SSOAR

Open Access Repository
www.ssoar.info

Wie empfinden Teilnehmer die Fragen in Online-Befragungen? Entwicklung eines Diktionärs für die automatische Codierung freier Antworten

Kaczmirek, Lars; Baier, Christian; Züll, Cornelia

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Kaczmirek, L., Baier, C., & Züll, C. (2010). Wie empfinden Teilnehmer die Fragen in Online-Befragungen? Entwicklung eines Diktionärs für die automatische Codierung freier Antworten. In M. Welker, & C. Wünsch (Hrsg.), *Die Online-Inhaltsanalyse: Forschungsobjekt Internet* (S. 191-223). Köln: Halem. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-46650-7

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

LARS KACZMIREK / CHRISTIAN BAIER /
CORNELIA ZÜLL[1]

Wie empfinden Teilnehmer die Fragen in Online-Befragungen? Entwicklung eines Diktionärs für die automatische Codierung freier Antworten

Online-Befragungen gewinnen sowohl in der kommerziellen als auch in der wissenschaftlichen Sozialforschung immer mehr an Bedeutung. Im Zuge dieser Entwicklung steigt auch die Popularität offener Fragen wieder, da diese in Online-Befragungen mit geringerem Aufwand gestellt und erhoben werden können. Die Vorteile offener Fragen liegen in der Erfassung neuer, nicht vorgegebener Dimensionen. Offene Fragen können im Vergleich zu geschlossenen Antworten besonders gehaltvolle Inhalte liefern. Sie eignen sich deshalb besonders, wenn die Antwortmöglichkeiten mit der Untersuchung exploriert werden sollen und im Vorhinein nicht bekannt sind. Die Bewertung eines Fragebogens durch die Teilnehmer ist solch ein Anwendungsgebiet. Zur Qualitätssicherung von Befragungen bietet es sich an, am Ende jeder Befragung Urteile über die Befragung selbst zu erheben. Zur Vergleichbarkeit zwischen den Befragungen wird in der Praxis allerdings häufig eine Notenskala verwendet. Dies erlaubt es, auf einem allgemeinen Niveau Schwankungen in der Akzeptanz von Befragungen zu erkennen. Andererseits können im Falle auffälliger Bewertungen, seien es die Häufung negativer

1 Dank: Wir bedanken uns bei allen Personen für die Kooperation bei der Durchführung der verschiedenen Studien, insbesondere bei Wolfgang Bandilla, Michael Blohm, Michael Braun (Studien 1 und 2), Wolfgang Neubarth, Julia Rector (Studien 1 bis 3), den Mitarbeitern von DerZweiteFrühling (Studie 4) und den Mitarbeitern der Respondi AG, besonders Otto Hellwig und Tom Wirth (Studien 1, 3, 5 und 6).

oder besonders positiver Rückmeldungen, keine Rückschlüsse auf das ›Warum‹ gezogen werden. Damit bleibt die Möglichkeit verborgen, die Kriterien besonders gelungener Befragungen oder Faktoren, die zu negativen Bewertungen führen, zu identifizieren. Eine offen gestellte Frage zur Bewertung des Fragebogens bzw. der Fragen löst dieses Problem, da mit offen gestellten Fragen bedeutende und hervorstechende Themen erfasst werden können (GEER 1991). Als Teil der Qualitätssicherung können solche Fragen auch dazu dienen, die Belastung der Teilnehmer zu erfassen (HEDLIN/DALE/HARALDSEN/JONES 2005). Weiterhin können solche Fragen bei der Entwicklung von Fragebögen innerhalb von Standard-Pretests verwendet werden, um die Suche nach Problembereichen zu erleichtern, zum Beispiel als Teil des Zwei-Phasen-Pretesting (vgl. PRÜFER/REXROTH 2000).

Der vorliegende Beitrag beinhaltet die Ergebnisse aus 6 Studien, in denen insgesamt 4150 Personen eine offene Bewertungsfrage beantwortet haben. Am Ende der Befragungen wurden die Teilnehmer gebeten, den unvollständigen Satz »Ich empfand die Fragen als ...« zu beenden. Mithilfe der Antworten wurde ein Kategoriensystem (d.h. ein Diktionär für die automatische Codierung) erstellt, mit dem es nun möglich ist, eine computerunterstützte Inhaltsanalyse durchzuführen. Dieses Diktionär wurde so entwickelt, dass damit auch entsprechende offene Fragen in zukünftigen Befragungen codiert werden können. Die automatische Codierung klassifiziert die Antworten in positive, negative und ambivalente Bewertungen, wodurch ähnlich der Notenskala eine allgemeine Aussage zur Qualität von Befragungen möglich ist. Darüber hinaus stehen jedoch auch die Antworten bzw. produzierten Wortlisten selbst zur Verfügung, wodurch sich Forscher leicht ein qualitatives Bild von der vorherrschenden Stimmung bei den Teilnehmern machen können. So wird schnell ersichtlich, ob ein Fragebogen einfach ›interessant‹ oder sogar ›wichtig‹ ist oder ob es sich eher um eine ›langweilige‹ oder gar ›zu persönliche‹ Befragung handelt. Der Forscher kann sich die Antworttexte aller Befragten ansehen, die einen Fragebogen zum Beispiel als negativ bewertet hatten, und überprüfen, was zu dieser Kritik führte. Damit wird es möglich, aus einer allgemein positiven oder negativen Bewertung auf spezifische Ursachen zu schließen. Die Vorteile gegenüber der alternativ möglichen Verwendung von Item-Batterien ist die höhere Durchführungsökonomie (kürzere Bearbeitungszeit) verbunden mit dem offenen Frageformat. Eine Festlegung auf bestimmte Bewertungskriterien im

Vorhinein ist dadurch nicht notwendig. Insbesondere bei kurzen Befragungen mit weniger als 30 Fragen würde eine Item-Batterie die Belastung der Teilnehmer unnötig erhöhen.

Nach einer kurzen Übersicht zu den Möglichkeiten der computerunterstützten Inhaltsanalyse, liegt der Schwerpunkt des Beitrages in der Darstellung der Konstruktion des Diktionärs. Erstens dient dies der Illustration eines typischen Vorgehens bei der computergestützten Inhaltsanalyse und veranschaulicht, wie Diktionäre auch mit Abwandlungen der hier vorgestellten Frage konstruiert werden können. Zweitens bietet es ein genaueres Verständnis des entwickelten Diktionärs, wodurch eine Weiterentwicklung bei anderen Forschungsfragen und die Verwendung in zukünftigen Fragebögen erleichtert wird.

1. Computerunterstützte Inhaltsanalyse

Zur systematischen Analyse von offenen Fragen bietet sich die Inhaltsanalyse an. Die Texte werden dabei durch Codierer von Hand verschiedenen Kategorien zugeordnet (FRÜH 2007). Diese Vorgehensweise ist allerdings sehr zeitaufwendig und kostenintensiv.

Die Analyse offener Fragen gestaltet sich im Vergleich zu geschlossenen Fragen aufwendiger, und sie werden daher seltener eingesetzt. Einfache Auswertungen beschränken sich darauf, die Texte durchzusehen und einzelne, interessante Zitate herauszugreifen.

Da bei Online-Umfragen häufig große Antwortmengen anfallen und die Texte schon maschinenlesbar vorliegen, bietet sich die computerunterstützte Inhaltsanalyse als eine vielversprechende Methode an, mit der nicht-nummerische Daten – beispielsweise Antworten zu offenen Fragen – strukturiert und der statistischen Bearbeitung und Auswertung zugänglich gemacht werden können.

In der computerunterstützten Inhaltsanalyse finden zwei verschiedene Techniken Anwendung (vgl. den Beitrag von SCHARKOW in diesem Band): zum einen die Co-Occurrence-Analyse und zum anderen das automatische Codieren auf der Basis eines Diktionärs. Beim Einsatz der Co-Occurrence-Analyse wird das gemeinsame Auftreten von Wörtern in einer Texteinheit gemessen und daraus ein Ähnlichkeitsmaß berechnet, das dann in einer Klassifikationsanalyse Verwendung findet. Diese Art der Analyse findet vor allem im Bereich des Text Minings Anwendung.

Der Vorteil dieses explorativen Herangehens an den Text ist, dass keine Vorkenntnisse über die Texte erforderlich sind: Es ist keine Kategorisierung notwendig. Der Nachteil ist allerdings, dass man, um gute Ergebnisse zu erhalten, die Texte aufwendig vorbereiten muss. Zunächst muss eine Rechtschreibprüfung erfolgen, die gerade bei online erhobenen Texten dringend erforderlich ist. In einem zweiten Schritt sollten die Wörter lemmatisiert[2] werden und in einem dritten Schritt empfiehlt sich die Festlegung von Synonymen. In einem letzten Schritt müssen die inhaltlich relevanten Wörter festgelegt werden. Das Ergebnis der Co-Occurrence-Analyse sind dann Themenprofile oder Wortcluster, die in der Umfrage angesprochen wurden. Ein Beispiel für diese Art der Analyse von offenen Fragen findet sich in Kronberger/Wagner (2002).

Die zweite Form der computerunterstützten Inhaltsanalyse ist das diktionärbasierte automatische Codieren. Automatisches Codieren ermöglicht es, auch große Mengen nicht-nummerischer Daten effizient zu bearbeiten und zugleich qualitativ gehaltvolle Ergebnisse zu erzielen. Diese Art des automatischen Codierens ist nicht neu, erscheint aber im Zusammenhang mit Online-Befragungen sehr attraktiv: Die Texte liegen direkt verwendbar vor, es können große Datenmengen verarbeitet werden und die einmal erstellten Diktionäre eignen sich gut für die Wiederverwendung in verschiedenen Studien. Zudem können die Codierungen direkt mit den anderen in der Umfrage erhobenen Daten zusammengefügt und gemeinsam analysiert werden. Der Aufwand bei diesen Verfahren liegt in der Konstruktion des Diktionärs. Es gibt Versuche, die Entwicklung eines Diktionärs durch linguistische Optionen zu unterstützen (siehe z. B. SPSS INC. 2007). Linguistische Optionen, die die Konstruktion des Diktionärs unterstützen, sind u. a. die Analyse der Wortart. Dabei wird für jedes Wort bestimmt, ob es sich zum Beispiel um ein Substantiv, ein Verb, ein Adjektiv, ein Adverb oder um eine Präposition handelt. Auf diese Information kann dann bei der Konstruktion des Diktionärs zugegriffen werden. Allerdings erweisen sich solche Techniken gerade bei Online-Umfragen als wenig hilfreich, denn die Antworten sind in der Regel nicht in grammatikalisch korrekten Sätzen formuliert. Oft werden nur Stichwörter oder Halbsätze geschrieben und Rechtschreibung und

2 Lemmatisieren nennt man den Vorgang, bei dem alle Flexionsformen eines Wortes auf den jeweiligen Wortstamm reduziert werden: zum Beispiel haben ›gehen‹, ›ging‹ und ›gegangen‹ alle das Lemma ›gehen‹.

Grammatik spielen für den Befragten keine Rolle, was die automatische Identifikation von z. B. Subjekten oder Verben fast unmöglich macht.

Aufgrund dieser Überlegungen haben wir uns bei der im Folgenden beschriebenen Analyse für die diktionärbasierte Variante der computerunterstützten Inhaltsanalyse entschieden. Die Texte müssen dazu nicht korrigiert werden, denn falsch geschriebene Wörter können im Diktionär entsprechend aufgenommen werden. Codiert werden hier nur Wörter oder Phrasen, d. h., die Grammatik spielt keine Rolle. Zudem stehen die Codierungen fallweise zur Verfügung und können zusammen mit den anderen Angaben des Befragten analysiert werden.

2. Zusammenfassung des Vorgehens

2.1 *Definition der Bewertungskategorien*

Ziel der Arbeit an den Studien war die Entwicklung eines Diktionärs, das zum automatischen Codieren der Antworten auf folgende offene Frage verwendet werden kann: »Vervollständigen Sie bitte folgenden Satz: Ich empfand die Fragen als ...«. Die Antwort auf die Frage ergibt eine Phrase, die auch mehrere Wörter enthalten kann. Ein Diktionär besteht aus einer Textdatei und enthält die Kategorien, die in einem Text codiert werden sollen. Bestimmte Wörter oder Wortfolgen werden unter jeweils einer dieser Kategorien aufgelistet und ihnen dadurch als Indikatoren zugewiesen. Definiert man beispielsweise ›langweilig‹ als Indikator der Kategorie *negative Bewertung*, so wird beim automatischen Codieren jede Antwort, die das Wort ›langweilig‹ enthält, als *negative Bewertung* markiert. Die Antwort eines Teilnehmers kann dadurch auch mehrmals markiert werden. Ist die Antwort etwa »langweilig und zu lang« so erfolgt die negative Codierung zweimal.

Im vorliegenden Fall werden Aussagen zur Qualität der Befragung codiert. Von Interesse ist in diesem Zusammenhang vor allem die Valenz der Antworten: Wird die Befragung positiv oder negativ bewertet oder fällt die Bewertung ambivalent aus? Die Kategorien des Diktionärs sollten also die Valenz der Antworten erfassen.

Die Definition von Kategorien stellt einen wichtigen ersten Schritt im Prozess der Konstruktion eines Diktionärs dar. Da Änderungen an den Kategorien eines bereits bestehenden Diktionärs relativ aufwendig sind, wurde vor dem eigentlichen Beginn der Diktionärskonstruktion eine

Stichprobe der zu codierenden Daten gesichtet, um das darin enthaltene inhaltliche Spektrum zu erfassen. So konnte sichergestellt werden, dass sich die Kategorien auf Inhalte beziehen, die in den Daten auch tatsächlich vorhanden sind. Auf der Basis dieser Sichtung wurden die folgenden Kategorien definiert: positive Antwort, negative Antwort, ambivalente Antwort, neutraler Ausfall, nicht codierbar.

Positive Antworten: Das Spektrum positiver Angaben beginnt mit fast neutralen Ausdrücken, die auf eine Annahme der Befragung hindeuten, die also ausdrücken, dass die Befragung ›in Ordnung‹ bzw. ›normal‹ ist. Deutlich positive Ausdrücke sind ›sehr gut‹, ›sehr interessant‹ und ›wichtig‹. Weitere Beispiele für typisch positive Angaben sind: ›angemessen‹, ›angenehm‹, ›einfach zu beantworten‹ und ›verständlich‹.

Negative Antworten: Negative Ausdrücke deuten auf Probleme im Fragebogen oder bei der Fragestellung hin. Der Fragebogen kann ›zu lang‹ oder ›einseitig‹, ›oberflächlich‹, ›redundant‹, ›zu allgemein‹ gehalten sein. Andere Teilnehmer mögen die Fragen als ›langweilig‹ oder sogar ›nervig‹ empfinden. Für den Forscher außerdem problematisch dürften Fragebögen sein, in denen sich Nennungen von ›teilweise schwer zu beantworten‹ und ›unangenehm‹ häufen.

Ambivalente Antworten: Bei Vorhandensein von mindestens einer positiven und einer negativen Angabe in der Antwort handelt es sich um eine ambivalente Antwort. Das heißt, dass Antworten zunächst sowohl positiv als auch negativ codiert werden. Da eine Antwort dann sowohl der Kategorie ›positiv‹ als auch ›negativ‹ zugewiesen ist, wird sie als ambivalent bezeichnet. Hierzu zählen Antworten wie ›interessant, oft aber sinnlos‹, ›teilweise verbesserbar, sonst gut‹ und ›leicht zu beantworten, aber etwas durcheinander‹.

Neutrale Ausfälle: Neutrale Ausfälle sind alle Angaben, mit denen Teilnehmer klarstellen, dass sie keine inhaltliche Antwort geben möchten. Beispiele für solche Antworten sind ›kein Kommentar‹, ›...‹, ›xxx‹ oder Ähnliches. Für inhaltliche Analysen werden neutrale Ausfälle daher wie »keine Angabe« behandelt.

Nicht codierbare Antworten: Fällt eine Antwort in keine der obigen Kategorien, so ist sie nicht codierbar. Bei der computerunterstützten Inhaltsanalyse ist dies gleichbedeutend mit der Tatsache, dass sich für die Antwort kein passender Eintrag im Diktionär befindet. Ein wesentlicher Aspekt bei der Erweiterung eines Diktionärs ist es, die nicht codierbaren Antworten durchzusehen und den bestehenden Kategorien zuzuweisen.

2.2 *Aufbau des Diktionärs*

Für den Aufbau des Diktionärs und das automatische Codieren verwendeten wir das Textanalyse-Programm TEXTPACK (MOHLER/ZÜLL 2002). Dieses Programm kann mithilfe eines Diktionärs automatisch Texte codieren. Die Vorformatierung des Textes legt dabei fest, in welche Einheiten der Text gegliedert ist. In unserem Fall war jede Antwort auf die Frage »Ich empfand die Fragen als...« eine solche Texteinheit; in anderen Kontexten können einzelne Sätze, Absätze, Kapitel, Zeitungsartikel usw. als Einheiten definiert werden. Das Diktionär übernimmt beim automatischen Codieren die Rolle des Codierleitfadens. Es enthält die Kategorien sowie Indikatoren, die den Kategorien zugeordnet werden. Bei den Indikatoren kann es sich um Wortstämme, ganze Wörter oder Wortfolgen handeln. Die Kategorie ›positive Antwort‹ kann also den Wortstamm ›unterhaltsam‹, das Wort ›Bereicherung‹ sowie die Wortfolge ›gut durchdacht‹ enthalten. Das Programm sucht nun in den Texteinheiten nach den Indikatoren aus dem Diktionär. Sofern eine Antwort einen oder mehrere Indikatoren enthält, wird sie entsprechend codiert. Die Antwort ›unterhaltsam, gut durchdacht‹ würde demnach zweimal als positiv codiert, da sie zwei positive Indikatoren enthält. Ergebnis der Codierung ist ein Datensatz, der auflistet, wie häufig jede einzelne Antwort entsprechend den verschiedenen Kategorien codiert wurde. Dabei ist es auch möglich, dass Antworten überhaupt nicht codiert werden, weil sie keinen der Indikatoren aus dem Diktionär enthalten.

Beim Aufbau des Diktionärs kam es darauf an, Wortstämme, Wörter und Wortfolgen zu finden, die eindeutig anzeigen, dass eine Antwort eine positive oder negative Bewertung enthält. Solche Wörter oder Wortfolgen bezeichnen wir im Folgenden als codierbar. Die Antworten enthalten häufig auch nicht codierbare Wörter, die für sich betrachtet nicht auf eine bestimmte Bewertung schließen lassen. Ein Beispiel ist das Wort ›ein‹ in der Antwort ›ziemlich unterhaltsam, manchmal ein wenig zu lang‹. Hier wurde sinnvollerweise nur das Wort ›unterhaltsam‹ und die Wortfolge ›zu lang‹ codiert. ›Unterhaltsam‹ wurde als Wortstamm ins Diktionär eintragen, damit würden auch Wörter wie ›unterhaltsame‹, ›unterhaltsamer‹ usw. erfasst. ›Zu lang‹ wurde als Wortfolge eingetragen. Zwar wäre es auch möglich, stattdessen ›ein wenig zu lang‹ oder sogar ›manchmal ein wenig zu lang‹ ins Diktionär zu integrieren. Diese Vorgehensweise empfiehlt sich jedoch nicht, da Wortfolgen nur codiert wer-

den, sofern sich die Formulierungen in der Antwort und im Diktionär exakt entsprechen. Je mehr Wörter eine Wortfolge enthält, desto geringer ist die Wahrscheinlichkeit, dass genau diese Formulierung auch in anderen Antworten auftaucht. Die Verwendung von ›zu lang‹ als Indikator ermöglicht es dagegen, auch Antworten wie ›viel zu lang‹, ›bei Weitem zu lang‹ usw. zu erfassen. Um die Reichweite des Diktionärs zu maximieren, sollten daher die Wortfolgen möglichst nur die Wörter enthalten, die mindestens nötig sind, um eine bestimmte Bewertung auszudrücken.

Ein Diktionär besteht also aus Wörtern und Wortfolgen, die als Indikatoren dienen und bestimmten Kategorien zugeordnet sind. Unser erstes Diktionär entwickelten wir anhand der Antworten aus der ersten Studie. Mit TEXTPACK wurde eine Liste aller vorkommenden Wörter erstellt. Diese Liste wurde zunächst auf diejenigen Wörter und Wortfolgen reduziert, die im Sinne einer positiven oder negativen Bewertung codierbar sind. Daraufhin wurden diese Wörter und Wortfolgen jeweils den oben beschriebenen Kategorien zugeordnet. Am Schluss dieses Prozesses stand ein erstes Diktionär, das zum automatischen Codieren der Daten verwendet werden kann. Das Diktionär wurde zunächst an dem Datensatz getestet, aus dem es entwickelt wurde.

2.3 *Weiterentwicklung des Diktionärs*

Darauf folgte ein iterativer Prozess der Überprüfung und Ergänzung, der die generelle Grundlage einer Diktionärskonstruktion bildet: Nach der ersten Codierung des Datensatzes wurden die nicht codierten Antworten gesichtet; das Diktionär wurde um zwar codierbare, aber bisher nicht erfasste Indikatoren ergänzt. Im Anschluss an diese Ergänzungen wurde eine zweite Codierung mit dem erweiterten Diktionär durchgeführt. Die Ergebnisse dieses zweiten Durchgangs wurden auf falsche Codes überprüft. Dafür codierte der Ersteller des Diktionärs die Antworten manuell und verglich die automatische Codierung mit der manuellen Codierung. Je nach der Ursache der fehlerhaften Codierung wurden in einigen Fällen weitere Ergänzungen oder Korrekturen am Diktionär vorgenommen. Da die manuelle Überprüfung auf fehlerhafte Codes sehr aufwendig ist, wurde sie nur für die ersten drei Studien durchgeführt. Ziel war es dabei, grobe Fehlentwicklungen in der Grundstruktur des Diktionärs zu vermeiden. Am Ende der Arbeit wurde die Qualität des entstandenen

Diktionärs überprüft. Hierzu codierte ein weiterer Forscher manuell die Antworten. Diese manuelle Codierung wurde dann mit der automatischen Codierung verglichen (s. Abschnitt 9, S. 214).

Schon die Codierung des ersten Datensatzes förderte verschiedene methodische Möglichkeiten und Probleme zutage, die sich im weiteren Verlauf der Analyse in verschiedener Form wiederholten und daher als methodische Grundfragen der Diktionärskonstruktion betrachtet werden können. Es handelt sich dabei um den Zielkonflikt zwischen möglichst vollständiger und möglichst fehlerfreier Codierung, die damit verbundene Verwendung von heuristischen Indikatoren, den Umgang mit negativen Konstruktionen und mit Rechtschreibfehlern.

Ein Diktionär zum automatischen Codieren verfolgt zwei allgemeine Ziele: Erstens soll es möglichst alle codierbaren Antworten erfassen, zweitens soll es die Codes so zuweisen, dass sie die Antworten valide messen. Das Ziel einer möglichst vollständigen Codierung stellt die größere Herausforderung dar, denn das Diktionär gilt nicht nur für den Datensatz, an dem es entwickelt wurde. Mit dem entwickelten Diktionär sollen im Idealfall auch Antworten aus anderen Befragungen möglichst vollständig codiert werden. Nun können sich aber Befragungen hinsichtlich ihrer Themen, Kontexte und Zielgruppen deutlich unterscheiden. Damit gehen häufig auch Unterschiede im Sprachgebrauch einher sowie verschiedene thematische Elemente, die in den Bewertungen der jeweiligen Befragung auftauchen. Diese Differenzen stellen das automatische Codieren vor Probleme, denn das sprachliche Spektrum eines Diktionärs bildet grundsätzlich nur den Sprachgebrauch und die Themen der Befragungen ab, an denen das Diktionär bisher entwickelt und weiterentwickelt wurde.

2.4 *Verwendung heuristischer Indikatoren*

Eine Möglichkeit, die Reichweite eines Diktionärs über den Horizont der bisher integrierten Datenbasis hinaus zu gewährleisten, ist die Verwendung heuristischer Indikatoren. Darunter sind Wörter oder Wortfolgen zu verstehen, die für sich betrachtet keine bestimmte Valenz ausdrücken, aber häufig in Ausdrücken vertreten sind, die eine spezifische (positive oder negative) Bewertung enthalten. Der erste Kandidat für einen heuristischen Indikator war in unserem Fall das Wörtchen ›zu‹. Sehr viele

Antworten enthalten Formulierungen wie ›zu lang‹, ›zu allgemein‹, ›zu ausführlich‹, ›zu viele‹ oder ›zu sehr‹. Da das Wort ›zu‹ sehr häufig in Verbindung mit solchen negativen Aussagen vorkommt, ergibt sich die Möglichkeit, es der Kategorie ›negative Antwort‹ als heuristischen Indikator hinzuzufügen. Diese Methode bietet den Vorteil, die inhaltliche Reichweite des Diktionärs über die bisher darin verarbeiteten Daten hinaus auszudehnen. Die negativen Antworten in der Befragung, an der das Diktionär entwickelt wurde, könnten sich beispielsweise vor allem auf die Dauer und Ausführlichkeit der Befragung beziehen (›zu lang‹, ›zu viele‹ etc.). Dennoch könnte ein Diktionär mit dem heuristischen Indikator ›zu‹ auch viele negative Antworten erfassen, die einen ganz anderen Aspekt der Befragung kritisieren, etwa mangelnde Diskretion (›zu persönlich‹, ›zu direkt‹ etc.). Die Kehrseite dieser erweiterten Reichweite ist allerdings häufig eine zunehmende Zahl falscher Codierungen. Zwar enthalten viele negative Bemerkungen das Wort ›zu‹, aber es kann auch in vielen eindeutig positiven Antworten vorkommen, etwa ›leicht zu beantworten‹. Mit ›zu‹ als Indikator für negative Valenz würde diese Aussage als negativ codiert. Dieses Beispiel zeigt, dass Vollständigkeit und Fehlerfreiheit der Codierung gegeneinander abgewogen werden müssen. Die Verwendung heuristischer Indikatoren erhöht einerseits beträchtlich die Reichweite eines Diktionärs, macht aber zugleich das Auftreten falscher Codierungen wahrscheinlicher.

Die Frage, ob und in welchem Umfang heuristische Indikatoren verwendet werden, muss im Einzelfall auf der Basis einer abwägenden Betrachtung des Untersuchungsziels, der verwendeten Daten und der Eigenschaften des jeweiligen Indikators beantwortet werden. Nachdem sich herausgestellt hatte, dass die Verwendung des Wortes ›zu‹ eine beträchtliche Menge falscher Codes produziert, wurde es im vorliegenden Fall wieder aus dem Diktionär entfernt. Andere heuristischen Indikatoren sind aber weiterhin im Diktionär enthalten: ›Sollte‹ und ›teilweise‹ dienen als Indikatoren für negative Valenz, da sie nur in wenigen Einzelfällen falsche Codes produzieren.

2.5 *Negation von Indikatoren: das Wort nicht*

Eine weitere Herausforderung bei der Diktionärskonstruktion sind negative Formulierungen wie beispielsweise ›nicht aktuell‹, ›nicht detailliert

genug‹, oder ›nicht so leicht beantwortbar‹. Diese stellen die Diktionärskonstruktion vor mehrere Probleme: Zunächst ist klar, dass man es angesichts vorhandener negativer Konstruktionen nicht bei einer Codierung belassen kann, die sich nur auf Einzelworte stützt. Die drei obigen Beispiele würden in einem solchen Fall als positiv und damit falsch codiert. Um solche Fehler zu vermeiden, muss also das Diktionär angepasst werden. Eine mögliche Vorgehensweise hierbei ist es, neben den Adjektiven alle negativen Konstruktionen als Wortfolgen ins Diktionär zu integrieren. Diese Methode wurde im vorliegenden Fall ausprobiert, geriet jedoch schnell an ihre Grenzen. Eine Option in TEXTPACK macht es möglich, statt ganzen Wörtern nur die Wortstämme in einem Diktionär zu verwenden, wobei alle Wörter mit dem entsprechenden Stamm vom Programm erkannt und codiert werden. Das Diktionär enthält dann statt ›angenehme‹, ›angenehmer‹ und ›angenehmes‹ nur das Wort ›angenehm‹ Wenn hingegen Wortfolgen codiert werden sollen, steht diese Option nicht zur Verfügung. Wollte man also negative Konstruktionen direkt ins Diktionär integrieren, so müsste man viele verschiedene Variationen und Formulierungen aufnehmen: ›nicht so leicht‹, ›nicht so leichte‹, ›nicht so leichtes‹, ›nicht sehr leicht‹, ›nicht leicht‹ usw. Diese Vorgehensweise würde das Diktionär auf ein Vielfaches seiner Größe aufblähen, ohne dabei gewährleisten zu können, dass alle möglichen Varianten einer negativen Konstruktion erfasst werden. Das Diktionär würde zugleich sehr unübersichtlich, ohne dadurch wesentlich an Reichweite zu gewinnen.

Eine elegantere Methode im Umgang mit negativen Konstruktionen stellt die Definition des Wortes ›nicht‹ als eigene Kategorie im Diktionär dar. Bei dieser Vorgehensweise wird für jede Antwort neben den einzelnen Indikatoren für positive und negative Antworten auch das Wort ›nicht‹ eigens codiert. Im Anschluss daran muss der codierte Datensatz mittels logischer Transformationen (z. B. in SPSS mit einer Reihe von IF-Befehlen) so umgewandelt werden, dass Antworten, in denen etwa ein positives Adjektiv und das Wort ›nicht‹ vorkommen, als negativ codiert werden. Auf diese Weise bleibt die Übersichtlichkeit des Diktionärs erhalten. Außerdem können ohne weiteren Arbeitsaufwand automatisch alle negativen Formulierungen richtig erfasst werden, die im Diktionär vorhandene Adjektive verwenden. Wie gut diese Methode funktioniert, hängt vor allem von der logischen Struktur der verwendeten Transformation ab. Im Zuge der Arbeit am Diktionär wurde klar, dass auch diese

Methode in Einzelfällen zu falschen Codierungen führen kann, etwa wenn sich das ›nicht‹ nur auf eines von zwei Adjektiven in der Antwort bezieht. Dennoch sind auf diesem Weg Reichweite, Fehlerfreiheit und Übersichtlichkeit des Diktionärs am ehesten in Einklang zu bringen.

2.6 *Umgang mit alternativen Schreibweisen und Rechtschreibfehlern*

Eine weitere Herausforderung in der Arbeit am Diktionär war der Umgang mit abweichenden oder falschen Schreibweisen. Bei Online-Befragungen scheinen die Befragten auffallend wenig auf Rechtschreibung zu achten, was dazu führt, dass viele orthografisch falsche Antworten vorkommen. Eine hilfreiche Maßnahme in diesem Zusammenhang ist die Umwandlung des gesamten zu codierenden Textes in Großbuchstaben. Ohne diese Maßnahme müsste ein Diktionär für jedes einzelne Wort verschiedene Schreibweisen enthalten, die sich aus der Kombination von Groß- und Kleinbuchstaben ergeben: ›Sachlich‹, ›sachlich‹, ›sAchlich‹ ›sachlICH‹ usw. Durch eine Umwandlung des Textes in Großbuchstaben geht keine relevante Information verloren, wobei gleichzeitig die Anzahl zu berücksichtigender Schreibweisen drastisch reduziert wird. Entsprechend wird auch das Diktionär in Großbuchstaben angelegt.

Diese Maßnahme beseitigt zwar einen Großteil, aber eben nicht alle orthografischen Abweichungen, die für das automatische Codieren von Bedeutung sind. Eine weitere Art relevanter Fehler entstehen durch versehentliche, vertauschte oder fehlende Tastenanschläge, etwa ›angemehm‹, ›angenem‹ oder ›angenhem‹ statt ›angenehm‹. Grundsätzlich gibt es verschiedene Möglichkeiten, mit solchen Fehlern umzugehen, die jeweils unterschiedliche Konsequenzen für den Inhalt des Diktionärs haben. Wenn man nur orthografisch richtige Wörter ins Diktionär aufnehmen wollte, müsste man entweder alle fehlerhaften Wörter ignorieren oder den zu codierenden Text vor der Analyse einer Rechtschreibkorrektur unterziehen. Beide Möglichkeiten haben klare Nachteile: Im ersten Fall wird potenziell nutzbare Information verschenkt, im zweiten erhöht sich der Arbeits- und Zeitaufwand beträchtlich. Um diese beiden Nachteile zu vermeiden, haben wir uns für einen dritten Lösungsweg entschieden: Sofern die Wörter oder Wortfolgen trotz orthografischer Fehler noch eindeutig zu erkennen und zu interpretieren waren, wurden sie ins

Diktionär integriert. Das Diktionär enthält daher neben den eigentlichen Schlüsselwörtern auch abweichende und fehlerhafte Schreibweisen.

Ein potenzielles Problem ist die Notwendigkeit zu entscheiden, welche Wörter als noch interpretierbar beziehungsweise nicht mehr interpretierbar gelten sollen. Hier kann man sich kaum auf eine allgemeine Regel verlassen, sondern muss im Einzelfall abwägen. Im Verlauf der Diktionärskonstruktion wurde allerdings deutlich, dass solche Grenzfälle der Interpretierbarkeit nur sehr selten vorkommen. Das mag unter anderem daran liegen, dass nur eine einfache Interpretation vorgenommen wurde, nämlich die Valenz der jeweiligen Wörter oder Wortfolgen. In anderen Anwendungskontexten, wenn komplexere Inhalte erfasst werden sollen, können solche Interpretationsprobleme unter Umständen gravierender ausfallen. Im vorliegenden Fall spielten diese Probleme jedoch keine bedeutende Rolle. Sofern in seltenen Einzelfällen tatsächlich nicht entzifferbare Antworten gegeben wurden, wurden diese als nicht codierbar gewertet und wie ›keine Angabe‹ behandelt.

Auch die Befürchtung, dass sich durch diese Vorgehensweise die Einträge im Diktionär drastisch vermehren könnten, war in unserem Fall unbegründet. Der überwiegende Teil der Einträge lag nur in einer Schreibweise vor. Nur bei Wörtern, die besonders häufig geschrieben und daher öfter falsch geschrieben werden, enthält das Diktionär mehrere Schreibweisen (z. B. für ›interessant‹ und ›angenehm‹).

3. Studie 1: Anlegen des Diktionärs anhand einer quotierten Stichprobe

Ausgangspunkt des Diktionärs war der Fragebogen einer Online-Follow-Up-Befragung der allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (Allbus-Online-Follow-Up 2006, BANDILLA/KACZMIREK/BLOHM/NEUBARTH 2008). Der ALLBUS selbst wird alle zwei Jahre persönlich-mündlich mit unterschiedlichen Themenschwerpunkten in Deutschland durchgeführt. Die Stichprobe besteht dabei aus einer repräsentativen Auswahl der erwachsenen Wohnbevölkerung. Der thematische Schwerpunkt 2006 bestand in Einstellungen gegenüber ethnischen Gruppen. Vor dem eigentlichen ALLBUS Online-Follow-Up (Studie 2) mit Teilnehmern aus dem ALLBUS wurde Studie 1 mit Teilnehmern aus dem Panel der Respondi AG durchgeführt. Die Stichprobe war quotiert nach der

Verteilung im ALLBUS hinsichtlich der Merkmale Geschlecht, Alter und Schulbildung unter der Bedingung, dass das Internet genutzt wird. Das Durchschnittsalter der Teilnehmer betrug 40 Jahre, 53,6 Prozent waren Männer und 46,4 Prozent Frauen. Es wurden 1000 Personen zur Teilnahme an der Online-Befragung eingeladen, von denen 535 den Fragebogen beendeten. 506 Personen (50,6 %) beantworteten die Frage »Ich empfand den Fragebogen als...«.

Diese Antworten bilden die erste Grundlage zur Erstellung des Diktionärs. Dazu wurde mit TEXTPACK eine Liste aller Wörter erstellt, die in den Antworten vorkommen. Diese Wortliste wurde neu formatiert, jeder Buchstabe wurde in einen Großbuchstaben umgewandelt. Eine solche Liste eignet sich sehr gut als Ansatzpunkt bei der Konstruktion eines Diktionärs: Statt aus den Antworten jeweils die einzelnen codierbaren Wörter herauszusuchen und manuell ins Diktionär zu übertragen, kann man die Wortliste einfach auf diejenigen Wörter reduzieren, die codierbar sind. Im zweiten Schritt werden dann die codierbaren Wörter den einzelnen Kategorien des Diktionärs zugeordnet. Dadurch wird sichergestellt, dass alle codierbaren Wörter in allen vorkommenden Schreibweisen ins Diktionär eingefügt werden; darüber hinaus werden Schreibfehler vermieden, die bei der Übertragung vorkommen könnten.

Nachdem alle codierbaren Wörter und Wortfolgen integriert worden waren, wurde das Diktionär zum ersten Mal zum automatischen Codieren verwendet. Diese erste Anwendung war vergleichsweise trivial, da genau diejenigen Antworten codiert wurden, auf deren Basis das Diktionär entwickelt worden war. Dennoch erwies sich dieser erste Codierungsdurchgang hier und im weiteren Verlauf der Arbeit als nützlich.

Schon bei der Erstellung des Diktionärs fiel auf, dass das Wort ›zu‹ sehr häufig in negativen Antworten verwendet wurde. Daher wurde es schon in das erste Diktionär probeweise als heuristischer Indikator eingefügt. Nach dem ersten automatischen Codieren zeigte sich jedoch, dass durch die Verwendung von ›zu‹ sehr viele falsche Codierungen zustande kamen. Daher wurde ›zu‹ wieder aus dem Diktionär entfernt.

Des Weiteren wurde ebenfalls deutlich, dass negative Konstruktionen sowie (einfache und doppelte) Verneinungen in vielen Antworten enthalten sind und regelmäßig zu fehlerhafter Codierung führen. Als Lösungsstrategie wurde zunächst versucht, alle in den Daten enthaltenen negativen Formulierungen als Wortfolgen ins Diktionär einzufügen. Dazu gehörten beispielsweise: ›nicht leicht verständlich‹ und ›nicht unangenehm‹.

Nach diesen Änderungen am Diktionär wurde der Datensatz erneut codiert. Das automatische Codieren liefert verschiedene Statistiken, die über die Bewertung der Befragung Aufschluss geben und Hinweise auf die Qualität des Diktionärs und mögliche Verbesserungen liefern. Interessant sind vor allem die prozentualen Anteile der verschiedenen Kategorien sowie der nicht codierten Antworten im gegebenen Datensatz, wie sie Tabelle 1 zeigen.

TABELLE 1
Anteile codierter Kategorien in Studie 1

	Anzahl	Prozent	valide Prozent
negativ	90	9,0	17,8
positiv	348	34,8	68,8
ambivalent	23	2,3	4,5
nicht codiert	45	4,5	8,9
total 1	506	50,6	100,0
neutrale Ausfälle und ›keine Angabe‹	494	49,4	
total 2	1000	100,0	

Die im zweiten Durchgang vergebenen Codes wurden manuell auf Fehler überprüft. Dabei kamen beispielsweise folgende falsche Codierungen zum Vorschein:

»*zu wenig detailliert, eigene Antworten waeren wichtig*«: Die Adjektive ›detailliert‹ und ›wichtig‹ wurden als positiv codiert, ›waeren‹ und ›zu wenig‹ als negativ. Dies führte dazu, dass diese eindeutig negative Antwort als ambivalent codiert wurde. Dennoch erschien es uns nicht sinnvoll, das Diktionär zu ändern, da die manuelle Überprüfung gezeigt hat, dass außer in solchen Ausnahmefällen größtenteils die richtige Kategorie zugewiesen wurde.

»*beantwortenswert und man sollte die Problematik öfter behandeln*«: Hier wurde ›beantwortenswert‹ als positiv, das Wort ›sollte‹ als negativ codiert. Daraus ergab sich insgesamt ein ambivalenter Code, die Antwort drückt jedoch eine positive Bewertung aus. Obwohl ›sollte‹ im Großteil aller Fälle ein verlässlicher heuristischer Indikator für negative Bewertung ist, führte es hier zu einer falschen Codierung. Dieses Beispiel verweist auf den oben diskutierten Zielkonflikt zwischen Reichweite und

Fehlerfreiheit des Diktionärs. Angesichts der Tatsache, dass ›sollte‹ nur sehr selten falsche Codes produzierte, wurde es im Diktionär belassen.

4. Studie 2: Ergänzung des Diktionärs anhand einer ALLBUS-Online-Follow-Up-Befragung

Das Diktionär wurde im nächsten Schritt auf eine zweite Stichprobe mit dem gleichen Fragebogen angewandt. Teilnehmer in Studie 2 waren Personen, die sich nach dem ALLBUS bereit erklärt hatten, an einer Online-Follow-Up-Befragung teilzunehmen (BANDILLA/KACZMIREK/BLOHM/NEUBARTH 2008). Von den 517 eingeladenen Personen beendeten 345 den Fragebogen, wovon 327 (63,2 %) die offene Frage zur Fragebogenbewertung beantworteten. Das Durchschnittsalter betrug 40,6 Jahre. 57,1 Prozent der Teilnehmer waren Männer, 42,9 Prozent Frauen.

Diese Anwendung stellte nur eine moderate Herausforderung an das Diktionär dar, da Thema, Kontext und Zielgruppe der Befragung gleich blieben. Größere Änderungen im verwendeten Wortschatz, im Sprachgebrauch oder den angesprochenen Themen waren daher nicht zu erwarten. Der erste Codierungsdurchgang lieferte folgende Ergebnisse: Insgesamt wurden 27,1 Prozent der Antworten als negativ, 56,0 Prozent als positiv und 6,5 Prozent als ambivalent codiert. 10,5 Prozent der Antworten konnten im ersten Durchgang nicht codiert werden.

Wiederum wurden die nicht codierten Antworten durchgesehen und auf eventuell codierbare Wörter oder Wortfolgen untersucht. Einige neue Einträge wurden dem Diktionär hinzugefügt. Anschließend wurde der Datensatz mit dem erweiterten Diktionär ein zweites Mal codiert, die Codes wurden auf Fehler überprüft. Es folgen einige Beispiele für falsche Codes und daraus resultierende Überlegungen und Änderungen am Diktionär.

Die Antwort »teilweise überraschend« wurde wegen des heuristischen Indikators ›teilweise‹ als negativ codiert. Die eigentliche Valenz der Antwort ist allerdings unklar, denn ›überraschend‹ kann sowohl eine positive als auch eine negative Eigenschaft einer Befragung sein (positiv im Sinne von hohem Unterhaltungswert, negativ im Sinne von Irritation oder Verstörung). Es wurden schließlich keine Änderungen am Diktionär vorgenommen, einerseits wegen dieser Ambivalenz, andererseits wegen der insgesamt hohen Verlässlichkeit des heuristischen Indikators ›teil-

weise‹. Auch dies ist ein Beispiel für die Abwägungsentscheidungen, die bei der Konstruktion eines Diktionärs ständig anfallen.

Die Antwort »nicht genau genug...« wurde wegen ›genau‹ als positiv codiert. Als Konsequenz wurde ›nicht genau‹ ins Diktionär integriert.[3] In ähnlicher Weise wurde die Antwort »manchmal so nicht richtig zu beantworten« aufgrund des Wortes ›richtig‹ als positiv codiert und konsequenterweise wurde ›nicht richtig‹ ins Diktionär aufgenommen.

Die Antwort »sachlich, interessant, zuviel interbretationsspielraum« (sic.) wurde wegen ›sachlich‹ und ›interessant‹ als positiv codiert. Eigentlich handelt es sich aber um eine ambivalente Bewertung. Entsprechend wurden ›zu viel‹ und ›zuviel‹ ins Diktionär integriert.

Die Bemerkung »nicht ausreichend für eine ordentliche Analyse« wurde wegen des Wortes ›ordentlich‹ als positiv codiert. Entsprechend wurde ›nicht ausreichend‹ in das Diktionär aufgenommen, was im Ergebnis zu einer Codierung als ambivalent führt. Um eine richtige Codierung zu gewährleisten, müsste das Wort ›ordentlich‹ aus dem Diktionär entfernt werden. Dagegen spricht, dass die Adjektive, die in den Antworten vorkommen, sich in der Regel auf die Qualität der Befragung beziehen (im Sinne von »Ich empfand die Fragen als ordentlich«). Die obige Antwort bildet hier eine seltene Ausnahme. Aus diesem Grund wurde trotz der falschen Codierung ›ordentlich‹ als Indikator für eine positive Bewertung beibehalten.

In einigen der obigen Beispiele wird deutlich, wie häufig negative Konstruktionen zu falschen Codierungen führen. Bis zu diesem Punkt unserer Arbeit hatten wir die Strategie verfolgt, vorhandene negative Konstruktionen als Wortfolgen ins Diktionär zu integrieren. An dieser Stelle wurde jedoch deutlich, dass diese Methode aus zwei Gründen nicht befriedigend ist: Erstens kann man auf diesem Weg die Reichweite eines Diktionärs nur in sehr kleinen Schritten erhöhen. Da praktisch jede positive Bewertung durch ein hinzugefügtes ›nicht‹ in eine negative verwandelt werden kann, würde es mit dieser Methode sehr lange dauern, bis alle relevanten negativen Konstruktionen ins Diktionär Eingang finden. Eine Möglichkeit, diesen Sättigungsprozess zu beschleunigen, wäre es, alle denkbaren negativen Konstruktionen ins

3 Bei der Codierung räumt TEXTPACK Wortfolgen den Vorrang vor einfachen Wörtern ein. Wenn das Diktionär also die Einträge ›genau‹ (positiv) und ›nicht genau‹ (negativ) enthält, so wird die Antwort ›nicht genau genug‹ als negativ codiert.

Diktionär aufzunehmen, ohne darauf zu warten, bis sie in einem der Datensätze auftauchen. Dazu müsste man erstens alle positiven Adjektive mit einem ›nicht‹ versehen und als Indikatoren für negative Valenz ins Diktionär integrieren, zweitens alle negativen Adjektive mit einem vorangehenden ›nicht‹ als positive Indikatoren aufnehmen. Damit wären zwar einige, aber bei Weitem nicht alle negativen Konstruktionen abgedeckt. Denn viele Antworten enthalten zwischen dem ›nicht‹ und dem jeweiligen Adjektiv weitere Wörter, etwa: ›nicht zu langweilig‹, ›nicht besonders unterhaltsam‹, ›nicht so gut‹ etc. Um wirklich alle Konstruktionen mit ›nicht‹ zu erfassen, müsste man daher auch alle diese Varianten ins Diktionär integrieren. Diese Überlegung führte uns zum zweiten gravierenden Nachteil der bisher verfolgten Strategie: Das Diktionär würde auf ein Vielfaches seiner bisherigen Größe ausgeweitet. Dadurch würden Korrekturen und Modifikationen, die schon bei mittleren Diktionären mit einigen hundert Einträgen aufwendig sind, praktisch unmöglich gemacht. In der Flut der negativen Konstruktionen würden Übersichtlichkeit und Nachvollziehbarkeit der Diktionärsstruktur zu sehr abnehmen.

Aufgrund dieser Bedenken entschlossen wir uns, den oben bereits erwähnten Ansatz zu testen: Das Wort ›nicht‹ wurde als eigenständige Kategorie ins Diktionär integriert. Dies hat zur Folge, dass bei jeder Antwort neben den bisherigen Codes auch für ein eventuell vorhandenes ›nicht‹ ein eigener Code vergeben wird. Im resultierenden Datensatz sind damit alle Antworten, die das Wort ›nicht‹ enthalten, klar zu identifizieren. Dies ermöglicht es, mithilfe einer logischen Transformation den Datensatz so umzuformen, dass etwa bei gleichzeitigem Vorhandensein eines positiven Indikators und des ›Nicht‹-Indikators ein negativer statt eines positiven Codes vergeben wird und umgekehrt.

Die ›Nicht‹-Kategorie wurde zum Diktionär hinzugefügt, im Gegenzug wurden alle bisher enthaltenen Wortfolgen mit ›nicht‹ entfernt. Dieses modifizierte Diktionär verwendeten wir nun zur Codierung der ersten beiden Datensätze. Dabei stellte sich heraus, dass die neue Methode nur wenige und größtenteils erfreuliche Abweichungen produzierte: Im ersten Datensatz wurden zehn, im zweiten Datensatz nur acht Antworten anders als bisher codiert. Von diesen Antworten waren die meisten zuvor falsch oder gar nicht codiert worden und konnten nun mithilfe der neuen Kategorie richtig erfasst werden. Diese Ergebnisse sprachen für die Vorteile des Ansatzes und veranlassten uns, ihn beizubehalten.

Die abschließende Codierung der zweiten Studie lieferte folgende Ergebnisse: 27,1 Prozent negative, 56,9 Prozent positive und 4,6 Prozent ambivalente Antworten. 11,4 Prozent der Antworten konnten nicht codiert werden.

5. Studie 3: Erste Anwendung des Diktionärs in einer Wertestudie unter Studierenden

Die nun folgende Anwendung des Diktionärs auf eine neue Befragung stellte einen ersten Härtetest dar, denn hier handelte es sich um eine Befragung, deren Thema, Kontext und Zielgruppe sich deutlich von den ersten beiden Studien unterschied. Dementsprechend waren auch Änderungen in den angesprochenen Qualitätsmerkmalen und Themen der Befragung sowie im Sprachgebrauch und Wortschatz der Befragten zu erwarten. Bei Studie 3 wurden Studierende zu ihren Werturteilen bezüglich Berufstätigkeit, Kinder und Familie befragt. Die Teilnehmer ließen daher einen höheren Anteil von Jugend- und Umgangssprache erwarten. Die Befragung wurde offen per E-Mail beworben, eine Responserate kann daher nicht sinnvoll berechnet werden. Insgesamt begannen 776 Personen mit der Teilnahme, wovon 536 Teilnehmer ein Urteil zur Fragenbewertung abgaben (37,7 % Männer, 62,3 % Frauen, Durchschnittsalter 28,5 Jahre).

Die Frage war nun, wie gut das Diktionär in seiner bisherigen Form auch die Antworten dieser neuen Befragung erfassen könnte. Der erste Codierungsdurchgang lieferte 17,2 Prozent negative, 59,5 Prozent positive und 2,6 Prozent ambivalente Antworten. 20,7 Prozent der Antworten konnten nicht codiert werden.

Der vergleichsweise hohe Anteil nicht codierter Antworten verweist auf die angesprochenen Probleme: Differenzen im Sprachgebrauch und Wortschatz führten dazu, dass viele potenziell codierbare Antworten vom Diktionär in seiner bisherigen Form nicht erfasst werden konnten. Dieses zunächst wenig erfreuliche Ergebnis stellt bei genauerer Betrachtung eine wichtige Chance zur Verbesserung des Diktionärs dar. Denn durch Integration der bisher nicht codierten Antworten wird die Reichweite, die sich bisher auf den durchschnittlichen Wortschatz und Sprachgebrauch der ALLBUS-Befragten beschränkte, auf die Sprech- und Schreibgewohnheiten von Studenten und Jugendlichen ausgedehnt. Der Vorteil der diktionärbasierten Codierung ist, dass jede solche Erweiterung nur

einmal vorgenommen werden muss, dann aber bei jeder weiteren Anwendung des Diktionärs zu dessen Effektivität beiträgt.

Die nicht codierten Antworten wurden durchgesehen und – soweit codierbar – ins Diktionär integriert. Es folgen einige Beispiele für Antworten mit ungewöhnlichem Sprachgebrauch: ›aus dem vorigen Jahrhundert!‹, ›cool‹, ›blöd‹, ›doppelt gemoppelt‹, ›doof‹, ›innordnung‹ (sic).

Nach den Änderungen am Diktionär wurde der Datensatz erneut codiert. Die Codierung lieferte folgende Ergebnisse: 27,2 Prozent negative, 63,2 Prozent positive und 3,0 Prozent ambivalente Antworten. 6,4 Prozent der Antworten konnten nicht codiert werden. Die Abnahme der nicht codierten Antworten von 20,7 Prozent auf 6,4 Prozent ist ein deutliches Zeichen für die maßgebliche Erweiterung des Diktionärs.

6. Studie 4: Zweite Anwendung in einer Studie über Partnerschaften

Der Fragebogen zu Studie 4 behandelte das Thema Zufriedenheit in der Partnerschaft. Die Teilnehmer wurden per Newsletter und über die Webseite von dem Partnervermittlungsportal *DerZweiteFrühling* eingeladen. Von den 1420 Personen, die mit der Teilnahme begannen, beendeten 537 Teilnehmer den Fragebogen. 511 Teilnehmer beantworteten die Frage »Ich empfand die Fragen als...«, davon waren 43,4 Prozent Männer und 56,6 Prozent Frauen. Das Durchschnittsalter lag bei 45,1 Jahren.

Auch bei der ersten Codierung dieser Befragung war der Anteil der nicht codierten Antworten zunächst relativ hoch. Ursache dafür war die bisher nicht erfasste Befragtengruppe und das besonders persönliche Thema der Befragung. Beide Aspekte lassen einen vom Durchschnitt abweichenden Sprachgebrauch erkennen, der zu zahlreichen noch nicht im Diktionär enthaltenen Antworten führte. Die erste Codierung lieferte 16,9 Prozent negative, 55,9 Prozent positive und 1,6 Prozent ambivalente Antworten. 25,7 Prozent der Antworten konnten im ersten Durchgang nicht codiert werden.

Folgende Antworten können als Beispiele für die Besonderheiten im Sprachgebrauch dieser Befragtengruppe betrachtet werden: »wenig auf Ist-Zustand ausgerichtet«, »positive Übung zum Nachdenken«, »in meine Spiegelbild sehen«, »etwas abgedroschen«. Wiederum wurden die nicht codierten Antworten durchgesehen, das Diktionär wurde entspre-

chend erweitert. Anschließend wurde der Datensatz mit dem erweiterten Diktionär erneut codiert. Die Codierung führte zu folgenden Ergebnissen: 23,4 Prozent negative, 60,6 Prozent positive und 2,0 Prozent ambivalente Antworten. Mit dem erweiterten Diktionär blieben nur noch 14,1 Prozent nicht codierbare Antworten übrig.

Die Anzahl der nicht codierten Antworten konnte also um mehr als die Hälfte verringert werden. Dennoch bleiben auch in diesem Fall einige Antworten übrig, die aus verschiedenen Gründen nicht codiert werden können oder bei denen eine Aufnahme ins Diktionär wenig sinnvoll erscheint. Nach den Erfahrungen mit den bisherigen vier Studien waren wir nun in der Lage, vier verschiedene Typen nicht codierbarer Antworten zu identifizieren:

Erstens gibt es Ausdrücke, die zwar eine Bewertung enthalten, jedoch nicht eindeutig interpretiert werden können: ›ausführlich‹, ›ziemlich direkt‹, ›umfangreich‹, ›sehr intensiv‹, ›harmlos‹, ›eher unverbindlich‹, ›Überraschung‹. Bei diesen Antworten kann man davon ausgehen, dass der Befragte eine bestimmte Bewertung der Befragung ausdrücken wollte. Allerdings kann jede Antwort sowohl negativ als auch positiv gemeint sein (›ausführlich‹ und ›umfangreich‹ im Sinne von Vollständigkeit oder von Überlänge, ›sehr intensiv‹ im Sinne von Unterhaltsamkeit oder von emotionalem Stress). Diese Ambivalenz macht Antworten dieser Art uncodierbar.

Zweitens gibt es Antworten, die sehr eng mit dem Thema der Befragung zusammenhängen. In diesen Fällen ist eine Aufnahme in das Diktionär nicht sinnvoll, da die enthaltenen Bewertungen nur in eingeschränktem Maße oder überhaupt nicht auf andere Befragungen übertragbar sind. Beispiele: »beziehungsrelevant«, »Aufrischung, die Beziehung zu reflektier« (sic).

Drittens kommen Antworten vor, die unvollständig sind oder in keinem klaren Bezug zur Bewertung der Umfrage stehen, beispielsweise: »Mann i.d. sogenannten Auszeit lebend«, »bin Singel u. liebe einen verh. Mann«, »bin immer herliech«.

Viertens gibt es Antworten, die zwar eine Bewertung enthalten, aber ungewöhnlich formuliert sind, sodass klare Schlüsselwörter fehlen, die ins Diktionär aufgenommen werden könnten, beispielsweise: »sehr auf den ›Jetztzustand‹ projeziert«, »für zu alte Zielgruppe gestellt«. Es ist kaum sinnvoll, solche Antworten als Wortfolgen ins Diktionär zu integrieren, weil die Wahrscheinlichkeit sehr gering ist, dass genau diese Formulierungen in späteren Befragungen erneut vorkommen.

7. Studie 5: Dritte Anwendung in einer Themenbefragung über Handball

Studie 5 war eine offene Themenbefragung der Respondi AG zum Thema Handball. Von den 2352 Teilnehmern, die die Befragung begonnen, beendeten 1349 Teilnehmer den Fragebogen. Davon gaben 878 Teilnehmer eine Beurteilung der Fragen ab. 57,6 Prozent der Teilnehmer waren Männer, 42,4 Prozent Frauen. Der Median des Alters lag in der Gruppe der 30- bis 39-Jährigen. (Da in den Studien 5 und 6 das Alter in Kategorien erhoben wurde, kann hier kein Mittelwert berichtet werden.)

Zunächst wurde auch dieser Datensatz mit dem bisher entwickelten Diktionär codiert. Dies führte zu 5,3 Prozent negativen, 81,1 Prozent positiven und 0,8 Prozent ambivalenten Antworten. 12,7 Prozent der Antworten konnten nicht codiert werden. Auffällig war in dieser Studie der relativ geringe Prozentsatz nicht codierter Antworten schon im ersten Durchgang der Codierung. Bei den vorhergehenden Studien waren im ersten Durchgang bis zu 25 Prozent der Antworten nicht codiert worden. Wir interpretierten dieses Ergebnis als Hinweis darauf, dass das Diktionär sich einer gewissen Sättigungsstufe näherte und dass daher in folgenden Studien immer weniger Ergänzungen nötig sein würden. Diese zunehmende Sättigung zeigte sich, obwohl sich im Vergleich zu Studie 4 der Kontext (Online-Panel) und das Thema der Befragung (Handball) deutlich gewandelt hatten.

Dennoch gab es auch unter den nicht erfassten Antworten dieses Datensatzes einige codierbare Wörter und Wortfolgen, die ins Diktionär aufgenommen wurden. Anschließend wurde der Datensatz ein zweites Mal codiert. Die Codierung lieferte 6,8 Prozent negative, 86,1 Prozent positive und 0,8 Prozent ambivalente Antworten. Nur 6,3 Prozent der Antworten konnten mit dem erweiterten Diktionär nicht codiert werden.

8. Studie 6: Vierte Anwendung in einer Themenbefragung über Hunde und Katzen

Studie 6 war eine offene Themenbefragung der Respondi AG zum Thema Hunde und Katzen. Von den 3830 Teilnehmern, die die Befragung begonnen, beendeten 3152 Teilnehmer den Fragebogen. Davon gaben 1391 Teilnehmer eine Beurteilung der Fragen ab. 19 Prozent der Teilnehmer waren

Männer, 81 Prozent Frauen. Der Median des Alters lag in der Gruppe der 25- bis 29-Jährigen.

Auch die letzte Studie wurde zunächst mit dem bisherigen Diktionär codiert und ergab 6,4 Prozent negative, 80,9 Prozent positive und 1,2 Prozent ambivalente Antworten. 11,4 Prozent der Antworten wurden im ersten Durchgang nicht codiert. Nach der Durchsicht der nicht codierten Antworten und einigen neuen Einträgen im Diktionär wurde der Datensatz ein zweites Mal codiert. Die Codierung lieferte folgende Ergebnisse: 8,2 Prozent negative, 85,1 Prozent positive und 2,0 Prozent ambivalente Antworten. Lediglich 5,0 Prozent der Antworten blieben uncodiert.

Nach Einarbeitung dieser letzten Befragung scheint das Diktionär die Sättigungsstufe, die sich in der vorherigen Studie bereits andeutete, erreicht zu haben. Dafür sprechen zwei Hinweise.

Erstens waren ein Großteil der neu einzutragenden Wörter andere Schreibweisen oder fehlerhafte Versionen bereits vorhandener Wörter. Zum Teil waren die nicht codierten Antworten so voller Rechtschreibfehler, dass man nur mit Mühe ihre Bedeutung erahnen konnte. Dies war besonders bei häufig verwendeten Wörtern der Fall. Inzwischen enthält das Diktionär beispielsweise dreizehn Schreibweisen von ›interessant‹ und acht Schreibweisen von ›angenehm‹. Es ist abzusehen, dass bei zukünftigen Erweiterungen des Diktionärs bald eine Grenze erreicht würde, bei der ein Großteil der nicht codierten Antworten orthografisch soweit degeneriert ist, dass man sie nicht mehr ins Diktionär aufnehmen kann. Dieser Befund verweist darauf, dass die eingangs gestellte Frage nach dem Umgang mit Rechtschreibfehlern im Verlauf der Diktionärskonstruktion an Bedeutung zunimmt.

Zweitens lagen die neu aufgenommenen Wörter immer weiter vom durchschnittlichen Sprachgebrauch entfernt. Es handelte sich dabei entweder um Wortkreationen der Befragten (›hilfsweisend‹, ›nachdenkenswert‹ etc.), um Umgangssprache (›cool‹; ›scheiße und behindert‹) oder um ausgefallene, kaum gebräuchliche Wörter (›variantenreich‹, ›lapidar‹, ›holzschnittartig‹). Auch hier stellt sich die Frage, wie man auf lange Sicht mit diesen Antworttypen umgehen kann. Insbesondere die Neologismen scheinen problematisch, denn Wörter wie ›hilfsweisend‹ können ohne Weiteres auch als Rechtschreib- oder Denkfehler betrachtet werden. Nimmt man sie in das Diktionär auf, werden solche Antworten ebenso gewertet wie etwa ›hilfreich‹ (was hier wohl gemeint ist).

Insgesamt sind diese Sättigungserscheinungen eindeutig positiv zu bewerten. Sie sind deutliche Indizien dafür, dass das Diktionär inzwi-

schen einen Großteil dessen enthält, was ein Befragter mit durchschnittlichem Sprachgebrauch zu dieser Frage sagen kann. Angesichts dieser Ergebnisse kann man mit Recht erwarten, dass die Leistungsfähigkeit des Diktionärs von Veränderungen in Themen, Kontexten oder Zielgruppen von Befragungen kaum mehr beeinträchtigt ist.

9. Qualität des Diktionärs: Reliabilität und Validität der Codierung

Nach Abschluss der Arbeiten am Diktionär wurden noch einmal alle Studien mit dem nun fertiggestellten Diktionär codiert. Danach wurde die Reliabilität und Validität der Codierung überprüft. Unter der Reliabilität versteht man die Zuverlässigkeit der Einordnungen der Texte in die vorgegebenen Kategorien. Bei einer computerunterstützten Inhaltsanalyse ist die Reliabilität immer zu 100 Prozent gegeben, denn Kategorien und Codierregeln sind im Diktionär festgelegt, und der Computer codiert immer genau nach diesen Regeln, d. h., auch bei mehrfacher Codierung eines Textes mit demselben Instrument wird immer dasselbe Ergebnis erreicht.

Reliabilitätstests sagen nichts über die Qualität der Kategorien aus, sondern nur über die Messvorschriften und deren Anwendung. Dagegen ist die Validität einer Codierung ein »inhaltsanalytischer Qualitätsstandard, der angibt, ob die Codierungen (also die produzierten Daten) den in der Forschungsfrage anvisierten Bedeutungsgehalt (das zu messende theoretische Konstrukt) tatsächlich messen« (FRÜH 2007: 196). Es muss möglich sein, eine eindeutige Beziehung zwischen den codierten Daten und der Forschungsfrage herzustellen (*face validity*).

Während der Entwicklung und Verbesserung des Diktionärs in den verschiedenen Entwicklungsstufen wurde die Qualität des Diktionärs, d. h. dessen Gültigkeit, ständig überprüft und verbessert. Nach Abschluss der Codierung muss das Ergebnis jedoch validiert werden, um die Qualität der Codierungen zu bewerten. Früh (2007: 197f.) schlägt vor, dass an der Überprüfung der Validität der Codierung des Textmaterials der Forscher selbst beteiligt wird, denn er weiß selbst am besten, was seine Kategorien messen sollen. Geprüft wird also die Übereinstimmung der Codierungen des Forschers mit denen der computerunterstützten Inhaltsanalyse. Der Grad der Übereinstimmung kann mithilfe der glei-

chen Methoden überprüft werden, wie sie auch für die Berechnung von Reliabilitätskoeffizienten verwendet werden. Da die automatische Codierung mit der intendierten ›wahren‹ Codierung verglichen wird, handelt es sich nun um Maße der Validität anstelle der Reliabilität.

Dazu haben wir das folgende einfache Ähnlichkeitsmaß, wie es Früh (2007) vorschlägt, verwendet

$$CR = \frac{2 \times \text{Anzahl der Übereinstimmungen}}{\text{Anzahl der Codierungen Forscher} + \text{Anzahl der Codierungen cui}}$$

Codiert wurde, ob eine Antwort negativ, positiv oder ambivalent ist oder ob sie nicht codierbar oder ein neutraler Ausfall ist. Codiert wurden jeweils 4153 Antworten. Hierbei erfolgte die Erstellung des Diktionärs durch einen Forscher und ein anderer Forscher führte eine manuelle Codierung der unterschiedlichen Antworten durch. Hierbei entscheidet der Codierer für jede Antwort, welcher der oben definierten Kategorien die Antwort zuzuordnen ist. Automatisch und manuell wurden übereinstimmend 3680 Antworten codiert. Dies ergibt einen Validitätskoeffizienten von 0,886. Ein Wert größer als 0,85 spricht für eine gute Validität der Codierung bei solch relativ einfachen Kategoriendefinitionen wie im vorliegenden Fall. Tabelle 2 zeigt die Übereinstimmung der beiden Codierungen bei der Zuweisung der verschiedenen Kategorien. Dabei zeigt sich bei der Codierung positiver Antworten eine sehr hohe Übereinstimmung (98,2 %). Auch bei der Codierung von negativen Antworten wird noch eine Übereinstimmung von 87,0 Prozent erreicht. Es zeigt sich aber auch, dass viele der bei der computerunterstützten Inhaltsanalyse nicht codierbaren Antworten bei einer manuellen Codierung zugeordnet werden können.

Überprüft man nur Codierungen, die zu ›negativen‹ oder ›positiven‹ Codierungen führten, und ignoriert alle anderen Codierungen, z. B. ›nicht codierbar‹ oder ›neutraler Ausfall‹, erhält man bei 3986 Codierungen des Forschers und 3708 automatisch codierten Antworten einen Validitätskoeffizienten von 0,951. Das bedeutet, dass für die beiden zentralen Kategorien die Validität höher ausfällt, als wenn nachrangige Kategorien mit berücksichtigt werden.

Trotz dieser guten Ergebnisse lohnt es sich, die Fehlcodierungen genauer zu überprüfen. Probleme machen hier vor allem Negationen. Im Diktionär wurde – wie oben beschrieben – eine Antwort als positiv

TABELLE 2

Validitätskontrolle. Vergleich der Codierungen des Forschers mit der automatischen Codierung

		Manuelle Codierung durch Forscher				
Automatische Codierung		negativ	positiv	ambivalent	andere	total
negativ	Anzahl	517	52	13	12	594
	% Forscher	87,0%	8,8%	2,2%	2,0%	100%
positiv	Anzahl	32	3058	24	0	3114
	% Forscher	1,0%	98,2%	0,8%	0,0%	100%
ambivalent	Anzahl	18	20	51	2	91
	% Forscher	19,8%	22,0%	56,0%	2,2%	100%
andere	Anzahl	84	205	11	54	354
	% Forscher	23,7%	57,9%	3,1%	15,3%	100%

und/oder negativ codiert und zusätzlich codiert, ob eine Negation vorliegt. In einem zweiten Schritt wurde dann als Regel ›positiv‹ + ›Negation‹ ergibt ›negativ‹ bzw. ›negativ‹ + ›Negation‹ ergibt positiv angewendet. Wird aber in einer Antwort sowohl ›positiv‹ als auch ›negativ‹ zusammen mit ›Negation‹ codiert, kann nicht mehr eindeutig festgestellt werden, zu welcher Aussage die Negation gehört. Die Antwort wird ›nicht codierbar‹. Ein menschlicher Codierer dagegen ist in der Lage, die Antwort zu verstehen und entsprechend zu codieren. Ein Beispiel für solch eine Antwort ist: »sehr angenehm [positiv] und nicht [Negation] zeitaufwendig [negativ]«. Zur Verbesserung der Codierung besteht u.a. die Möglichkeit, die Antwort eines Befragten in Aussagen aufzuteilen und getrennt zu codieren. Die Antwort »sehr angenehm und nicht zeitaufwendig« würde dann aufgeteilt in zwei Aussagen des Befragten ›sehr angenehm‹ und ›nicht zeitaufwendig‹, die nun auf der Basis des Diktionärs eindeutig codiert werden könnten.

Betrachtet man weitere Unterschiede in der Codierung, so zeigt sich, dass relativ häufig die Antwort »sehr persönlich« gegeben wurde. Diese Aussage wurde in der computerunterstützten Inhaltsanalyse als ›negativ‹ codiert, bei der manuellen Codierung aber als ›positiv‹. Man kann sich natürlich fragen, was ein Befragter empfindet, wenn er antwortet, dass eine Frage ›sehr persönlich‹ sei. Empfindet er es als freundlich, dass man ihn sehr persönlich anspricht, oder fühlt er sich zu persönlich angesprochen? Ähnlich verhält es sich mit der Antwort, dass die Fragen

›durchschnittlich‹ seien. Auch diese Antwort wurde manuell als ›positiv‹ codiert, im Diktionär aber als ›negativ‹ eingeordnet. Hieran lässt sich grundsätzlich erkennen, dass ein Diktionär den Interpretationen der Forscher unterliegt. Die hier genannten Einzelfälle, in denen die Interpretation unklar bleiben muss, wurden daher im abschließenden Diktionär als nicht codierbar markiert und tragen damit zu einer kleineren Validitätsverbesserung bei.

Nach der erfolgreichen Erstellung, Überprüfung und Validierung des Diktionärs werden im nächsten Abschnitt die Ergebnisse der Codierungen im Ganzen betrachtet.

10. Ergebnisse: Wie wird kommentiert, von wem und was?

In diesem Abschnitt wird zunächst die Frage beantwortet, wie kommentiert wurde bzw. welche Anteile positive, ambivalente und negative Angaben ausmachen. Danach werden die Bewertungen von Teilnehmersubgruppen nach Geschlecht, Alter und Bildungsgrad anhand eines Zufriedenheitsindexes dargestellt. Abschließend zeigt ein Überblick, welche Bewertungen erfolgten bzw. was bewertet wird.

Bereits in den vorherigen Abschnitten wurde deutlich, dass Teilnehmer mehrheitlich positive Phrasen zur Satzergänzung von »Ich empfand die Fragen als...« verwendeten. Tabelle 3 zeigt die Anteile der computerunterstützten Codierung in positive, ambivalente und negative Angaben. Die Anzahl fehlender Codierungen (Missings) ergibt sich aus der Summe der nicht codierbaren Angaben (uncodiert) und der korrekt identifizierten neutralen Ausfälle. Deutlich erkennbar ist der hoch signifikante Unterschied im Anteil positiver und negativer Angaben zwischen den Studien, $\chi^2(df=5, N=5)=2{,}7$, $p<0{,}001$. Betrachten wir die Zielsetzungen, Inhalte und Stichproben der Studien, so werden die Unterschiede verständlich. Der höchste Grad positiver Angaben lag in den Studien 5 und 6 mit den Themen *Handball* sowie *Hunde und Katzen*. Beide Studien waren Themenumfragen, bei denen der Unterhaltungswert für die Teilnehmer bei der Konstruktion der Fragen besondere Beachtung geschenkt wurde. Die Teilnahme war offen für alle Interessierten. Demnach werden diese beiden Studien vorwiegend Teilnehmer angelockt haben, die auch an den jeweiligen Themen interessiert waren. Entsprechend positiv fiel

die Bewertung aus. Die Stichprobe zu Studie 1 beruhte auf dem gleichen Panel wie die Studien 5 und 6. Demgegenüber wurden in Studie 1 jedoch nicht nur die Teilnehmer aktiv rekrutiert, sondern das Thema war mit *Einstellungen gegenüber Ausländern* auch weniger unterhaltungsorientiert. Beide Aspekte könnten zu dem geringeren Anteil positiver Kommentare beigetragen haben. Studie 2, mit dem geringsten Anteil von 64 Prozent positiver Nennungen, verwendete denselben Fragebogen wie Studie 1 und ebenfalls eine aktive Rekrutierung. Die Stichprobe in Studie 2 bestand jedoch aus Personen, die im ALLBUS einer Folgebefragung zugestimmt hatten. Da im ALLBUS eine repräsentative Auswahl der erwachsenen Wohnbevölkerung befragt wird, ist hier der selektive Effekt des Interesses am Thema im Vergleich zu den anderen Studien deutlich geringer. Die Teilnahme an den Studien 3 (68 % positive Antworten) und 4 (71 % positive Antworten) wiederum war offen für alle Teilnehmer. Beide Studien waren, wie Studie 2 auch, wissenschaftliche Erhebungen mit zahlreichen, teilweise sehr persönlichen Fragen.

TABELLE 3
Verteilung der Kommentare über die Studien

Studie	positiv	ambivalent	negativ	total 1	neutrale Ausfälle	uncodiert	total 2
1	364	19	86	469	8	30	507
	77,6%	4,1%	18,3%	100%			
2	192	18	90	300	1	23	324
	64,0%	6,0%	30,0%	100%			
3	344	18	144	506	1	29	536
	68,0%	3,6%	28,5%	100%			
4	332	12	122	466	6	39	511
	71,2%	2,6%	26,2%	100%			
5	752	7	61	820	3	58	881
	91,7%	0,9%	7,4%	100%			
6	1130	17	91	1238	5	151	1394
	91,3%	1,4%	7,4%	100%			
total	3114	91	594	3799	24	330	4153
% innerhalb	82,0%	2,4%	15,6%	100%			

Um im Folgenden die Darstellung zur Verteilung der Kommentare zu vereinfachen, wird ein Zufriedenheitsindex gebildet. Dieser berechnet sich als Anteil positiver Angaben zum Anteil positiver plus negativer Angaben. Der Zufriedenheitsindex gibt damit das Ausmaß positiver Angaben in Prozent an und eignet sich als zusammenfassendes Maß einer Studie, da die Anzahl ambivalenter Antworten in allen Studien sehr gering war.

Um die Frage zu beantworten, inwiefern sich die Teilnehmer in ihren Bewertungen unterscheiden, wurde der Zufriedenheitsindex für alle Subgruppen hinsichtlich Geschlecht, Altersgruppen und Schulbildung gebildet. Tabelle 4 zeigt, dass insgesamt betrachtet Frauen die Fragen als positiver empfinden, $\chi^2(df=1, N=3657)=8{,}1$, $p=0{,}004$. Der geringere Zufriedenheitsindex bei Frauen in den Studien 3 und 4 ist nicht signifikant unterschiedlich von dem der Männer, $\chi^2(df=1, N=481)=0{,}08$, $p=0{,}78$. bzw. $\chi^2(df=1, N=454)=0{,}57$, $p=0{,}45$.

Mit zunehmendem Alter verringert sich der Zufriedenheitsindex von zunächst 94 Prozent auf 73 Prozent mit jeder älteren Gruppe, $\chi^2(df=6, N=3692)=96{,}8$, $p<0{,}001$.

Die Schulbildung wurde nur in den ersten drei Studien erhoben. Um einen Vergleich über die Studien zu erlauben, wurden drei Kategorien gebildet. Zur Kategorie Hauptschule gehört auch die Volksschule und die Hauptschule bis zur 9. Klasse. Die Kategorie Realschule umfasst auch den Fachabschluss und die Handelsschule. In der Gruppe Abitur sind auch die Kategorien Gymnasium, Oberschule, Fachabitur enthalten. In den Studien 1 und 2 zeigen sich signifikante Unterschiede in Bezug auf die Schulbildung, $\chi^2(df=2, N=442)=10{,}9$, $p=0{,}004$ bzw. $\chi^2(df=2, N=279)=9{,}7$, $p=0{,}008$. Die Unterschiede in Studie 3 hingegen sind nicht signifikant, $\chi^2(df=2, N=415)=1{,}8$, $p=0{,}4$. Teilnehmer mit gymnasialer Schulbildung empfinden die Fragen im Fragebogen demnach als negativer als Teilnehmer mit niedrigerer Schulbildung.

Abschließend gehen wir auf die Frage ein, was bewertet wird bzw. welche Empfindungen Teilnehmer bezüglich der Fragen im Fragebogen angaben. Tabelle 5 listet hierzu die zehn häufigsten Nennungen positiver und negativer Angaben. Hieran zeigen sich bereits die verschiedenen verwendeten Bewertungsdimensionen. Deutlich erkennbar sind Angaben zu den ersten drei Aspekten der Teilnehmerbelastung (›*respondent burden*‹) nach Bradburn (1978): Länge der Befragung, erforderliche Anstrengung der Teilnehmer bzw. Schwierigkeit der Fragen und induzierter Stress aufgrund unangenehmer oder persönlicher Fragen. Der vierte Aspekt, die

TABELLE 4
Zufriedenheitsindex über Teilnehmergruppen und Studien

	1	2	3	4	5	6	gesamt
Männer	79,7	64,7	71,0	75,0	91,9	91,7	81,7
Frauen	81,7	72,6	69,8	71,8	93,2	92,6	85,3
14 bis 18	83,3	-	-	--	93,3	95,6	94,1
19 bis 24	90,0	71,4	--	67,6	92,9	92,9	89,3
25 bis 29	80,0	67,7	-	69,0	95,5	91,4	87,7
30 bis 39	78,6	67,9	--	65,2	95,5	93,3	87,2
40 bis 49	83,8	66,7	--	69,4	89,2	90,4	82,4
50 bis 59	78,1	67,9	72,7	79,2	86,2	93,9	81,8
60+	73,9	63,6	70,4	78,7	94,6	92,0	73,1
Hauptschule	83,5	73,0	67,9	----	----	----	----
Realschule	86,9	79,5	74,8	----	----	----	----
Abitur	73,3	60,4	68,0	----	----	----	----
gesamt	80,9	68,1	70,5	73,1	92,5	92,5	84,0

Zeichenerklärung: - keine Teilnehmer in dieser Gruppe, -- weniger als 10 Fälle, ---- keine Angaben. Alle Angaben sind Prozentwerte.

Häufigkeit von Befragungen, befindet sich als Eigenschaft, die sich nicht auf die Fragen bezieht, erwartungsgemäß nicht darunter. Bezüglich der Anstrengung der Teilnahme geben die Angaben ein differenziertes Bild zu deren Ursachen. So sehen einige Teilnehmer die Fragen als einfach, angemessen und verständlich. Andere Teilnehmer wiederum empfinden die Fragen als nicht differenziert genug, zu allgemein, redundant, zu oberflächlich; Probleme, denen sich vor allem wissenschaftliche Befragungen stellen müssen. Teilnehmerstress lässt sich ebenfalls auf verschiedene Aspekte der Fragen zurückführen. Fragen können einerseits sehr gut, angenehm, interessant, angemessen sein, andererseits aber auch die Teilnehmer belasten im Sinne von zu persönlichen und nervigen Fragen. Diese Beispiele zeigen, dass sich die Angaben der Teilnehmer in konkreten Ausprägungen unterschiedlicher Aspekte der zunächst theoretischen Konzeption von Teilnehmerbelastung niederschlagen.

Die Analysen, die wir im Rahmen dieses Beitrags durchführten, zeigen einige der Möglichkeiten, wie die Frage »Ich empfand die Fragen als...« zusammen mit dem entwickelten Diktionär zur Qualitätssicherung verwendet werden kann.

TABELLE 5
Die zehn meistgenannten positiven und negativen Phrasen

Position	positiv	Anzahl	Prozent	negativ	Anzahl	Prozent
1	interessant	376	12,0	(zu) oberflächlich	19	3,2
2	angenehm	190	6,1	langweilig	14	2,4
3	gut	151	4,8	(zu) persönlich	10	1,6
4	ok	124	4,0	zu allgemein	9	1,5
5	normal	57	1,8	zu lang	6	1,0
6	einfach	48	1,5	einseitig	5	0,8
7	sehr gut	47	1,5	uninteressant	5	0,8
8	angemessen	42	1,3	nervig	4	0,7
9	verständlich	41	1,3	nicht differenziert genug	4	0,7
10	in Ordnung	38	1,2	redundant	4	0,7
	total	3114	100,0		594	100,0

Diskussion

Die Qualitätssicherung und -kontrolle von Fragebögen ist ein wesentlicher Aspekt bei der Konstruktion und Durchführung von Befragungen. Mithilfe nur einer zusätzlichen Frage im Fragebogen stellt dieser Beitrag ein Instrument vor, mit dem es möglich ist, einen Zufriedenheitsindex zu berechnen und über verschiedene Studien und Teilnehmergruppen hinweg zu vergleichen. Der Einsatz einer leicht zu beantwortenden, kurzen offenen Frage erlaubt darüber hinaus jedoch auch die Identifikation spezifischer Faktoren erfolgreicher und guter Fragebögen sowie negativer Aspekte in Fragebögen. Der Beitrag konzentrierte sich hierbei auf die Erstellung des Diktionärs zur computerunterstützen Inhaltsanalyse dieser Frage, wobei die verschiedenen Problemlösungen im Zuge der Erstellung über die Studien hinweg systematisch dargestellt wurden. Das Diktionär wurde anhand von 6 verschiedenen Studien mit unterschiedlichen Themen und Stichproben entwickelt und erweitert. Die abschließende Validitätsprüfung ergab für die beiden zentralen Bewertungen positiver und negativer Nennungen einen sehr hohen Validitätskoeffizienten von 0,951.

Beim Vergleich der Anteile positiver Nennungen zwischen den Studien zeigten sich deutliche Unterschiede. Aufgrund der hier verfolgten explorativen Forschungsstrategie lassen sich zwar einzelne Ursachen

nicht direkt zuordnen, einige Wirkungsmuster erscheinen jedoch naheliegend. Zunächst ist hier der Effekt der Selbstselektion zu nennen, der zu einer erhöhten Rate an positiven Bewertungen führt. Dadurch addieren sich zwei Effekte: Erstens beginnen eher Personen mit der Teilnahme, die der Umfrage gegenüber positiv eingestellt sind. Zweitens führt die Position der Bewertung am Ende des Fragebogens dazu, dass bereits demotivierte Abbrecher gar nicht zur Beantwortung gelangen. Aber auch das Thema und die Zielsetzung der Befragungen erzeugen unterschiedliche Bewertungen. Wissenschaftliche Untersuchungen mit dem Hang zu komplexeren Fragen und schwierigeren, oft auch innerhalb einer Befragung wechselnden Themen haben hier deutlich das Nachsehen. Sichtbar werden diese Zusammenhänge besonders an der Abnahme positiver Bewertungen, beginnend mit den offenen Themenumfragen bis zu den wissenschaftlichen Umfragen mit aktiver Rekrutierungsstrategie. Diese Erklärungen deuten wir als Beleg für die Validität und Praxistauglichkeit des hier vorgestellten Bewertungsinstruments.

Die Konsistenz in den Unterschieden zwischen den Teilnehmergruppen bestätigt diesen Eindruck. Frauen empfanden die Fragen generell als etwas positiver, wohingegen mit zunehmendem Alter und höherer Schulbildung die Fragen kritischer bewertet wurden.

Die inhaltlichen Angaben der Teilnehmer lassen sich der theoretischen Konzeption der Teilnehmerbelastung (BRADBURN 1978) zuordnen. Darüberhinaus helfen die Angaben aufgrund ihrer Konkretheit, Differenziertheit und Fülle, die Ursachen konkreter Teilnehmerbelastung in Fragebögen zu identifizieren.

Der Beitrag verfolgte das Ziel, auf neue Anwendungsmöglichkeiten der computerunterstützten Inhaltsanalyse im Zusammenspiel mit Online-Befragungen hinzuweisen. Die heutige Internet- und Computertechnik ermöglicht die Kombination qualitativer und quantitativer Methodik, die Vorteile beider Ansätze vereint, ohne dass Aufwand und Kosten dabei ein vernünftiges Maß übersteigen. Das entwickelte Codierschema (Diktionär) kann in verschiedenen Studien verwendet werden. Mit dieser Methode werden bisher ungenutzte Informationen analytisch verwertbar. Dieses Verfahren kann zur Evaluation, Qualitätssicherung und Überprüfung von Online-Fragebögen (u.a. in Pretests) eingesetzt werden. Darüberhinaus ist die hier vorgestellte Methode der Diktionärserstellung auch auf andere inhaltliche Kontexte übertragbar, in denen offene Fragen eine Rolle spielen.

Literatur

BANDILLA, W.; L. KACZMIREK; M. BLOHM; W. NEUBARTH: Coverage- und Nonresponse-Effekte bei Online-Bevölkerungsumfragen. In: JACKOB, N.; H. SCHOEN; T. ZERBACK (Hrsg.): *Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung*. [VS Verlag für Sozialwissenschaften] 2008

BRADBURN, N. M.: Respondent Burden. In: *Proceedings of the American Statistical Association, Survey Research Methods Section*. [American Statistical Association] 1978, S. 35-40

FRÜH, W.: *Inhaltsanalyse: Theorie und Praxis*. [UVK] 2007

GEER, J. G.: Do Open-Ended Questions Measure »Salient« Issues? In: *Public Opinion Quarterly*, 55, 1991, S. 360-370

HEDLIN, D.; T. DALE; G. HARALDSEN; J. JONES (Hrsg.): *Developing Methods for Assessing Perceived Response Burden*. [Statistics Sweden/Statistics Norway/ Office for National Statistics] 2005

KRONBERGER, N.; W. WAGNER: Keywords in Context: Statistical Analysis of Text Features. In: BAUER, M. W.; G. GASKELL (Hrsg.): *Qualitative Researching with Text, Image and Sound: A Practical Handbook*. [Sage] 2002, S. 299-317

PRÜFER, P.; M. REXROTH: Zwei-Phasen-Pretesting. In: MOHLER, P. P.; P. LÜTTINGER (Hrsg.): *Querschnitt: Festschrift für Max Kaase*. [ZUMA] 2000, S. 203-219

MOHLER, P. P.; C. ZÜLL: *TEXTPACK User's Guide*. [ZUMA] 2002

SPSS INC.: *SPSS Text Analysis for Surveys 2.1. Benutzerhandbuch*. [SPSS Inc.] 2007