MICHAEL SCHARKOW

Crowdsourcing von Inhaltsanalysen im World Wide Web?

Größere Inhaltsanalysen sind im Forschungsalltag zumeist mit sehr hohem Aufwand und entsprechenden Kosten verbunden. Umfangreiche Stichproben und komplexe Codepläne können nur mithilfe mehrerer Codierer umgesetzt werden, deren Leistung entscheidenden Einfluss auf die Reliabilität und Validität der Erhebung hat. Bei der Planung einer jeden Inhaltsanalyse muss der Forscher darüber entscheiden, ob eine Untersuchungseinheit stets nur einmal von einem Codierer (Einzelcodierung) oder mehrmals von verschiedenen Codierern (Mehrfachcodierung) erfasst wird. Aus Kostengründen wird das Codiermaterial jedoch zumeist aufgeteilt und nur einer Einzelkodierung unterzogen. Die Qualität der Analyse steht und fällt dann mit der individuellen Codierung, die zumeist nur stichprobenartig kontrolliert werden kann.

Auch bei wissenschaftlichen Untersuchungen ist es plausibel, anzunehmen, dass viele Augen mehr sehen als wenige. Wenn diese Annahme stimmt, dann hätte die Mehrfachcodierung von Medieninhalten deutliche Vorteile gegenüber der gängigen Einzelcodierung (vgl. SCHEUFELE 2001). Individuelle Fehler bei der Codierung könnten so diagnostiziert und durch die Aggregation der Codierentscheidungen mehrerer Codierer minimiert werden. Die ›Weisheit der Vielen‹ (SUROWIECKI 2004) würde so den geschulten Blick eines einzelnen Codierers ersetzen. Die Frage ist jedoch, ob und wie diese ›Weisheit‹ wissenschaftlich nutzbar gemacht werden kann.

Eine im Verhältnis zu konventionellen Inhaltsanalysen kostengünstige und einfache Möglichkeit, die Arbeitskraft vieler Codierer zu nutzen, stellt das sogenannte *Crowdsourcing* (PAPSDORF 2007) dar, d. h. das

Auslagern von Arbeiten an viele Freiwillige im Internet. Da bislang nur recht wenige sozialwissenschaftliche Erfahrungen mit dem Crowdsourcing der Datenerhebung vorliegen, soll dieser Beitrag konzeptionell und anhand zweier kleiner Machbarkeitsstudien die Möglichkeiten und Grenzen des Crowdsourcings von Inhaltsanalysen prüfen. Konkret wird dabei die Codierung von einer Vielzahl anonymer, untrainierter Helfer des Dienstes *Amazon Mechanical Turk* (AMT) durchgeführt. Speziell in den Blick genommen werden dabei (1) der organisatorische Aufwand und (2) die Qualität (konkret: Verfügbarkeit und Reliabilität) der Ergebnisse. Abschließend werden die Vor- und Nachteile des Einsatzes der Crowdsourcing-Technik für die Forschungspraxis der Inhaltsanalyse diskutiert.

Inhaltsanalysen und die Rolle der Codiererschulung

Obwohl der Codierprozess als »gelenkte Rezeption« (WIRTH 2001) oder »kontrollierte Interpretationsleistung« (FRÜH 2007) theoretisch gut fundiert ist, werden die Rolle und der Umgang mit den Codierern – sofern der Forscher nicht selbst der Codierer ist – in der Literatur nur recht vage beschrieben. Selbst wenn Einigkeit darüber herrscht, dass Codierer den symbolischen Gehalt von Mitteilungen hinsichtlich zuvor definierter Kriterien in (nummerische) Codes übertragen sollen, bleibt weiterhin sowohl theoretisch als auch empirisch unklar, was einen guten Codierer auszeichnet und welche Codierer-Eigenschaften maßgeblich zur Qualität des Forschungsprozesses beitragen.

Shapiro (1997) unterscheidet zwei unterschiedliche idealisierte Rollenverständnisse von Codierern, die folgenreich für den Umgang mit diesen sind: Zum einen kann ein externer Helfer als Ersatz für den Forscher dienen, der im Idealfall selbst die Codierung übernehmen würde. Zum anderen kann der Codierer aber auch schlicht als ›Instrument‹ verstanden werden, das aufgrund seiner Rezeptionskompetenz verbale oder audiovisuelle Inhalte effektiv und effizient in Bezug auf vorgegebene Kategorien verarbeiten kann. Brosius, Koschel und Haas (2009: 162) sprechen in diesem Zusammenhang vom Codierer als »vernunftbegabte Zählmaschine des Forschers«. Beide Rollenverständnisse lassen sich zugespitzt als Experten- bzw. Laien-Codierer bezeichnen. Während bei ersteren domänenspezifisches Wissen und aufwendige Schulung notwendig ist, geht man im zweiten Fall davon aus, dass vor allem die

Fähigkeit, Texte systematisch zu verarbeiten, einen guten Codierer ausmacht (vgl. WIRTH 2001).[1]

Da die Ressourcen bei allen inhaltsanalytischen Forschungsprojekten begrenzt sind, kann und muss stets eine abwägende Entscheidung zwischen der Anzahl an Codierern und dem veranschlagten Schulungsaufwand getroffen werden. Die Schulung der Codierer erfüllt dabei mehrere Funktionen: Erstens kann durch ein intensives Training der gelenkte Rezeptionsprozess so eingeübt werden, dass die Codierer auch bei komplexen Stimuli nicht von einer systematischen zu einer weniger belastenden heuristischen Verarbeitung umschwenken (WIRTH 2001). Zweitens kann im Rahmen eines gemeinsamen Probecodierens eine Koorientierung bei der Interpretation der Inhalte stattfinden und zwar nicht nur zwischen Codierern und dem Forscher, sondern auch unter den Codierern. Drittens ist anzunehmen, dass Mitarbeiter, die intensiv geschult wurden und/oder selbst inhaltliche Expertise besitzen, ein besseres Verständnis von der forschungsleitenden Frage und der Operationalisierungsstrategie haben als Laien. Viertens kann man bei speziell geschulten Mitarbeitern vermuten, dass ihre Motivation, viel Zeit und kognitive Anstrengungen in die Codierung zu investieren, die Reliabilität der Daten erhöht. Diese Vermutung hat sich mittlerweile auch empirisch bestätigt, etwa in der vergleichenden Studie von Hsueh, Melville und Sindhwani (2009), bei der eine deutliche höhere Übereinstimmung zwischen Experten als zwischen Laien bei der Codierung von Politikerbewertungen festgestellt wurde. Der Studie zufolge fiel diese Differenz umso höher aus, je komplexer die Kategorien angelegt waren.

Bei allen unbestrittenen Vorteilen, die intensiv geschulte Codierer für die Inhaltsanalyse besitzen, sind auch methodologische Nachteile einer umfangreichen Schulung zu konstatieren. Ein zentrales Problem bezieht sich auf die Reliabilität der Codierung, ein weiteres liegt im Kosten-Nutzen-Verhältnis der Schulung.

1 Ich unterscheide im Folgenden inhaltsanalytische Codierer von Teilnehmern einer Rezeptionsanalyse (KEPPLINGER 2009) oder rekrutierten Ratern, etwa für die Attraktivitätsbeurteilung (KLEIN/ROSAR 2005). Dabei liegt der Unterschied nicht primär in deren Arbeitsweise oder der Einbindung dieser Personen in die Datenerhebung, sondern im Erkenntnisinteresse: Bei Inhaltsanalysen wird durch die Vorgabe eines Interpretationsmusters der Rezeptionsprozess so weit gelenkt, dass jede Nicht-Übereinstimmung zwischen den Codierern zwingend als Messfehler gewertet werden muss. Bei subjektiven Urteilen wird die hingegen die rater-spezifische Varianz als (potenziell) erklärbare Größe im Rezeptionsprozess verstanden, die nicht durch entsprechende Anweisungen kontrollierbar ist.

Auch wenn die meisten Lehrbücher sich im Reliabilitätskapitel auf die Bestimmung von Übereinstimmungen zwischen verschiedenen Codierern beschränken, sind diese doch nicht gleichbedeutend mit der Reliabilität der Codierung. Krippendorff (2004) betont, dass Reliabilität vor allem auf Reproduzierbarkeit basiert und dass aus der Intercoderreliabilität lediglich der *Inferenzschluss* gezogen werden kann, dass die Analyse mit anderen Codierern zu denselben Ergebnissen führen würde. Je mehr Expertise zur Codierung jedoch notwendig ist und je spezifischer die Codierer geschult werden müssen, desto geringer ist die Wahrscheinlichkeit, dass die Ergebnisse der Inhaltsanalyse tatsächlich von Dritten reproduzierbar sind. Im Extremfall könnte sogar nur der Forscher selbst in der Lage sein, zuverlässig nach seinem Kategorienschema zu codieren. Experten-Codierer sind folglich schwer austauschbar – eine weitere Annahme, die dem Reliabilitätstest zugrunde liegt – und das Codierverfahren selbst ist intersubjektiv nur schwer nachprüfbar. Hinzu kommt, dass gerade die Auswahl und Schulung der Codierer in den meisten Fällen gar nicht oder nur unzureichend dokumentiert sind, sodass andere Forscher das Vorgehen des Codierprozesses letztlich nur – wenn überhaupt – anhand des Codebuchs rekonstruieren könnten.

Angesichts dieser Problematik kann man, insbesondere im Hinblick auf die Möglichkeiten des Crowdsourcings der Codierung, fragen: Wie viel Schulung ist überhaupt notwendig für eine Inhaltsanalyse? Die Antwort kann darauf nur lauten: gerade so viel, dass der Rezeptionsprozess der Codierer durch die Kategorienvorgabe und die Codieranweisungen so gelenkt wird, dass hinsichtlich der interessierenden Konstrukte eine reliable und valide Messung gelingt. Jedoch gilt aus forschungsökonomischer und methodologischer Perspektive auch das Sparsamkeitsprinzip: Je eindeutiger das Codierschema und je weniger individuelle Schulung notwendig ist, desto wahrscheinlicher ist es, dass andere die publizierten Ergebnisse auch mit eigenen Codierern replizieren könnten.

Konzeptionell ergeben sich aus den oben dargestellten Überlegungen plausible Einsatzfelder für den Einsatz von Laien-Codierern. Erstens kann man durch den Rückgriff auf ungeschulte Mitarbeiter empirisch bestimmen, wie sich die (a) Vorgabe der Kategorien, (b) die erklärenden Codieranweisungen und (c) die konkrete Schulung an Beispielen auf die Reliabilität der Erhebung auswirken. Da Kategorienvergabe, Codieranweisungen und Schulung einerseits mit (absteigender) Reproduzierbarkeit und andererseits mit (aufsteigenden) Kosten verbunden sind, ist es

wünschenswert, diese in der gegebenen Reihenfolge bis zu einer ausreichenden Codierleistung zu optimieren. Anders formuliert: Wenn sich empirisch zeigt, dass mit den vorgegebenen Interpretationsanweisungen eine zuverlässige und gültige Messung möglich ist, kann man weitere aufwendige Schulungen einsparen.

Zweitens ergeben sich durch den Einsatz von vielen ggf. weniger geschulten Codierern auch für die eigentliche Codierung Vorteile, die es u. U. als attraktiver erscheinen lassen, zusätzliche Ressourcen in die Mehrfachcodierung statt in weitere Schulungen zu investieren. Gerade bezogen auf die Sicherung der Reliabilität und vor allem auf den Umgang mit tendenziell weniger zuverlässigen Codierungen bieten (individuell schlechtere) Mehrfachcodierungen mehrere Vorteile, etwa durch Aggregation und Ex-post-Kontrolle. Selbst wenn einzelne Laien-Codierer im Durchschnitt unzuverlässiger sind und die Codierqualität auch eine größere Variabilität aufweist als die von Experten-Codierern, steigt bei der Aggregation mehrerer Codierungen (z. B. mittels Mehrheitsentscheidung) die Reliabilität des Instruments, d. h. der Kategorien, in der Regel an. Snow et al. (2008) kommen in diesem Zusammenhang zu dem Ergebnis, dass für ihre Aufgabenstellungen bereits vier ungelernte Codierer eine so gute Leistung erbringen wie ein ausführlich geschulter Experte.[2] Mehrfachcodierungen haben zudem den Vorteil, dass die Reliabilität mit den Daten aller Einzelcodierungen berechenbar wird, sodass der Inferenzschluss vom Reliabilitätstest auf das eigentliche Sample entfällt. Damit wird auch das Problem der Auswahl passenden Testmaterials umgangen, das häufig für verzerrte Reliabilitätsschätzungen verantwortlich ist (KRIPPENDORFF 2004). Zudem kommt man beim Einsatz von Mehrfachcodierungen ohne die Annahme aus, dass die Codierer bei der eigentlichen Arbeit genauso gut arbeiten wie im Reliabilitätstest.

Liegt zudem für jede Untersuchungseinheit nicht nur ein Datenpunkt vor, sondern mehrere Codierungen, lassen sich mit entsprechenden statistischen Modellen die Leistungen der Codierer und deren Auswirkungen auf die gesamte Inhaltsanalyse differenziert schätzen und ggf. korrigieren (CARPENTER 2008). Diese Möglichkeit bleibt dem Forscher bei

2 Diese Erkenntnis findet sich auch in anderen Gebieten empirischer Forschung, etwa in der klassischen Testtheorie, die postuliert, dass mit der Länge eines Tests dessen Reliabilität steigt (MOOSBRUGGER 2007).

der Einmalcodierung versagt, d. h. ein gewisses Maß an Vertrauen in und damit Abhängigkeit von den Codierern ist hier unabdingbar.

Zusammenfassend lässt sich sagen, dass ein Verzicht auf intensive Schulung und (stattdessen) der Einsatz von Laien-Codierern einige methodische und forschungsökonomische Vorzüge aufweisen, sodass es sich zumindest lohnt, die Möglichkeiten einer solchen Inhaltsanalyse empirisch zu untersuchen.

Datenerhebung durch Crowdsourcing

Verteilt man Codierarbeit auf viele Laien-Codierer, stellt sich die Frage, wie dies konkret organisiert werden kann. Da der Aufwand einer solchen Arbeitsteilung mit der Anzahl der Beteiligten steigt, bietet es sich an, dies so weit wie möglich zu automatisieren und die Kommunikation mit den Codierern elektronisch zu organisieren. In den letzten Jahren hat das Prinzip des ›Crowdsourcings‹ im Internet an Bedeutung gewonnen. Die deutsche Wikipedia definiert Crowdsourcing als »Auslagerung auf die Intelligenz und die Arbeitskraft einer Masse von Freizeitarbeitern im Internet« (http://de.wikipedia.org/wiki/Crowdsourcing) und ist dabei selbst das prominenteste Beispiel für dieses Prinzip: Statt von wenigen ausgewiesenen Experten und Redakteuren wird die Online-Enzyklopädie von tausenden interessierten Laien gepflegt und erweitert (VOSS 2005). Wikipedia setzt konzeptionell und technisch den Gedanken der Freien Software fort, ohne jedoch deren vergleichsweise hohe Anforderungen an die Expertise der Teilnehmer zu teilen. Auf die Formen, Motive und Organisationsprinzipien von Crowdsourcing kann an dieser Stelle nicht vertieft eingegangen werden (vgl. zusammenfassend PAPSDORF 2009).

Auch im wissenschaftlichen Kontext wurde bereits vor einiger Zeit damit begonnen, die Möglichkeiten des Crowdsourcings zu evaluieren. Dabei sind vor allem die Arbeiten der Open-Mind-Initiative (STORK 1999; SINGH 2002) involviert, die sich mit Fragen von künstlicher und menschlicher Intelligenz befasst und dabei die Antworten auf hunderttausende ›Common-Sense-Tasks‹ sammelt. Die Erhebung erfolgt dabei zentral auf der Plattform des Projekts http://openmind.org. Von Ahn (2004, 2006) entwickelt den Gedanken von Crowdsourcing dahingehend weiter, dass mit dem Bearbeiten von Aufgaben gleich mehrere Vorteile verbunden sind, unter anderem auch für die freiwilligen Teilnehmer.

Dies gelingt zum Beispiel durch den Einsatz von kleinen Aufgaben im Rahmen von Spielen. Im ESP Game geht es beispielsweise vordergründig darum, gemeinsam mit einem (virtuellen) Partner verschiedene Bilder zu beschreiben. Je mehr übereinstimmende Schlagworte ein Team hat, desto höher ist die Punktzahl. Dies ist nichts anderes als die spielerische Variante eines Reliabilitätstests mit vielen tausend Bildern als Stimulusmaterial. Nach demselben Prinzip werden auf http://www.espgame.org Lieder, Videos und Texte spielerisch annotiert.

Ein weiteres Beispiel für die Lösung verschiedener Problemstellungen durch Crowdsourcing stellt der Dienst ReCaptcha (VON AHN et al. 2008) dar. Bei CAPTCHAs (Completely Automated Public Turing Test To Tell Computers and Humans Apart) handelt es sich um Aufgaben, die in Web-Anwendungen zum Schutz gegen automatische Anmeldungen und andere Angriffe durch softwaregesteuerte Bots eingesetzt werden. Ein klassisches Beispiel ist die verzerrte Darstellung eines oder mehrerer Worte, die von Menschen relativ leicht, von Computern jedoch nur schwer identifiziert werden können. Der zusätzliche Nutzen von ReCaptcha liegt nun in der Tatsache, dass nicht neue Wörter generiert werden, sondern Ausschnitte aus bereits gescannten Texten, die jedoch nicht vom Computer erkannt wurden, dem Nutzer vorgelegt werden (vgl. Abb. 1). So wird einerseits gewährleistet, dass nur echte Menschen Zugang zu bestimmten Diensten erhalten, gleichzeitig wird die Digitalisierung von Büchern Wort für Wort vorangetrieben.

ABBILDUNG 1
ReCaptcha

Quelle: http://recaptcha.net

Amazon Mechanical Turk als Plattform für verteilte Inhaltsanalysen

Die bisherigen Beispiele für Crowdsourcing sind auf freiwillige, unbezahlte Teilnehmer ausgelegt. Dies hat zur Folge, dass erstens ein vergleichsweise großer Aufwand in die Rekrutierung und Incentivierung der Teilnehmer investiert werden muss, zweitens der Rücklauf nur schlecht prognostiziert und noch schlechter beeinflusst werden kann. Drittens müssen erhebliche Ressourcen in die Entwicklung und Wartung der Crowdsourcing-Plattform gesteckt werden. Dies macht Crowdsourcing für einzelne Wissenschaftler mit begrenzten technischen und finanziellen Möglichkeiten relativ unattraktiv. Eine Alternative zum traditionellen ›Freizeit‹-Crowdsourcing bietet die kommerzielle Plattform Amazon Mechanical Turk, die als Vermittler zwischen Auftraggebern (Requester = Forscher) und einem großen Pool an bezahlten freien Mitarbeitern (Worker = Coder) dient. Wie funktioniert nun AMT für das Crowdsourcing von Inhaltsanalysen?

1. Der Auftraggeber definiert einen oder mehrere sogenannte ›Human Intelligence Tasks‹ (HIT), zum Beispiel eine Codieraufgabe, für die eine kurze Beschreibung und ggf. Kriterien für die Auswahl an Workern definiert werden. Er muss gleichzeitig die Bezahlung pro HIT festlegen, z. B. pro Variabler und Codierung 0,05 EUR.
2. Das Eingabe-Interface für den HIT wird vom Requester gestaltet. Dazu stellt AMT eine Vielzahl an Vorlagen bereit, mit denen auch ohne HTML-Kenntnisse benutzerfreundliche Aufgabenseiten erstellt werden können. Eine Aufgabe enthält immer einen Aufgabentext, ein Formular zur Dateneingabe sowie ggf. Platz für das Stimulusmaterial .
3. Der Requester lädt eine Liste mit *k* Stimulusobjekten (Bilder, Texte, eingebettete Multimedia-Inhalte) hoch und legt fest, wie viele HITs pro Stimulusobjekt gelöst werden sollen. Es empfiehlt sich, aus den o. g. Gründen der Reliabilitätssicherung, jede Aufgabe von mindestens zwei, besser aber *n* Workern bearbeiten zu lassen.
4. AMT erstellt aus der Vorlage und Stimulusliste *n* x *k* HITs, die dann freigeschaltet werden können. Gleichzeitig wird der für die Bezahlung der Worker fällige Betrag von Amazon eingezogen.
5. Nach Freischaltung der HITs werden diese den als Worker registrierten Personen angezeigt, die dann beliebig viele Aufgaben

erfüllen können und dafür Geld gutgeschrieben bekommen. Auf die Auswahl der Worker hat der Auftraggeber nur relativ wenig Einfluss, kann aber bei schlechter Arbeit die Bezahlung verweigern.

6. Nachdem alle HITs abgearbeitet sind, werden die Ergebnisse dem Requester zur Begutachtung vorgelegt. Dabei wird eine csv-Datei generiert, die neben den eigentlichen Codes auch Metadaten zu Workern und Bearbeitungszeit enthält.

Das Mechanical-Turk-Prinzip bietet für die Online-Forschung und insbesondere für die Online-Inhaltsanalyse eine Reihe von Vorteilen: Die komplette Logistik von der Stellung und Verteilung der Aufgaben, die Rekrutierung und Bezahlung der Teilnehmer und das Datenmanagement werden zentral von AMT übernommen. Die Overhead-Kosten dafür sind zumeist so niedrig, dass auch einzelne Wissenschaftler oder Studierende diese bezahlen können. Zudem kann die Menge an Crowdsourcing fast beliebig skaliert werden: von wenigen schwierigen Aufgaben mit wenigen Workern bis hin zu tausenden kleinerer HITs, die von hunderten von Workern gelöst werden.

Allerdings sind mit dieser Art des Crowdsourcings auch Nachteile verbunden: Die Kommunikation mit den Workern ist aufwendig und beschränkt sich zumeist auf kurze Anweisungen einerseits und Kommentarfelder für Feedback andererseits. Die nachträgliche Bewertung einzelner Worker lohnt zumeist den Aufwand nicht, sodass man ggf. mithilfe statistischer Modelle schlechte Codierer nachträglich entfernen sollte (CARPENTER 2008). Nachteilig ist ferner, dass es bislang nur wenige Informationen zur Reliabilität und Validität von Mehrfachcodierungen in den Sozialwissenschaften gibt. Die meisten Evaluationsstudien zu AMT erfolgten in den Gebieten der Bilderkennung und einfachen Wortannotation (SNOW et al. 2008; SOROKIN et al. 2008). Im folgenden Abschnitt sollen im Rahmen von zwei Fallstudien folgende Forschungsfragen beantwortet werden: Wie lange dauert die Codierung von binären Variablen und welche Reliabilität weist diese Codierung auf? Eignet sich der englischsprachige Dienst AMT auch für die Codierung deutscher Inhalte?

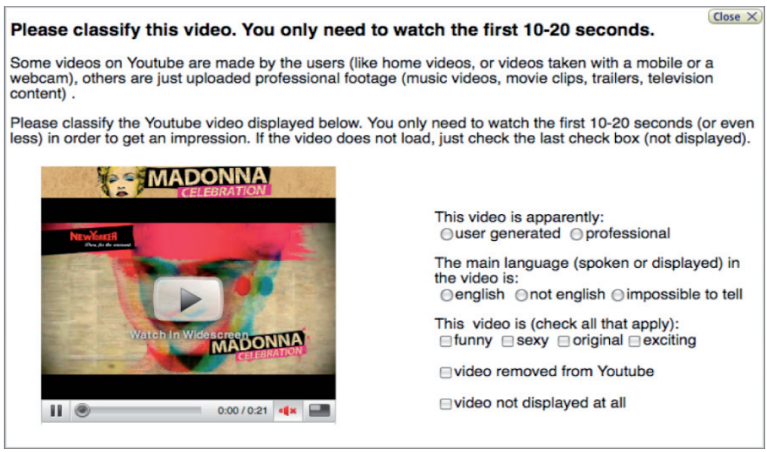
Zwei Fallstudien zur Anwendung von Mechanical Turk

Mit den folgenden zwei Fallstudien soll ein erster Schritt zur Evaluation von AMT für das Crowdsourcing von Online-Inhaltsanalysen unternommen

werden. Die beiden Beispiele dienen primär der Illustration des Vorgehens und sollen helfen, einen ersten Eindruck von der Umsetzbarkeit und der erwartbaren Reliabilität einer AMT-Codierung zu erhalten. Sie sind nicht zuletzt aus Gründen der Darstellung weniger komplex als eine durchschnittliche Inhaltsanalyse und erheben daher auch nicht den Anspruch, repräsentativ für sozialwissenschaftliche Forschungsfragen zu sein.

In der ersten Evaluationsstudie sollten, im Anschluss an ein Forschungsprojekt zur Erfolgsmessung bei YouTube-Videos (vgl. ausführlich dazu SCHARKOW 2008), ausgewählte Videos dahingehend klassifiziert werden, ob der Inhalt – erstens – professionell produziert oder nutzergeneriert und – zweitens – in welcher Sprache diese Videos erstellt worden sind. Es wurde dafür von der forschungspraktisch reizvollen Möglichkeit Gebrauch gemacht, multimediale Inhalte – in diesem Fall Videos der Plattform YouTube – direkt in das Codier-Interface einzubauen (vgl. Abb. 2).

ABBILDUNG 2
Eingabemaske für die Codierung von YouTube-Videos



Um den Aufwand für die Worker und damit die Kosten so gering wie möglich zu halten, wurden diese aufgefordert, nur die ersten 10 bis 20 Sekunden des Videos anzusehen und dann eine Codierentscheidung zu treffen. In einem Pretest mit studentischen Codierern hatte sich eine Zeit von 15 Sekunden als ausreichend für eine reliable Messung herausgestellt. Insgesamt sollten 100 Videos von je zwei Workern annotiert

werden, wobei 0,02 EUR pro HIT gezahlt wurden (dies entspricht kalkulierten 3 EUR pro Stunde). Nach 41 Minuten waren alle 200 HITs erfüllt, die Gesamtkosten inkl. der AMT-Gebühren betrugen 5 EUR. Insgesamt wurden die Aufgaben von 19 Workern bearbeitet, wobei die fleißigsten drei Codierer zusammen 150 HITs lösten. Angesichts der äußerst knappen Codieranweisungen und der kurzen Bearbeitungszeit konnte eine gute Intercoder-Reliabilität für die Klassifikation der Videos erreicht werden (Inhalt: CR = .86, Krippendorffs Alpha = .70, Sprache: CR = .92, Krippendorffs Alpha = .77).

Mit der zweiten Fallstudie sollte der Frage nachgegangen werden, ob auch deutschsprachige Medieninhalte mithilfe von AMT codiert werden können. Da bislang nur englischsprachige Studien vorliegen und AMT selbst auch nur in englischer Sprache verfügbar ist, stellt sich die Frage, ob und wie viele deutschsprachige Worker überhaupt zur Verfügung stehen. Um diese Frage zu beantworten, wurde eine Codierung von 100 deutschen Nachrichtenschlagzeilen aus dem Jahr 2008 durchgeführt. Die einzige zu codierende Variable war das Thema der Nachrichtenschlagzeile. Die Themenvariable mit zwölf Ausprägungen wurde von Quandt (2008) übernommen. Um nur deutschsprachige Worker zu rekrutieren, wurde die Aufgabe in deutscher Sprache formuliert (zudem waren nur Worker zugelassen, die bei der Anmeldung zu AMT angegeben hatten, deutsch zu sprechen). Da es bei der Machbarkeitsstudie um eine simple Einschätzung der Durchführung deutschsprachiger Inhaltsanalysen mittels AMT ging, wurde die Themenvariable ohne weitere Codieranweisungen als Einfachauswahl mit Radio-Buttons präsentiert. Mit anderen Worten: Als einziger Mechanismus zur Lenkung der Rezeption der Teilnehmer diente die Kategorienbezeichnung. Dies stellt sozusagen eine Minimalvariante von Codieranweisungen dar, wobei davon ausgegangen wird, dass die Themenbezeichnungen zumindest intersubjektiv gleich verstanden werden.

Im Ergebnis der zweiten Studie zeigen sich die Grenzen des deutschsprachigen Crowdsourcings mit AMT: Die Codierung von 2 x 100 Schlagzeilen dauerte mit 3,5 Stunden deutlich länger als die englische Aufgabe, und dies trotz leicht besserer Bezahlung (0,05 EUR pro HIT). Insgesamt nahmen nur drei Worker an der Codierung teil. Die Reliabilität der Themencodierung (ohne weitere Codieranweisungen) lag bei Krippendorffs Alpha = .54 (CR = .88). Dieser Reliabilitätswert ist zwar keineswegs selten in der inhaltsanalytischen Forschung (vgl. die ausführliche Diskussion von zufallsbereinigten Reliabilitätswerten für die Forschungspraxis bei

RAUPP/VOGELGESANG 2009), gleichwohl aber ungenügend. Leider liegen keine Informationen zur Reliabilität der Themencodierung (Experten-Codierung) bei Quandt (2008) vor, sodass bis auf Weiteres offen bleiben muss, wie viel Varianz in der Reliabilität durch den Codeplan und wie viel durch die Art der Codierung erklärt werden kann. Hier sind weitere Evaluationsstudien notwendig, in deren Rahmen z. B. auch verschiedene Codieranweisungen verglichen werden könnten (vgl. WIRTH/HARDEN 2005). Angesichts der wenigen Codierer ist zum jetzigen Zeitpunkt allerdings nicht davon auszugehen, dass für solche Methodenexperimente immer genügend Probanden vorhanden sind.

Zusammenfassung und Diskussion

Im Rahmen dieses Beitrags wurden zwei Fallstudien vorgestellt, bei denen der Einsatz von Laien-Codierern bei Crowdsourcing-Inhaltsanalysen evaluiert wurde. Es zeigte sich, dass simple inhaltsanalytische Klassifikationsaufgaben zeit- und kostensparend mithilfe des Mechanical-Turk-Prinzips reliabel von Laien-Codierern gelöst werden können. Außerdem ergab die zweite Studie, dass es nur bedingt möglich ist, den englischsprachigen Dienst AMT auch für die Codierung deutschsprachiger Inhalte zu nutzen.

Angesichts der ersten empirischen Ergebnisse ergeben sich vor allem zwei Einsatzmöglichkeiten für AMT im Kontext sozialwissenschaftlicher Inhaltsanalysen: einerseits als Ergänzung für konventionelle manuelle Analysen, vor allem nonverbaler oder englischsprachiger Stimuli, andererseits als Plattform für Methodenexperimente. Die mit dem Crowdsourcing einhergehende Laien-Codierung ist vor allem als eine Ergänzung zu Experten-Codierungen zu verstehen. Allein die Tatsachen, dass der Raum für Codieranweisungen begrenzt ist, umfangreiche Aufgaben in kleinere Einheiten zerteilt werden müssen und die Kommunikation mit den Codierern schwerfällt, machten Crowdsourcing für komplexe Inhaltsanalysen unattraktiv. Der Einsatz von Laien-Codierern, der mit dem Mechanical-Turk-Prinzip einhergeht und der durch AMT forschungspraktisch mit leichter Hand realisiert werden kann, ermöglicht es dem Forscher jedoch, herauszufinden, wie gut ein Untersuchungsinstrument ohne die Schulung von Codierern funktioniert. Messtheoretisch kann man dies als Reliabilitäts-Baseline interpretieren. Ist diese ausreichend hoch, spricht nichts dagegen, Teile der Codierung vollständig auf AMT auszulagern.

Durch das Crowdsourcen von Inhaltsanalysen kann außerdem eine in der Regel unzureichend dokumentierte Phase im Forschungsprozess, die Entwicklung des Codebuchs und die Schulung der Codierer, mit geringem Zeit- und Kostenaufwand systematisch erforscht werden. Durch den Einsatz von Methodenexperimenten können auch der Schulungs- und Codierprozess empirisch untersucht und relevante Einflussfaktoren identifiziert werden. Denkbar und teilweise empirisch belegt sind neben dem Umfang des Codeplans (WIRTH/HARDEN 2005) auch die Höhe der Bezahlung und die Anzahl der Codierer, die beide die Reliabilität der Messung deutlich erhöhen (HSUEH et al. 2009; FENG/ZAJAC 2009). Diese Größen lassen sich in AMT problemlos manipulieren, um die Effektivität und Effizienz der Analysen zu optimieren. Mit kleinen Feldexperimenten lassen sich zudem auch verschiedene Operationalisierungsentscheidungen anhand der resultierenden Reliabilität miteinander vergleichen. Ob die eigentliche Codierung dann auch nach dem Crowdsourcing-Prinzip erfolgt, könnte somit anhand empirischer Pretest-Ergebnisse getroffen werden.

Schließlich eignet sich AMT nicht nur für einfache Inhaltsanalysen, sondern ist auch für die schnelle und unkomplizierte Umsetzung anderer empirischer Studien sinnvoll – etwa für die bereits angesprochenen Ratings und Rezeptionsanalysen sowie für die ersten Schritte einer Skalenentwicklung. Dies und nicht zuletzt die geringen Kosten macht das Crowdsourcing für die kommunikationswissenschaftliche Forschung, aber auch für den Einsatz in der Lehre interessant.

Literatur

BROSIUS, H.-B.; F. KOSCHEL; A. HAAS: *Methoden der empirischen Kommunikationsforschung.* Wiesbaden [vs Verlag] 2009

CARPENTER, B.: *Multilevel Bayesian Models of Categorical Data Annotation*. 2008 Unpublished manuscript, http://lingpipe.files.wordpress.com/2009/01/anno-bayes- entities-09.pdf

FENG, D.; S. ZAJAC: *Acquiring High Quality Non-Expert Knowledge from On-demand Workforce.* Singapur (Proceedings of the 2009 Workshop on the People's Web Meets NLP) 2009

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

HSUEH, P.; P. MELVILLE; V. SINDHWANI: Data quality from crowdsourcing: A study of annotation selection criteria. In: *Proceedings of the NAACL HLT 2009.* Boulder, Colorado 2009, S. 27-35

KEPPLINGER, H. M.: *Politikvermittlung*. Wiesbaden [vs Verlag] 2009

KLEIN, M.; U. ROSAR: Physische Attraktivität und Wahlerfolg. Eine empirische Analyse am Beispiel der Wahlkreiskandidaten bei der Bundestagswahl 2002. In: *Politische Vierteljahresschrift*, 46(2), 2005, S. 263-287

KRIPPENDORFF, K.: Reliability in content analysis. Some Common Misconceptions and Recommendations. In: *Human Communication Research*, 30(3), 2004, S. 411-433

MOOSBRUGGER, H.: Klassische Testtheorie (KTT). In: MOOSBRUGGER, H.; A. KELAVA (Hrsg.): *Testtheorie und Fragebogenkonstruktion*. Heidelberg [Springer] 2007, S. 100-112

PAPSDORF, C.: *Wie Surfen zu Arbeit wird. Crowdsourcing im Web 2.0*. Frankfurt/M. [Campus] 2009

QUANDT, T.: Neues Medium, alter Journalismus? Eine vergleichende Inhaltsanalyse tagesaktueller Print- und Online-Nachrichtenangebote. In: QUANDT, T.; W. SCHWEIGER (Hrsg.): *Journalismus Online – Partizipation oder Profession?* Wiesbaden [vs Verlag] 2008, S. 131-155

SCHARKOW, M.: *Einschaltquoten im Social Web – Möglichkeiten der Erhebung und Analyse von Publikumsdaten am Beispiel YouTube*. (Vortrag auf der Jahrestagung der Fachgruppe ›Methoden der Publizistik- und Kommunikationswissenschaft‹ in der DGPuK) 2008

SCHEUFELE, B.: Notwendigkeit, Nutzen und Aufwand von Mehrfach- und Sondercodierungen. In: WIRTH, W.; E. LAUF (Hrsg.): *Inhaltsanalyse. Perspektiven, Probleme, Potentiale.* Köln [Herbert von Halem] 2001, S. 82-97

SHAPIRO, G.: The future of coders: Human judgments in a world of sophisticated software. In: ROBERTS, C. W. (Hrsg.): *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts.* Mahwah, NJ [Lawrence Erlbaum] 1997, S. 225-238

SNOW, R.; B. O'CONNOR; D. JURAFSKY; A. Y. NG: Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, S. 254-263

SOROKIN, A.; D. FORSYTH: Utility data annotation with Amazon Mechanical Turk. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008, S. 1-8

STORK, D.: The OpenMind Initiative. In: *IEEE Intelligent Systems & their applications*, 14(3), 1999, S. 19-20

SUROWIECKI, J.: *The Wisdom Of Crowds: Why The Many Are Smarter Than The Few And How Collective Wisdom Shapes Business, Economies, Societies And Nations*. NY [Little, Brown Book Group] 2004

VON AHN, L.; L. DABBISH: Labeling Images with a Computer Game. In: *Proceedings of the ACM. Conference on Human Factors in Computing Systems*. 2004, S. 319-326

VON AHN, L.; R. LIU; M. BLUM: Peekaboom: A game for locating objects in images. In: *ACM CHI*, 2006

VON AHN, L.; B. MAURER; C. MCMILLEN; D. ABRAHAM; M. BLUM: reCAPTCHA: Human-Based Character Recognition via Web Security Measures. In: *Science*, September, 2008

VOSS, J.: *Measuring Wikipedia*. Stockholm (Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics) 2005

WIRTH, W.: Der Codierprozeß als gelenkte Rezeption. Bausteine für eine Theorie des Codierens. In: WIRTH, W.; E. LAUF (Hrsg.): *Inhaltsanalyse. Perspektiven, Probleme, Potentiale*. Köln [Herbert von Halem] 2001, S. 157-182

WIRTH, W.; L. HARDEN: *Heuristisches und systematische Codieren. Eine empirische Analyse zum Codierprozess.* Düsseldorf (Vortrag auf der 7. Tagung der Fachgruppe »Methoden der Publizistik- und Kommunikationswissenschaft« in der DGPuK) 2005