ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

STL SYSTEMTECHNIKLABOR

Qualitätssicherung im Croudsourcing mittels automatisch generierter Wissensfragen

Yves Hary

Technical Report – STL-TR-2016-06 – ISSN 2364-7167

stl.htwsaar.de

Technische Berichte des Systemtechniklabors (STL) der htw saar
Technical Reports of the System Technology Lab (STL) at htw saar
ISSN 2364-7167

Yves Hary: Qualitätssicherung im Croudsourcing mittels automatisch generierter Wissensfragen
Technical report id: STL-TR-2016-06

First published: September 2016
Last revision: September 2016
Internal review: Klaus Berberich

For the most recent version of this report see: https://stl.htwsaar.de/

Title image source: Ove Tøpfer, http://www.freeimages.com/photo/light-bulb-1416824

htw saar – Hochschule für Technik und Wirtschaft des Saarlandes (University of Applied Sciences)
Fakultät für Ingenieurwissenschaften (School of Engineering)
STL – Systemtechniklabor (System Technology Lab)
Prof. Dr.-Ing. André Miede (andre.miede@htwsaar.de)
Goebenstraße 40
66117 Saarbrücken, Germany
https://stl.htwsaar.de

ingenieur
wissenschaften
htw saar

Hochschule für
Technik und Wirtschaft
des Saarlandes
University of
Applied Sciences

Bachelor-Thesis

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)
an der Hochschule für Technik und Wirtschaft des Saarlandes
im Studiengang Praktische Informatik
der Fakultät für Ingenieurwissenschaften

Qualitätssicherung im Croudsourcing mittels automatisch generierter Wissensfragen

vorgelegt von
Yves Hary

betreut und begutachtet von
Prof. Dr.-Ing. Klaus Berberich

Saarbrücken, 14. September 2016

Zusammenfassung

„Crowdsourcing“ ist ein neues Modell der Datenverarbeitung, das die Vorzüge von Computern und Menschen miteinander verbinden soll. Menschen bearbeiten Aufgaben, die für Computer schwierig zu erledigen sind, wie z.B. Bilder zu kategorisieren oder Texte einzuordnen. Dabei werden über Internet-Plattformen die Verteilung der Aufgaben und das Sammeln der Ergebnisse vorgenommen. Menschen arbeiten allerdings nicht immer fehlerfrei. Daher müssen Maßnahmen ergriffen werden, um unzuverlässige oder stark fehlerproduzierende Arbeiter aus dem Arbeitsprozess ausschließen zu können. Die vorliegende Arbeit untersucht, inwiefern sich Wissensfragen, auch als Quizfragen bezeichnet, in Form von Multiple-Choice-Fragen dazu eignen, die Arbeitsqualität im Crowdsourcing-Prozess zu erhöhen. Dabei werden diese Multiple-Choice-Fragen, die von einem existierenden Software-Tool generiert werden, zusammen mit den Aufgaben eines Crowdsourcing-Jobs präsentiert und zur Bearbeitung freigegeben. Die Ergebnisse werden mit verschiedenen Maßzahlen untersucht und mit einem populären Qualitätssicherungsmechanismus des Crowdsourcings verglichen. Es zeigt sich, dass diese automatisch generierten Wissensfragen ihre Berechtigung als Qualitätssicherungsmechanismus im Crowdsourcing besitzen.

Inhaltsverzeichnis

1 Einleitung

Dieses Kapitel gibt einen kurzen Überblick darüber, was unter „Crowdsourcing“ zu verstehen ist und auf welche Weise es genutzt wird. Es wird anschließend das zentrale Problem des „Crowdsourcing“ erörtert und daraus die Motivation und das Thema dieser Arbeit abgeleitet. Es wird der in dieser Arbeit verfolgte Lösungsweg und zum Schluss der weitere Aufbau der Arbeit beschrieben.

1.1 Hintergrund

„Crowdsourcing“ ist *das Konzept*, das derzeit durch die IT-Landschaft geistert. So ähnlich wie „Gamification“ vor einigen Jahren für Aufsehen sorgte, ist nun „Crowdsourcing“ der neue Stern am „Informatik-Himmel“. Es lassen sich damit viele Probleme der Datenverarbeitung und Datenbeschaffung lösen und das zu einem Bruchteil der Kosten herkömmlicher Methoden. Beim „Crowdsourcing“ arbeitet eine große, anonyme Masse von menschlichen Arbeitern an in eine kleine Häppchen aufgeteilten Problemen und erhält dafür kleine monetäre Beträge. Diese Zusammenarbeit wird durch Online-Plattformen ermöglicht, welche die Arbeit koordinieren und verteilen. Die Tatsache, dass in unserer Zeit quasi *jeder* Mensch mindestens einen Computer besitzt, macht diese Zusammenarbeit und damit das Konzept von „Crowdsourcing“ erst möglich.

Aber von welchen Problemen sprechen wir hier? Eine auf solchen Internet-Plattformen beliebte Problemstellung ist z.B. das Aussortieren von Beiträgen auf Social-Media-Plattformen wie *Facebook*[1]. Ein Mensch, irgendwo auf der Welt, bekommt einen einen geschriebenen Beitrag präsentiert und darf dann entscheiden, ob der Beitrag etwa sexistische oder rassistische Äußerungen enthält. Für jeden auf diese Weise gekennzeichneten Beitrag bekommt der sogenannte „Worker“ etwas Geld, meist einige Cent. Auf diese Weise könnte z.B. *Facebook* nun ganz einfach die unpassenden Beiträge ausfiltern und seine Nutzer schützen. Eine andere beliebte Art von Arbeit, die man durch „Crowdsourcing“ zu lösen versucht, sind Übersetzungen. Einen ausgebildeten Übersetzer zu engagieren ist teuer und das professionelle Übersetzen von Texten dauert lange. Wie wäre es also, die Kenntnisse von Unmengen von *nicht professionellen* Übersetzern zu nutzen? Die Worker bekommen einen kleinen Textausschnitt präsentiert und übersetzen diesen. Andere Worker wiederum bekommen die Arbeit von Kollegen gezeigt und versuchen darin Fehler zu finden. Die Masse an Menschen kontrolliert und verbessert sich also gegenseitig.

Man kann nun einwenden, dass diese Arbeit doch auch voll-automatisiert von Computern übernommen werden könnte. Ja, das könnte sie – doch leider liegen Computer in diesen Bereichen noch weit hinter dem menschlichen Verstand zurück. Ein Computer kann auf jeden Fall genauer und schneller große Zahlenkolonnen addieren, subtrahieren oder multiplizieren. Aber zu entscheiden, ob ein Text unangemessen ist, oder eine gute Übersetzung anzufertigen, das können Menschen immer noch besser und zuverlässiger. Genauso ist es mit anderen Problemen, die für Computer schwierig, für Menschen aber sehr einfach sind, wie Inhalte auf Bildern erkennen, Texte zu übersetzen oder Gesprochenes niederzuschreiben.

[1] http://www.facebook.com (zuletzt abgerufen am 14. September 2016).

„Crowdsourcing“ ist also ein Kompromiss aus den Vorzügen beider Welten: Die Unermüdlichkeit von Computern wird durch die Kreativität des Menschen ergänzt. Dadurch ergibt sich ein ganz besonderes Potential.

Diese besonderen Eigenschaften von „Crowdsourcing“ haben aber auch unerwünschte Effekte auf Wirtschaft und Gesellschaft. Da bei „Crowdsourcing“ die große Masse an Menschen die Expertise eines Einzelnen kompensieren soll und gerade dadurch kostengünstiger ist, verdrängt „Crowdsourcing“ in einigen Bereichen die klassische Arbeitsweise von Spezialisten. So hat z.B. die Erfindung von Stock-Foto-Plattformen im Internet die klassische Stock-Fotografie einzelner Fotografen unrentabel gemacht und diese Einkommensquelle für Fotografen vernichtet [14]. Wegen der Vernetzung durch das Internet hat man so die Möglichkeit auf viele talentierte Arbeiter aus aller Welt zurückzugreifen, die nicht zwangsweise Profis auf ihrem Gebiet sein müssen. Dementsprechend wird diesen Menschen auch viel weniger gezahlt als ausgebildeten Experten.

1.2 Motivation

So wie die Computertechnik aber den entscheidenden Nachteil hat, nicht wie der Mensch Probleme lösen zu können, so hat auch der Mensch den Nachteil, nicht bis in alle Ewigkeit Leistung mit der gleichen Qualität zu erbringen. Menschen werden müde, Menschen sind faul, Menschen machen Fehler. Diese menschlichen Eigenschaften passen aber nicht mit den professionellen Ansprüchen zusammen, die auch an „Crowdsourcing“ gestellt werden. Eine Firma, die Geld dafür bezahlt, dass Texte übersetzt werden, möchte dies mit den geringsten Kosten und der besten Qualität erledigt sehen. Dieser Wunsch ist verständlich und der Ausgangspunkt dafür, dass Systeme und Konzepte entwickelt wurden, um die Arbeiter auf Crowdsourcing-Plattformen dabei zu unterstützen qualitativ hochwertige Arbeit abzuliefern.

Diese Arbeit setzt genau bei der Frage an, wie man die Qualität der Arbeit im Crowdsourcing sichern kann. In der Vergangenheit wurden einige Möglichkeiten entwickelt, die es zum Ziel haben, die Arbeitsqualität zu verbessern. So werden oft von Hand Aufgaben kreiert, deren Lösung man bereits kennt. Diese werden dann zufällig unter die noch zu bearbeitenden Aufgaben gemischt und fungieren als „Honeypots“, also als Fallen, in welche weniger bemühte Arbeiter tappen sollen. Diese bewusst dafür eingesetzten Aufgaben nennt man auch „Golddaten“. Eine andere Möglichkeit ist es, mehreren Arbeitern die gleiche Aufgabe zu geben und zu untersuchen, wie sehr diese in ihrer Lösung übereinstimmen. Eine hohe Übereinstimmung deutet auf die beste Lösung hin. Diesen Ansatz nennt man „Mehrheitsvotum“. Die derzeit genutzten Qualitätssicherungsmechanismen sind in Abschnitt 2.4 ausführlich beschrieben.

1.3 Thema der Arbeit

Ein weiterer Ansatz, die Qualität der Arbeit auf Crowdsourcing-Plattformen zu verbessern, wird nun im folgenden dargestellt. Weiterhin wird das Thema der Arbeit konkretisiert.

1.3.1 Lösungsansatz dieser Arbeit

Um den Nachteilen der in Abschnitt 2.4 genauer beschriebenen Qualitätssicherungsmechanismen zu entgehen, wird in dieser Arbeit ein neuartiger Ansatz untersucht:

Ein existierendes Software-Tool generiert Wissensfragen bzw. Quizfragen mit mehreren Antwortmöglichkeiten, deren Lösungen bekannt sind. Diese Quizfragen werden mit den

zu bearbeitenden Aufgaben präsentiert. Anhand der Übereinstimmung der Lösung einer Quizfrage und der Antwort eines Arbeiters sollen Rückschlüsse gezogen werden können, wie zuverlässig die Arbeiter die eigentlichen Aufgaben bearbeiten. So könnten Arbeiter von der weiteren Bearbeitung ausgeschlossen werden, wenn sie eine festgelegte Anzahl von Quizfragen falsch beantwortet haben. Diese Vorgehensweise hat mehrere Vorteile:

- Die Bereitstellung der Quizfragen mit ihren Antworten erfolgt automatisch und ist somit nicht mit weiterem Zeitaufwand oder Kosten verbunden.
- Das aufwändige manuelle Anfertigen von „Honeypots“ kann entfallen.
- Man kann die Anzahl der Arbeitern pro Aufgabe reduzieren, da man Qualitätssicherungsmechanismen wie das „Mehrheitsvotum“ nicht mehr benötigt. Das spart finanzielle und zeitliche Ressourcen.
- Die Quizfragen sind für Bot-Software schwer zu beantworten und filtern diese zuverlässig heraus.

Daraus ergibt sich die zentrale Frage dieser Arbeit: *Wie gut funktionieren automatisch generierte Quizfragen als Qualitätssicherungsmechanismus im Crowdsourcing?*

1.3.2 Vorgehen

Um diese Frage zu klären, wurden für diese Arbeit drei typische Crowdsourcing-Jobs erstellt, die auf der Plattform *Crowdflower* zur Bearbeitung frei gegeben wurden. Die Ergebnisse der Jobs wurden darauf hin untersucht, ob die erfolgreiche Bearbeitung der Quizfragen Rückschlüsse auf die Arbeitsqualität der Arbeiter zulassen und, ob ein Ausschluss der Arbeiter, die bei den Quizfragen schlecht abschneiden, zu besseren Arbeitsergebnissen führt, als der Ausschluss anhand von falsch bearbeiteten Golddaten. Für diese Untersuchung wurde ein Arbeiter als ausgeschlossen betrachtet, wenn er drei oder mehr Quizfragen oder drei oder mehr Golddaten falsch beantwortet bzw. bearbeitet hatte. Die angefallenen Daten wurden statistisch auf Signifikanz überprüft, um belastbare Ergebnisse zu erhalten.

Außerdem wurde darauf geachtet, dass die Jobs den Qualitätskriterien des aktuellen Forschungsstandes entsprechen. Entsprechende Literatur wurde untersucht und die Mechanismen auf die gewählten Crowdsourcing-Jobs übertragen.

1.3.3 Aufbau der Arbeit

In diesem Kapitel wird beschrieben, wie die nachfolgende Arbeit aufgebaut ist. Sie lässt sich in folgende Kapitel einteilen:

Technische Grundlagen In diesem Kapitel werden alle Aspekte erläutert, die zum Verständnis der Arbeit notwendig sind. „Crowdsourcing“ wird definiert und durch nachvollziehbare Beispiele dem Leser erläutert. Außerdem wird die Crowdsourcing-Plattform „Crowdflower“ beschrieben und der Umgang damit vorgestellt. Zum Schluss wird die Qualitätsproblematik dargestellt, mit der sich Crowdsourcing-Plattformen konfrontiert sehen, sowie bestehende Lösungsansätze für dieses Problem aufgezeigt und diskutiert.

Task Design In diesem Kapitel wird beschrieben, welche Aufgabentypen nach welchen Kriterien für die Versuche in dieser Arbeit ausgewählt wurden. Es wird erläutert,

welche grundlegenden Gedanken und Überlegungen das Design der Tasks bestimmten. Anschließend wird die Umsetzung der einzelnen Tasks auf Crowdflower präsentiert und durch Screenshots veranschaulicht.

Technische Umsetzung Es wird beschrieben, was beim Erstellen der Ausgangsdatensätze zur Durchführung der Versuche zu beachten war. Anschließend wird präsentiert, welche Werkzeuge zur Evaluation der Ergebnisse eingesetzt wurden und wie die Generierung der Quizfragen abläuft. Zum Schluss wird beschrieben, wie die durchgeführten Versuche technisch umgesetzt wurden.

Evaluation Hier wird dargelegt, nach welchen Werten hin die Ergebnisse der Versuche untersucht wurden und wie die Berechnung der Simulation mit Golddaten umgesetzt wurde. Weiterhin werden die Ergebnisse der einzelnen Versuche evaluiert und besprochen. Weiterhin wird die in diesem Kapitel erarbeitete Frage beantwortet.

Probleme und Hindernisse Dieses Kapitel beschäftigt sich mit den Problemen, die Crowdsourcing als Möglichkeit der Datenverarbeitung mit sich bringt – im Bezug auf Entwickler, Gesellschaft und Wirtschaft. Weiterhin werden Hindernisse besprochen, die im Umgang mit der Crowdsourcingplattform „Crowdflower“ sowie bei der Generierung der Quizfragen auftraten.

2 Technische Grundlagen

In diesem Kapitel werden alle Aspekte erläutert, die zum Verständnis der Arbeit notwendig sind. Es wird definiert, was Crowdsourcing bedeutet und durch Beispiele verständlich gemacht. Weiterhin wird die Qualitätsproblematik dargestellt, mit der sich Crowdsourcingplattformen konfrontiert sehen, sowie bestehende Lösungsansätze für dieses Problem diskutiert.

2.1 Was ist Crowdsourcing?

Nachfolgend wird der Begriff „Crowdsourcing“ erläutert sowie seine Herkunft und Bedeutung geklärt. Beispiele aus der Praxis werden dazu genutzt, die Definitionen besser verständlich zu machen.

2.1.1 Begriffsherkunft und Definition

Der Begriff „Crowdsourcing“ wurde zum ersten Mal im Jahr 2006 von Jeff Howe, Redakteur für den Bereich „Media and Entertainment Industry“ des *wired*-Magazines, verwendet [5, S. 17]. Dabei setzt sich der Begriff „Crowdsourcing“ als Portemanteau-Wort aus den Einzelteilen „Crowd“ und „Outsourcing“ zusammen – also „Menschenmenge“ und „Auslagerung“. „Crowdsourcing“ hat also etwas damit zu tun, dass verschiedene Aufgaben zur Bearbeitung an eine große Anzahl von Menschen abgegeben werden.

Jeff Howe selbst vertritt folgende Definition von „Crowdsourcing“:

> *Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call* [13].

Bei „Crowdsourcing“ wird demnach eine Arbeit, die üblicherweise von einem festgelegten Akteur, meistens einem Angestellten, ausgeführt wird, über einen öffentlichen Aufruf an eine undefinierte und normalerweise große Gruppe von Menschen ausgelagert.

David C. Brabham legt seinem Buch „Crowdsourcing“ eine etwas ausgedehntere Definition zugrunde:

> *Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge, and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage what the user has brought to the venture, whose form will depend on the type of activity undertaken* [5, S. 2].

Brabhams Definition besteht im Wesentlichen aus folgenden Kernaussagen:

1. Crowdsourcing ist eine Online-Aktivität.

2. Eine Organisation hat Aufgaben, die erledigt werden sollen.
3. Es gibt eine heterogene Gemeinschaft, die diese Aufgaben freiwillig bearbeiten möchte.
4. Das Bearbeiten der Aufgaben bringt einen Nutzen für die Organisation sowie für die Gemeinschaft.
5. Ein Nutzer erhält eine Entlohnung - Geld, Anerkennung oder das Erwerben neuer Kenntnisse.

Dabei ist es wichtig, Crowdsourcing von weiteren Begrifflichkeiten abzugrenzen, die etwas mit „Crowd“ zu tun haben. Nach Brabham gibt es mehrere Begriffe, die oft als Crowdsourcing bezeichnet werden, aber kein Crowdsourcing sind: *Commons-based Peer Production*, wie es sich in Initiativen wie Wikipedia oder Open-Source-Software-Projekten niederschlägt, baut zwar auf einer Menge von Menschen auf, die zu einem Projekt gemeinsam etwas beitragen, kann aber wegen der fehlenden hierarchischen Struktur von Auftraggeber und Auftragnehmer nicht als Crowdsourcing bezeichnet werden. Auch *Marktforschung und Meinungsanalysen*, wie man sie über Crowdsourcing-Plattformen durchführen kann, sind kein Crowdsourcing, weil die hierarchische Struktur zu eng ist und der Kreativität der Auftragnehmer zu wenig Rechnung getragen wird [5, S. 7-8].

2.1.2 Beispiele aus der Praxis

Beispiele aus der Praxis sollen einen besseres Verständnis dafür geben, was Crowdsourcing konkret bedeutet.

threadless.com Threadless ist eine Online-Plattform, auf welcher Designer und Käufer von T-Shirts zusammengebracht werden. Interessenten können Vorlagen für T-Shirt-Designs herunterladen und ihre eigenen Designs auf die Plattform hochladen. Dort werden die Designs von potentiellen Käufern bewertet und die am besten bewerteten T-Shirt-Designs von Threadless in Produktion gegeben [5, S. 23]. Durch dieses beständige Einbinden der Kunden in den Produktgestaltungsprozess fällt es der Firma sehr leicht ihre Produktion ganz und gar auf die Wünsche ihrer Kunden abzustimmen und viele Arbeitsschritte (Kundenwünsche erfragen, Prototypen erstellen) werden überflüssig.

FoldIt Bei dieser Anwendung handelt es sich um ein Computerspiel, mit dem im Hintergrund ernsthafte Probleme gelöst werden sollen. Bei FoldIt versucht der Spieler die optimale Faltung von Proteinen zu finden, indem er als Rätsel verpackte Aufgaben löst: Proteine bestehen aus hunderten bis tausenden Aminosäuren und die Art und Weise, wie diese angeordnet sind, haben Auswirkungen auf die Eigenschaften des Proteins. Da diese vielen Möglichkeiten auch mit leistungsstarken Computern nur sehr mühsam zu berechnen sind, liegt es an der großen Gemeinschaft von Spielern die optimale Faltung verschiedener Proteine herauszufinden [9]. Auf diese Weise wird ein für Computer kompliziertes Problem in kleine Teile zerlegt und von einer „Crowd“ bearbeitet. So gelang es 2010 hunderten von Spielern innerhalb von 3 Wochen die Struktur eines Proteins zu entschlüsseln, das eine wichtige Rolle bei der Vervielfältigung des AIDS-Virus spielt, was in 15 Jahren Forschung zuvor nicht gelungen war [6].

Microtask-Plattformen Unter einem Microtask versteht man eine kurze und wenig komplexe Aufgabe, die im Rahmen des Crowdsourcings von Menschen absolviert wer-

den. Auf Plattformen wie *Amazon Mechanical Turk* oder *Crowdflower* wird frei skalierbare Arbeitsleistung auf Abruf angeboten. Arbeitgeber bieten kleine Microtasks an, welche von Arbeitnehmern ausgewählt und absolviert werden [10]. Diese Microtasks werden dann mit kleinen Cent-Beträgen entlohnt, die über die Plattformen ausgezahlt werden können. Die Aufgabentypen sind sehr unterschiedlich und bedienen z.B. die Bedürfnisse von IT-Firmen oder Wissenschaftlern: Transkriptionen von Audio- oder Bilddateien, Kategorisieren von kurzen Texten (Kommentare oder Tweets), Suchrelevanz-Analyse oder Bereinigen von Datensätzen (Dopplungen finden) [7]. Generell werden Aufgaben angeboten, deren Bearbeitung durch Computer unrentabel oder unmöglich ist. Dabei sind diese Plattformen mittlerweile so gestaltet, dass sie sich nahtlos in die Abläufe traditioneller Software einfügen lassen. So bietet das Plug-In „Soylent“ für Microsoft Word die Möglichkeit, eigene Texte von Menschen auf Microtask-Plattformen Korrektur lesen zu lassen [4].

2.2 Was ist Crowdflower?

In diesem Abschnitt wird erläutert, was Crowdflower[1] ist und wofür es genutzt werden kann. Es wird kurz umrissen, wie die Arbeit mit dieser Plattform grundsätzlich abläuft.

2.2.1 Crowdflower - eine Crowdsourcingplattform

Bei Crowdflower handelt es sich um ein Internet-Portal, das sich auf die Bearbeitung von kleinsten Aufgaben im Bereich der Informationsaufbereitung und -beschaffung – sogenannten *Microtasks* – spezialisiert hat. Diese Aufgaben werden von einer anonymen Gruppe, die über die ganze Welt verteilt ist, bearbeitet. Dabei sind die Menschen dieser großen anonymen Gruppe ausschließlich durch Crowdflower als Internetplattform miteinander verbunden. Wie schon in Kapitel 2 erläutert, ist Crowdflower also eine typische Crowdsourcingplattform. Crowdflower wird dabei einerseits von Wissenschaftlern genutzt, die Umfragen oder Datenerhebungen durchführen möchten, als auch von Unternehmen, die Informationen bereinigen oder aufbereiten wollen, um sie z.B. als Trainingsdaten für selbst lernende Systeme zu verwenden. Dabei besticht Crowdflower durch einen einfach zu nutzenden Editor, mit dem die Microtasks (auf Crowdflower *Jobs* genannt) gestaltet und freigegeben werden können. Darüber hinaus lässt sich aber alles auch über eine API steuern, sodass Crowdflower nahtlos in andere Software integriert werden kann.

2.2.2 Crowdflower nutzen

Die Nutzung von Crowdflower läuft im Allgemeinen folgendermaßen ab:

Ein neuer Job wird angelegt. Dabei werden zum Einen die Arbeitsanweisungen erstellt, wie auch die Eingabemaske für den Job gestaltet. Dies erfolgt mittels HTML sowie Crowdflowers eigener Auszeichnungssprache, der *Crowdflower Markup Language (CML)*.

Datensätze werden hochgeladen. Diese werden im CSV-Format bereitgestellt und lassen sich auf einer Übersichtsseite noch einmal begutachten. Um die hochgeladenen Daten in den Job einzupflegen, können sie beim Job-Design mit bestimmten Tags aufgerufen werden. Es werden dann für jeden neuen Datensatz die passenden Daten in die Eingabemaske eingearbeitet.

Zentrale Einstellungen werden vorgenommen. So ist es möglich zu bestimmen, wie viele Worker an jeweils einem Datensatz arbeiten dürfen, für welche Nationen der Job zu Ver-

[1] http://www.crowdflower.com

fügung stehen wird, wie viele Datensätze bearbeitet werden oder welche Datensätze als Golddaten genutzt werden sollen.

Der Job wird freigegeben. Danach steht der Job den Workern zur Verfügung und kann bearbeitet werden. Die Ergebnisse werden in Echtzeit präsentiert, sodass es möglich ist bei Bedarf in den Ablauf einzugreifen.

Die Ergebnisse werden präsentiert. Diese stehen, wie die Ausgangsdaten, als CSV-Dateien zu Verfügung. Sie enthalten die bearbeiteten Daten und verschiedene weitere hilfreiche Informationen – woher die Worker stammen, wann eine Aufgabe begonnen und abgeschlossen wurde, den Trust-Score eines jeden Workers etc.

2.3 Weitere Begrifflichkeiten

Weitere zum Verständnis der Arbeit wichtige Begriffe sollen nun erläutert werden. Vorhergehend genannte Begrifflichkeiten werden ebenfalls noch einmal kurz zusammengefasst:

Microtask Kurze und wenig komplexe Aufgabe.

Worker Mensch, der auf einer Crowdsourcingplattform Jobs bearbeitet.

Job Sammlung von meist gleichen Microtasks auf einer Crowdsourcingplattform.

Golddaten Bereits bekannte Ergebnisse von Microtasks, die zum Zweck der Qualitätssicherung genutzt werden.

2.4 Qualitätsproblematik und Lösungen

Dieser Abschnitt geht auf die Problematik ein, die im Bezug auf die Qualität bei Crowdsourcing besteht. Weiterhin werden mehrere schon bekannte Lösungsansätze für das genannte Problem diskutiert.

2.4.1 Qualitätsproblematik

Wie oben beschrieben, ist Crowdsourcing eine moderne und flexible Möglichkeit, um Aufgaben lösen zu lassen, die für eine automatisierte Lösung zu schwierig oder zu teuer sind. Durch den Einsatz von menschlicher Arbeit lassen sich solche Probleme unkompliziert bearbeiten und die Kosten niedrig halten. Die Tatsache aber, dass an diesem Prozess Menschen beteiligt sind, führt dazu, dass sich Fehler in der Ausführung nicht verhindern lassen. Insbesondere auf Microtask-Plattformen kommt es immer wieder vor, dass Worker aufgrund verschiedener Faktoren (mangelnde Konzentration, ungenügende Kompetenzen etc.) nicht die gewünschte Zuverlässigkeit im Bezug auf die Qualität ihrer Arbeit erbringen, oder sogar bewusst Arbeitsergebnisse fälschen, um schneller die monetäre Entlohnung zu erhalten. Auch gezielte Angriffe auf Crowdsourcing-Plattformen sind bekannt, in denen Software eingesetzt wird, welche die Aufgaben in unzureichender Güte automatisiert absolvieren, um die menschliche Arbeitsleistung gar nicht erst erbringen zu müssen. Der Qualitätssicherung im Crowdsourcing kommt also eine besondere Bedeutung zu und hat die Aufgabe sicherzustellen, dass durch Crowdsourcing zuverlässig qualitativ hochwertige Daten produziert werden.

2.4.2 Bisherige Lösungsansätze

Um die oben genannten Probleme zu beherrschen, gibt es eine Reihe von Lösungsansätzen mit unterschiedlichen Vor- und Nachteilen. Diese sollen im Folgenden kurz vorgestellt werden.

Qualifikationstests Bevor ein Worker die eigentliche Aufgabe bearbeiten darf, muss er einen Test absolvieren, um seine Kompetenzen und seine Arbeitsmoral unter Beweis zu stellen. Dieser Test ist der eigentlichen Aufgabe sehr ähnlich [21, S. 309]. Wenn der Worker die im Qualifikationstest gestellten Aufgaben nicht zufriedenstellend löst, wird er zur Bearbeitung der eigentlichen Aufgabe nicht zugelassen. Vorteilhaft bei dieser Art der Qualitätssicherung ist, dass der Worker durch die im Qualifikationstest gestellten Aufgaben schon weiß, was ihn bei den kommenden Aufgaben erwartet. Nachteilig ist, dass die Anfertigung und Wartung solcher Testfragen zusätzliche Kosten verschlingt und sich ungünstig auf die Motivation der Worker auswirken kann [3, S. 4]. Ebenso muss ein solcher Test gut überlegt angefertigt werden, sodass ehrlichen Workern die Arbeit nicht zu sehr erschwert wird, er aber für Bots schwierig zu lösen ist [8, S. 3].

Honeypots Bei diesem Mechanismus nutzt man Aufgaben, für welche die Lösungen schon bekannt sind, und mischt sie unter die eigentlichen Aufgaben. Über die Abweichung bzw. Übereinstimmung der Lösungen eines Workers und den bekannten Lösungen versucht man Rückschlüsse auf die generelle Qualität der Arbeit eines Workers zu ziehen. So wird der Worker bei starker Abweichung von der weiteren Bearbeitung der Aufgabe ausgeschlossen oder die Ergebnisse dieses Workers im Nachhinein nicht gewertet. Vorteilhaft ist die einfache Umsetzung dieses Verfahrens und die Möglichkeit den Worker *während* der Bearbeitung der Aufgaben überprüfen zu können. Nachteilig ist, dass es viel Zeit kostet die passenden Datensätze zu erzeugen und ein „Honeypot“ gewartet werden muss, da sonst die Worker die Testfragen leicht identifizieren können [20].

Mehrheitsvotum Wenn eine Aufgabe auf einer Crowdsourcingplattform freigegeben wird, wird sie nicht nur einem sondern gleich mehreren Workern angeboten. Haben alle Worker diese Aufgabe bearbeitet, errechnet man den die Übereinstimmung der Antworten und wählt als Lösung jene Antwort, die unter den Workern die höchste Übereinstimmung erzielt hat. Dieses Verfahren ist mathematisch sehr einfach und in den meisten Crowdsourcingplattformen implementiert. Nachteilig ist, dass man an Arbeitskraft verliert und dadurch die Kosten steigen, da sich mehrere Worker mit der selben Aufgabe beschäftigen und dafür bezahlt werden müssen.

Review Bei diesem Qualitätssicherungsmechanismus löst ein Worker eine Aufgabe. Anschließend wird die gleiche Aufgabe mit der gewählten Lösung mehreren anderen Workern präsentiert, die über die angebotene Lösung abstimmen. So erhält man eine Einschätzung, von welcher Qualität die Lösung des ersten Workers ist. Dies ist besonders vorteilhaft, da es wesentlich einfacher ist eine bestehende Lösung zu beurteilen, als die Aufgabe von Grund auf neu zu lösen. Problematisch dabei ist jedoch, dass die Reviewer oftmals dazu neigen, der angebotenen Lösung vorbehaltlos zuzustimmen, was zu nicht identifizierbaren Lösungen führt [11]. Jedoch können mit dieser Vorgehensweise auch komplexe Antworten, wie sie z.B. bei Transkriptionen entstehen, auf ihre Qualität untersucht werden.

Qualitätsbeurteilung der Worker Auf den meisten Crowdsourcing-Plattformen wird einem Worker in Abhängigkeit seiner erbrachten Leistungen und der Zufriedenheit

seiner Auftraggeber ein numerischer Wert zugeordnet, der Aussagen über die Fähigkeiten, Zuverlässigkeit und die Arbeitsqualität des Workers treffen soll. Dieser Wert wird z.B. als „Trust Score“ bezeichnet. Auftraggeber können so im Vorfeld Worker ausschließen, die unterhalb eines gewünschten Schwellwertes liegen, und somit versuchen die Qualität der Arbeitsergebnisse zu erhöhen. Bedauerlicherweise hat ein einziger Wert nur eine begrenzte Aussagekraft. So kann es sein, dass ein Worker, der zuvor exzellente Kenntnisse im Kategorisieren von Bildern erworben hat, mit seinem hohen „Trust Score“ an eine Übersetzungsaufgabe gerät – für welche er jedoch nicht qualifiziert ist [18].

Aufgaben mit eindeutigen Lösungen Zusätzlich zu den eigentlichen Aufgaben werden Fragen eingeblendet, die eine eindeutige, „objektive“ Lösung besitzen. Dies können etwa Mathematik-Aufgaben oder einfache Multiple-Choice-Fragen sein. Solche Aufgaben werden auch „Objective Tasks“ genannt [2]. Da sich die Qualität der Antworten auf solche „Objective Tasks“ automatisiert überprüfen lässt, versucht man Rückschlüsse auf die Arbeitsqualität der Worker im Allgemeinen zu ziehen. Nachteilig ist, dass solche Fragen von Menschen angefertigt werden müssen. Das kostet viel Zeit und finanzielle Ressourcen. Greift man auf einen einmal erstellten, bestehenden Pool an Fragen zurück, besteht die Gefahr, dass die Worker die Fragen wiedererkennen oder die Lösung schon wissen. Diese Nachteile kann der Lösungsansatz, der in dieser Arbeit beschrieben ist, überwinden.

3 Task Design

In diesem Kapitel wird beschrieben, nach welchen Kriterien die Aufgabentypen für die Versuche in dieser Arbeit ausgewählt wurden. Weiterhin wird erläutert, welche grundlegenden Gedanken und Überlegungen das Design der Tasks bestimmten. Zum Schluss wird die Umsetzung der einzelnen Tasks auf Crowdflower präsentiert.

3.1 Auswahl der Aufgaben für Crowdflower

Es wurden verschiedene Arten von Microtasks ausgesucht, die als Basis für die Versuche in dieser Arbeit dienten. Dabei sollten die Microtasks gewissen Anforderungen genügen:

- Die gewählten Microtasks mussten die gängige Praxis auf Crowdsourcing-Plattformen widerspiegeln.
- Die Lösungen der Microtasks mussten dem Versuchsleiter bekannt sein, um anschließend die Arbeit der Worker mit den Lösungen vergleichen zu können.
- Die Microtasks mussten schwierig genug sein, um den Workern einen gewissen Spielraum bei der ihrer Arbeit zu lassen und dennoch so einfach, dass sie lösbar blieben.

3.1.1 Typische Microtasks

Um Microtasks zu finden, welche für Crowdsourcing-Plattformen typisch sind, wurden offizielle Dokumente, Literatur und die Task-Beispiele auf Crowdflower untersucht. Die häufigsten Arten von Microtasks auf Crowdflower sind der Literatur zufolge *Data Enhancement*, *Surveys*, *Categorization*, *Photo Moderation*, *Content Creation*, *Sentiment Analysis* und *Translation* [16].

In *Analyzing the amazon mechanical turk marketplace* [15] wurde die Plattform *Amazon Mechanical Turk* untersucht und eruiert, welche Schlagwörter am häufigsten in den Tasks dieser Plattform auftauchen. Nach der Anzahl der absolvierten Microtasks geordnet, fallen in den fünfzig häufigsten Schlagwörtern z.B. folgende auf: *categorization, collection, categorize, image, search, tagging, tag, photo, picture*. Das legt nahe, dass Kategorisierung als Task sehr beliebt ist, ebenso wie Microtasks, die etwas mit Bildern zu tun haben. Weiterhin scheinen Tasks populär zu sein, bei denen etwas im Internet gesucht werden muss sowie das Verschlagworten von Informationen.

Crowdflower bietet beim Erstellen eines neuen Jobs an, fertige Vorlagen als Ausgangsbasis zu nehmen. Auch diese können ein Hinweis sein, welche Arten von Microtasks typisch für diese Plattform sind. Von Crowdflower werden als Vorlagen angeboten: *Sentiment Analysis, Search Relevance, Data Categorization, Data Collection, Data Validation, Image Annotation, Transcription* und *Content Moderation*.

In Tabelle 3.1 sind die Bedeutungen der oben genannten Job-Typen erläutert.

Job-Typ	Beschreibung
Categorization	Entscheiden, welcher Kategorie eine Information, z.B. ein Text oder ein Bild - zugeordnet wird.
Content Creation	Erstellen neuer Informationen, z.B. eine Textzusammenfassung anfertigen.
Content Moderation	Z.B. Beiträge auf sozialen Plattformen einordnen.
Data Categorization	Informationen nach Kategorien ordnen.
Data Collection	Sammeln von Datensätzen; z.B. Bilder zu einem Schlagwort finden.
Data Enhancement	Vervollständigen von Datensätzen; z.B. zu einem Firmennamen eine Adresse finden. Wird auch „Data Enrichment“ genannt.
Data Validation	Daten auf ihre Richtigkeit überprüfen.
Image Annotation	Bilder mit Schlagworten versehen.
Photo Moderation	Beantworten von Fragen wie „Sind auf diesem Bild Personen zu sehen?“
Search Relevance	Entscheiden, ob Such-ergebnisse einer Suchmaschine zu einem Begriff passen.
Sentiment Analysis	Informationen nach ihrer vermittelten Stimmung zu bewerten.
Surveys	Durchführen von Umfragen.
Translation	Einzelne Wörter bis ganze Textpassagen übersetzen.

Tabelle 3.1: Übersicht der häufigsten Job-Typen auf Crowdflower.

3.1.2 Microtasks mit bekannten Lösungen

Die Microtasks für die durchgeführten Versuche wurden so ausgewählt, dass die Lösungen für die Aufgaben schon bekannt waren oder schnell generiert werden konnten. Dies ist nur dann möglich, wenn für die gewählte Aufgabe auch eindeutige Lösungen existieren. Aus diesem Grund wurden all jene Aufgabentypen abgelehnt, bei denen eine eindeutige Lösung nicht möglich ist. Dies betrifft:

Sentiment Analysis Hierbei wird von Workern die Stimmung eines Textes beurteilt. Eine solche Beurteilung ist natürlich subjektiv und besitzt keine eindeutige Lösung.

Surveys Bei Umfragen ist die subjektive Beurteilung von Fragen gerade erwünscht und damit für diese Arbeit nicht geeignet.

Content Creation Auch Jobs, bei denen die Worker zum weitgehend eigenständigen Arbeiten aufgefordert werden, lassen einen zu großen Spielraum bei möglichen Lö-

sungen. Daher ist diese Art von Jobs nicht geeignet.

Translation Da es für viele Wörter unterschiedliche Bedeutungen gibt, eignen sich Übersetzungen ebenso wenig als Job-Typen für die Versuche dieser Arbeit.

3.1.3 Lösungsschwierigkeit der Tasks

Die Schwierigkeit der Tasks sollte derart sein, dass der Lösungsspielraum eines Jobs klein aber doch vorhanden ist. Ein zu geringer Lösungsspielraum führt dazu, dass die Abweichungen in der Qualität der Arbeit zu gering sind, um sie zu untersuchen. Ein zu großer Lösungsspielraum führt zu unvorhersehbaren Lösungen, die nicht miteinander verglichen werden können. Neben der grundsätzlichen Art eines Jobs spielt es aber auch eine Rolle, wie der Arbeitsablauf einer Aufgabe gestaltet wird. Wird der Worker bei seiner Arbeit zu stark „an die Hand genommen“, können sich Unterschiede in der Arbeitsqualität nicht mehr niederschlagen. Lässt man dem Worker zu große Freiheiten, lassen sich wie eben erwähnt die Ergebnisse nicht mehr miteinander vergleichen.

3.1.4 Ausgewählte Tasks

Für die Versuche dieser Arbeit wurden drei verschiedene Arten von Tasks ausgewählt:

Transcription Die Worker sollen ein Wort, was als gesprochener Text in einer Audio-Datei vorliegt, erkennen und in eine Textbox schreiben. Wie oben erläutert, spielt dieser Aufgaben-Typ in der Praxis auf Crowdflower eine große Rolle. Die Audio-Dateien wurden mit einem Text-To-Speech-System erzeugt, sodass ein geringer Lösungsspielraum vorhanden ist. Ebenso ist auf diese Weise garantiert, dass die Lösungen im Vorfeld vorhanden sind.

Categorization Auch dieser Aufgaben-Typ ist Teil der Praxis auf Crowdflower. Die Worker sollen Bilder, die ihnen präsentiert werden, in vier verschiedene Kategorien einteilen (Car, House, Flower oder People). Die Bilder selbst entstammen einer Stockphoto-Seite und sind schon Kategorien zugeteilt – die Lösungen sind also vorhanden. Der Lösungsspielraum ist gegeben und bewegt sich zwischen den vier wählbaren Kategorien.

Data Enhancement Dieser Task wird auch „Data Enrichment“ genannt. Hierbei sollen die Worker zu einem gegebenen Filmtitel mithilfe einer Suchmaschine das Erscheinungsjahr des Films herausfinden und in ein Textfeld eintragen. Die Lösungen sind bekannt, da die Zuordnung von Filmtitel zu Erscheinungsjahr in einem früheren Job auf Crowdflower vorgenommen wurde. Der Lösungsspielraum ist vorhanden aber klein, da ausschließlich nach dem Erscheinungsjahr gefragt wurde.

3.2 Grundsätzliche Überlegungen beim Task-Design

In diesem Abschnitt werden Design-Entscheidungen erörtert, die unabhängig vom jeweiligen Task erfolgt sind. Es wird beschrieben, nach welchem Prozess ein Task gestaltet wurde und wie sich die Arbeitsanweisungen eines Tasks grundsätzlich aufbauen. Weiterhin wird auf die Aspekte des User-Interface, den Einsatz der Multiple-Choice-Frage, die Preisgestaltung und sprachliche Probleme eingegangen.

3.2.1 Review-Prozess

In *Practical Lessons for Gathering Quality Labels at Scale* [3] beschreibt Omar Alonso einen dreigeteilten Prozess, um Crowdsourcing-Tasks zu entwerfen. Dieser Prozess wurde für den Entwurf der Versuche dieser Arbeit adaptiert. In der ersten Phase wurde ein Prototyp entworfen und vier Mitstudenten zum Testen vorgelegt. Die Rückmeldungen und Hinweise der Tester wurden zur Optimierung des Task-Designs genutzt. Anschließend wurde das neue Task-Design mit wenigen Datensätzen auf Crowdflower freigegeben. Hier wurden zum Einen die Ergebnisse selbst als Hinweis auf mögliche Verbesserungen untersucht, als auch die Rückmeldungen in der auf Crowdflower angebotenen Umfragen. Abbildung 3.1 zeigt die Ergebnisse einer Worker-Befragung auf Crowdflower. „Overall“ spiegelt die Gesamtzufriedenheit der Worker mit dem Job wider, „Instructions Clear“ sagt aus, wie gut verständlich die Worker die Arbeitsanweisungen empfanden. „Test Questions Fair“ soll eine Rückmeldung zu den eingesetzten Golddaten geben (auch, wenn diese in dieser Arbeit nicht eingesetzt wurden), „Ease Of Job“ gibt die Einschätzung der Worker über die Schwierigkeit der bearbeiteten Aufgaben wider. Unter „Pay“ kann man die Zufriedenheit der Worker mit der Bezahlung nachvollziehen.

Nach weiteren Optimierungen des Task-Designs und des Arbeitsablaufs wurde die Tasks mit allen Datensätzen zur Bearbeitung freigegeben. Auch hier wurden zur Beurteilung der Güte des Task-Designs die Ergebnisse des Jobs ausgewertet sowie die Umfragen mit einbezogen. Nach diesen drei Phasen wurde das Task-Design als abgeschlossen betrachtet.

Abbildung 3.1: Umfrage auf Crowdflower zur Zufriedenheit der Worker.

3.2.2 Instruktionen der Tasks

Jeder Task auf Crowdflower besteht einmal aus den Arbeitsanweisungen – Instruktionen – und aus der Oberfläche zur Bearbeitung der präsentierten Aufgaben. Beide Teile sind in HTML bzw. CML umgesetzt. Bei den ersten Versuchen wurden die Instruktionen nach den Vorlagen auf Crowdflower gestaltet. Abbildung 3.2 zeigt in dieser Weise gestaltete Arbeitsanweisungen für den Task „Audio-Transcription“.

Die Arbeitsanweisungen bestanden zunächst aus folgenden Punkten, die in deutlich größerer Schrift und farblich hervorgehoben dargestellt wurden:

Task Eine kurze Beschreibung der Aufgabe.

Overview Eine Übersicht, was zu tun ist, welche Eingabemöglichkeiten es gibt und welches Equipment weiterhin benötigt wird.

Before you start Hinweise, was beachtet werden muss, bevor der Job begonnen wird.

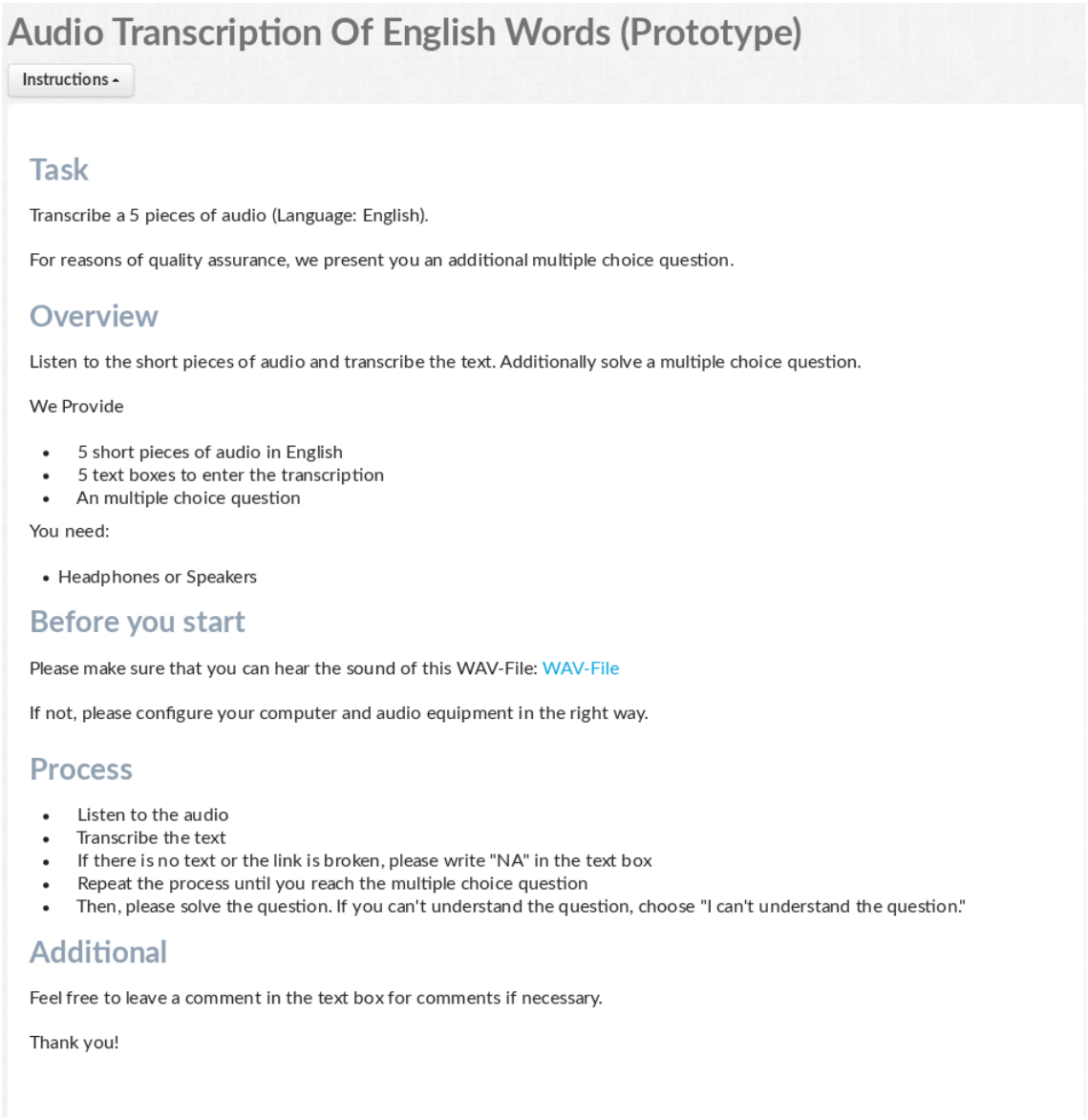


Abbildung 3.2: Umsetzung der Arbeitsanweisungen im Prototyp.

Process Eine Zusammenfassung in Stichpunkten, welche Handlungen in welcher Reihenfolge auszuführen sind.

Additional Zusätzliche Hinweise – z.B. über die Nutzung des Kommentarfeldes.

Im Verlauf des Review-Prozesses stellte sich allerdings heraus, dass die Tester kaum Zeit damit verbrachten die Instruktionen zu lesen, sondern sich entweder auf vorhandene Beispiele stürzten oder sich gleich an die Bearbeitung der Aufgaben machten. Damit war klar, dass das *Tun* wichtiger ist als das *Lesen*. Daraus wurden folgende Konsequenzen gezogen: Der Text wurde stark gestrafft und auf die nötigsten Hinweise reduziert. Wichtige Worte wurden fett hervorgehoben. Des weiteren wurde ein Bild an die Stelle im Text eingefügt, die als die Wichtigste erachtet wurde. Abbildung 3.3 zeigt den verbesserten Entwurf des Designs des Tasks „Audio-Transcription“.

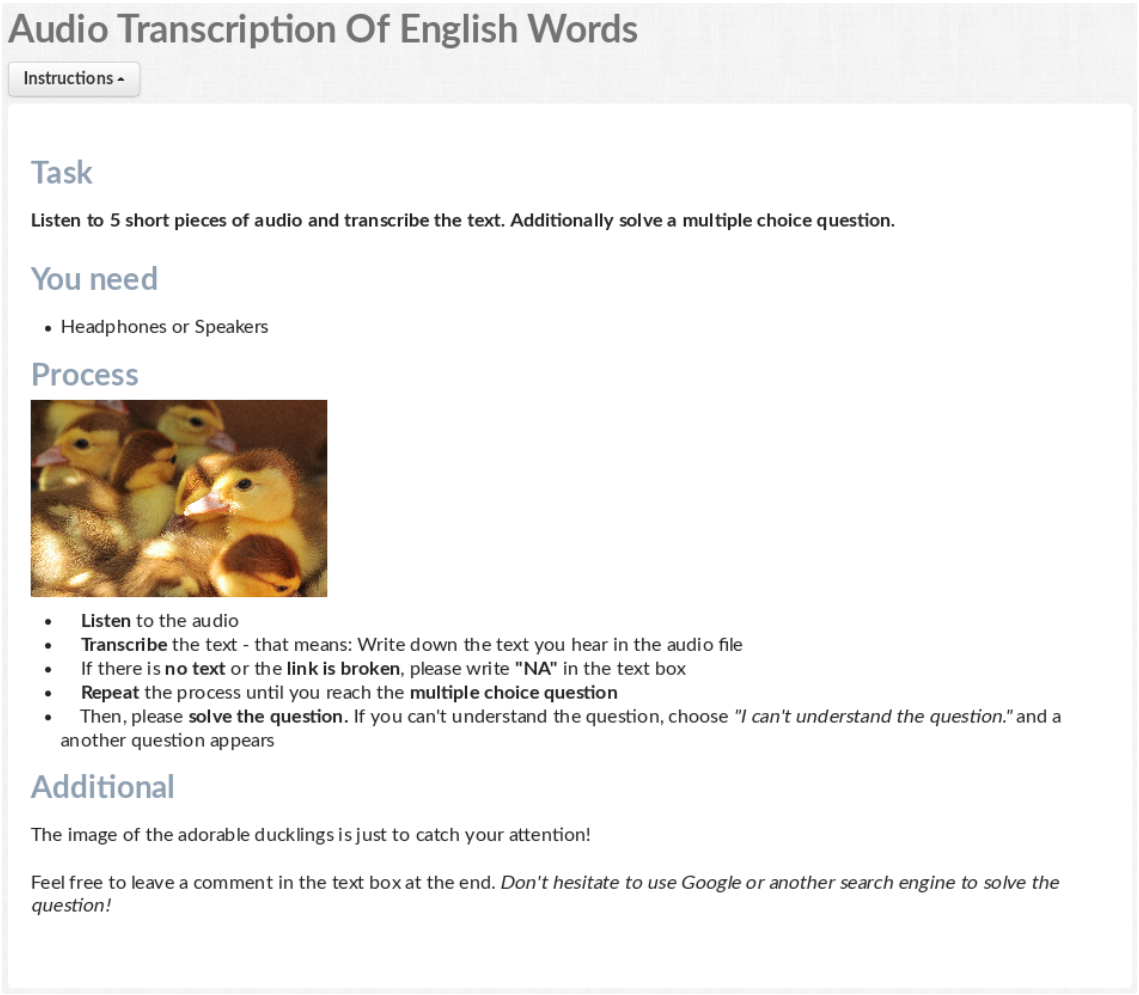


Abbildung 3.3: Verbesserte Arbeitsanweisungen des Prototyps.

3.2.3 Das User-Interface der Tasks

Die Eingabemaske wurde grundsätzlich so zweckmäßig wie möglich gestaltet und auf eine aufwendige optische Umsetzung verzichtet. Auch die Beschriftungen wurden so knapp wie möglich gehalten, um die Worker nicht von ihrer eigentlichen Aufgabe abzuhalten. Diese Umsetzung wurde auch nach dem Review-Prozess kaum verändert. Es wurden einfach nutzbare Steuerelemente verwendet, wie Links, Textboxen und Auswahlfelder. Um die Eingaben der Worker auf ein brauchbares Format zu beschränken, wurden für die Textboxen Validatoren verwendet. Ebenso wurden Validatoren genutzt, um sicherzustellen, dass jeder Link besucht wurde und in einem Auswahlfeld auch eine Möglichkeit ausgewählt wurde. Lediglich die Alternativfrage und das Kommentarfeld blieben optional.

3.2.4 Die Multiple-Choice-Frage

Das zentrale Element in diesen Versuchen stellt die am Ende eines jeden Tasks präsentierte Multiple-Choice-Frage dar. Die Frage selbst wurde mit Fettdruck und größerer Schrift hervorgehoben. Anschließend wurden vier verschiedene Antwortmöglichkeiten präsentiert, von denen nur eine als richtig gewertet wurde. Die Position der richtigen Antwort variierte von Frage zu Frage. Zusätzlich bekam jede Frage die Auswahlmöglichkeit „I can’t understand the question“, um Verständnisschwierigkeiten, die bei computergene-

rierten Fragen durchaus vorkommen, zu entgegnen. Auf eine Auswahlmöglichkeit, wie „Ich weiß die Antwort nicht“, wurde bewusst verzichtet, um die möglichen Auswahlmöglichkeiten nicht zu überfrachten und den ersten Eindruck einfach zu halten. Ein Beispiel für eine Multiple-Choice-Frage ist in Abbildung 3.4 zu sehen.

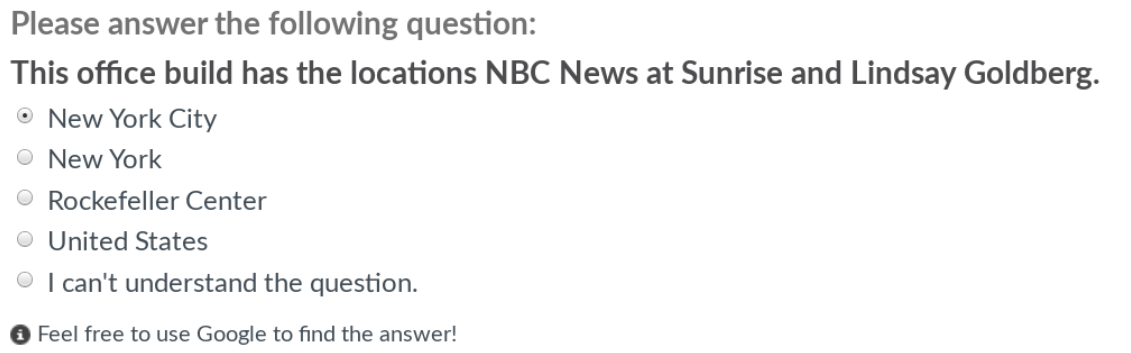


Abbildung 3.4: Eine Multiple-Choice-Frage, wie sie ein Worker zu sehen bekommt.

Die Arbeit mit den Testern im ersten Review-Schritt offenbarte, dass eine alternative Frage notwendig war. So wurde anschließend eine zweite Frage implementiert, die dann erschien, wenn ein Worker bei der ersten Frage die Auswahlmöglichkeit „I can’t understand the question“ markierte. Abbildung 3.5 zeigt, dass eine Auswahl von „I can’t understand the question“ bei der ersten Frage eine weitere Frage erscheinen lässt.

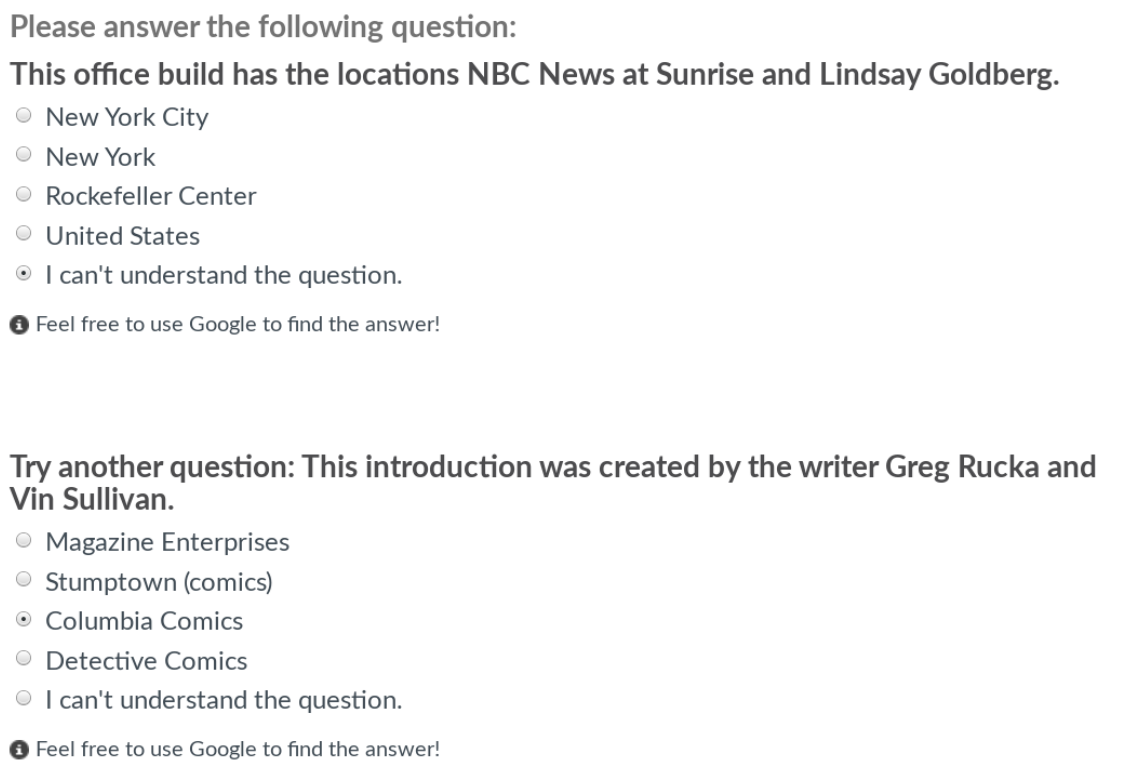


Abbildung 3.5: Die Alternativfrage öffnet sich.

Da das Lösen einer solchen Frage im Verhältnis zum Lösen der Aufgaben viel Zeit beansprucht, wurde nicht an jede Aufgabe eine Multiple-Choice-Frage angefügt. Nach

fünf Aufgaben war jeweils eine Frage zu lösen. Dies stellte sicher, dass die „zusätzliche Belastung“ so klein wie möglich war.

3.2.5 Preisgestaltung

Auf Croudsourcing-Plattformen ist es üblich, dass die Worker für ihre Arbeit bezahlt werden, dabei sollte die Bezahlung der Schwierigkeit und dem zeitlichen Einsatz angemessen sein. In den Versuchen mit den Testpersonen wurde die Zeit gemessen, die jeweils zur Bearbeitung von fünf Aufgaben und einer Multiple-Choice-Frage nötig war. Diese Messung diente als Basis für die Preisgestaltung. Es wurden aber auch die Rückmeldungen der Worker aus vorangegangenen Versuchen auf Crowdflower mit einbezogen. Es wurden drei Preisstufen verwendet:

10 Cent Dies war die Ausgangsbasis und die Vergütung für „Image-Categorization“. Fünf Aufgaben mit dem Lösen der Multiple-Choice-Frage wurden von den Testern in ca. *45 Sekunden* bearbeitet.

12 Cent Dieser Preis wurde für „Audio-Transcription“ gezahlt. Die Arbeit ist ein wenig aufwendiger, als bei „Image-Categorization“, da es nötig ist eine Audio-Datei ein oder mehrmals anzuhören und den Text in ein Textfeld zu schreiben. Die Bearbeitungszeit der Tester lag hier bei etwa *55 Sekunden*.

15 Cent Im Versuch „Data-Enhancement“ wurde am meisten gezahlt. Dieser Task wurde von den Testern in ca. *80 Sekunden* gelöst und war dadurch, dass die Suchmaschine Google genutzt werden musste, am schwierigsten.

3.2.6 Nationen teilnehmender Worker

In den ersten Versuchen mit Multiple-Choice-Fragen als Qualitätssicherungsmechanismus wurde beobachtet, dass viele Worker die Fragen nicht verstanden. Dies lag unter anderem daran, dass Worker aus Ländern einbezogen wurden, die die englische Sprache nicht auf einem solchen Niveau beherrschten, um die Fragen zu verstehen. Um dieses Problem zu umgehen, wurden die hier dokumentierten Versuche ausschließlich Workern aus dem *Vereinigten Königreich*, den *Vereinigten Staaten von Amerika*, aus *Irland* und *Kanada* zur Bearbeitung präsentiert.

3.3 Design der einzelnen Versuche

In diesem Abschnitt wird wird das Design der in der Arbeit durchgeführten Versuche vorgestellt und durch umfangreiche Screenshots visualisiert.

3.3.1 Audio-Transcription

Die Abbildungen 3.6, 3.7 und 3.8 zeigen den Task „Audio-Transcription“. Die Worker sollten fünf Audio-Dateien anhören und die Wörter, die in den Audio-Dateien zu hören sind, in die dafür vorgesehene Textbox eintragen. Danach musste eine Multiple-Choice-Frage beantwortet werden. Zum Schluss bekamen die Worker die Möglichkeit Anmerkungen in einem Kommentarfeld zu hinterlassen.

Ein aussagekräftiges Beispiel musste bei diesem Task ausgelassen werden, da die Weiterleitung von Crowdflower auf eine Audio-Datei auf dem Server des Software-Labors zwar im User-Interface, aber nicht in den Arbeitsanweisungen möglich war.

3.3.2 Image-Categorization

Die Abbildungen 3.9, 3.10 und 3.11 zeigen die Arbeitsanweisungen des Tasks. In den Abbildungen 3.12, 3.13 und 3.14 ist das User-Interface zu sehen, mit welchem die Worker den Task bearbeiteten. Die Worker sollten aus vier Kategorien diejenige auswählen, die am besten zum gezeigten Bild passte. Für den Fall, dass ein Bild nicht korrekt angezeigt würde, war die Antwortmöglichkeit „This image does not display correctly“ vorgesehen.

Das Beispiel nimmt in diesem Task viel Raum ein, da die entsprechenden Bilder angezeigt werden mussten. Um besser zu verdeutlichen, welche Kategorie zu welchem Bild ausgewählt werden sollte, wurde unter dem Bild ein schon markiertes Auswahlfeld eingefügt.

3.3.3 Data-Enrichment

Die Abbildung 3.15 zeigt die Arbeitsanweisungen des Tasks. In den Abbildungen 3.16 und 3.17 ist das User-Interface zu sehen. Die Worker sollten auf den angegebenen Link klicken, der eine Suchanfrage auf *www.google.com* öffnet. Danach sollten die Worker das Jahr suchen, in welchem der angezeigte Film veröffentlicht wurde, und es in die Textbox eintragen.

Audio Transcription Of English Words

Instructions

Task

Listen to 5 short pieces of audio and transcribe the text. Additionally solve a multiple choice question.

You need

- Headphones or Speakers

Process

- **Listen** to the audio
- **Transcribe** the text - that means: Write down the text you hear in the audio file
- If there is **no text** or the **link is broken**, please write **"NA"** in the text box
- **Repeat** the process until you reach the **multiple choice question**
- Then, please **solve the question.** If you can't understand the question, choose *"I can't understand the question."* and a another question appears

Additional

The image of the adorable ducklings is just to catch your attention!

Feel free to leave a comment in the text box at the end. *Don't hesitate to use Google or another search engine to solve the question!*

Abbildung 3.6: Arbeitsanweisungen von „Audio-Transcription“.

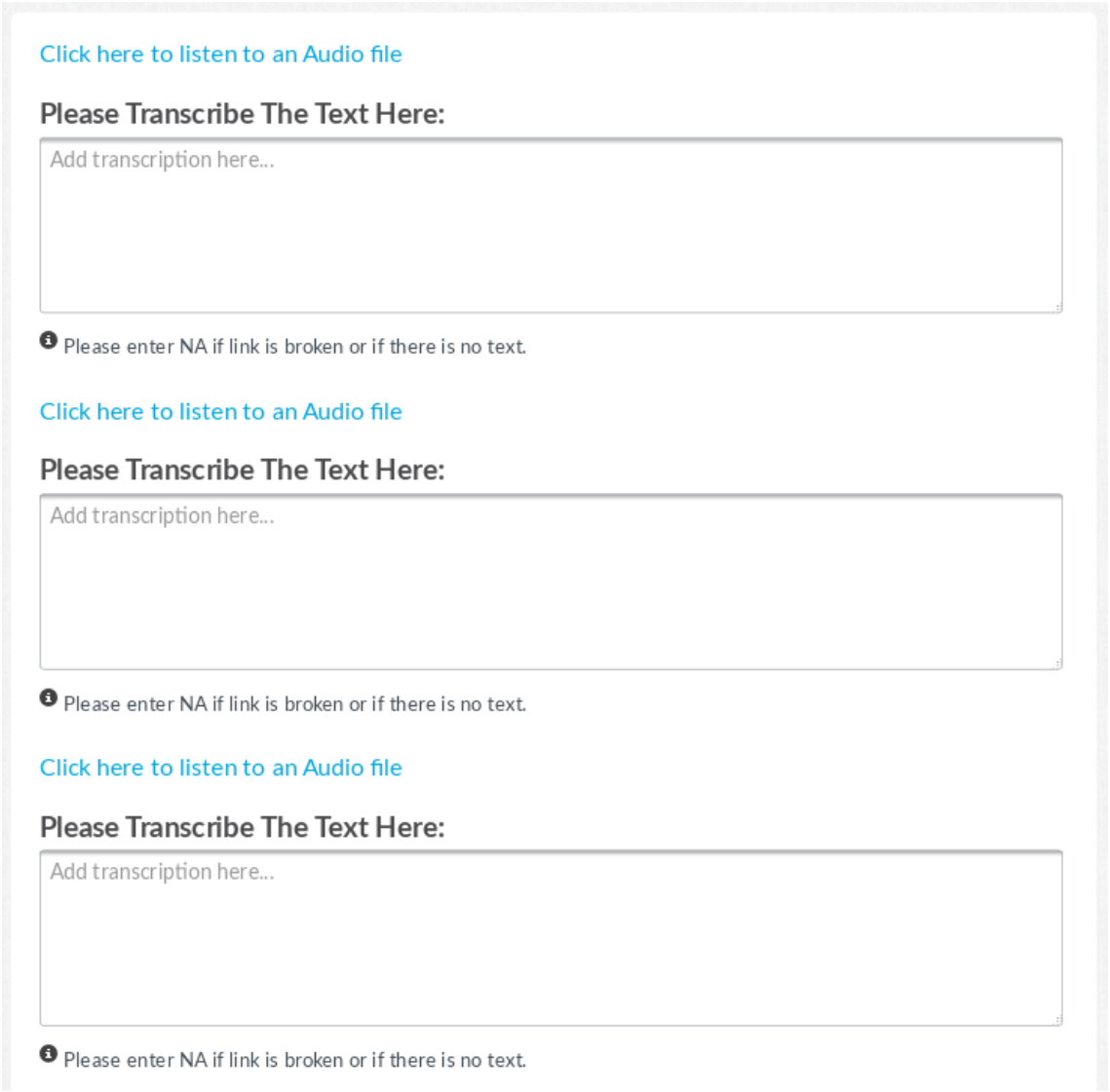


Abbildung 3.7: Erster Teil des User-Interfaces von „Audio-Transcription“.

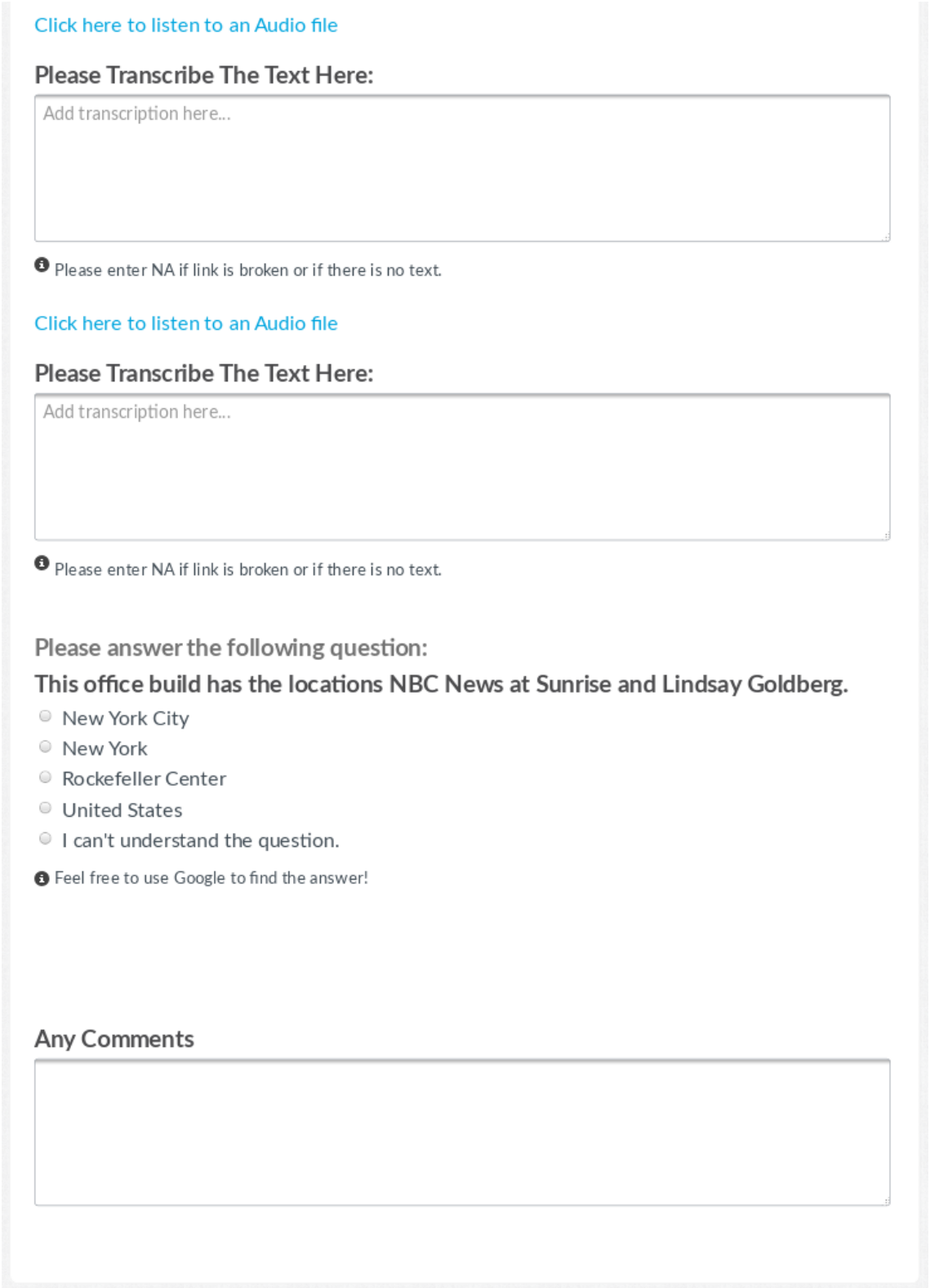


Abbildung 3.8: Zweiter Teil des User-Interfaces von „Audio-Transcription“.

Image Categorization

Instructions

Task

Look at an image and choose the best category for that image. Additional solve a multiple choice question.

Process

1. **Review** the image provided.
2. Click on the button of the **best match** category.
3. Answer the **multiple choice question.** If you can not understand the question, please choose *"I can't understand the question."* and another question appears.

Tips and Examples

Tips

- Feel free to leave a comment in the text box below if necessary.
- *Don't hesitate to use Google or another search engine to solve the multiple choice question*

Examples

Car

- House
- Flower
- People
- Car
- This image does not display correctly

- Choose category "Car" whenever the image shows one or more cars.
- Also choose category "Car", if the image shows one or more things related to a car (a tire, steering wheel etc.)

Abbildung 3.9: Erster Teil der Arbeitsanweisungen von „Image-Categorization“.

Flower

- House
- Flower
- People
- Car
- This image does not display correctly

- Choose category "Flower" whenever the image shows one or more flowers.
- Choose category "Flower", if the main focus of the image is on the flower themselves.

House

- House
- Flower
- People
- Car
- This image does not display correctly

- Choose category "House" whenever the image shows **one or more houses or buildings.**
- Choose category "House", if the image shows **things related to houses or buildings** and the main focus is on the house or building themselves.

Abbildung 3.10: Zweiter Teil der Arbeitsanweisungen von „Image-Categorization“.

People

- House
- Flower
- People
- Car
- This image does not display correctly

- Choose category "People" whenever the image shows **one person or many people.**
- Choose category "People", if the **main focus** of the image is on a person ore many people.

Link is broken?

If the link is broken or you just can't see the image, please choose the last option:

- House
- Flower
- People
- Car
- This image does not display correctly

Summary

For this task, you will be selecting the **best matching** category for an image.

Thank You!

Your careful attention on this task is greatly appreciated!

Abbildung 3.11: Dritter Teil der Arbeitsanweisungen von „Image-Categorization“.

Which of these categories does the above image fit into best?

- House
- Flower
- People
- Car
- This image does not display correctly

Which of these categories does the above image fit into best?

- House
- Flower
- People
- Car
- This image does not display correctly

Abbildung 3.12: Erster Teil des User-Interfaces von „Image-Categorization“.

Which of these categories does the above image fit into best?

- House
- Flower
- People
- Car
- This image does not display correctly

Which of these categories does the above image fit into best?

- House
- Flower
- People
- Car
- This image does not display correctly

Abbildung 3.13: Zweiter Teil des User-Interfaces von „Image-Categorization“.

Which of these categories does the above image fit into best?

- House
- Flower
- People
- Car
- This image does not display correctly

Please answer the following question:

This teacher is influenced by Franz Brentano influences Martin Buber.

- Aurel Kolnai
- Carl Stumpf
- Rudolf Steiner
- Christian von Ehrenfels
- I can't understand the question.

Feel free to use Google to find the answer!

Any Comments

Abbildung 3.14: Dritter Teil des User-Interfaces von „Image-Categorization“.

Data Enrichment

Instructions

Task

We present you the name of a film. Use the **given link** to a search query on *Google* and search for the **year the film was created**. Then write the year in the provided text box.

Additional, we present you an multiple choice question.

Process

1. **Click on the link** of the film
2. Find the **year the film was created** and write it in the text box
3. If you **can't find** the year, write "9999" in the text box
4. Answer the **multiple choice question**. If you can not understand the question, please choose "I can't understand the question." and another question appears.

Hints and Examples

Hints

- The image of the adorable ducklings is just to catch your attention.
- *Don't hesitate to use Google or another search engine to solve the multiple choice question*

Example

1. Film name: The Avengers
2. Search Query: "The Avengers" year
3. Result: 2012

Summary

We present you the name of a film. Please search on the internet for the year the film was created.

Thank You!

Your careful attention on this task is greatly appreciated!

Abbildung 3.15: Arbeitsanweisungen von „Data-Enrichment“.

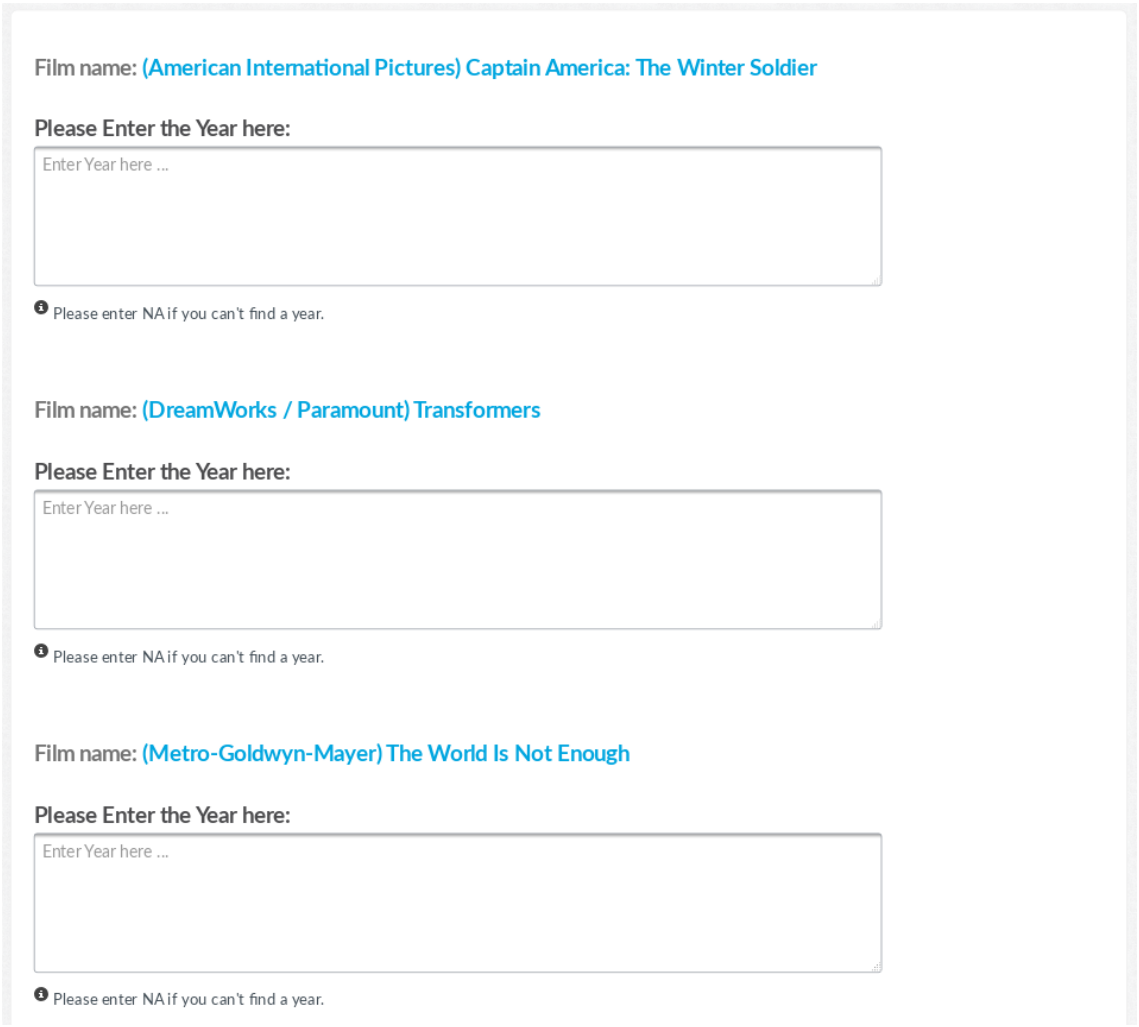


Abbildung 3.16: Erster Teil des User-Interfaces von „Data-Enrichment“.

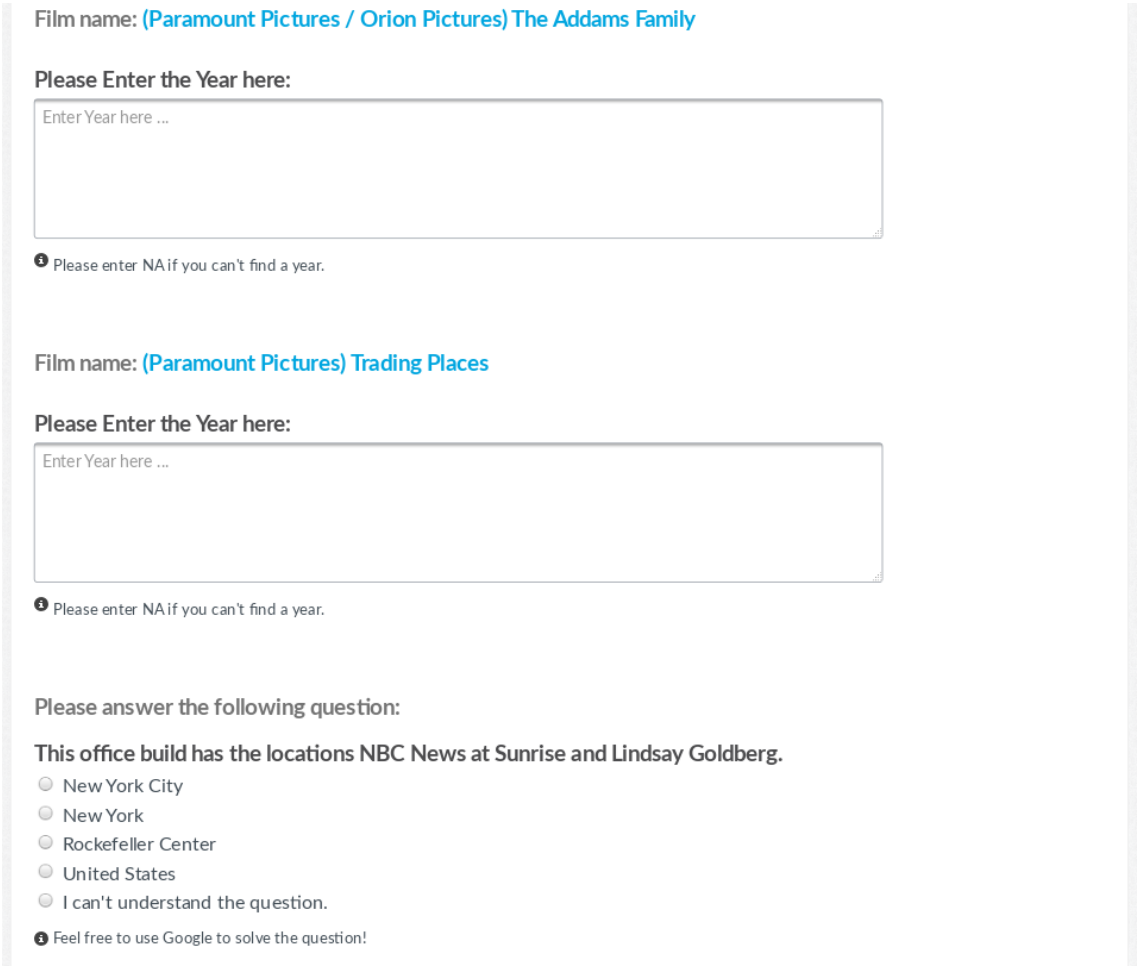


Abbildung 3.17: Zweiter Teil des User-Interfaces von „Data-Enrichment“.

4 Technische Umsetzung

Im folgenden Kapitel wird beschrieben, was beim Erstellen der Ausgangsdatensätze zu beachten war. Anschließend wird präsentiert, welche Werkzeuge zur Evaluation der Ergebnisse eingesetzt wurden und wie die Generierung der Quizfragen abläuft. Zuletzt wird ausgeführt, wie die durchgeführten Versuche technisch umgesetzt wurden.

4.1 Implementierung auf Crowdflower

Wie schon erwähnt, wird auf Crowdflower HTML genutzt, um die generelle Gestaltung der Arbeitsanweisungen und der Eingabemaske zu steuern. Dabei können die HTML-Anweisungen auch mit CSS kombiniert werden, was eine professionelle optische Aufbereitung ermöglicht. Die Gestaltung der Eingabeelemente in der Eingabemaske erfolgt mit der *CML*, eine speziell auf den Einsatzzweck zugeschnittene Auszeichnungssprache auf der Basis der Extensible Markup Language (XML). Um die Elemente der CML eindeutig zu kennzeichnen, befindet sich am Anfang eines jeden Elements das Namensraumpräfix `cml`, das mit einem Doppelpunkt vom eigentlichen Element getrennt ist. Einige Sprachelemente, die in den Versuchen dieser Arbeit genutzt wurden, sollen nun erwähnt werden:

<cml:textarea /> Mit diesem Element kann ein Texteingabefeld erzeugt werden, welches der Worker für seine Eingaben nutzen kann. Dabei sorgen einige Attribute für bequeme Anpassungen: Mit `label` kann das Textfeld mit einer Beschriftung versehen werden, `instructions` lässt zusätzliche Anweisungen unterhalb des Textfeldes anzeigen und `validates` ermöglicht den Einsatz von Validatoren. Diese werden nachfolgend besprochen.

<cml:radios /> Dieses Element erzeugt eine Auswahlliste mit Optionen (Radio-Buttons). Es kann immer nur eine Option zu einer Zeit ausgewählt sein. Auch hier lassen sich die eben beschriebenen Attribute nutzen.

<cml:html class="clicked"/> Hiermit werden alle Kindelemente in die Klasse `clicked` aufgenommen. Dies ermöglicht in Kombination mit Validatoren sicherzustellen, dass ein Worker ein bestimmtes Element angeklickt haben muss.

Validatoren Diese sind ein nützliches Hilfsmittel, um Nutzereingaben auf ein bestimmtes Format hin zu überprüfen und eine gewisse Ablauf-Logik in Gang zu setzen. Alle Validatoren, die genutzt werden sollen, werden in das Attribut `validates` eingetragen. Der Wert `required` sorgt dafür, dass eine Nutzereingabe bei einem bestimmten Element unbedingt notwendig ist. Mittels `alpha` oder `alphanum` lassen sich Texteingaben überprüfen, ob sie ausschließlich aus Buchstaben oder Buchstaben und Zahlen zusammengesetzt sind. In diesem Rahmen können auch beliebige reguläre Ausdrücke überprüft werden.

only-if Durch dieses Attribut, was jedem Element hinzugefügt werden kann, lässt sich in begrenztem Umfang eine Ablauf-Logik erstellen. Ein mit diesem Attribut ausgezeichnetes Element erscheint nur, wenn die Bedingungen in `only-if` erfüllt sind. Hinterlegt man darin den Namen eines anderen Elementes, so erscheint das neue

Element ausschließlich, wenn das genannte Element erfolgreich validiert wurde. Man kann aber auch eine Eingabe eines anderen Elementes abfragen (z.B. `question_one_answered:[did_not_understand]`, was in den Versuchen dieser Arbeit genutzt wurde, um abzufragen, ob ein Worker die Eingabe *I can't understand the question* gewählt hat).

4.2 Design der Datensätze

Um auf Crowdflower Daten von Workern bearbeiten zu lassen, benötigt man Ausgangsdatensätze. Diese werden in Form von CSV-Dateien in den jeweiligen Job geladen. Dabei werden die Daten gewünschter Spalten über ihren Spaltennamen aufgerufen, der in der CSV-Datei in der ersten Zeile angegeben sein muss. Grundsätzlich gilt eine Zeile in der CSV-Datei als ein Datensatz. Wird ein Job ausgeführt, so wird jedem Worker eine zufällige Auswahl an Zeilen des Datensatzes zur Bearbeitung präsentiert. Diese Zusammenstellung von Aufgaben nennt sich *Page*. Da in den Versuchen dieser Arbeit immer fünf Aufgaben zusammen mit einer Quizfrage abgearbeitet werden sollten, war es nötig diese fünf Aufgaben mit der Quizfrage zu einer Einheit zusammenzufassen. Somit sind in den Datensätzen der Versuche dieser Arbeit immer fünf Zeilen in einer Zeile vereint.

```
id,url0,trans original0,url1,trans original1,url2,trans original2,
   url3,trans original3,url4,trans original4,type,question_one,
   answer_one0,answer_one1,answer_one2,answer_one3,rightanswer_one,
   question_two,answer_two0,answer_two1,answer_two2,answer_two3,
   rightanswer_two
0,http://134.96.217.62:8080/~yhary/data/tts/information.wav,
   information,http://134.96.217.62:8080/~yhary/data/tts/available.
   wav,available,http://134.96.217.62:8080/~yhary/data/tts/
   copyright.wav,copyright,http://134.96.217.62:8080/~yhary/data/
   tts/university.wav,university,http://134.96.217.62:8080/~yhary/
   data/tts/management.wav,management,Q,This office build has the
   locations NBC News at Sunrise and Lindsay Goldberg.,New York
   City,New York,Rockefeller Center,United States,Rockefeller
   Center,This introduction was created by the writer Greg Rucka
   and Vin Sullivan.,Magazine Enterprises,Stumptown (comics),
   Columbia Comics,Detective Comics,Detective Comics
```

Listing 4.1: Kopfzeile und erster Datensatz aus den Daten für *Audio-Transcription*.

In Listing 4.1 sieht man zunächst die Kopfzeile, also die Namen der Spalten, über welche die Daten in einem Job angesprochen werden, dann den ersten Datensatz als Beispiel. Dieser enthält jeweils fünf mal eine URL zu einer Audio-Datei, sowie die richtige Transkription der Audio-Datei. Danach folgen zwei Multiple-Choice-Fragen, wobei für jede der Fragentext, vier Antwortmöglichkeiten und die richtige Antwort angegeben sind.

4.3 Werkzeuge zur Evaluation

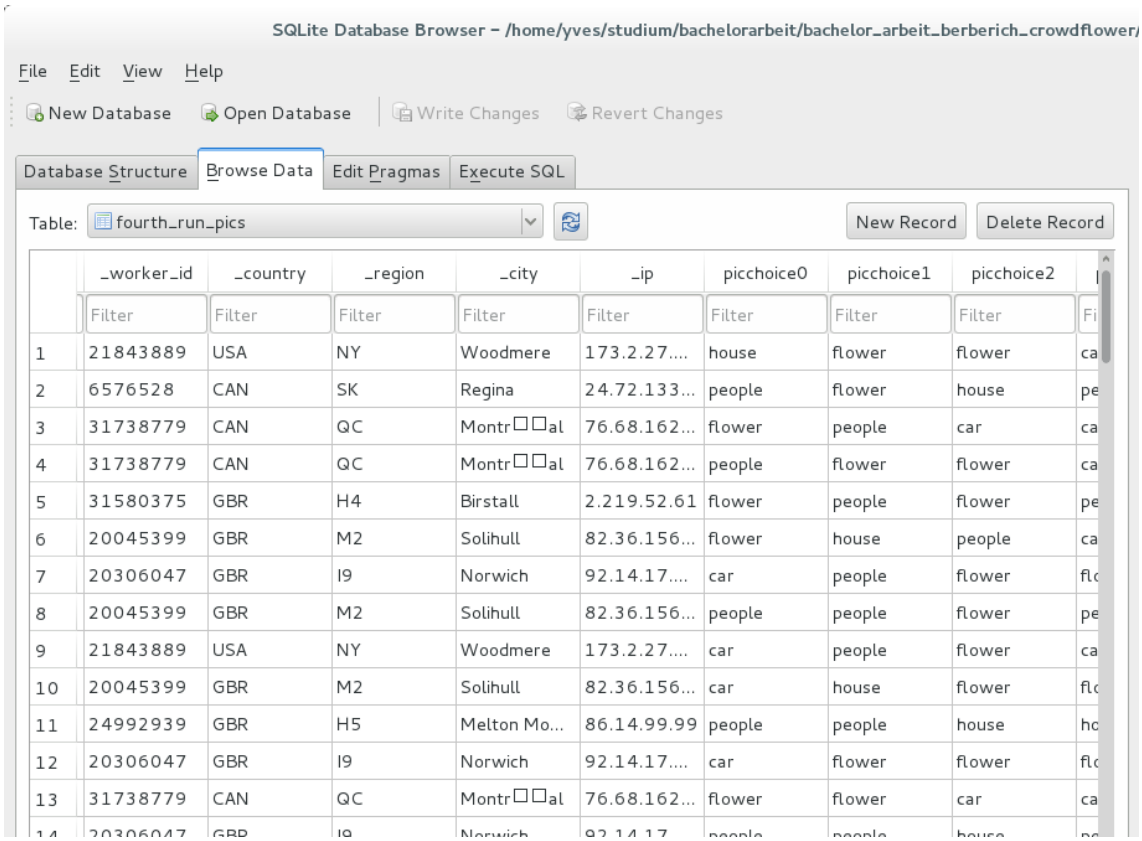
Um verschiedene Kennwerte zu ermitteln und Auswertungen der Ergebnisdaten vorzunehmen, kamen drei verschiedene Software-Werkzeuge zum Einsatz:

SQLite Dies ist ein leichtgewichtiges Datenbanksystem, welches ohne Server auskommt, und die Daten sowie die Datenbankinformationen in einer Datei ablegt. Dennoch werden alle Funktionen, die mit SQL geboten werden, von SQLite unterstützt. Dadurch eignet sich das Datenbanksystem besonders zum schnellen Testen oder für

Datenbanken, auf die wenig zugegriffen wird. Um die Arbeit mit SQLite zu erleichtern, wurde das Programm *SQLite-Browser* genutzt (siehe Abbildung 4.1), das eine bequeme und übersichtliche Arbeit mit den Daten ermöglicht. SQLite wurde hauptsächlich für das Speichern der Ergebnisse von Crowdflower genutzt sowie für einfache Analysen: Wie viele Worker haben den Task bearbeitet? Welche Worker haben die Multiple-Choice-Frage richtig oder falsch beantwortet? Was ist das Mittel der Richtigkeit aller bearbeiteten Aufgaben? Für weitergehende Untersuchungen wurde SQLite mit Python verbunden. Dadurch war es möglich kompliziertere Aggregationen und Auswertungen anzufertigen, z.B. den Einfluss von Golddaten auf die Ergebnisse zu simulieren.

Python In Verbindung mit SQLite wurde die Programmiersprache Python genutzt, um Auswertungen zu ermöglichen, die mit SQL nicht machbar sind. So wurde zum Beispiel für jeden Datensatz, der wie schon beschrieben jeweils 5 Fragen enthält, die gesamte Richtigkeit der gestellten Aufgaben berechnet und in die Datenbank als ein einziger Wert eingefügt. Auch die Frage, wie sich der Einsatz von Golddaten auf die Ergebnisse auswirken würde, konnte nur mit einem selbst programmierten Python-Skript beantwortet werden.

R Dies ist eine Statistikumgebung, mit der sich viele relevante statistische Probleme automatisch berechnen lassen. *R* wird unter der GNU-Lizenz angeboten und dient damit als Alternative zu kommerziellen Statistikumgebungen wie SPSS. In dieser Arbeit diente *R* dazu, die Ergebnisse auf Signifikanz hin zu untersuchen.



	_worker_id	_country	_region	_city	_ip	picchoice0	picchoice1	picchoice2	
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Fi
1	21843889	USA	NY	Woodmere	173.2.27....	house	flower	flower	ca
2	6576528	CAN	SK	Regina	24.72.133...	people	flower	house	pe
3	31738779	CAN	QC	Montr□□al	76.68.162...	flower	people	car	ca
4	31738779	CAN	QC	Montr□□al	76.68.162...	people	flower	flower	ca
5	31580375	GBR	H4	Birstall	2.219.52.61	flower	people	flower	pe
6	20045399	GBR	M2	Solihull	82.36.156...	flower	house	people	ca
7	20306047	GBR	I9	Norwich	92.14.17....	car	people	flower	fl
8	20045399	GBR	M2	Solihull	82.36.156...	people	people	flower	pe
9	21843889	USA	NY	Woodmere	173.2.27....	car	people	flower	ca
10	20045399	GBR	M2	Solihull	82.36.156...	car	house	flower	fl
11	24992939	GBR	H5	Melton Mo...	86.14.99.99	people	people	house	ho
12	20306047	GBR	I9	Norwich	92.14.17....	car	flower	flower	fl
13	31738779	CAN	QC	Montr□□al	76.68.162...	flower	flower	car	ca
14	20306047	GBR	I9	Norwich	92.14.17	people	people	house	pe

Abbildung 4.1: Ausschnitt aus SQLite Browser unter Linux.

4.4 Generierung der Quizfragen

Nachfolgend wird die Generierung der genutzten Quizfragen beschrieben und erläutert, wie diese in die Datensätze der Versuche eingeflossen sind.

4.4.1 Das System Q2G

Die Quizfragen, die als Qualitätssicherungsmechanismus in den Versuchen dieser Arbeit eingesetzt wurden, stammen aus dem System *Q2G*[1], das von Dominic Seyler, Mohamed Yahya und Klaus Berberich am Max-Planck-Institut für Informatik in Saarbrücken entwickelt wurde. Damit ist es möglich Quizfragen in menschenverständlicher Sprache mit mehreren Antwortmöglichkeiten zu erzeugen. Auf der Webseite gibt man eine gewünschte Entität in ein Textfeld ein und wählt wie viele Antwortmöglichkeiten die Quizfrage haben soll. Daraufhin wird die passende Frage generiert. Dies geschieht in drei Schritten: Zunächst wird die gewünschte Entität als richtige Antwort auf die Frage bestimmt und deren Oberkategorie ermittelt. Daraufhin wird eine Anfrage in der *SPARQL Protocol And RDF Query Language (SPARQL)* generiert, um aus einer Wissensdatenbank, welche in Wissensgraphen organisiert ist, einen bestimmten Teilgraphen auszuwählen. Dieser repräsentiert die Quizfrage in maschinenverständlicher Form. Anschließend werden die gewonnenen Informationen auf einfache Art in eine menschenverständliche Frage umgesetzt. Ein Beispiel aus *Generating Quiz Questions from Knowledge Graphs* [19] verdeutlicht den Prozess:

- Entität als Antwort wählen und Oberkategorie finden: `Elvis_Presley type American _rock_singers.`
- Verbindungen, die durch die SPARQL-Anfrage gefunden wurden: `Elvis_Presley diedIn Memphis,_Tennessee, Elvis_Presley created Jailhouse_Rock, Elvis_Presley created Heartbreak_Hotel`
- Verbalisierte Form der Frage: *This american rock singer died in Memphis, Tennessee and created Jailhouse Rock and Heartbreak Hotel.*

Zusätzlich zur Verbalisierung der Frage liefert das System einen Hinweis darauf, wie schwierig die generierte Frage ist.

4.4.2 Quizfragen nutzen

Um die Quizfragen nun in den Datensätzen für Crowdflower zu nutzen, wurde eine speziell dafür angefertigte Schnittstelle genutzt, die über einen URL-Request in einem Python-Skript angesprochen wurde. Ein direkter Zugriff auf das System war allerdings nicht möglich, da die benötigten Wissensdatenbanken und die Software ausschließlich auf den Servern des Max-Planck-Institutes verfügbar waren. Daher konnte auf den Parameter der Schwierigkeit einer Frage keinen Einfluss genommen werden und es war nötig solange Fragen erzeugen zu lassen, bis Fragen solcher Art generiert wurden, die den Anforderungen genügten. In den Versuchen dieser Arbeit war es wichtig, dass die Fragen den Schwierigkeitsgrad *Easy* und 3 weitere falsche Antwortmöglichkeiten besaßen. Um dieses Ziel zur erreichen, wurden aus einer weiteren Datenbank Entitäten extrahiert, die eine besonders hohe Popularität aufweisen. Diese wurden als Entitäten für die Fragengenerierung genutzt. Weiterhin wurden die Antwortmöglichkeiten zufällig gemischt und

[1] https://gate.d5.mpi-inf.mpg.de/q2g/q2G.htm

mit dem Fragentext in einer CSV-Datei abgelegt. Diese Datei diente als Fragenpool für die auf Crowdflower erstellten Jobs.

Listing 4.2 zeigt den Teil des Python-Skriptes, das die Anfrage an die Web-Schnittstelle beordert und entscheidet, ob die Frage geeignet ist. Ein `ValueError` tritt auf, wenn die JSON-Response des Q2G-Systems sich nicht parsen ließ – dies deutet daraufhin, dass für diese Entität keine Frage generiert werden kann. Sobald die `QuestionValidationException` gefangen wird, genügte die Frage einer der Kriterien nicht und wird an die Liste der Anfragen, die ausgeführt werden, wieder angehängt. Das gleiche passiert, wenn der Server nicht innerhalb einer vorgegebenen Zeitspanne eine Antwort liefert.

```
def fetchQuestionsFromEntityList(self, entityList, distractors):
        self._log('INFO','Requesting parsed Questions from
            EntityList ...')
        questionCount = 0
        parsedQuestions = list()
        for e in entityList:
            try:
                rawQuestion = self.fetchRawQuestion(e.value,
                    distractors)
                parsedQuestion = self.requestparser.
                    parseRequestToQuestion(rawQuestion)
                validateQuestion(parsedQuestion)
                parsedQuestions.append(parsedQuestion)
                questionCount += 1
            except ValueError:
                self._log('WARN', 'No question for entity {}
                    available'.format(e))

            except QuestionValidationException as ex:
                self._log('WARN', 'Drop Question! {}'.format(ex))
                self._log('INFO', 'I will try again later... ({})'.
                    format(e._trys))
                self.tryAgain(e, entityList)

            except requests.Timeout:
                    self._log('WARN', 'Request Timeout!')
                    self.tryAgain(e, entityList)

        self._log('INFO','Retrieved {} questions at all.'.format(
            str(questionCount)))
        return parsedQuestions
```

Listing 4.2: Python-Skript, das Fragen generieren lässt.

4.5 Technische Umsetzung der Versuche

In diesem Abschnitt wird erläutert, wie die durchgeführten Versuche vorbereitet und welche Software-Werkzeuge dafür genutzt wurden.

4.5.1 Allgemein

In allen Versuchen wurde Libre-Calc[2] genutzt, um die einzelnen Informationen zu einem Datensatz, der auf Crowdflower nutzbar ist, zusammenzusetzen. Dabei flossen die einzelnen Informationen für jeden Versuch und die vorher generierten Quizfragen ein. Es wurden pro Datensatz immer zwei Quizfragen eingefügt, damit eine alternative Frage möglich ist, falls die erste Frage von den Workern nicht verstanden werden kann.

4.5.2 Versuch 1: Audio-Transcription

Kurzbeschreibung des Versuchs: In diesem Versuch wurden die Worker dazu aufgefordert, einem Link zu einer WAV[3]-Datei zu folgen, sich diese anzuhören und den darin enthaltenen Text niederzuschreiben.

Die WAV-Dateien enthielten jeweils ein Wort in englischer Sprache, das von dem Text-To-Speech-System *espeak*[4] erzeugt wurde. Die passenden Wörter stammen aus einer englischen Wortdatenbank. Da bei einsilbigen Wörtern eine hohe Verwechslungsgefahr besteht[5], wurden nur Wörter einer Länge von 10 Buchstaben und mehr ausgewählt. Diese Wörter wurden zeilenweise aus der Wortdatenbank ausgelesen und mit *espeak* in WAV-Dateien mit gesprochenem Text umgewandelt. Anschließend wurden diese Dateien auf einen Webserver des Softwarelabors geladen, sodass sie aus dem Internet abrufbar waren. Zuletzt wurden die URL der Audio-Dateien sowie die dazugehörige erwartete Transkription in einer CSV-Datei abgelegt, um sie mit Crowdflower nutzen zu können. Listing 4.3 zeigt einen Ausschnitt des Python-Skriptes, das die Wörter aus der Wortdatenbank liest und mittels *espeak* umwandelt. Die Zuordnung, welche Audio- Datei welchem Wort entspricht, wird in einer CSV-Datei gespeichert.

```
def makeAudio(line, filename):
    print("Make word:", line, "to", filename)
    TTS_WRAPPER.runInFile(filename, line)

def readLinesAndMakeAudio():
    dicfile = open(DICTIONARY_FILE,'r')
    wordfile = open(WORD_FILE,'w')

    filenr = 0
    for line in dicfile:
        line.strip()
        if len(line) >= MINIMUM_WORD_LENGTH:
            line = cleanWhitespaces(line)
            name = next(iter(getRandomName()))
            makeAudio(line, TARGET_FOLDER + line + FILE_ENDING)
            wordfile.write(line + ',' + name + '\n')
            filenr += 1

        if filenr >= MAX_WORD_COUNT:
            break

    dicfile.close()
    wordfile.close()
    return filenr
```

[2]Ein Produkt aus der *Libre Office*-Suite. https://de.libreoffice.org/ (zuletzt abgerufen am 14. September 2016).
[3]Dies ist ein Datenformat zur digitalen Speicherung von Audiodaten.
[4]http://espeak.sourceforge.net/ (zuletzt abgerufen am 14. September 2016).
[5]Z.B. *meat* und *eat* oder *see* und *sea*.

Listing 4.3: Erzeugung von Audio-Dateien mittels eines Englischwörterbuchs.

4.5.3 Versuch 2: Image-Categorization

Kurzbeschreibung des Versuchs: Hier sollten die Worker ein ihnen präsentiertes Bild einer von vier Kategorien – *Car, Flower, House* oder *People* – zuordnen.

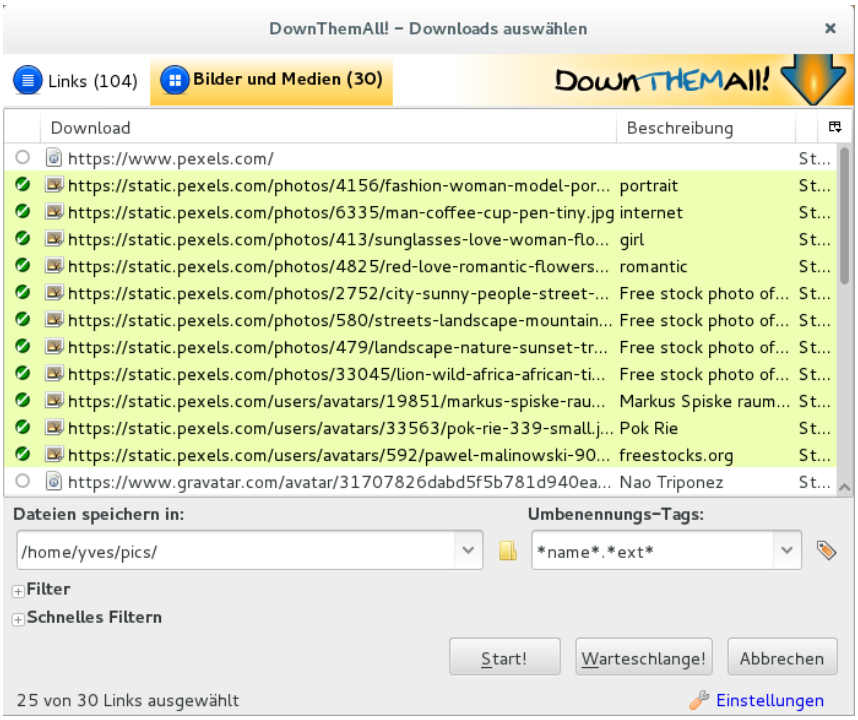


Abbildung 4.2: DownThemAll! versucht Bilder einer Webseite zu laden.

Um an Bilder zu gelangen, die schon einer Kategorie zugeordnet waren, wurde die Stockphoto-Seite *www.pexels.com* genutzt. Die Bilder dieser Seite sind alle unter der *Creative Commons Zero (CC0) license* verfügbar. Dies bedeutet, dass die Bilder für jeden legalen Zweck ohne vorherige Zustimmung des Urhebers verwendet werden dürfen. Es muss ebenso kein Hinweis auf die Urheberschaft angebracht werden, sodass rechtliche Hürden für den Versuch ausgeschlossen waren.

Mit dem Tool *DownThemAll!* wurden die Bilder der jeweils ausgewählten Kategorie von der Stockphoto-Seite geladen und anschließend den vier Kategorien zugeordnet auf den Webserver des Softwarelabors geschoben. In einer CSV-Datei wurden die URL der Bilder sowie die erwarteten Kategorien hinterlegt.

4.5.4 Versuch 3: Data-Enrichment

Kurzbeschreibung des Versuchs: Die Worker sollten über eine vorgefertigte Suchanfrage auf *www.google.com* zu einem Filmtitel das passende Erscheinungsjahr finden und dieses in einer Textbox hinterlegen.

Für diesen Versuch wurden Daten ausgewählt, die Crowdflower als „Data For Everyone“ frei zu Verfügung stellt. Die Datensätze enthielten den Filmtitel, die Produktionsfirma des Films und dessen Erscheinungsjahr. Die Suchanfrage wurde direkt in Crowdflower

generiert, indem in einem Link an die URL `https://www.google.com/search?q=` der Filmtitel und `+year` angehängt wurde.

5 Evaluation

Dieses Kapitel beschreibt und erklärt die Vorgehensweise bei der Untersuchung der Ergebnisse der in Kapitel 3 vorgestellten Tasks. Es werden alle berücksichtigten Werte vorgestellt und auf Besonderheiten in der Verarbeitung der Daten eingegangen. Danach wird erläutert, nach welchem statistischen Test die Signifikanzüberprüfung der Ergebnisse stattfand. Zuletzt werden die Ergebnisse der einzelnen Versuche präsentiert und diskutiert.

5.1 Untersuchte Werte

Nach der Durchführung der Versuche wurden verschiedene Werte untersucht. Zentraler Wert ist der *Richtigkeitsquotient*: Um festzustellen, wie gut ein Worker eine Serie von Aufgaben bearbeitet hat, wurden die jeweils fünf bearbeiteten Aufgaben mit den fünf vorhandenen Lösungen verglichen und die Anzahl der bearbeiteten Aufgaben, die den vorhandenen Lösungen entsprachen, durch fünf geteilt. Dies ergibt eine Maßzahl, die für die Qualität der Bearbeitung einer Aufgabe durch einen Worker steht.

$$RQ_{proAufgabenserie} = \frac{\sum_{i=1}^{5} x_i}{5} \text{ mit } \begin{cases} x_i = 0, & \text{falls } \textit{Teilaufgabe}_i \text{ falsch} \\ x_i = 1, & \text{falls } \textit{Teilaufgabe}_i \text{ richtig} \end{cases} \tag{5.1}$$

Bei dem Versuch „Audio-Transcription“ wurden beim Vergleich von Antwort und vorhandener Lösung keine Rücksicht auf führende oder abschließende Leerzeichen genommen; ebenso wurden Abweichungen in der Groß- bzw. Kleinschreibung ignoriert. Hat ein Worker z.B. eine Audio-Datei mit `automatioN` transkribiert, wurde dies als gleich mit der vorhandenen Lösung `Automation` gewertet. Bei dem Versuch „Image-Categorization“ waren keine Freitextantworten der Worker gefordert und bei „Data-Enrichment“ wurde das Freitextfeld auf numerische Eingaben begrenzt. Weiterhin wurden folgende Werte untersucht:

- *Anzahl der Worker (AW)*: Die Anzahl der Worker, die beim jeweiligen Versuch teilgenommen haben.
- *Mittel des Richtigkeitsquotienten (MRQ)*: Das arithmetische Mittel[1] der Richtigkeitsquotienten aller bearbeiteten Aufgaben.
- *Maximum des Richtigkeitsquotienten (MaxRQ)*: Der maximale Wert aller Richtigkeitsquotienten. Dieser Wert setzt das arithmetische Mittel in einen Kontext.
- *Minimum des Richtigkeitsquotienten (MinRQ)*: Der minimale Wert aller Richtigkeitsquotienten. Dieser Wert setzt das arithmetische Mittel in einen Kontext.
- *Mittel des Richtigkeitsquotienten bei richtiger Multiple-Choice-Frage (MRQr)*: Das Mittel des Richtigkeitsquotienten aller Aufgaben derjenigen Worker, die die Multiple-Choice-Frage richtig beantwortet haben.

[1] Nachfolgend ist mit „Mittel“ immer das arithmetische Mittel gemeint.

- *Mittel des Richtigkeitsquotienten bei falscher Multiple-Choice-Frage (MRQf)*: Das Mittel des Richtigkeitsquotienten aller Aufgaben derjenigen Worker, welche die Multiple-Choice-Frage falsch beantwortet haben.
- *Mittel der richtig beantworteten Multiple-Choice-Fragen bei maximalem Richtigkeitsquotienten (MMC)*: Die Anzahl aller Aufgaben, bei welcher die Multiple-Choice-Frage richtig beantwortet wurde *und* der maximale Richtigkeitsquotienten erreicht wurde, geteilt durch die Anzahl aller Aufgaben mit maximalem Richtigkeitsquotienten. Dieser Wert gibt Aufschluss darüber, wie aussagekräftig die richtige Beantwortung der Multiple-Choice-Frage für die Aufgaben mit hoher Qualität ist.
- *Anzahl eliminierter Worker durch Multiple-Choice-Fragen (AeWMC)*: Die Anzahl derjenigen Worker, die durch falsche Beantwortung der Multiple-Choice-Fragen aus dem Ergebnispool ausgeschlossen würden. Dabei werden diejenigen Worker als ausgeschlossen betrachtet, welche mindestens drei Fragen falsch beantworten.
- *Mittel des Richtigkeitsquotienten bei Ausschluss der Worker (MRQa)*: Das Mittel der richtig bearbeiteten Aufgaben unter der Bedingung, dass Worker, welche mindestens drei Multiple-Choice-Fragen falsch beantwortet haben, aus dem Ergebnispool ausgeschlossen werden.
- *Maximale Anzahl der bearbeiteten Aufgaben pro Worker (MaxAAW)*
- *Minimale Anzahl der bearbeiteten Aufgaben pro Worker (MinAAW)*
- *Mittel der Anzahl der bearbeiteten Aufgaben pro Worker (MAAW)*
- *Mittel des Richtigkeitsquotienten von Workern mit minimaler Anzahl bearbeiteter Aufgaben (MRQWmin)*: Dieser Wert gibt – mit dem nachfolgenden Pendant – Aufschluss, ob die Qualität der Bearbeitung einer Aufgabe von der Anzahl der bearbeiteten Aufgaben abhängen könnte.
- *Mittel des Richtigkeitsquotienten von Workern mit mehr als 75 Prozent der maximalen Anzahl bearbeiteter Aufgaben (MRQWmax)*
- *Richtigkeitsquotient multipliziert mit Golddatenwahrscheinlichkeit (RQsim)*: Der Richtigkeitsquotient multipliziert mit der in Abschnitt 5.2 errechneten Wahrscheinlichkeit. Dies gewichtet den Richtigkeitsquotienten, sodass der Ausschluss der Worker anhand von Golddaten mit dem Ausschluss der Worker anhand der Quizfragen verglichen werden kann.
- *Mittel des Richtigkeitsquotienten mit simulierten Golddaten (MRQsim)*: Dieser Wert dient zum Vergleich mit MRQr und hilft bei der Entscheidung, ob der Einsatz von Multiple-Choice-Fragen zu besseren Arbeitsergebnissen führt, als der bisherige Einsatz von Golddaten.

Um zu entscheiden, ob der Einsatz von Multiple-Choice-Fragen zu einer Verbesserung der Arbeitsqualität führt, werden insbesondere folgende Werte miteinander verglichen:

- *MRQ mit MRQr*: Schneidet die Gruppe der Worker, die die Multiple-Choice-Fragen richtig beantwortet haben, besser ab als alle Worker gemeinsam?
- *MRQ mit MRQf*: Schneidet die Gruppe der Worker, welche die Multiple-Choice-Fragen falsch beantwortet haben, schlechter ab als alle Worker gemeinsam?

- *MRQr mit MRQsim*: Schneidet die Gruppe der Worker, die die Multiple-Choice-Frage richtig beantwortet haben, besser ab, als die Worker unter dem Einsatz von Golddaten?

Diese Messwerte wurden jeweils paarweise auf Signifikanz untersucht – d.h. wie wahrscheinlich es ist, dass die Unterschiede in den Messwerten durch Zufall entstanden sind.

5.2 Auswirkungen von Golddaten

Um vergleichen zu können, ob der Einsatz von Golddaten zu besserer Arbeitsqualität führt, als der Einsatz der Multiple-Choice-Fragen, wurde nach der Durchführung der Versuche der Einsatz von Golddaten simuliert. Dabei wurde angenommen, dass ein Worker aus dem Ergebnispool ausgeschlossen wird, wenn er drei Aufgaben, die als Golddaten markiert sind, falsch beantwortet hat. Ebenso wurde angenommen, dass Golddaten genau 20 Prozent aller Aufgaben ausmachen – bei jeder Serie von fünf Aufgaben ist *eine* Aufgabe eine Aufgabe mit Golddaten. Mit diesen Voraussetzungen wurde ein Wert errechnet, welcher die Wahrscheinlichkeit angibt, mit der ein Worker im Ergebnispool bleibt. Diese Wahrscheinlichkeit wurde mit den entsprechenden Richtigkeitsquotienten einer jeden Aufgabe multipliziert, sodass die tatsächlichen Ergebnisse gewichtet betrachtet werden und mit den Ergebnissen der Gruppe der Worker, welche die Multiple-Choice-Frage richtig beantwortet hatten, verglichen werden konnte. Auf die genaue Berechnung soll nachfolgend eingegangen werden.

5.2.1 Berechnung der Gewichtung

Um zu berechnen, wie wahrscheinlich es ist, dass ein Worker im Ergebnispool verbleiben darf, wurde zunächst die Gegenwahrscheinlichkeit des Ereignisses berechnet – wie wahrscheinlich ist es, dass ein Worker aus dem Ergebnispool ausgeschlossen wird. Voraussetzung für einen Ausschluss ist, dass drei oder mehr Aufgaben, die als Aufgaben mit Golddaten markiert sind, falsch beantwortet wurden. Da die Golddaten aber von der Crowdsourcing-Plattform in der Regel *zufällig* ausgewählt werden, mussten alle Möglichkeiten der Verteilung von Golddaten berücksichtigt werden. Zunächst wurden dazu alle Ergebnisse eines einzelnen Workers extrahiert (Tabelle 5.1 zeigt einen beispielhaften, aber auf drei Aufgaben pro Serie begrenzten Ausschnitt aus den Ergebnissen eines Workers). Diese Ergebnisse können auch als Matrix betrachtet werden, indem jedes Ergebnis eines Workers, das dem Original entspricht als 1, und jedes Ergebnis, das vom Original abweicht als 0 gewertet wird. Anschließend soll aus jeder Zeile der Matrix jeweils ein Element ausgewählt werden (bei fünf Aufgaben einer Serie entspricht das den gewünschten 20 Prozent Golddaten). Es werden nun alle Möglichkeiten, ein Element aus jeweils einer Zeile auszuwählen, berechnet und nacheinander abgearbeitet. Dabei wird gezählt, wie viele Nullen bei einer Kombination vorkommen. Sind es mehr als drei, so wäre der Worker bei dieser Verteilung von Golddaten aus dem Ergebnispool ausgeschlossen. Diese Fälle werden gezählt und am Schluss durch die Anzahl aller Möglichkeiten der Verteilung von Golddaten geteilt.

Listing 5.1 zeigt die Umsetzung des eben beschriebenen Algorithmus in Python. In `_computeFailureOneCombination(...)` werden die Abweichungen der Arbeitsergebnisse mit den vorhandenen Lösungen gezählt. Dabei wird dies immer für eine bestimmte „Kombination“, also eine Verteilung der Golddaten, vorgenommen. Mit der Funktion `_computeFailuresAllCombinationsWithThreshhold(...)` werden alle Kombinationen abgearbeitet. Überschreiten die gewerteten Fehlergebnisse den in `self.failure_threshold` hin-

Transcription 1	**Original 1**	**Transcription 2**	**Original 2**	**Transcription 3**	**Original 3**
Information	information	Available	available	Copyright	copyright
Department	department	Description	description	Insurance	insurance
Understand	understand	Publication	publications	Worldwide	worldwide
Beautiful	beautiful	Location	locations	Significance	significant

Tabelle 5.1: Ausschnitt der Ergebnisse eines Workers in „Audio-Transcription“.

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \tag{5.2}$$

Passende Matrix zu Tabelle 5.1

terlegten Wert – in unserem Fall *drei* – so wäre der Worker, dessen Ergebnisse untersucht werden, in diesem Fall ausgeschlossen. Dies wird durch eine Erhöhung der Variable `failuresAll` vermerkt. All diese Funktionen werden angestoßen durch den Aufruf von `computeProbability(...)`; in dieser Funktion wird die Wahrscheinlichkeit für einen Verbleib des Workers berechnet und zurückgegeben.

```
class golddataSimulator(object):

    permutation_failure = 0
    failure_over_theshold_value= 0
    def __init__(self, valueMatrix, width, failure_threshold=3):
        self.valueMatrix = valueMatrix
        self.half = int(width / 2)
        self.elementCreator = golddataElementCreator(len(
            valueMatrix), self.half)
        self.failure_threshold = failure_threshold

    def computeProbability(self):
        returnTuple = self.
            _computeFailuresAllCombinationsWithThreshhold()
        failuresAll = returnTuple[0]
        number_combinations_all = returnTuple[1]
        probability = (number_combinations_all - failuresAll) /
            number_combinations_all

        return probability

    def _computeFailuresAllCombinationsWithThreshhold(self):
        failuresAll = 0
        combinationCount = 0
        for c in self.elementCreator.getIter():
            failures =  self._computeFailureOneCombination(c)
            if failures >= self.failure_threshold:
                failuresAll += 1
            combinationCount += 1

        return (failuresAll, combinationCount)
```

```
def _computeFailureOneCombination(self, combination):
    failures = 0
    combinationIndex = 0
    for i in self.valueMatrix:
        value_to_check0 = i[combination[combinationIndex]]
        value_to_check1 = i[combination[combinationIndex] +
            self.half]
        if compareStrings(value_to_check0,value_to_check1) is
            False:
             failures += 1
        combinationIndex += 1
    return failures
```

Listing 5.1: Simulation von Golddaten über einer Ergebnis-Tabelle.

5.3 Signifikanztests

Die einzelnen Messreihen wurden daraufhin untersucht, ob die Unterschiede *signifikant* sind. Dazu wurden die Messreihen zunächst auf Normalverteilung untersucht, um einen geeigneten Signifikanztest auswählen zu können. Die Normalverteilung wurde mit dem *Shapiro-Wilk-Test* überprüft. Dieser Test überprüft die Hypothese, ob Beobachtungen einer Zufallsstichprobe einer normalverteilten Zufallsvariable zugeordnet werden können. Dabei wird das Verhältnis des Quadrates einer kleinsten Fehlerquadratschätzung und der Stichprobenvarianz betrachtet. Liegt eine Normalverteilung vor, so wird das Verhältnis von eins erreicht. Kleine Werte sprechen wider eine Normalverteilung der beobachteten Werte [12]. Nachdem sich herausgestellt hat, dass die entstandenen Messwerte nicht normalverteilt sind, wurde der *Wilcoxon-Mann-Whitney-Test* zur Überprüfung signifikanter Unterschiede zweier Stichproben ausgewählt. Dieser Test wird auch „U-Test nach Wilcoxon, Mann und Whitney“ genannt. Es handelt sich dabei um einen Rangsummentest zweier unabhängiger Stichproben für nicht-normalverteilte Grundgesamtheiten. Er überprüft, ob sich die Verteilungen der beiden Stichproben unterscheiden und demnach nicht derselben Grundgesamtheit angehören. Ist letzteres der Fall, gilt auch, dass die Unterschiede signifikant sind [12]. Als Signifikanzniveau wurde in allen Fällen $p = 0.05$ gewählt.

5.4 Versuch 1: Audio-Transcription

Wie schon in Abschnitt 3.3.1 beschrieben, sollten die Worker englische Wörter, die sie über Audio-Dateien vorgespielt bekamen, in die Textform übertragen. Anschließend wurde eine Multiple-Choice-Frage beantwortet. Die Ergebnisse der Worker wurden anschließend mit den bekannten Lösungen verglichen.

5.4.1 Diskussion der Ergebnisse

In Tabelle 5.2 und 5.3 sind zunächst die Ergebnisse der untersuchten Werte aufgeführt und weiterhin die Ergebnisse der Signifikanzuntersuchungen, wie in Abschnitt 5.3 beschrieben. In diesem Versuch bearbeiteten 18 Worker die angegebenen Aufgaben. Im Mittel erbrachten die Worker dabei eine Qualität von 0,8711. Die maximale Leistung erreichte den Höchstwert von 1 (alle Aufgaben einer Serie richtig bearbeitet), der schlechteste Wert liegt

Untersuchter Wert	Ergebnis
AW	18
MRQ	0,871
MaxRQ	1,000
MinRQ	0,400
MRQr	0,875
MRQf	0,858
MMC	0,725
AeWMC	3
MRQa	0,883
MaxAAW	30
MinAAW	5
MAAW	25
MRQWmin	0,564
MRQWmax	1,000
MRQsim	0,829

Tabelle 5.2: Evaluation des Versuches „Audio-Transcription“.

immerhin bei 0,4. Auffällig in diesen Ergebnissen ist, dass sich MRQ, das Mittel aller Richtigkeitsquotienten, kaum von MRQf (0,8757) oder MRQr (0,8583) unterscheidet. Es gibt also nur geringfügige Unterschiede zwischen den Workern, welche die Multiple-Choice-Frage falsch und denen, die sie richtig beantwortet hatten. Die in Tabelle 5.3 aufgeführten p-Werte zeigen deutlich, dass diese geringen Unterschiede nicht signifikant und damit nicht von Bedeutung sind. Es gibt also keinen Hinweis darauf, dass sich die Arbeitsqualität der Gruppe der Worker, welche die Multiple-Choice-Frage richtig beantwortet haben, von der Gruppe der Worker mit falscher Antwort der Multiple-Choice-Frage unterscheiden. Bei einem Ausschluss von Workern aufgrund falscher Multiple-Choice-Frage zeigt sich aber ein leichter Unterschied zum Ausschluss aufgrund falsch bearbeiteter Golddaten: Die Arbeitsqualität, MRQa ist im Mittel ein wenig erhöht. Der p-Wert mit $p = 0,01069$ bestätigt, dass dieser Unterschied signifikant ist. Es ist also anzunehmen, dass die Methode der Multiple-Choice-Fragen als Qualitätssicherungsmechanismus in diesem Fall Vorteile gegenüber dem Einsatz von Golddaten hat. In diesem Fall würden durch den Einsatz der Fragen nur drei Worker aus dem Ergebnispool ausgeschlossen. Von einem einzelnen Worker wurden maximal 30 und minimal 5 Aufgaben bearbeitet. Besonders ist, dass das Mittel der bearbeiteten Aufgaben, MAAW, mit 25 nahe am Maximum liegt. Das bedeutet, dass wenige Worker einen großen Teil der Aufgaben bearbeiten. Interessant ist in diesem Zusammenhang auch, dass die Worker, die den Großteil der Aufgaben beantwortet haben, die beste Arbeitsqualität (MRQWmax) lieferten.

5.4.2 Fazit

Es kann in diesem Versuch nicht mit Sicherheit gesagt werden, dass Worker, welche die Multiple-Choice-Frage richtig beantwortet haben, eine bessere Arbeitsqualität lieferten, als Worker, die die Fragen falsch beantwortet hatten. Es ist aber ersichtlich, dass ein Ausschluss aufgrund von falschen Multiple-Choice-Fragen eine signifikante Verbesserung

Untersuchte Werte	p-Wert
Normalverteilung von MRQ	2,422E-11
Normalverteilung von MRQr	1,940E-09
Normalverteilung von MRQf	3,045E-05
Normalverteilung von MRQa	1,766E-10
Normalverteilung von MRQsim	1,813E-09
Signifikanztest von MRQ und MRQr	0,947
Signifikanztest von MRQ und MRQf	0,897
Signifikanztest von MRQr und MRQsim	0,032
Signifikanztest von MRQa und MRQsim	0,010

Tabelle 5.3: Signifikanz-Untersuchungen des Versuches „Audio-Transcription“.

der Arbeitsqualität zur Folge hat, als der Ausschluss aufgrund von falsch bearbeiteten Golddaten.

5.5 Versuch 2: Image-Categorization

In diesem Versuch sollten die Worker Bilder, die ihnen präsentiert wurden, in eine von vier verschiedenen Kategorien einordnen und zum Abschluss eine Multiple-Choice-Frage beantworten. Die Ergebnisse der Worker wurden anschließend mit den bekannten Lösungen verglichen.

Untersuchter Wert	Ergebnis
AW	33
MRQ	0,910
MaxRQ	1,000
MinRQ	0,400
MRQr	0,916
MRQf	0,901
MMC	0.666
AeWMC	11
MRQa	0,909
MaxAAW	30
MinAAW	5
MAAW	26,360
MRQWmin	0,800
MRQWmax	0,913
MRQsim	0,901

Tabelle 5.4: Evaluation des Versuches „Image-Categorization“.

Untersuchte Werte	p-Wert
Normalverteilung von MRQ	2,200E-16
Normalverteilung von MRQr	1,315E-14
Normalverteilung von MRQf	1,374E-11
Normalverteilung von MRQa	5,594E-15
Normalverteilung von MRQsim	2,200E-16
Signifikanztest von MRQ und MRQr	0,908
Signifikanztest von MRQ und MRQf	0,879
Signifikanztest von MRQr und RQsim	0,075
Signifikanztest von MRQa und RQsim	0,122

Tabelle 5.5: Signifikanz-Untersuchungen des Versuches „Image-Categorization“.

5.5.1 Diskussion der Ergebnisse

Tabelle 5.4 und 5.5 zeigen die Ergebnisse der untersuchten Werte und die Ergebnisse der Signifikanzuntersuchungen, wie in Abschnitt 5.3 ausgeführt. In diesem Versuch bearbeiteten 33 Worker die präsentierten Aufgaben. Im Mittel betrug die Arbeitsqualität dabei 0,9103 (MRQ), was als sehr hoch einzustufen ist. Die maximale Leistung erreicht auch hier den Höchstwert von 1, die minimale Leistung liegt bei 0,4. In den Ergebnissen dieses Versuchs fällt auf, dass die Werte MRQ, MRQr, MRQf, MRQa und MRQsim sehr nahe beieinander liegen und sich nur um Bruchteile unterscheiden. MRQ beschreibt dabei das Mittel des Richtigkeitsquotienten, MRQr das Mittel bei richtiger Multiple-Choice-Frage, MRQf das Mittel bei falscher Multiple-Choice-Frage und MRQa das Mittel des Richtigkeitsquotienten bei Ausschluss der Worker, welche die Frage falsch beantwortet hatten. MRQsim beschreibt das gewichtete Mittel der Richtigkeit beim simulierten Einsatz von Multiple-Choice-Fragen. Die Signifikanzuntersuchungen zeigen, dass die Unterschiede nicht von Bedeutung sind – der p-Wert liegt immer über dem gewählten Signifikanzniveau. Der Einsatz von Multiple-Choice-Fragen zur Qualitätssicherung trägt hier nicht zu einer signifikanten Veränderung der Arbeitsqualität bei – auch nicht im Vergleich zum Einsatz von Golddaten. Auch die Werte MRQWmin und MRQWmax liegen nicht weit voneinander entfernt. Alle Worker scheinen also in ihrer Arbeitsqualität gleich zu sein, unabhängig, ob sie die Multiple-Choice-Fragen lösen konnten oder von der Anzahl der bearbeiteten Aufgaben. Dieses Phänomen könnte auf den geringeren Antwortspielraum im Versuch „Image-Categorization“ zurückzuführen sein.

5.5.2 Fazit

In diesem Versuch ist der Einsatz von Multiple-Choice-Fragen mit dem Einsatz von Golddaten im Bezug auf die Arbeitsqualität gleich auf. Auch gibt es keinen Unterschied zwischen den Workern, die die Fragen richtig, und denen, die sie falsch beantwortet hatten. Dieses Ergebnis kann mit dem geringen Antwortspielraum der Art des genutzten Jobs zusammenhängen.

5.6 Versuch 3: Data-Enrichment

Hierbei wurden die Worker beauftragt zu Filmtiteln, die ihnen präsentiert wurden, das Jahr der Entstehung des Films mithilfe der Suchmaschine *Google* herauszufinden. Dabei wurde als Hilfestellung die Suchanfrage in einem Link kodiert, der nur noch aufgerufen werden musste.

Untersuchter Wert	**Ergebnis**
AW	26
MRQ	0,892
MaxRQ	1,000
MinRQ	0,600
MRQr	0,895
MRQf	0,884
MMC	0,745
AeWMC	19
MRQa	0,950
MaxAAW	25
MinAAW	5
MAAW	26,360
MRQWmin	0,950
MRQWmax	0,913
MRQsim	0,887

Tabelle 5.6: Evaluation des Versuches „Data-Enrichment“.

Untersuchte Werte	**p-Wert**
Normalverteilung von MRQ	1,869E-12
Normalverteilung von MRQr	1,220E-10
Normalverteilung von MRQf	2,133E-05
Normalverteilung von MRQa	1,815E-11
Normalverteilung von RQsim	1,429E-10
Signifikanztest von MRQ und MRQr	0,890
Signifikanztest von MRQ und MRQf	0,785
Signifikanztest von MRQr und RQsim	0,079
Signifikanztest von MRQa und RQsim	0,073

Tabelle 5.7: Signifikanz-Untersuchungen des Versuches „Data-Enrichment“.

5.6.1 Diskussion der Ergebnisse

In Tabelle 5.6 und Tabelle 5.7 sind zum Einen die Ergebnisse der untersuchten Werte und zum Anderen die Ergebnisse der Signifikanzuntersuchungen dargestellt. 26 Worker ar-

beiteten an diesem Versuch. Die Arbeitsqualität betrug im Mittel 0,892 (MRQ) und liegt damit zwischen den beiden oben erläuterten Versuchen. Die maximale Leistung betrug 1, das Minimum liegt bei 0,6. Die Ergebnisspanne ist damit weniger breit als in den anderen Versuchen. MRQ, MRQr und MRQf unterscheiden sich hier nur unwesentlich. Diese Unterschiede besitzen keine Signifikanz, wie in Tabelle 5.7 zu sehen ist. Bemerkenswert ist hier der Unterschied zwischen MRQa und MRQsim. Die Signifikanzuntersuchung überschreitet allerdings in diesem Fall das Signifikanzniveau knapp, sodass anzunehmen ist, dass die Unterschiede zufälliger Art sind. Auch hier trägt der Einsatz von Multiple-Choice-Fragen nicht zu einer signifikanten Veränderung der Arbeitsqualität bei.

5.6.2 Fazit

In diesem Versuch gibt es keinen Unterschied in der Arbeitsqualität beim Einsatz von Multiple-Choice-Fragen oder Golddaten. Auch unterscheiden sich die Worker nicht in ihrer Arbeitsqualität – egal, ob sie die Frage richtig oder falsch beantwortet haben.

5.7 Abschließendes Fazit

In einem von drei Versuchen erwies sich der Einsatz von Multiple-Choice-Fragen dem Einsatz von Golddaten überlegen. In keinem der Versuche hatte der Einsatz von Wissensfragen alleine einen Einfluss auf die Arbeitsqualität der Worker. Daraus muss der Schluss gezogen werden, dass der Einsatz von Wissensfragen als Qualitätssicherungsmechanismus im Crowdsourcing die Arbeitsqualität *nicht verbessert*. Andererseits ist aber ersichtlich, dass der Einsatz von Wissensfragen mindestens so gut funktioniert, wie der Einsatz von Golddaten, und damit als weiterer Qualitätssicherungsmechanismus seine Berechtigung besitzt. Dies bedeutet, dass Wissensfragen durchaus als Alternative zu Golddaten dienen können. Dabei haben Wissensfragen einige Vorteile gegenüber Golddaten und anderen Qualitätssicherungsmechanismen:

- Wissensfragen können durch das beschriebene Softwaretool automatisch erzeugt werden. Ein menschliches Eingreifen ist nicht mehr nötig – dies spart Zeit und Geld.
- Wissensfragen können in wahnsinnig hoher Anzahl generiert werden. Das Wiedererkennen von Fragen ist nahezu ausgeschlossen. Dies bedeutet, dass keine Wartung der Fragen erfolgen muss.
- Durch den Einsatz von Wissensfragen kann man auf andere Qualitätssicherungsmechanismen verzichten. Dadurch verbessert sich das Verhältnis von Workern zu Aufgaben, da z.B. das „Mehrheitsvotum“ nicht benötigt wird.

Automatisch generierte Wissensfragen funktionieren also so gut wie der Einsatz von Golddaten und können daher in die Reihe der Qualitätssicherungsmechanismen im Crowdsourcing aufgenommen werden.

6 Probleme und Hindernisse

Dieses Kapitel beschäftigt sich mit den Problemen von Crowdsourcing im Allgemeinen und den Hindernissen, die bei der Arbeit mit Crowdflower und dem Fragengenerierungstool Q2G aufgetreten sind.

6.1 Crowdsourcing

Crowdsourcing ist eine zukunftsträchtige Möglichkeit der Datenverarbeitung und wird sicher in den kommenden Jahren noch mehr an Bedeutung gewinnen. Leider hat diese Art und Weise Informationen zu bearbeiten einige entscheidende Nachteile, die auch in dieser Arbeit zu Tage getreten sind.

Auf der Seite desjenigen, der Jobs auf Crowdsourcing-Plattformen erstellt und anbietet, besteht der größte Nachteil darin, dass das „Debugging“ – die Fehlersuche – eines Jobs auf nur wenigen Fakten und viel Spekulationen beruht. Während man bei der Arbeit mit Testpersonen vor Ort die Möglichkeit hat genau nachzufragen, was an welcher Stelle unklar ist, und im Zweifelsfall gemeinsam mit der Testperson herausfinden kann, was das Problem ist, hat man genau diese Möglichkeiten mit einer anonymen Masse von Arbeitern nicht. Diese Anonymität, die einerseits große Vorteile bietet, ist aber selbst das größte Hindernis: Es gibt keine Anhaltspunkte, welche Eigenschaften die Menge an Workern hat, die gerade einen Task bearbeitet. Es ist fast unmöglich herauszufinden, welche Kenntnisse eine Gruppe von Workern hat, was der beste Arbeitsablauf für diese Gruppe ist oder in welcher Komplexität ein Job angeboten werden muss, damit er am besten zur Masse der Workern passt. In der Praxis nutzt man die Übereinstimmung von Antworten mehrerer Worker in Form von Übereinstimmungsmaßen und setzt sie ins Verhältnis zu den erwarteten Lösungen. Daraus kann man zwar Rückschlüsse ziehen, ob die Form, in welcher der Job oder ein Teil davon dargeboten wird, in der Praxis geeignet ist, aber man *weiß* immer noch nicht, was genau das Problem ist. Somit kann man sich nie ganz sicher sein, warum ein Job gute Datenqualität liefert oder warum nicht.

Des weiteren muss man wissen, dass die Arbeiter auf Crowdsourcing-Plattformen aus allen Teilen der Erde kommen. Amerika, Indien, China, Russland, die Türkei usw. – alle Nationen sind auf diesen Plattformen vertreten. Das bedeutet auch, dass man mit den unterschiedlichsten Kulturen, insbesondere *Arbeits*kulturen konfrontiert ist, ohne genau zu wissen, was dies bedeutet. In den meisten Fällen ist Englisch die Sprache, in der sich alle diese Nationen verständigen sollen. Die Versuche in dieser Arbeit haben aber auch gezeigt, dass viele Worker keine ausreichende Sprachfähigkeiten besitzen, um die sehr einfachen englischen Multiple-Choice-Fragen zu verstehen. So haben in Testläufen ohne Einschränkung der Nationalität der Worker z.B. über die Hälfte der Arbeiter als Antwort auf die Frage „I can’t understand the question“ gewählt. In den Versuchen hingegen, die ausschließlich mit Workern aus den vereinigten Staaten von Amerika, dem vereinigten Königreich, Irland und Kanada durchgeführt wurden, hat so gut wie *keiner* diese Antwortoption gewählt. Ein klares Zeichen dafür, dass die Sprachfähigkeiten zum Verständnis der Multiple-Choice-Fragen nicht ausreichten. Und wenn es schon dafür nicht reicht, muss man auch davon ausgehen, dass die Arbeitsanweisungen und das User-Interface auch nicht vollständig verstanden werden.

Zuletzt hat Crowdsourcing einige bemerkenswerte Auswirkungen auf Gesellschaft und Wirtschaft. Crowdsourcing hat das Potenzial bestehende Unternehmungen zu untergraben und letztlich überflüssig zu machen. Jeff Howe berichtet in seinem Artikel *The Rise of Crowdsourcing* [14] davon, wie ein freiberuflicher Stock-Fotograf[1] seine Arbeitsgrundlage verliert, weil Crowdsourcing-Plattformen – in diesem Fall *iStockphoto*[2] – Preise für Bilder anbieten, die er nicht unterbieten kann. Dabei verzichten Kunden ganz bewusst auf professionelle Arbeit, denn auf solchen Plattformen bieten vorwiegend Amateure ihre Bilder an. Da jedoch der Preis pro Bild unschlagbar ist und die Qualität mindestens *ausreichend* ist, verdrängt diese Art des Crowdsourcing den ursprünglichen Geschäftszweig. Crowdsourcing wirkt also disruptiv auf die Unternehmenslandschaft ein. Dieser Effekt ist in vielen Bereichen zu beobachten, wo Experten-Arbeit des Preises wegen durch die Arbeit einer anonymen Masse ersetzt wird. Dieser Preisvorteil für den Kunden führt aber auf der anderen Seite zu Problemen. Auch die Worker einer Crowdsourcing-Plattform wollen bezahlt werden. Leider scheint es völlig legitim zu sein, für die Arbeit der Worker Preise zu bezahlen, welche die Bezeichnung „Vergütung“ nicht verdienen. Im Falle der Anwendungen, die als „Games with a purpose“ bekannt geworden sind [1], ist es sogar üblich für die menschliche Arbeitsleistung überhaupt nichts zu zahlen. Demgegenüber steht die Tatsache, dass für einige Worker die Arbeit auf Crowdsourcing-Plattformen eine wichtige Einkommensquelle darstellt [17]. Somit hat die Bezahlung der Worker einen direkten Einfluss auf deren Lebensumstände und man kann bei den niedrigen Preisen, die für die meiste Arbeit gezahlt wird, durchaus von Ausbeutung sprechen. Die Zukunft wird zeigen, ob sich dieses Problem weiter verschärfen wird.

6.2 Crowdflower

Im Laufe der Versuche, die im Rahmen dieser Arbeit durchgeführt wurden, offenbarten sich auch einige Schwächen der Crowdsourcing-Plattform *Crowdflower*, die ein professionelles Arbeiten und die nahtlose Einbettung in andere Systeme erschweren.

```
<p>
  <a target="_blank" href="{{url0}}" id="">Click here to listen to an Audio file</a>
</p>
<cml:textarea label="Please Transcribe The Text Here:" name="transcribed_text0" default="Add transcripti
on here..." aggregation="agg" instructions="Please enter NA if link is broken or if there is no text."
class="" validates="required alpha"></cml:textarea>

<p>
  <a target="_blank" href="{{url1}}" id="">Click here to listen to an Audio file</a>
</p>
<cml:textarea label="Please Transcribe The Text Here:" name="transcribed_text1" default="Add transcripti
on here..." aggregation="agg" instructions="Please enter NA if link is broken or if there is no text."
class="" validates="required alpha"></cml:textarea>

<p>
  <a target="_blank" href="{{url2}}" id="">Click here to listen to an Audio file</a>
</p>
<cml:textarea label="Please Transcribe The Text Here:" name="transcribed_text2" default="Add transcripti
on here..." aggregation="agg" instructions="Please enter NA if link is broken or if there is no text."
class="" validates="required alpha"></cml:textarea>

<p>
  <a target="_blank" href="{{url3}}" id="">Click here to listen to an Audio file</a>
</p>
<cml:textarea label="Please Transcribe The Text Here:" name="transcribed_text3" default="Add transcripti
on here..." aggregation="agg" instructions="Please enter NA if link is broken or if there is no text."
class="" validates="required alpha"></cml:textarea>
```

Abbildung 6.1: Editor-Fenster auf Crowdflower.

Der Editor, mit welchem die Jobs erstellt und bearbeitet werden, ist von einfachster Art und genügt höchstens dann, wenn man gelegentlich Jobs anlegt, die nicht besonders

[1] Ein Fotograf, der Bilder auf Vorrat anfertigt, um sie später an Kunden zu vermitteln.
[2] http://www.istockphoto.com (zuletzt abgerufen am 14. September 2016).

verändert oder optimiert werden müssen. Alle Prinzipien, die in der Softwareentwicklung gängig sind und zu guter Softwarequalität beitragen, sollten auch bei der Entwicklung von Crowdsourcing-Jobs möglich sein. Leider lässt der einfache Editor dies nicht zu. Es ist nicht möglich den Code in hierarchischen Projekten zu organisieren, Code-Verdopplung kann nicht vermieden werden, es ist nicht möglich Softwareteile auszulagern und mehrfach zu verwenden. Auch die bequemen Möglichkeiten, die eine ausgereifte integrierte Entwicklungsumgebung bietet – Autovervollständigung, Refactoring, Quellcode-Navigation – können nicht genutzt werden. Das macht das Erstellen des Quellcodes mühsam und fehleranfällig. Abbildung 6.1 zeigt den Editor, mit dem das User-Interface der Jobs auf Crowdflower bearbeitet werden kann.

Die übrige Arbeitsumgebung von Crowdflower lässt für den professionellen Anwender ebenfalls viele Wünsche offen. So können die angelegten Jobs zwar mittels Tags kategorisiert werden, durch die Listendarstellung aller Jobs ist es aber dennoch schwierig den Überblick zu behalten. Weiterhin sorgt die als Webseite umgesetzte Arbeitsumgebung dafür, dass die Navigation zwischen den einzelnen Seiten vergleichsweise langsam verläuft. Dies kostet Zeit und sorgt für Frust. Für einen professionellen Einsatz wäre es notwendig, dass Crowdflower einige Arbeitsabläufe vereinfacht und automatisiert, und die Arbeitsumgebung mit den auf dem Markt der Entwicklungsumgebungen üblichen Möglichkeiten ausstattet.

6.3 Fragengenerierung mittels Q2G

Die automatisierte Generierung der Fragen durch das Q2G-System ist zweifelsohne beeindruckend und liefert in der Regel verständliche und nutzbare Fragen. Die Autoren des Systems geben allerdings zu bedenken, dass lediglich eine sehr einfache Umwandlung der Informationen in menschenverständliche Sprache implementiert wurde [20]. Dies hat zur Folge, dass die Einfachheit der genutzten Sprache manchmal nicht ausreicht, um grammatikalisch richtige und für Menschen verständliche Fragen zu generieren. Dies wurde auch in den Testläufen mit den Testpersonen vor Ort bemängelt. Einige vom System generierte Fragen sahen z.B. so aus:

- This country has the location Muay Lao and has event Vietnam War.
- This musician has the musical role of wordnet guitar 103467517 and created So Excited.
- This Subject?verb?object language and Languages of Italy is the official language of Andorra.

Bei den obigen Fragen kann man sich die Bedeutung der Frage durchaus erschließen. Es wäre jedoch besser, wenn die Fragen einheitlich verständlich generiert werden würden. Für einen weiterreichenden Einsatz wäre es also notwendig, die Umsetzung der Informationen aus den Wissensdatenbanken in menschliche Sprache zu verbessern.

Literatur

[1] Laura Dabbish Luis von Ahn. „Designing Games With A Purpose“. In: *Communications of the ACM* 58.8 (2008), S. 58 –67.

[2] Ahmet Aker, Mahmoud El-haj, M-dyaa Albakour und Udo Kruschwitz. „Assessing Crowdsourcing Quality through Objective Tasks“. In: *International Conference on Language Resources and Evaluation (LREC)* (2012), S. 1456–1461. URL: http://cswww.essex.ac.uk/staff/udo/papers/LREC2012Assessing.pdf.

[3] Omar Alonso. „Practical Lessons for Gathering Quality Labels at Scale“. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15* (2015), S. 1089–1092. DOI: 10.1145/2766462.2776778.

[4] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell und Katrina Panovich. „Soylent“. In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*. New York, New York, USA: ACM Press, 2010, S. 313. ISBN: 9781450302715. DOI: 10.1145/1866029.1866078.

[5] Daren C. Brabham. *Crowdsourcing*. 2013, S. 176. ISBN: 9780262314251. URL: https://www.timeshighereducation.co.uk/books/crowdsourcing-by-daren-c-brabham/2005865.article.

[6] Michael J. Coren. *Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks*. URL: http://www.scientificamerican.com/article/foldit-gamers-solve-riddle/ (besucht am 08. 06. 2016).

[7] CrowdFlower Inc. *How CrowdFlower can work for you*. URL: https://www.crowdflower.com/use-cases/ (besucht am 08. 06. 2016).

[8] Djellel Eddine Difallah, Gianluca Demartini und Philippe Cudre-Mauroux. „Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms“. In: *Proceedings of the 1st International Workshop on Crowdsourcing Web Search* (2012), S. 20–25. ISSN: 16130073.

[9] FoldIt Project. *The Science behind FoldIt*. URL: http://fold.it/portal/info/about (besucht am 08. 06. 2016).

[10] Samuel Greengard. „Following the crowd“. In: *Communications of the ACM* 54.2 (2011), S. 20. ISSN: 00010782. DOI: 10.1145/1897816.1897824.

[11] Derek L. Hansen, Patrick Schone, Douglas Corey, Matthew Reid und Jake Gehring. „Quality Control Mechanisms for Crowdsourcing: Peer Review, Arbitration, & Expertise at Familysearch Indexing“. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (2013), S. 649–660. DOI: 10.1145/2441776.2441848.

[12] J. Hedderich und L. Sachs. *Angewandte Statistik: Methodensammlung mit R*. Springer Berlin Heidelberg, 2011. ISBN: 9783642244001.

[13] Jeff Howe. *Crowdsourcing - Why the Power of the Crowd is driving the Future of the Business*. URL: www.crowdsourcing.com (besucht am 03. 06. 2016).

[14] Jeff Howe. „The Rise of Crowdsourcing“. In: *Wired Magazine* 14.06 (2006), S. 1–5. ISSN: 10006788. DOI: 10.1086/599595.

[15] P. G. Ipeirotis. „Analyzing the amazon mechanical turk marketplace“. In: *XRDS: Crossroads* 17.2 (2010), S. 16–21. ISSN: 15284972. DOI: 10.1145/1869086.1869094.

[16] Tim Matthews. *The State of Enterprise Crowdsourcing 2013*. 2013. URL: https://www.crowdflower.com/the-state-of-enterprise-crowdsourcing-2013/.

[17] Joel Ross. „Who are the Turkers? Worker Demographics in Amazon Mechanical Turk“. In: *CHI 2010: 28th ACM Conference on Human Factors in Computing Systems*. 2010, S. 2863–2872.

[18] Thimo Schulze, Dennis Nordheimer und Martin Schader. „Worker Perception of Quality Assurance Mechanisms in Crowdsourcing and Human Computation Markets“. In: *19th Americas Conference on Information Systems 2013: AMCIS 2013 Proceedings* (2013), S. 1–11.

[19] Dominic Seyler, Mohamed Yahya und Klaus Berberich. „Generating Quiz Questions from Knowledge Graphs“. In: *Proceedings of the 24th International Conference on World Wide Web* (2015), S. 113–114. DOI: 10.1145/2740908.2742722.

[20] Dominic Seyler, Mohamed Yahya, Klaus Berberich und Omar Alonso. „Automated Question Generation for Quality Control in Human Computation Tasks“. In: *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*. New York, New York, USA: ACM Press, 2016, S. 360–362. ISBN: 9781450342087. DOI: 10.1145/2908131.2908210.

[21] Web Recommender Systems und Klaus Berberich. „Crowdsourcing and Human Computation, Introduction“. In: *Encyclopedia of Social Network Analysis and Mining (ESNAM)* 1995 (2014), S. 304–315. DOI: 10.1007/978-1-4614-6170-8.

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

Abkürzungsverzeichnis

CML	Crowdflower Markup Language
SPARQL	SPARQL Protocol And RDF Query Language
AW	Anzahl der Worker
MRQ	Mittel des Richtigkeitsquotienten
RQsim	Richtigkeitsquotient multipliziert mit Golddatenwahrscheinlichkeit
MaxRQ	Maximum des Richtigkeitsquotienten
MinRQ	Minimum des Richtigkeitsquotienten
MRQr	Mittel des Richtigkeitsquotienten bei richtiger Multiple-Choice-Frage
MRQf	Mittel des Richtigkeitsquotienten bei falscher Multiple-Choice-Frage
MMC	Mittel der richtig beantworteten Multiple-Choice-Fragen bei maximalem Richtigkeitsquotienten
AeWMC	Anzahl eliminierter Worker durch Multiple-Choice-Fragen
MRQa	Mittel des Richtigkeitsquotienten bei Ausschluss der Worker
MaxAAW	Maximale Anzahl der bearbeiteten Aufgaben pro Worker
MinAAW	Minimale Anzahl der bearbeiteten Aufgaben pro Worker
MAAW	Mittel der Anzahl der bearbeiteten Aufgaben pro Worker
MRQWmin	Mittel des Richtigkeitsquotienten von Workern mit minimaler Anzahl bearbeiteter Aufgaben
MRQWmax	Mittel des Richtigkeitsquotienten von Workern mit mehr als 75 Prozent der maximalen Anzahl bearbeiteter Aufgaben
MRQsim	Mittel des Richtigkeitsquotienten mit simulierten Golddaten
XML	Extensible Markup Language

Kolophon

Dieses Dokument wurde mit der LaTeX-Vorlage für Abschlussarbeiten an der htw saar im Bereich Informatik/Mechatronik-Sensortechnik erstellt (Version 2.1). Die Vorlage wurde von Yves Hary und André Miede entwickelt (mit freundlicher Unterstützung von Thomas Kretschmer, Helmut G. Folz und Martina Lehser). Daten: (F)10.95 – (B)426.79135pt – (H)688.5567pt