FQS FORUM: QUALITATIVE SOZIALFORSCHUNG SOCIAL RESEARCH

Volume 9, No. 2, Art. 17
Mai 2008

Schnellere Transkription durch Spracherkennung?

Thorsten Dresing, Thorsten Pehl & Claudia Lombardo

Keywords: Transkription, Spracherkennung, Software

Zusammenfassung: Ist Spracherkennung dazu in der Lage, Forschenden eine schnellere Transkription ihrer Interview-Audioaufnahmen zu ermöglichen als die bisher etablierte Form der manuellen Transkription? Die Untersuchung, über die hier berichtet wird, wirft einen ersten empirischen Blick auf die Möglichkeit, die Transkription von Interviews im sozialwissenschaftlichen Kontext durch den Einsatz von Spracherkennungssoftware zu bewältigen. Hierzu wurden von 20 Personen unter gleichen Bedingungen Transkripte sowohl manuell als auch unter Zuhilfenahme von Spracherkennungssoftware erstellt. Die Erfahrungen hiermit wurden qualitativ und quantitativ ausgewertet. Dabei lässt sich feststellen, dass Spracherkennung und manuelle Transkription etwa gleiche Bearbeitungszeiten benötigen, die Spracherkennung aber hinsichtlich ihrer Präzision und Bedienbarkeit deutliche Schwächen aufweist.

Inhaltsverzeichnis

1. Zum Hintergrund der Studie

In angeregten Diskussionen in den Mailinglisten qual-software und QSF-L (Mailingliste für Qualitative Sozialforschung) Ende 2006 tauschten sich Wissenschaftlerinnen und Wissenschaftler in verschiedenen Beiträgen über ihre Erfahrungen aus, inwieweit Spracherkennung eine sinnvolle Unterstützung bei der Transkription von Interviews darstellt. Spracherkennung bezeichnet Verfahren, die gesprochene Sprache computergestützt in Schriftform übersetzen. [1]

Transkription ist ein wesentlicher Arbeitsschritt in qualitativen Forschungsprojekten. Das zu transkribierende Material kann aus Audioaufnahmen oder Videomitschnitten bestehen. Je nach Forschungsgegenstand oder Disziplin werden diese Aufnahmen nach unterschiedlichsten Transkriptionsregeln verschriftlicht (vgl. KUCKARTZ 2007;

KOWAL & O'CONNELL 1995). Grob gesagt unterscheiden sich die Transkriptionssysteme in der Differenziertheit der Darstellung lautsprachlicher und nonverbaler Elemente. Einen Überblick über eine Auswahl verschiedener Transkriptionssysteme liefern EHRLICH und SWITALLA (1973; siehe auch DITTMAR 2004). [2]

Die in einigen Transkriptionskonventionen nötige Partiturschreibweise oder Intonationszeichen, wie z.B. bei der Analysesoftware Praat (BOERSMA & WEENIK 1996) und dem Gesprächsanalytischen Transkriptionssystem GAT (SELTING et al. 1998), sind mit dem Einsatz von Spracherkennungssoftware grundsätzlich unvereinbar. Spracherkennung kann eine so detaillierte Verschriftlichung als Partitur und mit angemessenen Sonderzeichen nicht leisten. Die Diskussionen in den Mailinglisten zum Einsatz von Spracherkennung bezogen sich dagegen meist auf einfache Transkriptionsregeln, wie sie beispielsweise für die Themen- oder Inhaltsanalyse in den Sozialwissenschaften verwendet werden. Nachfolgend ein kleiner Auszug aus einem Beispieltranskript nach Transkriptionsregeln von KUCKARTZ, DRESING, RÄDIKER und STEFER (2007, S.27f.):

> B7: Ich habe, also ich habe so eine Lerngruppe mit meinem Freund. Das heißt, ich erkläre ihm alles zweimal und dann sitzt es bei mir auch. Und dann noch, ja, habe ich mich noch mal mit, mit einem aus meiner Arbeitsgruppe da von Statistikgruppe getroffen.
>
> I: Und wie, wie fühlst du dich dabei? Also, hast du positive oder negative Einstellungen gegenüber der Statistik oder (...)
>
> B7: Ich mag das ganz gerne. Hätte ich am Anfang auch nicht gedacht, aber ich mochte auch Mathe, und deshalb finde ich das ganz okay.
>
> I: Und hat sich das im Laufe des Semesters verändert? (B7: Ja!) Und wenn ja, wie? [3]

Bisher ist der typische Weg zur Transkription von Interviews die manuelle Transkription, d.h. das vollständige Abhören und Abtippen des Interviewtextes nach angemessenen Transkriptionsregeln. Unter etabliertem, manuellem Transkribieren verstehen wir folgendes Vorgehen: Eine Audioaufnahme wird mit einem Transkriptionsgerät oder einer Transkriptionssoftware über Kopfhörer angehört. Transkriptionssoftware unterscheidet sich von reinen Abspielprogrammen wie dem WindowsMediaPlayer, Winamp, iTunes oder anderen: Sie ermöglicht, die Wiedergabe in der Geschwindigkeit zu regulieren und durch einen Fußschalter oder Tastenkürzel zu steuern. Beim Stoppen der Wiedergabe wird die Aufnahme automatisch um einige Sekunden zurückgespult, um beim Wiedereinstieg ein gutes Anknüpfen an das vorher Gehörte zu ermöglichen. Moderne Transkriptionsprogramme bieten darüber hinaus die Möglichkeit, Zeitmarken zu setzen und wiederkehrende Begriffe (z.B. "Interviewer:") über Textbausteine einzufügen. Einen Überblick über verfügbare Software gibt das Gesprächsanalytische Informationssystem des Instituts für Deutsche Sprache Mannheim. [4]

Spracherkennungssoftware ermöglicht die automatische Verschriftlichung von Diktaten. Zurzeit sind Dragon Naturally Speaking oder IBM ViaVoice erhältlich (beide vertrieben durch die Firma Nuance). Auch die Software Voice Pro (Vertrieb durch Firma Linguatec) wird teilweise noch beworben, wurde jedoch seit 2004 nicht weiterentwickelt. [5]

Voraussetzung einer möglichst fehlerfreien Erkennung und Verschriftlichung des Gesprochenen ist, dass die Software zuvor auf die Sprechweise der diktierenden Person trainiert wurde. Unterschiedliche Stimmen in einer Audioaufnahme können Spracherkennungssysteme nicht unterscheiden. Die Software benötigt für eine möglichst fehlerfreie Erkennung und Verschriftlichung der Sprache zudem eine sehr deutliche und einheitliche Aussprache und Betonung. Spracherkennungssoftware ist daher nicht geeignet, Audioaufnahmen normaler Interviewsituationen mit annehmbarem Ergebnis direkt zu verschriftlichen, da hier gleichzeitig gesprochen wird und Umgangssprache, Dialekte oder Nebengeräusche vorkommen. [6]

In der qual-software-Mailingliste wurde jedoch von einer Methode berichtet, wie diese Hürden umgangen und Spracherkennung zur Transkription von Interviews wirksam eingesetzt werden könnte: Die Person, auf die die Spracherkennung trainiert wurde, könne die Audioaufnahme abhören und das Abgehörte in den Computer diktieren. Dies wird dann von der Spracherkennungssoftware in Text umgesetzt. Der Vorteil dieser Methode soll eine deutliche Zeiteinsparung gegenüber der manuellen Transkription sein. Zum Vergleich: Für die manuelle Transkription einer Interviewstunde nach einfachen Transkriptionsregeln (beispielsweise HOFFMANN-RIEM 1984, S.301) fallen etwa vier bis acht Stunden Bearbeitungs- und Korrekturzeit an (vgl. KUCKARTZ et al. 2007, S.29). Bei diesem zeitaufwändigen, aber in vielen Forschungsvorhaben unerlässlichem Verfahren liegt der Wunsch nach Zeitersparnis auf der Hand. Die in qual-software angegebene Zeitspanne für die Transkription mit Hilfe der Spracherkennung hingegen läge bei etwa zwei bis drei Stunden pro Interviewstunde: "I can transcribe and edit three hours of multi-person conversations in about eight hours" (Auszug aus einem Mailinglistenbeitrag, qual-software, 15.12.2006). Dies wäre eine erhebliche Erleichterung und würde Forschenden unzählige Stunden Arbeitszeit ersparen. Wir waren zunächst aufgrund eigener negativer Erfahrungen mit Spracherkennung sehr skeptisch. Nach unserem Eindruck ist bei Spracherkennungsprogrammen, trotz propagierter Fehlerrate von nur einem Prozent, viel Übungszeit zu investieren, es sind viele Korrekturen nötig und es ist auf eine disziplinierte und genaue Aussprache zu achten. [7]

Um nun bisherige Aussagen und Empfehlungen aus dem Bereich der individuellen Erfahrung in den Bereich beleg- und diskutierbarer Daten zu heben, wäre eine empirische Untersuchung nötig. Solche Untersuchungen gibt es bisher nicht. Da wir uns seit 2004 intensiv mit digitaler Aufnahme und Transkription beschäftigen, möchten wir hier einen ersten empirischen Beitrag leisten. [8]

Unsere Frage lautet: Lassen sich in einer nicht-repräsentativen Stichprobe von Studierenden Hinweise finden, dass Spracherkennung dazu in der Lage ist, eine schnellere Transkription von Interview-Audioaufnahmen zu ermöglichen? [9]

2. Setting

Wir wollten Proband(inn)en jeweils einmal manuell und einmal per Spracherkennung Audioaufnahmen transkribieren und Korrekturlesen lassen. Der Vergleich sollte sich nicht ausschließlich auf die verbrauchte Zeit stützen, sondern auch subjektiv eingeschätzte Vorteile und Schwierigkeiten berücksichtigen. [10]

Um Spracherkennung und manuelle Transkription zu vergleichen, benötigten wir jeweils gleiche Ausgangsvoraussetzungen bei der Durchführung. Gleichbleibend und einheitlich musste daher die Audioaufnahme, die PC-Hard- und Software (Spracherkennung, Transkriptionssoftware[1], Fußschalter, Prozessor usw.) und die Reihenfolge, also der Ablauf der Bearbeitung sein. Wir haben deshalb von uns erhobenes Interviewmaterial als mp3-Audiodatei vorgegeben und einen Computer mit Fußschalter und Headset vorbereitet; außerdem eine genaue Arbeitsanleitung für die Versuchsteilnehmer(innen). Da wir einen Zusammenhang zwischen der Tippgeschwindigkeit und der Versuchszeit vermuteten, haben wir zudem die Tippgeschwindigkeit mittels eines Internetprogramms erhoben. Die Proband(inn)en schrieben hierfür für einige Minuten einen vorgegebenen Text ab, während die Anschläge pro Minute gezählt wurden.[2] So ermittelten wir einen Wert für die individuelle Tippgeschwindigkeit. [11]

Als Transkriptionsregeln haben wir bewusst eine stark reduzierte Form gewählt, die lediglich auf eine wortgenaue Darstellung mit sichtbaren Sprecher(innen)wechseln achtet. Dabei ist zu bedenken, dass Zeichen gängiger Transkriptionsregeln, wie z.B. /em/ oder /hm/ für Fülllaute, prinzipiell eher ungünstig durch Spracherkennung darstellbar sind. Das gewählte Setting stellt also für Spracherkennung tendenziell Optimalbedingungen her. (Folglich wären bei der Anwendung erweiterter Regelsysteme jeweils Zeitspannen hinzuzuaddieren.) Alle Texte sollten von den Proband(inn)en nach der Transkription abschließend noch einmal Korrekturgelesen werden, um Orthografiefehler zu beseitigen, ohne dabei die Audioaufnahme erneut anzuhören (aus zeitökonomischen Gründen). Die Proband(inn)en sollten während des Versuchs präzise die jeweilige Tipp- und Nachkorrekturzeit getrennt erfassen, um für uns die jeweiligen Gesamtzeiten besser interpretierbar zu machen. Zudem sollten sie auf einem Fragebogen[3] Angaben zu ihren Eindrücken

1 Software zur Unterstützung der Transkription durch einstellbare Wiedergabegeschwindigkeit, automatisches Rückspulintervall und Steuerung der Wiedergabe über einen Fußschalter.

2 Durch seine Kürze ermöglicht der Test ein schnelles Ergebnis. Es ist nicht davon auszugehen, dass die ermittelte Tippgeschwindigkeit über einen längeren Zeitraum, bei dem normale Ermüdungserscheinungen auftreten, aufrechtzuerhalten ist. Dennoch lieferte uns der Wert eine Möglichkeit, die Schreibgeschwindigkeit der Proband(inn)en miteinander zu vergleichen.

3 Die Durchführung der Transkription fand aufgrund des hohen Zeitaufwands ohne unsere Anwesenheit statt. Daher die Rückmeldung auf einem Fragebogen.

und möglichen Problemen im Umgang mit der jeweiligen Transkriptionsmethode festhalten. [12]

Als Basismaterial, das allen Proband(inn)en am gleichen Notebook (zeitversetzt) zur Verfügung gestellt wurde, dienten zwei leitfadengestützte Interviews á acht Minuten, die wir in einem ruhigen Seminarraum während der Tagung der Deutschen Gesellschaft für Soziologie 2006 in Kassel geführt hatten. Die Aufnahme wurde mit einem iPod Video (5. Generation) und dem Belkin TuneTalk Aufnahmeadapter aufgenommen und nach Übertragung auf den PC in eine mp3-Datei konvertiert[4]. Als Basismaterial wurden zwei unterschiedliche Interviews mit ähnlichem Thema und Komplexitätsgrad gewählt, damit der erste Durchgang "manuelle Transkription" nicht den Durchgang "Spracherkennung" durch Erinnerung an den Gesprächsverlauf beeinflusst. [13]

Als Proband(inn)en wählten wir Studierende am FB Erziehungswissenschaften der Philipps-Universität Marburg, zum Teil Seminarteilnehmende des Onlineseminars "Einführung in die computergestützte Text- und Inhaltsanalyse", die mit dem Thema Transkription und Interpretation qualitativer Daten bereits erste Vorerfahrungen gesammelt hatten. Insgesamt haben wir 20 Studierende für die Untersuchung gewinnen können. [14]

Zur Spracherkennung wurde die zurzeit aktuelle Version 9 des Programms Naturally Speaking eingesetzt, welches zu den verbreitetsten Spracherkennungsprogrammen gehört. Zum Abhören und Diktieren wurde das mitgelieferte Headset genutzt. Um zu gewährleisten, dass die Ergebnisse nicht durch unterschiedliche Verarbeitungszeiten aufgrund verschiedener Rechnerkonfigurationen beeinflusst werden, haben alle Proband(inn)en auf dem gleichen Notebook[5] gearbeitet. Zur manuellen Transkription stand die Transkriptionssoftware f4 samt Fußschalter zur Steuerung der Wiedergabe zur Verfügung. Die Teilnehmenden erhielten schriftliche Informationen über die Arbeit mit Tastenkürzeln, zu den geforderten Transkriptionsregeln, zudem einen Fragebogen mit offenen und geschlossenen Fragen und eine Anleitung zum Versuchsablauf. [15]

Der Ablauf im Überblick:

1. Die Proband(inn)en erhielten die nötige Ausrüstung (Notebook, CD mit Audiodateien, Headset) und ein schriftliches "Manual". Dessen Anweisungen wurden mit den Transkribierenden besprochen, um eventuelle Fragen zu klären.
2. Alle Teilnehmenden führten dann einen Test der eigenen Tippgeschwindigkeit durch und notierten das Ergebnis.

4 Der Aufnahmeadapter ermöglicht, Audioaufnahmen mit dem Musikabspielgerät iPod zu erstellen. Um diese Dateien in der Transkriptionssoftware wiederzugeben, muss die Aufnahme in das mp3-Format konvertiert werden.

5 Sony Vaio VGN-FS415M, Intel(R) Pentium(R) M processor 1.73GHz, 1GB RAM, Microsoft Windows XP Home Edition, Soundkarte: NVIDIA GForce Go 6400.

3. Der erste Text wurde manuell mit Fußschalterunterstützung transkribiert. Benötigte Schreib- und Korrekturzeit wurden notiert und der Fragebogen wurde ausgefüllt. Zur besseren Nachvollziehbarkeit wurden die Ergebnisse jeweils vor und nach der Korrektur gespeichert.
4. Die Spracherkennung wurde auf die Stimme der Proband(inn)en trainiert (ein Standardtraining ist laut Herstellerangaben ausreichend, um 98%ige Erkennungsgenauigkeit zu erreichen).
5. Das zweite Interview wurde abgehört und über ein Headset-Mikrofon der Spracherkennungssoftware diktiert. Die Wiedergabe wurde ebenfalls über Fußschalter gesteuert. Die Korrektur des Textes wurde per Hand durchgeführt. Auch für diesen Durchgang wurden Diktier- und Korrekturzeit notiert und der Fragebogen wurde ausgefüllt. [16]

3. Auswertung

Nachdem alle Teilnehmenden nach der Durchführung der Transkription ihre Erfahrungen und Bearbeitungszeiten notiert hatten, wurden ihre Aussagen im Textanalyseprogramm MAXQDA inhaltsanalytisch ausgewertet (vgl. MAYRING 1983, 2000). Dabei wurden die Aussagen nach den benannten Vor- und Nachteilen manueller Transkription und Spracherkennung kodiert. Kodiert wurden jeweils Sinneinheiten. Alle Kodes wurden im Team besprochen und dann thematisch zusammengefasst. Um in der Beschreibung der Ergebnisse die Aussagen der Proband(inn)en deutlich werden zu lassen, haben wir zentrale Zitate herausgesucht und in den folgenden Text eingeflochten (vgl. KUCKARTZ et al. 2007). Die Beschreibung der Ergebnisse ist zunächst deskriptiv. Die Schlussfolgerungen stellen dann unsere Interpretation der Beschreibung dar. [17]

3.1 Qualitative Auswertung

3.1.1 Manuelle Transkription – easy to start, präzise und regelkonform

Benannte Vorteile

Die manuelle Transkription wurde als "präziser" (Teilnehmer [TN] 8, Absatz 20) erlebt und es gab "weniger Fehler als bei der Sprachversion, da man selber schreibt" (TN14, 7). Sie wurde als eine "sichere Methode alles Gesagte zu erfassen" (TN5, 24) empfunden. Ein wesentlicher Vorteil sei die leichte Formatierung und Einhaltung von Transkriptionsregeln, man sei "schneller in Bezug auf Format und Interpunktion" (TN11, 6). "Generell brauchte ich im Vergleich zur Transkription mit Hilfe der Spracherkennung wenig Zeit, um mich auf die Situation einzustellen und konnte direkt loslegen" (TN12, 4). [18]

Einige Teilnehmer(innen) äußerten sich positiv zur Unterstützung durch die Transkriptionssoftware. Das Programm f4 sei "übersichtlich aufgebaut" (TN3, 4), "funktioniert nicht schlecht" (TN9, 4) und sei "leicht zu lernen" (TN7, 4). Man könne damit "gut arbeiten" (TN1, 6) und "die Geschwindigkeit selber bestimmen" (TN7, 4). Besonders hilfreich sei die "Wiederholung der letzten Worte nach dem

erneuten Treten des Fußschalters" (TN5, 6 + TN4, 9), denn "dass man immer wieder kurz zurückspulen konnte" (TN2, 7), habe den Schreibfluss erleichterte: "Der Fußschalter macht das Ganze sehr bequem" (TN12, 6) und sei "sozusagen eine 3. Hand" (TN19, 9), "was sicherlich einiges an Zeit eingespart hat" (TN15, 5). [19]

Benannte Nachteile

Als nachteilig wurde aufgrund der insgesamt "langen Bearbeitungszeit" (TN14, 9) die nachlassende "Konzentration" empfunden, die es erschwert habe, "fließend und zügig zu schreiben" (TN6, 20). [20]

3.1.2 Spracherkennung – cool aber mangelhaft

Benannte Vorteile

Das Arbeiten mit der Spracherkennung wird als "sehr bequem im Gegensatz zu der manuellen Transkription" (TN15, 23) und "sehr viel angenehmer" (TN17, 19) beschrieben, gerade auch, weil die "Hände frei sind" (TN7, 73). Zudem kam bei einigen ein "gutes Gefühl auf, Gesprochenes direkt auf dem Bildschirm zu sehen" (TN19, 40): Das ist "richtig cool, wenn man redet und es erscheint auf dem Bildschirm" (TN1, 18). Ein Proband empfand es als "praktisch und viel angenehmer 'nur' zu sprechen und nicht sehr angestrengt alles abzutippen" (TN6, 79). Für einige Befragte war es "in Verbindung mit gleichzeitiger Kontrolle am Bildschirm ein sehr effizientes und schnelles System zum Erfassen von Texten bzw. Interviews" (TN11, 17). Das "Nachsprechen geht in der Regel sehr schnell" (TN5, 69) und "macht Spaß" (TN14, 25). Es sei vor allem dann bequemer und schneller, "solange ich nichts korrigieren musste" (TN3, 24). [21]

Benannte Nachteile

Aussprache muss deutlich sein: Eine Voraussetzung für die genaue Umsetzung des gesprochenen Wortes in Schrift ist eine deutliche Aussprache. "Man muss immer sehr, sehr genau sprechen. Das ist mir ziemlich anstrengend …" (TN7, 73). Vor allem "am Ende …" war ein Proband "… nicht mehr so konzentriert und das hat dann … die deutliche Aussprache beeinträchtigt" (TN16, 5). Spracherkennung sei zudem "nicht geeignet, wenn [die] sprechende Person erkältet [ist]", vermutet ein Teilnehmer (TN8, 74). Und "die Leute, die nicht so gut deutsch sprechen können, können fast nicht dabei arbeiten, der Text von mir ist … gar kein Text" (TN9, 20), resümiert eine Probandin aus Asien. [22]

Erkennungsschwierigkeiten: Trotz der vom Hersteller versprochenen Genauigkeit der Spracherkennung ist klar: das "System erkennt nicht immer Wörter richtig" (TN4, 35). Im Umgang mit dem Programm sei dies in der Anwendung "sehr nervig" (TN1, 18) und "bringt einen auf die Palme" (TN17, 22), denn "viele Wörter [werden] erst nach mehrmaligem Nennen erkannt" (TN17, 22). Bezüglich der "vielen Worterkennungsschwierigkeiten" (TN17, 22) stellt ein Proband fest, "Wörter, die z.B. mit H oder F anfangen, kann der Computer nicht gut erkennen"

(TN7, 73). "Besonders schwierig und nervenaufreibend war das Buchstabieren" (TN16, 7), welches nötig wird, wenn einzelne Wörter nicht erkannt werden. Fülllaute oder Bestätigungslaute wie: "och, ach, ja oder ähnliches erkennt er gar nicht" (TN14, 23). [23]

Korrektur und Rechtschreibung mangelhaft: "Der Aufwand für das Korrigieren ist … sehr hoch, … nicht nur bei 'Fremdwörtern', sondern auch bei einfachen Konstellationen und Wortfolgen" (TN13, 22). Besonders schwierig sei die Korrektur der Erkennungsfehler, wenn "Wörter durch die Spracherkennung teilweise so entfremdet waren, dass der logische Zusammenhang nicht mehr ersichtlich war" (TN16, 7). So unterscheiden sich die Fehler der Spracherkennung deutlich von denen, die bei manueller Transkription getippt werden, da es ohne zweites Abhören des Textes teilweise nicht mehr möglich war, "das Geschriebene bzw. Gesprochene inhaltlich nachzuvollziehen und zu korrigieren" (TN18, 19). Hinzu kommen "Probleme mit der Interpunktion" (TN11, 17): "Wenn man einen Punkt setzt, erkennt das Programm nicht automatisch, dass es groß schreiben soll" (TN14, 22). Insgesamt sei die "Korrekturzeit fast so lange wie das Nachsprechen des Interviews" (TN5, 59). [24]

Hohe Konzentration: Die Transkription per Spracherkennung ist im Vergleich zur manuellen Transkription "interessanter, aber auch wesentlich anstrengender" (TN16, 4). Die Anstrengung bestehe darin, dass die "eigene Aussprache sehr deutlich sein musste" (TN18, 18). Hierzu sei eine "sehr hohe Konzentration" (TN19, 40) erforderlich, denn es sei "sehr schwierig das Gehörte wiederzugeben" (TN18, 18). Ein Proband stellt fest: "Ich kann nicht gleichzeitig hören und sprechen" (TN14, 21). Nachlassende Konzentration nach längerem Diktieren wirke sich schließlich auch negativ auf die Erkennungsgenauigkeit aus, "was mich am Ende ziemlich genervt hat" (TN16, 5). [25]

Installation, Performance und Bedienbarkeit: Bezüglich der Ergonomie des Programms sind sowohl zur Installation und Einrichtung als auch zur Bedienbarkeit negative Äußerungen festzuhalten. Die ersten Einstellungen und das Training sei "zeitaufwendig" (TN13, 19) und Rechner-"ressourcenvebrauchend" (TN10, 16). Das Speichern der Benutzerdaten "hat den PC für fast 5 Minuten blockiert" (TN13, 18). Ist das Programm dann installiert, muss man "sich lange einarbeiten" (TN10, 16) und "man muss darin sehr geübt sein, um einen zeitlichen Vorteil zu erlangen" (TN10, 21). Die Sprachbefehle werden als "problematisch" (TN10, 17) und "hinderlich im Fluss des Diktierens" (TN19, 40) beschrieben. Zudem beansprucht die Spracherkennung viel Rechnerkapazität. Die "Langsamkeit des Programms" (TN17, 20) störe unter anderem dann, "wenn man immer auf den PC 'warten' muss" (TN17, 20). Als besondere Anforderung an die Transkription per Spracherkennung stellt eine Person darüber hinaus fest: "Man braucht einen eigenen Raum, da man sehr laut wird beim Sprechen" (TN20, 40). [26]

Ein Teilnehmer resümiert seine Erfahrung mit Spracherkennung folgendermaßen: "An sich eine großartige Idee, aber die Durchführung ist katastrophal" (TN1, 16). [27]

3.2 Quantitative Auswertung

3.2.1 Vergleich der Transkriptionszeiten

Im quantitativen Vergleich der Transkriptionszeiten (inklusive Korrekturen) lässt sich im arithmetischen Mittel kein signifikanter Unterschied zwischen Spracherkennung und manueller Transkription feststellen (Abbildung 1). Im Schnitt benötigen die Proband(inn)en etwa 77 Minuten für die Erstellung des fertigen Textes inklusive Korrekturen. Das entspricht etwa dem 9,6-fachen der Länge der Interviews (je 8 Minuten)[6].

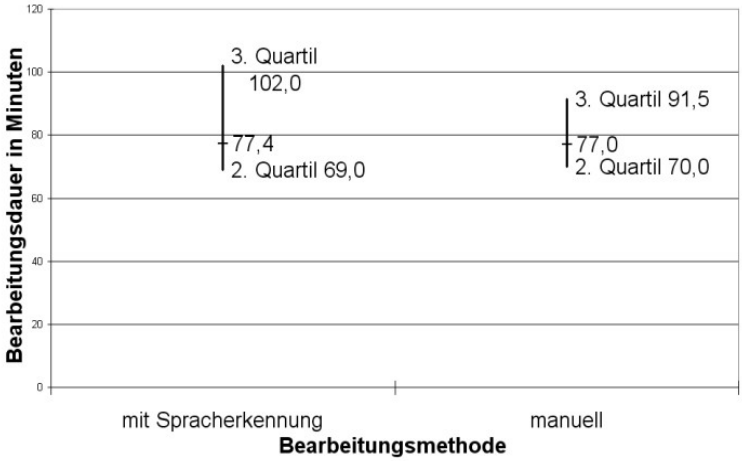

Abbildung 1: Mittlere Bearbeitungszeit nach Bearbeitungsmethode (Quartilsdarstellung)[7] [28]

Wenn man alle Einzelergebnisse in Abbildung 2 überblickt, ist zu erkennen, dass einige Proband(inn)en von Spracherkennung profitiert haben, andere mit manueller Transkription schneller waren. Die Abweichungen liegen jedoch deutlich außerhalb eines Signifikanzniveaus, sodass auf Basis dieser Daten keine Methode als grundsätzlich schneller bewertet werden kann.

6 Wir vermuten, dass gerade bei sehr kurzen Audiodateien die Transkriptionszeit länger ist, da man erst nach einer gewissen "Aufwärmphase" zu einer schnelleren Geschwindigkeit kommt.

7 In dieser Darstellung wird neben dem Mittelwert außerdem die Spanne angezeigt, innerhalb derer sich 50% der Gesamtheit der Werte befinden (zweites und drittes Quartil).

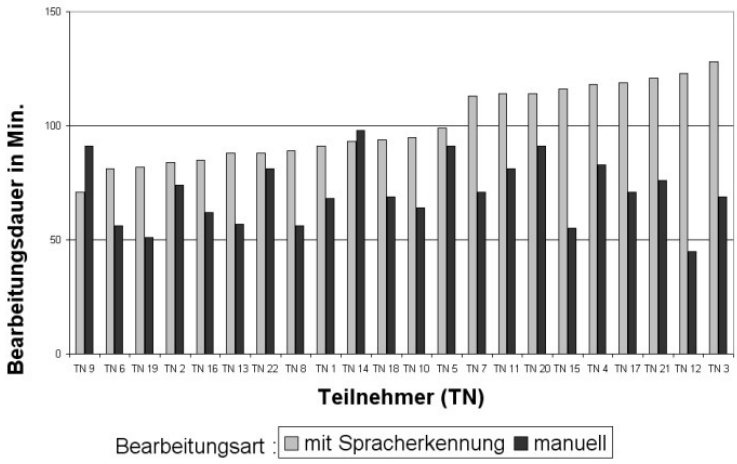

Abbildung 2: Bearbeitungszeiten der Proband(inn)en inklusive Korrekturen (Teilnehmende sind nach Tippgeschwindigkeit sortiert) [29]

3.2.2 Unterschiede zwischen langsamen und schnellen "Tippern"

Auch wenn Spracherkennung im Mittelwert aller Proband(inn)en ähnliche Bearbeitungszeiten wie die manuelle Transkription erfordert, lässt sich die Vermutung aufstellen, dass Proband(inn)en mit langsamer Schreibgeschwindigkeit eher von der Spracherkennung profitieren als Proband(inn)en mit einer schnellen Schreibgeschwindigkeit. Die Spracherkennung würde in diesem Fall gegebenenfalls die Nachteile einer langsamen Tippgeschwindigkeit kompensieren können. Zur Prüfung dieser Annahme teilten wir die Proband(inn)en in drei Gruppen entsprechend ihrer Tippgeschwindigkeit auf:

1. eine schnelle Gruppe (Tippgeschwindigkeiten von 240 bis 327 Anschläge/Min.)
2. eine mittlere Gruppe (Tippgeschwindigkeiten von 174 bis 221 Anschläge/Min.)
3. eine langsame Gruppe (Tippgeschwindigkeiten von 114 bis 173 Anschläge/Min.) [30]

In Abbildung 3 sind die Bearbeitungszeiten der Gruppen gegenübergestellt. Für jede Gruppe ist die mittlere Bearbeitungszeit für die Transkription mit Spracherkennung und manuell in Quartilsdarstellung sichtbar. Deutlich wird, dass die schnelleren Gruppen auch insgesamt weniger Zeit benötigten.

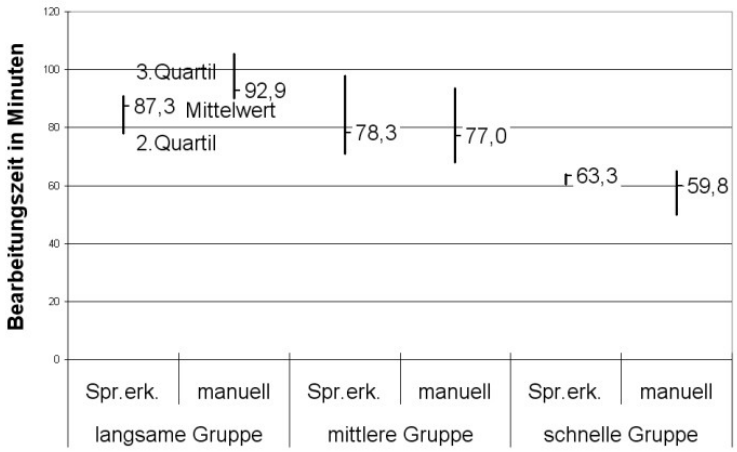

Abbildung 3: Transkriptionszeiten gruppiert nach Tippgeschwindigkeit der Proband(inn)en [31]

Der optisch erkennbare Zusammenhang zwischen Bearbeitungsdauer und Tippgeschwindigkeit lässt sich auch anhand der Daten belegen. Sowohl für manuelle Transkription ($r = -.56$, $p = .005$), als auch für Transkription mit Spracherkennung ($r = -.44$, $p = .027$) gibt es einen negativen Zusammenhang von Bearbeitungszeit und Tippgeschwindigkeit. Das heißt, tippt jemand schnell, so ist seine oder ihre Bearbeitungszeit tendenziell sowohl bei Spracherkennung als auch bei manueller Transkription kürzer. Es scheint, als sei nicht die Spracherkennung das entscheidende Merkmal für schnelles oder langsames Transkribieren, sondern die Tippgeschwindigkeit der jeweiligen Person. Das zeigt auch die Korrelation zwischen der Bearbeitungszeit manueller Transkription und Spracherkennung mit $r = .54$, $p = .007$. Tippt eine Person langsam, so wird unabhängig vom Spracherkennungseinsatz ihre Transkriptionsdauer länger sein als bei einer Person, die schnell tippt. Spracherkennung bewirkt im Transkriptionsprozess unter unseren Proband(nn)en und in diesem speziellen Setting keinen signifikanten Zeitvorteil oder -nachteil. Zugespitzt lässt sich sagen: Wer schnell tippt, transkribiert auch mit Spracherkennung schnell (aber nicht schneller), wer langsam tippt, transkribiert auch mit Spracherkennung entsprechend langsam. [32]

4. Schlussfolgerung

Spracherkennung hat eine besondere Anziehungskraft und das vor allem durch den Spaß am Effekt, Gesprochenes automatisch auf dem Bildschirm angezeigt zu bekommen. Dieser Effekt wurde im Falle unserer Untersuchung mit positiven Attributen wie "cool" und "faszinierend" verbunden. Bezüglich der Funktionalität und Bedienbarkeit erntet die Spracherkennung im Bereich wissenschaftlicher Transkription in unserer Untersuchung jedoch viele Kritikpunkte. Genannt wurden vor allem die fehlende Genauigkeit und Flexibilität und die anstrengende Nutzung, die eine sehr hohe Konzentration erfordere, wobei die hohe Konzentration bei beiden Methoden als Voraussetzung genannt wurde. [33]

Es lässt sich auf unserer Datenbasis kein Zeitvorteil durch die Nutzung von Spracherkennung nachweisen. Zudem scheint nach wie vor die individuelle Tippgeschwindigkeit ausschlaggebend für die Bearbeitungsdauer zu sein, eventuell aufgrund des hohen Korrekturaufwandes. Ein dreistündiges Interview mit Hilfe von Spracherkennung in acht Stunden zu verschriftlichen, wie in der qual-software-Mailingliste berichtet, ist unserer Ansicht nach ein Ausnahmephänomen. Bearbeitungszeiten von 1 zu 10 bei langsamem Tippen (mit bis zu 173 Anschlägen pro Minute) oder 1 zu 6 bei schnellem Tippen (mit bis zu 327 Anschlägen pro Minute) scheinen eher realistisch. [34]

Die Möglichkeiten und Auswirkungen einer längeren Nutzung von Spracherkennung wurden aufgrund unseres Forschungssettings zunächst nicht berücksichtigt. Hier ist zu erwarten, dass sich die Fehlerquote aufgrund von Trainings verbessern lässt. Trotz aller Hoffnungen bezüglich der Leistungsfähigkeit von Spracherkennungssoftware und deren damit verbundener Anziehungskraft: Die Forschungsfrage "Ist Spracherkennung dazu in der Lage, Forschenden eine schnellere Transkription ihrer Interview-Audioaufnahmen zu ermöglichen als die bisher etablierte Form der manuellen Transkription?" muss aufgrund der vorliegenden Ergebnisse zumindest für unsere Studie verneint werden. [35]

Literatur

Boersma, Paul & Weenik, David (1996). *PRAAT, a system for doing phonetics by computer, version 3.4.* Institute of Phonetic Sciences of the University of Amsterdam, Report 132, http://www.praat.org/ [Datum des Zugriffs: 30.08.2007].

Dittmar, Norbert (2004). *Transkription – Ein Leitfaden mit Aufgaben für Studenten, Forscher und Laien* (Reihe: Qualitative Sozialforschung Bd. 10). Wiesbaden: VS-Verlag.

Ehrlich, Konrad & Switalla, Bernd (1976). Transkriptionssysteme – Eine exemplarische Übersicht. *Studium Linguistik*, *2*, 78-105.

Hoffmann-Riem, Christa (1984). *Das adoptierte Kind. Familienleben mit doppelter Elternschaft*. München: Fink.

Kowal, Sabine & O'Connell, Daniel (1995). Transcription systems for spoken discourse. In Jef Verschueren, Jan-Ola Oestmann & Jan Blommaert (Hrsg.), *Handbook of pragmatics* (S.646-656). Amsterdam: John Benhamins.

Kuckartz, Udo (2007). *Einführung in die computergestützte Analyse qualitativer Daten*. Wiesbaden: VS-Verlag.

Kuckartz, Udo; Dresing, Thorsten; Rädiker, Stefan & Stefer, Claus (2007). *Qualitative Evaluation – Der Einstieg in die Praxis*. Wiesbaden: VS-Verlag.

Mayring, Philipp (1983). *Qualitative Inhaltsanalyse*. Weinheim: Psychologie Verlagsunion.

Mayring, Philipp (2000). Qualitative Inhaltsanalyse [28 Absätze]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(2), Art. 20, http://www.qualitative-research.net/fqs-texte/2-00/2-00mayring-d.htm [Datum des Zugriffs: 30.08.2007].

Selting, Margret; Auer, Peter; Barden, Birgit; Bergmann, Jörg; Couper-Kuhlen, Elizabeth; Günthner, Susanne; Meier, Christoph; Quasthoff, Uta; Schlobinski, Peter & Uhmann, Susanne (1998). Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, *173*, 91-122.

Zu den Autoren und zur Autorin

Dr. *Thorsten DRESING* ist wissenschaftlicher Mitarbeiter am Fachbereich Erziehungswissenschaften der Philipps-Universität Marburg. Er leitet zudem als Geschäftsführer der dr. dresing & pehl GmbH die Entwicklung des Portals audiotranskription.de und Fortbildungen zur computergestützten qualitativen Datenanalyse.

Dipl. Päd. *Thorsten PEHL* leitet als Geschäftsführer der dr. dresing & pehl GmbH die Entwicklung und den Vertrieb des Portals audiotranskription.de

Claudia LOMBARDO studiert Erziehungswissenschaften an der Philipps-Universität Marburg (Diplom).

Kontakt:

dr. dresing & pehl GmbH

Uferstrasse 3
D-35037 Marburg

Tel.: 06421-933426
Fax: 06421-983893

E-Mail: info@audiotranskription.de
URL: http://www.audiotranskription.de/

Zitation

Dresing, Thorsten; Pehl, Thorsten & Lombardo, Claudia (2008). Schnellere Transkription durch Spracherkennung? [35 Absätze]. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *9*(2), Art. 17, http://nbn-resolving.de/urn:nbn:de:0114-fqs0802174.