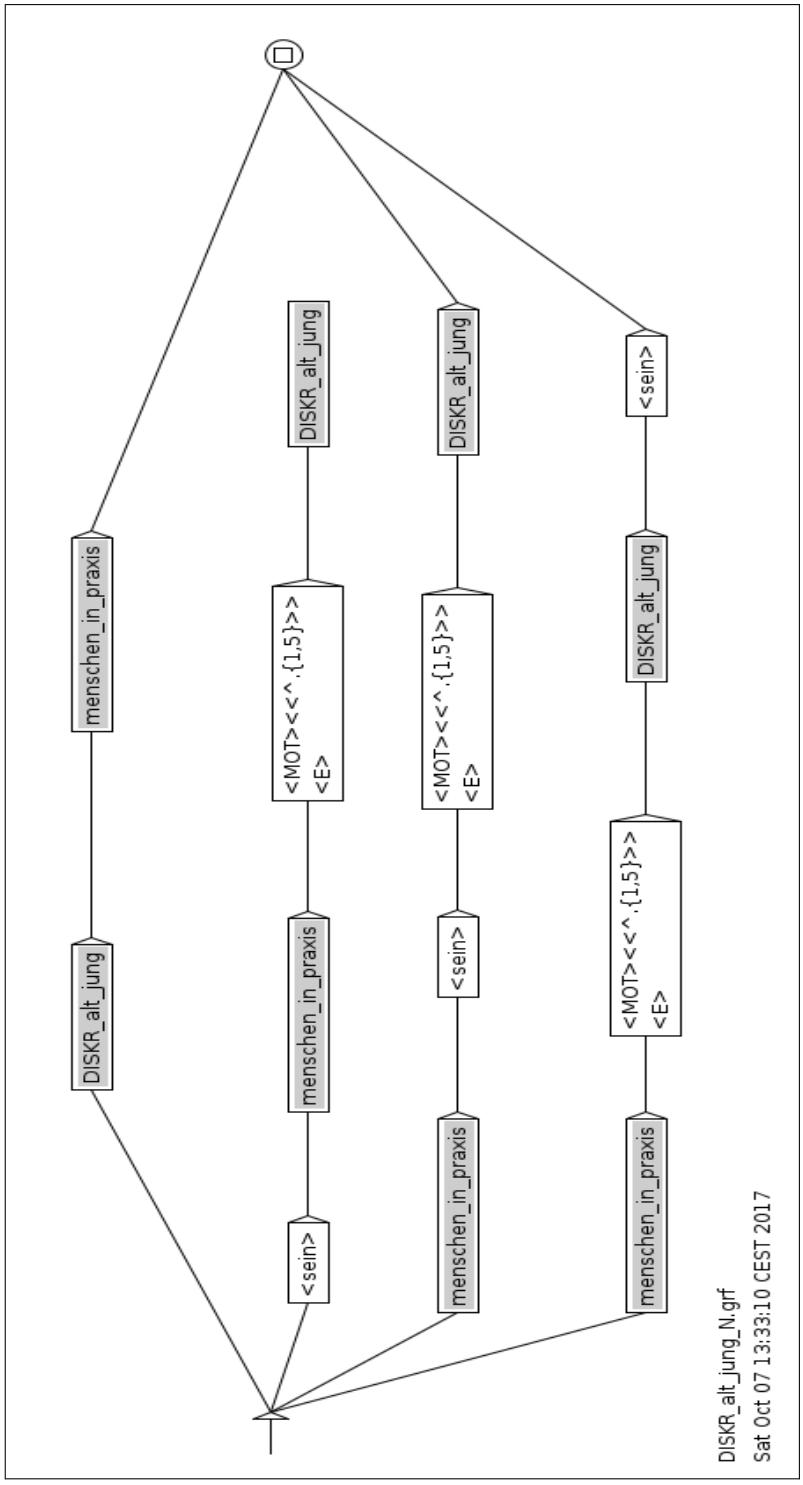

Abbildung C.7: Graph zur Erkennung der Phrasen zu Diskriminierung (bzgl. des Alters)

Anhang D

Linguistische Muster zu Bewertungsdimensionen

1. alten praxis
2. alte zeitschriften
3. angenehme atmosphäre
4. arzt meines vertrauens
5. auch alternative therapien
6. auch mit naturheilverfahren
7. auch pflanzliche mittel
8. auch traditionelle chinesische medizin
9. barrierefreiheit ist nicht gegeben
10. bekam ich einen termin
11. bekommt man sofort einen termin
12. Betreuung durch seine Angestellte fühlt man sich bei ihm persönlich sehr gut beraten und sehr freundlich behandelt
13. den Arzt sehr gut, freundlich und kompetent, ausreichende Aufklärung, es blieb ein Gefühl des Vertrauens

14. die praxis ist hell
15. die praxis ist sauber
16. die praxis ist schön eingerichtet
17. die praxis ist sehr sauber
18. die praxisräume sind hell
19. eigene parkplätze
20. erklärt alles
21. es gibt wasser
22. es keine termine gibt
23. freundliche betreuung
24. Freundliches Personal
25. freundliches team
26. freundlich und nett
27. für gehbehinderte nicht zu erreichen
28. geduldig
29. gibt es eine spielecke
30. gibt es kaffee
31. gut ausgestattete kinderspielecke
32. gut versorgt
33. habe mich sehr wohl gefühlt
34. hat sich viel zeit genommen
35. hervorragend betreut
36. hervorragender arzt

37. ich fühle mich sehr wohl
38. immer sehr gut betreut
39. immer telefonisch zu erreichen
40. interesse an alternativen heilmethoden
41. ist kein weg zu weit
42. jetzt weiter weg
43. keine aufklärung
44. keine parkplätze
45. kein gespräch
46. kein kinderfreundlicher
47. kinderfreundlicher augenarzt
48. kinderfreundlicher mediziner
49. klasse arzt
50. kleine praxis
51. kompetent behandelt
52. kurzen Öffnungszeiten
53. kurze wartezeiten
54. längere wartezeiten
55. mit kinderwagen oder karre dank einer rampe und eines fahrstuhls gut zu erreichen
56. mit kinderwagen zu erreichen
57. moderne geräte
58. nahm sich sehr viel zeit

59. nettes praxisteam
60. nicht kinderfreundlich
61. nimmt sich zeit
62. ohne komplikationen
63. ohne lange wartezeiten
64. ohne schmerzen
65. optimal betreut
66. perfekte betreuung
67. praxis ist angenehm
68. sehr engagiert, großzügige Öffnungszeiten
69. sehr großzügige Öffnungszeiten
70. sehr gut betreut
71. Sehr gute Behandlung
72. sehr gute beratung
73. sehr gute betreuung
74. sehr gute erreichbarkeit
75. sehr gute Öffnungszeiten
76. sehr gute sprechstundenzeiten
77. sehr kinderfreundliche
78. sehr kinderfreundliches personal
79. sehr kompetente beratung
80. sehr kompetenter arzt
81. Sehr kompetente und freundliche Ärztin

82. sehr kompetent und freundlich betreut
83. sehr netter und guter arzt
84. sehr netter und kompetenter arzt
85. sehr schnelle terminvergabe
86. sehr schöne sprechzeiten
87. sich zeit für den patienten nimmt
88. super ärztin
89. telefonische erreichbarkeit sehr gut
90. telefonisch gut zu erreichen
91. telefonisch immer zu erreichen
92. termin sofort
93. toller arzt
94. unfreundliche praxis
95. verständnisvoller arzt
96. vertrauenswürdiger arzt
97. weiter weg
98. wenige parkplätze
99. zugang zur praxis ist nur über eine treppe zu erreichen
100. zu längeren wartezeiten

Anhang E

Abkürzungen

Abkürzungen	**Erläuterungen**
ADJNAT	Adjektiv zur Nationalität
ADJPOS / ADJNEG	Adjektiv mit positiver /negativer Polarität
ah / AH	Alternative Heilmethoden
ak / AK	Aufklärung
AUSR	Ausreißer
BEST_FEHL	Bestätigungsfehler
bf / BF	Barrierefreiheit
bh / BH	Behandlung
bt / BT	Betreuung
CISLEX_SENTIWS	Lexikon, das mittels zwei Wörterbücher CISLEX und SentiWS (korpusbasiert) erstellt wurde
CognIEffect	Das Identifikationssystem für kognitive Effekte
DISKR	Diskriminierungen
et / ET	Entertainment
FN	False negatives
FP	False positives
fr / FR	Freundlichkeit
G.Std.	Goldstandard
gz / GZ	Genommene Zeit
IE	Informationsextraktion
kfr / KFR	Kinderfreundlichkeit
LAND(KF)	Land Kurzform

LAND(VF)	Land Vollform
LEITDIM	Leitdimension
NAT	Nationalität
neg / NEG	negativ
oee / OEE	Öffentliche Erreichbarkeit
pa / PA	Praxisausstattung
PHRASE_LEX	Phrasenlexikon
pm / PM	Parkmöglichkeiten
pos / POS	positiv
P	Precision
R	Recall
SA	Stimmungsanalyse
sz / SZ	Sprechstundenzeiten
te / TE	Telefonische Erreichbarkeit
TN	True negatives
TP	True positives
UGC	User Generated Content
vv / VV	Vertrauensverhältnis
WZ(POS) / WZ(NEG)	Wartezeit (positiv / negativ)
wzp / WZP	Wartezeit (Praxis)
wztneg	Wartezeit (Termin), negative Polarität
wztpos	Wartezeit (Termin), positive Polarität
wzt / WZT	Wartezeit (Termin)

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Anderson, J. R. (2001). *Kognitive Psychologie*, Kapitel 2,6,9,11,12, Seiten 57–65, 173–187, 198–202, 379–422. Spektrum Akademischer Verlag, Heidelberg, Berlin, 3 Auflage.

Archak, N., Ghose, A. und Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, 57(8):1485–1509. `http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1110.1370`, 16.04.2017.

Bagheri, A., Saraee, M. und de Jong, F. (2013). An unsupervised aspect detection model for sentiment analysis of reviews. In *International Conference on Application of Natural Language to Information Systems*, Seiten 140–151. Springer. `http://link.springer.com/chapter/10.1007/978-3-642-38824-8_12`, 16.04.2017.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B. und Belyaeva, J. (2013). Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*, Seiten 2216–2220. `https://arxiv.org/abs/1309.6202`, 20.05.2017.

Bamman, D. und Smith, N. A. (2015). Contextualized Sarcasm Detection on Twitter. In *ICWSM*, Seiten 574–577. Citeseer. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.738.5739&rep=rep1&type=pdf`, 29.04.2017.

Biechele, B. (2006). Film / Video / DVD in Deutsch als Fremdsprache - Bestandsaufnahme und Perspektiven. In Barkowski, H. und Wolff, A. (Hrsg.), *Umbrüche. Materialien Deutsch als Fremdsprache*, Band 76, Seiten 309–328. Fachverband Deutsch als Fremdsprache (FaDaF), Regensburg.

Bornebusch, F. und Cancino, G. e. a. (2014). Aspekt-basierte Sentiment Analysis. In *INFORMATIK 2014 Big Data – Komplexität meistern*, Band 232, Seiten 2389–2400, Stuttgart. Köllen Druck+Verlag GmbH, Bonn. http://subs.emis.de/LNI/Proceedings/Proceedings232/2389.pdf, 30.04.2015.

Bretschneider, U. (2015). Vorhersagen und Stimmungsanalyse im Web 2.0. http://wcms.uzi.uni-halle.de/download.php?down=26782&elem=2621200, 20.03.2017.

Broß, J. (2013). *Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques*. Dissertation, Freie Universität Berlin, Germany. http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000094711?lang=de, 20.05.2015.

Buttler, G. (1996). Ein einfaches Verfahren zur Identifikation von Ausreißern bei multivariaten Daten. Technischer Bericht, Diskussionspapiere Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie. http://www.econstor.eu/handle/10419/29594, 09.03.2016.

Carstensen, K.-U. (2017). Informationsextraktion (IE). In *Sprachtechnologie – Ein Überblick*, Seiten 57–72. Carstensen, Kai-Uwe. http://www.kai-uwe-carstensen.de/Publikationen/Sprachtechnologie.pdf, 03.09.2017.

Caverni, J.-P., Fabre, J.-M. und Gonzalez, M. (1990). *Cognitive Biases*, Band 68, Seiten 7–12, 59–67. Elsevier Science Publishers B.V., Amsterdam, New York, Oxford, Tokyo.

Chen, H. (2012). Sentiment Analysis. In Sharda, R. und Voß, S. (Hrsg.), *Dark Web*, Band 30, Seiten 171–201. Springer New York, New York, NY. http://link.springer.com/10.1007/978-1-4614-1557-2_10, 20.05.2015.

Clauß, G. und Ebner, H. (1977). Der Zusammenhang zwischen Meßwerten (Maßkorrelation); Kombinatorik. In *Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen*, Seiten 116–124, 136– 161. Harri Deutsch, Thun.

Cowie, J. und Wilks, Y. (2000). Information Extraction. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.6480&rep=rep1&type=pdf`, 19.04.2015.

Czeschik, J. C. (2015). Kognitive Verzerrung: Fehlerhafte Wahrnehmung der Realität (Vortrag). `https://www.youtube.com/watch?v=woug36Y4_y8`, 22.03.2016.

Ding, X., Liu, B. und Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, Seiten 1–9. ACM. `http://dl.acm.org/citation.cfm?id=1341561`, 04.04.2017.

Droste, H. W. (2013). Dunning-Kruger-Effekt: So lässt sich Inkompetenz bändigen! `http://pr4punktnull.de/wp-content/uploads/downloads/2013/09/Droste_DunningKrugerSyndrom092013.pdf`, 20.01.2014.

Elspass, S. und Maitz, P. (2011). Sprache und Diskriminierung. *Der Deutschunterricht*, 6:2–6.

Garg, C. und Goyal, L. (2014). Automatic Extraction of Idiom, Proverb and its Variations from Text using Statistical Approach. *An International Journal of Engineering Sciences*, 10. `http://ijoes.vidyapublications.com/paper/Vol10/02-vol10.pdf`, 20.05.2017.

Gehrig, M. und Breu, M. (2013). Controlling hilft, strategische Denkfehler zu vermeiden. *Gabler Verlag*, 57(3):46–53.

Geierhos, M. (2010). *BiographIE – Klassifikation und Extraktion karrierespezifischer Informationen*, Band 5 aus *Linguistic Resources for Natural Language Processing*. Lincom, München. ISBN 978-3-86288-013-3.

Geierhos, M., Bäumer, F. S., Schulze, S. und Stuß, V. (2015a). Filtering Reviews by Random Individual Error. In Ali, M., Kwon, Y. S., Lee, C.-H., Kim, J. und Kim, Y. (Hrsg.), *Current Approaches in Applied Artificial Intelligence*, Band 9101 aus *Lecture Notes in Computer Science*, Seiten 305–315. Springer International Publishing Switzerland. ISBN: 978-3-319-19065-5.

Geierhos, M., Bäumer, F. S., Schulze, S. und Stuß, V. (2015b). I grade what I get but write what I think. Inconsistency Analysis in Patients' Reviews. In *ECIS 2015 Completed Research Papers*. Paper 55.

Geierhos, M. und Stuß, V. (2015). DHd2015. Von Daten zu Erkenntnissen. Book of Abstracts. Vorträge. `http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege`, 03.02.2017.

Gigerenzer, G. und Gaissmaier, W. (2006). Denken und Urteilen unter Unsicherheit: Kognitive Heuristiken. In *Fast and frugal heuristics: The tools of bounded rationality*. Blackwell, Oxford. `https://www.psychologie.uni-heidelberg.de/ae/allg/enzykl_denken/Enz_06_Heuristiken.pdf`, 22.03.2016.

Goerke, B. (2016). *Essays zur Förderung und Bewertung von Innovationen*. Dissertation, Christian-Albrechts Universität Kiel. `http://macau.uni-kiel.de/receive/dissertation_diss_00020026`, 12.06.2017.

Grams, T. (2006). Denkfallen: Klug irren will gelernt sein. Vortrag zur MIND AKADAMIE 2006 vom 5.-8. Oktober in Marburg. `http://www2.hs-fulda.de/~grams/Denkfallen/KlugIrren.pdf`, 04.03.2016.

Grcic, J. (2008). The halo effect fallacy. *Electronic Journal for Philosophy*, 15:1–6. `http://nb.vse.cz/kfil/elogos/mind/grcic08.pdf`, 06.07.2015.

Gross, M. (1997). The Construction of Local Grammars. In Roche, E. und Schabés, Y. (Hrsg.), *Finite-state language processing*, Seiten 329–354. MIT Press.

Gross, M. (1999). A bootstrap method for constructing local grammars. In *Proceedings of the Symposium on Contemporary Mathematics*, Seiten 229–250. University of Belgrad. `https://halshs.archives-ouvertes.fr/halshs-00278319/`, 07.11.2015.

Guenthner, F. und Maier, P. (1994). Das CISLEX-Wörterbuchsystem. CIS-Bericht 76, Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwig-Maximilians-Universität München (LMU), München. `http://www.cis.lmu.de/download/cis-berichte/94-076.pdf`, 06.11.2015.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):1–21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/, 17.12.2016.

Hallmann, K., Kunneman, F., Liebrecht, C., van den Bosch, A. und van Mulken, M. (2016). Sarcastic Soulmates: Intimacy and irony markers in social media messaging. *LiLT (Linguistic Issues in Language Technology)*, 14:1–23. http://csli-lilt.stanford.edu/ojs/index.php/LiLT/article/view/50, 29.04.2017.

Hammann, M., Jördens, J. und Schecker, H. (2014). Übereinstimmung zwischen Beurteilern: Cohens Kappa (κ). *Methoden in der naturwissenschaftsdidaktischen Forschung*, Seiten 1–6. http://static.springer.com/sgw/documents/1426183/application/pdf/Cohens+Kappa.pdf, 16.12.2016.

Harris, Z. S. (1979). Sublanguages. In *Mathematical structures of language*, Seiten 152–155. Krieger, Huntington, NY.

Hatzivassiloglou, V. und McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, Seiten 174–181. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=979640, 23.07.2015.

Hernandez, I. und Preston, J. L. (2012). Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology*, 49(1):1–5. http://linkinghub.elsevier.com/retrieve/pii/S002210311200176X, 06.09.2015.

Heyd, G. (1997). *Aufbauwissen für den Fremdsprachenunterricht (DaF): Ein Arbeitsbuch : Kognition und Konstruktion*, Seiten 85–88. Narr Studienbücher. G. Narr, Tübingen.

Hu, N., Koh, N. S. und Reddy, S. K. (2013). Ratings Lead You To The Product. Reviews Help You Clinch It: The Dynamics and Impact of Online Review Sentiments on Product Sales. *Decision support systems*, 57(42):1–31. http://ink.library.smu.edu.sg/lkcsb_research/3509/, 27.02.2016.

Höer, R., Galliker, M., Huerkamp, M., Wagner, F., Weimar, D. und Graumann, C. F. (1996). Implizite sprachliche Diskriminierungen: Eine facettentheoretische Modellvalidierung. Arbeiten aus dem Sonderforschungsbereich 245 „Sprache und Situation“ Heidelberg/Mannheim 103, Psychologisches Institut der Universität Heidelberg, Heidelberg. http://www.psychologie.uni-heidelberg.de/institutsberichte/SFB245/SFB103.pdf, 20.02.2017.

Johnson, R. E., Conlee, M. C. und Tesser, A. (1973). Effects of Similarity of Fate on Bad News Transmission: A Reexamination. In *Midwestern Psychological Association*, Seiten 1–19, Chicago.

Kahneman, D. und Tversky, A. (1984). Choices, Values, and Frames. *American Psychologist*, 39(4):341–350.

Kaiser, C. (2012). *Business Intelligence 2.0*. Gabler Verlag, Wiesbaden. http://link.springer.com/10.1007/978-3-8349-3990-6, 02.07.2015.

Kim, S.-M. und Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, Seite 1367. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1220555, 23.07.2015.

Kruger, J. und Dunning, D. (1999). Unskilled and Unaware of It. How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134.

Kühne, R. (2013). Emotionale Framing-Effekte auf Einstellungen: Ein integratives Modell. *M&K Medien & Kommunikationswissenschaft*, 61(1):5–20.

Leonhart, R. (2013). Kombinatorik; Produkt-Moment-Korrelation. In *Lehrbuch Statistik Einstig und Vertiefung*, Seiten 133–138, 261–271. Huber Hans.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B. und Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, Seiten 939–948. ACM. http://dl.acm.org/citation.cfm?id=1871557, 23.05.2017.

Linke, A. (2003). Spam oder nicht spam? *c't*, 17:150–153.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:1 – 38.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167. http://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016, 20.05.2015.

Manning, C. D., Raghavan, P. und Schütze, H. (2009). Evaluation in information retrieval. In *An Introduction to Information Retrieval*, Seiten 151–175. Cambridge University Press, Cambridge, England. http://www.informationretrieval.org/, 20.06.2017.

Medhat, W., Hassan, A. und Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113. http://linkinghub.elsevier.com/retrieve/pii/S2090447914000550, 20.04.2017.

Meifort, B. (2002). Personenbezogene Dienstleistungen im Wandel: neue Unternehmensund Wirtschaftsformen–neue berufliche Anforderungen–neue Berufe. *Berufsbildung in Wissenschaft und Praxis (BWP) Heft*, 1:34–35.

Michalkiewicz, M. (2015). Wie Heuristiken uns helfen Entscheidungen zu treffen. *Kognitionspsychologie im Alltag. Teil 2: Lernen und Gedächtnis*, 4:1–3. http://de.in-mind.org/article/wie-heuristiken-uns-helfen-entscheidungen-zu-treffen?page=3, 22.03.2016.

Mikheev, A. (1996). Learning part-of-speech guessing rules from lexicon: Extension to non-concatenative operations. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, Seiten 770–775. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=993302, 24.08.2016.

Mukherjee, A., Liu, B. und Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, Seiten 191–200. ACM. http://dl.acm.org/citation.cfm?id=2187863, 23.05.2017.

Muzny, G. und Zettlemoyer, L. S. (2013). Automatic Idiom Identification in Wiktionary. In *EMNLP*, Seiten 1417–1421. http://anthology.aclweb.org/D/D13/D13-1145.pdf, 20.05.2017.

Nagel, S. (2008). *Lokale Grammatiken zur Beschreibung von lokativen Sätzen und ihre Anwendung im Information Retrieval*. Dissertation, LMU München. http://edoc.ub.uni-muenchen.de/10965/4/Nagel_Sebastian.pdf, 20.05.2015.

Nardi, A. (2006). *Der Einfluss außersprachlicher Faktoren auf das Erlernen des Deutschen als Fremdsprache*. Dissertation, Universität Zürich.

Neumann, G. (2004). Informationsextraktion. In Carstensen, K.-U., Ebert, C., Endriss, C., Jekat, S., Klabunde, R. und Langer, H. (Hrsg.), *Computerlinguistik und Sprachtechnologie. Eine Einführung*, Kapitel 5.5, Seiten 502–510. Spektrum Akademischer Verlag, Heidelberg.

Pang, B., Lee, L. und Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Seiten 79–86. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1118704, 07.04.2017.

Park, S., Kang, S., Chung, S. und Song, J. (2009). NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Seiten 443–452. ACM. http://dl.acm.org/citation.cfm?id=1518772, 07.03.2016.

Patil, M. S. und Bagade, A. M. (2012). Online review spam detection using language model and feature selection. *International Journal of Computer Applications*, 59(7):33–36. http://search.proquest.com/openview/932270328fcc0c65bb32b7e65c1d49f4/1?pq-origsite=gscholar&cbl=136216, 23.05.2017.

Paumier, S. (2015). Unitex 3.1. beta User Manual. http://igm.univ-mlv.fr/~unitex/UnitexManual3.1.pdf, 04.10.2015.

Pollock, J. (2012). The Halo effect: The influence of attractiveness on perceived promiscuity. *FALL 2012*, 7:34–37. https://sites.google.com/a/umn.edu/sentience/Pollock_2012.pdf, 20.01.2017.

Prestin, E. (2003). Theorien und Modelle der Sprachrezeption. In Rickheit, G., Herrmann, T. und Deutsch, W. (Hrsg.), *Psycholinguistik - Ein Internationales Handbuch*, Seiten 491–499. Walter de Gruyter, Berlin / New York.

Remus, R. und Ahmad, K. (2010). Stimmungen in deutschsprachigen Nachrichten, Blogs und dem DAX. *eDITion*, 10(01):24–27.

Remus, R., Quasthoff, U. und Heyer, G. (2010). SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, Seiten 1168–1171.

Sager, N. (1986). Sublanguage: Linguistic Phenomenon, Computational Tool. In Grishman, R. und Kittredge, R. (Hrsg.), *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Seiten 1–12. Psychology Press.

Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3):261–377. `http://dl.acm.org/citation.cfm?id=1498845`, 05.11.2015.

Schieber, A., Hilbert, A. und Stillich, C. (2012). Identifikation und Analyse von ironischen und sarkastischen Kundenrezensionen im Web. In *Multikonferenz Wirtschaftsinformatik*, Seiten 1157–1168, Berlin. GITO mbH Verlag. `http://digisrv-2.biblio.etc.tu-bs.de:8081/docportal/servlets/MCRFileNodeServlet/DocPortal_derivate_00027689/Beitrag271.pdf`, 06.11.2015.

Schneider, J., Yemane, R. und Weinmann, M. (2014). Diskriminierung am Ausbildungsmarkt Ausmaß, Ursachen und Handlungsperspektiven. Studie, Forschungsbereich beim Sachverständigenrat deutscher Stiftungen für Integration und Migration (SVR), Berlin.

Schneider, M. und Bauhoff, F. (2013). Stellenanzeigen und AGG : von Geschlechtsneutralität noch weit entfernt. *Personal quarterly*, 65(3):15–20.

Schneider, S. (2013). *Das 3-Dimensionsmodell der Wissensrekonstruktion: A priorische Sicherstellung der Güte generierten Wissens*. Forschungsbericht. Fachhochschule Kiel, Fachbereich Wirtschaft, Institut für Wirtschaftsinformatik.

Schuhmacher, R. und Schwegler, D. (2007). Rechtspsychologie: „Natürlich sind Richter überfordert“. *Plädoyer*, 1:8–10. http://www.decisions.ch/dissertation/diss_schweizer_streitgespraech.pdf, 03.04.2016.

Schweizer, M. (2005). *Kognitive Täuschungen vor Gericht Eine empirische Studie*. Dissertation, Universität Zürich. http://www.decisions.ch/dissertation.html, 05.06.2016.

Schöberl, S. (2012). *Verbraucherverhalten bei Bio-Lebensmitteln: Analyse des Zusammenhangs zwischen Einstellungen, Moralischen Normen, Verhaltensabsichten und tatsächlichem Kaufverhalten*. Dissertation, Technische Universität München.

Shailesh, K. Y. (2015). Sentiment Analysis and Classification: A Survey. *International Journal of Advance Research in Computer Science and Management Studies*, 3(3):113–121.

Stadler, D., Schrank, J., Walm, R., Meißner, T., Flöter, C., Becker, D., Höhne, K., Esch, M. und Wanninger, L. (2007). Vergleich von Diensten für automatische GPS-Datenauswertung. *AVN*, 1:34–37. https://tu-dresden.de/bu/umwelt/geo/gi/gg/ressourcen/dateien/veroeffentlichungen/avn07_34-37.pdf?lang=de, 12.06.2017.

Stotz, S. C. (2018). *Informationsextraktion aus deutschsprachigen Wirtschaftsnachrichten über Unternehmenszusammenschlüsse mit lokalen Grammatiken*. Dissertation, LMU München. Im Druck.

Stürmer, S. (2009). *Sozialpsychologie*, Kapitel 4, 8, Seiten 69–90, 165–216. Ernst Reinhardt, GmbH & Co KG, München.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4:25–29.

Tsang, A. S. und Prendergast, G. (2009). Is a “star” worth a thousand words?: The interplay between product-review texts and rating valences. *European Journal of Marketing*, 43(11/12):1–23. http://www.emeraldinsight.com/doi/10.1108/03090560910989876, 03.06.2017.

Turney, P. D. (2002). Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, Seiten

417–424. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1073153, 24.08.2016.

Tversky, A. und Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Vakili Samiyan, L. (2014). The Comparison Between Contextual Guessing Strategy vs. Memorizing a List of Isolated Words in Vocabulary Learning Regarding Long Term Memory. *International journal of Science Culture and Sport*, 2(1):12–18. http://dergipark.gov.tr/doi/10.14486/IJSCS41, 24.08.2016.

Vázquez, S. und Bel, N. (2013). A classification of adjectives for polarity lexicons enhancement. *CoRR*, abs/1303.1931:3557–3561. http://arxiv.org/abs/1303.1931, 08.04.2013.

Verma, R. und Vuppuluri, V. (2015). A New Approach for Idiom Identification Using Meanings and the Web. In *RANLP*, Seiten 681–687. https://www.aclweb.org/anthology/R/R15/R15-1087.pdf, 20.05.2017.

Vázquez, S., Padró, M., Bel Rafecas, N. und Gonzalo, J. (2012). Automatic extraction of polar adjectives for the creation of polarity lexicons. In *Kay M, Boitet C, editors. Proceedings of COLING 2012: Posters: 24th International Conference on Computational Linguistics COLING 2012; 2012 December 8-15; Mumbai, India. Mumbai: The COLING 2012 Organizing Committee; 2012. p. 1271-1280.*, Seiten 1271–1280. ACL (Association for Computational Linguistics). https://repositori.upf.edu/handle/10230/20422, 04.04.2017.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140. http://www.tandfonline.com/doi/abs/10.1080/17470216008416717, 26.10.2015.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. und Martin, M. (2004). Learning subjective language. *Computational linguistics*, 30(3):277–308. http://www.mitpressjournals.org/doi/abs/10.1162/0891201041850885, 06.02.2016.

Wiegand, M. und Klakow, D. (2009). The Role of Knowledge-based Features in Polarity Classification at Sentence Level. In

FLAIRS Conference. https://pdfs.semanticscholar.org/001b/f3c08d546ecfa73dedec89d998348bfca231.pdf, 07.04.2017.

Wilkening, R. (2008). Denkfehler erkennen. http://www.psychotherapie-davos.ch/Kontakt/Service/Download_Materialien/ABC_Denkfehler.pdf, 20.04.2015.

Wolff, D. (1993). Der Beitrag der kognitiv orientierten Psycholinguistik zur Erklärung der Sprach- und Wissensverarbeitung. In Gienow, W. und Hellwig, K. (Hrsg.), *Prozeßorientierte Mediendidaktik im Fremdsprachenunterricht*, Seiten 27–41. Lang, Frankfurt am Main.

Wolff, D. (2002). *Fremdsprachenlernen als Konstruktion: Grundlagen für eine konstruktivistische Fremdsprachendidaktik*, Seiten 15–45, 60–70, 103–145. Lang, Frankfurt am Main.

Wolfgruber, M. (2015). *Sentiment Analyse mit lokalen Grammatiken: wissensbasierter Ansatz zur Extraktion von Sentiments in Hotelbewertungen*, Band 3 aus *Dissertationen der LMU München*. Verlagshaus Monsenstein und Vannerdat OHG, Münster.

Wu, G., Greene, D. und Cunningham, P. (2010). Merging multiple criteria to identify suspicious reviews. In *RecSys '10: Proceedings of the forth ACM conference on Recommender systems*, Seiten 1–8. ACM Press. http://portal.acm.org/citation.cfm?doid=1864708.1864757, 23.05.2017.

Xie, S., Wang, G., Lin, S. und Yu, P. S. (2012a). Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Seiten 823–831. ACM. http://dl.acm.org/citation.cfm?id=2339662, 23.05.2017.

Xie, S., Wang, G., Lin, S. und Yu, P. S. (2012b). Review spam detection via time series pattern discovery. In *Proceedings of the 21st International Conference on World Wide Web*, Seiten 635–636. ACM. http://dl.acm.org/citation.cfm?id=2188164, 23.05.2017.

Zick, A. (2004). Soziale Einstellungen. In Sommer, G. und Fuchs, A. (Hrsg.), *Krieg und Frieden: Handbuch der Konflikt- und Friedenspsychologie*, Seiten 129–142. Beltz/Psychologie Verlags Union, Weinheim.

Zimmer, H. D. (1988). Gedächtnispsychologische Aspekte des Lernens und Verarbeitens von Fremdsprache. *Informationen Deutsch als Fremdsprache*, 15(2):149–163.