universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Skype Translator.
Reaktionen des maschinellen Dolmetschsystems auf problemauslösende Faktoren in der Spracherkennung im Englischen

verfasst von / submitted by

Elisabeth Barfuß, BA

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2017 / Vienna 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 065 342 351
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Dolmetschen
Betreut von / Supervisor:	ao. Univ.-Prof. Mag. Dr. Franz Pöchhacker

Danksagung

Hätte die Entstehung dieser Arbeit allein an mir gelegen, wäre sie wohl noch lange nicht gebunden. Daher ist es Zeit, mehr als nur ein paar abgedroschene Dankesworte niederzuschreiben.

Zuallererst danke ich meinem Betreuer ao. Univ-Prof. Mag. Dr. Franz Pöchhacker, der mir immerzu mit wertvollen Tipps, Enthusiasmus und Ideenreichtum beiseite stand und meine Arbeit in rekordverdächtiger Schnelligkeit bis ins kleinste Detail durcharbeitete. Verblüffenderweise schaffen Sie es, die hohe Qualität Ihrer Betreuung trotz der gewaltigen Flut an Studierenden und Forschungsprojekten stets aufrechtzuerhalten.

Daneben, so gut wie an erster Stelle, stehen meine Familie und meine Freunde. Ihr habt mir Schreibasyl gewährt, Motivationsreden geschwungen, die Struktur meiner Arbeit über den Haufen und diese direkt ins Feuer geworfen, um mir dann wie der Phönix aus der Asche zur glänzenden Neuerstehung zu verhelfen; meine Arbeit so eifrig gelesen, kommentiert und verbessert wie keiner sonst; mich abgelenkt, wenn es nicht mehr ging; mit mir frühmorgens mit Kaffee in der Hand die neuesten Hindernisse zermalmt; mir eure Stimmen und euer Wissen geschenkt und eure Kontakte genutzt, damit ich auch deren Stimmbänder ausbeuten kann – wie nur ihr das tut, habt ihr eure Energie eingesetzt, um mir zu diesem Abschluss zu verhelfen. Ihr seid wondrous!

Ein großes Danke gebührt auch „meinen“ SprecherInnen, die mir ihre Stimmen liehen und deren Lebenserfahrungen und unterschiedlichen Hintergründe die gewisse Würze ist, die diese Arbeit einzigartig macht.

Zu allerletzt: Alex, mit dir würde ich jederzeit wieder ein Forschungsprojekt entwickeln.

„Nun lehne ich mich nicht zurück, sondern dem, was mich erwartet, entgegen.“

(Martin Barfuß-Benavente)

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

0 Einleitung

„Language is a wild beast. It changes all the time, it comes in many flavors and varieties […]“ (Skype 2014). Dieses Zitat beschreibt sehr gut, dass Sprache ein sich ständig veränderndes Phänomen mit vielen verschiedenen Facetten und schwer zu bändigen ist. Obwohl auch die Technologie sich rasant weiterentwickelt, scheint es wohl immer so, als wäre ihr die Komplexität der Sprache immer einen Schritt voraus (vgl. Och & Ney 2004: 417).

Weltweite Kommunikation ist ein essenzielles Gut in einer Gesellschaft, deren Schnelllebigkeit und rasante Entwicklung kein Ende zu nehmen scheint. Und so macht die Diskussion, ob Maschinen bald einen Großteil der Arbeitsplätze ersetzen können, auch vor dem Translationsbereich nicht Halt. Dementsprechend sind maschinelle Dolmetschungen keine fantastische Vorstellung mehr aus Star Trek.

Eines der frei verfügbaren und bereits vielfältig eingesetzten Dolmetsch-Programme ist *Skype Translator*. Die Frage, die sich dabei ganz automatisch zu ergeben scheint, lautet: „Kann man Maschinen wirklich trauen?“ Die Verlässlichkeit von maschinellen Dolmetschsystemen am Beispiel von *Skype Translator* ist genau jener Forschungsgegenstand, mit dem sich diese Arbeit beschäftigt. Es wäre jedoch naiv zu glauben, dies mit Ja oder Nein beantworten zu können, da auch diese Frage, wie alles in der Wissenschaft, aus vielerlei Perspektiven betrachtet werden kann.

So muss hierbei im Hinterkopf behalten werden, dass die Entwicklung des *Skype Translator* noch lange kein Ende erreicht hat und sich das System erst in den Kinderschuhen seines Daseins befindet.

Skype Translator beruht auf einer Wort-für-Wort-Basis, die von Vergleichen mit anderen Texten abhängig ist und durch die Nutzung von *Big Data* seine statistisch wahrscheinlichsten Ergebnisse berechnet, während es ständig mithilfe neuer Daten wächst. Beim Verwenden von *Skype Translator* fiel daher in der Praxis auf, dass trotz der gut programmierten Software immer wieder inhaltliche Schwierigkeiten in den Dolmetschleistungen auftreten.

Während bei HumandolmetscherInnen bestimmte sprachliche Elemente als besonders herausfordernd in der Dolmetschleistung gelten, soll mit dieser Arbeit an der Oberfläche des Programms gekratzt werden, um herauszufinden, wie das maschinelle Dolmetschsystem *Skype Translator* auf unterschiedliche problemauslösende Elemente

reagiert. Damit überhaupt etwas gedolmetscht werden kann, muss es zunächst verstanden werden. Deswegen liegt der Fokus dieser Arbeit auf der Phase der Spracherkennung und kulminiert in der Beantwortung der naheliegenden Forschungsfrage: Wie reagiert *Skype Translator* in der Spracherkennung auf problemauslösende Faktoren?

Im Rahmen einer Untersuchung wird ein Ausgangstext von 32 SprecherInnen aufgenommen und von *Skype Translator* gedolmetscht. Die im Text befindlichen problemauslösenden Elemente werden auf Ebene der Spracherkennung miteinander verglichen und analysiert. Dadurch soll festgestellt werden können, auf welche dieser problemauslösenden Faktoren *Skype Translator* wie reagiert und ob das Programm daraus lernt.

Im ersten Kapitel werden die Grundlagen von *Skype Translator* und dem dahinter befindlichen System erläutert. Das zweite Kapitel widmet sich dem Bereich der Spracherkennung und dem Hörverständnis sowie den damit verbundenen Problemauslösern. Im Methodenkapitel wird das Forschungsdesign im Detail präsentiert und die Problemauslöser im Ausgangstext sowie die SprecherInnen, deren Aufnahmen für den Vergleich der Spracherkennungsphase von *Skype Translator* herangezogen wurden, präsentiert. Anschließend werden die Ergebnisse diskutiert und Schlussfolgerungen daraus gezogen.

Diese Arbeit wurde in Zusammenarbeit mit Alexander Wonisch geschrieben, der sich in seiner Parallelarbeit „Die Funktionsweise des Programms und eine Analyse der Qualität der Dolmetschleisung in der Sprachrichtung Englisch – Deutsch“ auf die Phase der maschinellen Übersetzung konzentriert und dort im theoretischen Teil sehr viele weiterführende Grundsatzdiskussionen ausführlich behandelt.

1 Skype Translator

Skype (vgl. Skype 2016), ein 2003 veröffentlichtes, kostenloses Kommunikations-Programm, gehört seit 2011 zur *Microsoft* Corporation. Mittels Sprach- und Videoanrufen, Sofortnachrichten wie auch Gruppenkonferenzen über das Internet soll es vor allem Privatpersonen die Möglichkeit bieten, überall auf der Welt miteinander zu kommunizieren. Es gestattet außerdem, für relativ geringe Gebühren vom Skype-Programm direkt Mobil- oder Festnetztelefone anzurufen und Nachrichten zu schicken und empfangen.

Als eine der neuesten Erweiterungen von Skype wurde *Skype Translator* Mitte Dezember 2014 von Gurdeep Pall in einer der ersten Phasen des *Skype Translator* Preview Programms eingeführt. Diese „Preview", also die vorläufige Variante des *Skype Translator*s, war aufgrund der Möglichkeit des automatisierten Lernens essenziell für die Entwicklung des maschinellen Dolmetschsystems.

BenutzerInnen konnten sich in dieser Phase online für die damals noch separat verfügbare Applikation registrieren, um in eine Gruppe aufgenommen zu werden, die berechtigt war, Skype mit dem Betriebssystem Windows 8.1. oder Windows 10 zu verwenden.

Skype Translator war zu diesem Zeitpunkt als maschinelles Dolmetschprogramm nur für die Sprachen Englisch und Spanisch sowie für Übersetzungen von Instant Messages in über 40 Sprachen verfügbar.

Während jedes Skype-Gesprächs mit dem *Dolmetschsystem* (dieser Begriff wird nachstehend näher beschrieben) wird automatisch auch eine Transkription zur besseren Nutzung und zum erleichterten Verständnis angefertigt und bietet damit auch die Möglichkeit, Fehler schriftlich auszubessern. Das Programm arbeitet mit Sinnabschnitten, so genannten Segmentierungen, die nach einer 2-sekündigen Latenzzeit mit einer maximalen Länge von 30 Sekunden wiedergegeben werden. Laut *Microsoft* (2014: 5) liefern Sätze mit einer Wortanzahl zwischen 7 und 15 Wörtern die besten Ergebnisse in der Dolmetschleistung, da genug aber nicht zu viele Schlüsselwörter miteinander in Beziehung gesetzt werden können.

Selbststudien für den empirischen Teil dieser Arbeit zeigten, dass *Skype Translator* nach sechs Minuten (hierbei in Monologform) mit der Dolmetschleistung

aufhörte und nur nach Beenden und neuerlichem Beginnen des Gespräches weitergeführt werden konnte. In der Parallelarbeit von Wonisch (2017) werden die Benutzeroberfläche sowie Besonderheiten zur Anwendung eingehender betrachtet.

Anfang April 2015 wurden die Sprachen Chinesisch (Mandarin) und Italienisch als weitere Sprachen hinzugefügt (vgl. Skype 2015), kurz darauf folgten Französisch und Deutsch Mitte Juni (vgl. Skype 2015a).

Außerdem kamen weitere Anwendungsmöglichkeiten hinzu: Textnachrichten konnten von da an auch in jeder beliebigen Sprache wiedergegeben werden. Des Weiteren wurde eine automatische Lautstärkeregelung eingeführt, also die Möglichkeit, die Dolmetschung in voller Lautstärke zu hören, selbst wenn die GesprächspartnerIn schon weiterspricht und diese/dieser somit leiser zu hören ist, sowie die Möglichkeit, die Stimme des Dolmetschsystems abzudrehen, um nur die geschriebene „Dolmetschung" zu lesen. Außerdem wurden für die Sofortnachrichten weitere Sprachen wie Serbisch, Bosnisch, Kroatisch, Maya und Otomi hinzugefügt.

Mitte Oktober 2015 kam brasilianisches Portugiesisch hinzu (vgl. Skype 2015b) und Anfang Dezember 2015 war *Skype Translator* sogar ohne Anmeldung für Windows 8.1. beziehungsweise Windows 10 für PC oder Tablet online verfügbar (vgl. Skype 2015c).

Ab Mitte Jänner 2016 wurde *Skype Translator* für sämtliche Betriebssysteme von Windows verfügbar gemacht (vgl. Skype 2015d). Seit Anfang März wird *Skype Translator* auch für modernes Standard Arabisch (MSA), also Arabisch, das besonders im Mittleren Osten und Nordafrika gesprochen wird, angeboten (vgl. Skype 2016a). Im Oktober 2016 wurde Russisch als mögliche „Arbeitssprache" hinzugefügt (vgl. Skype 2016b).

Seit April 2017 (vgl. Skype 2017) ist *Skype Translator* mit Japanisch also in 10 Sprachen für Dolmetschungen verfügbar und in über 50 Sprachen für Sofortnachrichten. *Skype Translator* wird vor allem damit beworben, Sprachbarrieren zu überbrücken und alltägliche Situationen stark zu vereinfachen. Insgesamt ist laut Skype die beliebteste Sprachkombination derzeit Englisch und Französisch – und die beliebteste Verbindung von Deutschland nach Ghana (vgl. Skype 2016c).

Inzwischen ist *Skype Translator* auch als Programm für Drittanbieter verfügbar und wird beispielsweise bereits vom Mobilfunkanbieter *Tele2* für eigene Kundendienstleistungen genutzt (vgl. Microsoft 2016).
Um die Qualität des Gespräches evaluieren zu können, wird ähnlich wie bei normalen Skype-Gesprächen um ein Feedback von den KundInnen gebeten. Im Fall von *Skype Translator* wird mithilfe eines 5-Sterne Bewertungssystems danach gefragt, wie gut die „Qualität der Übersetzung", die „Übersetzungsgeschwindigkeit" sowie die „allgemeine Qualität des Anrufs" waren. Das Spektrum der 5 Sterne reicht von „perfekt, deutlich, keine Probleme" bis hin zu „Probleme waren so schlimm, dass der Anruf nicht möglich war" (vgl. Skype 2014).

Skype Translator wirbt damit, bei nahezu allen Gelegenheiten über sprachliche Barrieren helfen zu können. Diese Mentalität wird auch in Werbevideos propagiert, die sich auf folgende Erlebnisberichte stützen sollen: Eine multikulturelle Familie kann wöchentlich im Kontakt bleiben. InhaberInnen eines kleinen Unternehmens können sich mit ihren besten LieferantInnen durch Sofortnachrichten austauschen. Ein/e PhD-StudentIn, der/die seine/ihre Forschung verbessern will, kann sich mühelos mit ExpertInnen aus aller Welt beratschlagen. Non-Profit ArbeiterInnen sammeln Spendengelder und australische Weltreisende erleichtert sich ihren Weg quer über die Kontinente, indem sie sich Schlüsselphrasen übersetzen lassen (vgl. Skype 2015e).

Spätestens das Beispiel des/der PhD-StudentIn wirft die Frage auf, ob *Skype Translator* tatsächlich so umfassend, genau und verlässlich arbeitet, um etwas Wesentliches wie die eigene wissenschaftliche Forschung darauf bauen zu können.

Skype teilt die Kommunikationsarten, die in *Skype Translator* verwendet werden, in vier Kommunikationsmodi ein. Erstens in „remote conversation", also Gespräche mit zwei Personen, zwei Sprachen und zwei Geräten. Beispielhaft wäre hierzu ein Skype-Gespräch mit einem Freund anderer Muttersprache zu nennen. Zweitens ein „brief exchange", also ein kurzer Austausch. Hierbei sind zwei Personen involviert, jedoch nur ein Gerät verwendet Skype. Ein Szenario, das zu dieser Kategorie zählt, wäre beispielsweise eines, in dem jemand Essen bestellt. Drittens „extended social conversations", also Gruppengespräche, in denen viele Personen mit vielen Geräten und unterschiedlichen Spachen teilnehmen. Dies könnte ein Szenario sein, in dem eine Familie gemeinsam per Videotelefonie zu Abend isst. Viertens ein

unidirektionales Gespräch, also ein Gespräch, in dem eine Person zu vielen gleichzeitig spricht. Das Szenario könnte hier ein Klassenzimmer oder eine Vorlesung an der Uni sein (vgl. Wendt 2016: 11).

1.1 Maschinelles Dolmetschen

Maschinelles Dolmetschen (MD) ist der Prozess, bei dem gesprochene Äußerungen einer Sprache verwendet werden, um diese mithilfe einer Maschine in einer anderen Sprache wiederzugeben. Die computerwissenschaftliche Literatur und Forschung zu diesem Thema ist hauptsächlich in Englisch verfasst, weshalb in der Literatur und in dieser Arbeit entweder von Maschinellem Dolmetschen oder von *speech-to-speech translation* gesprochen wird, wie es auch von Kitano (1994) und Wahlster (2000) gehandhabt wird.

Der Vorgang ist im Fall von *Skype Translator* in drei beziehungsweise in vier Phasen unterteilt: In der ersten Phase erfolgt die automatische Spracherkennung. Als weiterer Schritt, und dieser wird hier als Teilschritt der Spracherkennung gesehen, werden Sprachkorrekturen (Speech Correction) (vgl. Skype 2014a) durchgeführt, um sprachliche Verzögerungslaute auszublenden (näher behandelt in Abschnitt 2.1).

In der nächsten Phase werden mithilfe einer maschinellen Übersetzung (Machine Translation – MT) die in Text umgewandelten Daten der Spracherkennung in eine andere Sprache übersetzt, um dann in der letzten Phase mithilfe von Sprachsynthese (Text-to-speech synthesis) besagte Übersetzung von einer Maschine akustisch wiederzugeben (vgl. Waibel & Fügen 2008: 71). Bei *Skype Translator* findet dasselbe maschinelle Übersetzungssystem Verwendung, das auch von *Microsoft* für den Bing Translator angewandt wird und seit 2001 für Kundensupportdienste und seit 2007 für die Öffentlichkeit verfügbar ist. Das System bietet die Möglichkeit, durch Humantranslationen hinzuzufügen, zu bewerten und zu bestätigen. Auch die dadurch gewonnenen Daten werden für das automatisierte Lernen verwendet (vgl. Wendt 2010).

Beim Maschinellen Übersetzen wird in regelbasierte und korpusbasierte Ansätze unterschieden, also Systeme, die sich entweder auf semantische Repräsentation der Wörter beziehen, oder Systeme, die ihre Lösungen rein auf Paralleltexte bauen. Dies wurde nur der Vollständigkeit halber hier angebracht und wird bezüglich genauerer Funktionsweisen in der Parallelarbeit von Wonisch (2017: 1f) reichlich erklärt. *Skype*

Translator verwendet ein System, das aus einer Mischform besteht und statistikbasiert in Paralleltexten Entsprechungen sucht, welche dann in einem weiteren Schritt mit einer syntaktischen Komponente nachbearbeitet werden (vgl. Wendt 2010: 1; Och & Ney 2004: 417).

Abschließend wird im Zuge der Sprachsynthese der Zieltext der MÜ wieder in mündliche Form umgewandelt. Dabei werden einzelne Segmente der Syntax herangezogen, um die Intonation der künstlich erzeugten Sprache möglichst menschlich klingen zu lassen (vgl. Härtel 2016: 94; Black & Lenzo 2004: 761). Auch dieses Thema wird von Wonisch (2017) detaillierter behandelt.

Diese Phasen werden durch ein zusammenhaltendes System miteinander verknüpft, welches in dieser Arbeit automatisches Dolmetschsystem genannt wird. Dadurch, dass jede einzelne dieser vom System zusammengeführten Phasen eigene Prozesse mit individuellen Herausforderungen sind, kann davon ausgegangen werden, dass die maschinelle Dolmetschleistung sehr fehleranfällig ist.

Hierbei soll erwähnt werden, dass maschinelles Dolmetschen entstand, als die technologische Entwicklung der Spracherkennung sowie der statistischen maschinellen Übersetzung sich unabhängig voneinander effektiv entwickelt hatten und daraufhin erst die Idee eines „automatic spoken language translation"-Systems (SLT) also einem Programm, welches die gesprochene Sprache automatisch übersetzen könnte, aufkam. Eines der Vorreiterprojekte ist bekannt unter dem Namen „Verbmobil" (vgl. Wahlster 2000). Dieses Dolmetschsystem spezialisierte sich auf den Bereich der Reiseplanung und Terminvereinbarung und sollte darin Prozesse erleichtern. Darüber hinaus gab es auch andere Projekte wie C-STAR oder NESPOLE!, welche sich auf Online-Anwendungen spezialisierten. Auch das so genannte DARPA TransTac schuf eine weitere Ebnung für fruchtbaren Boden in der Mensch-zu-Mensch Kommunikation (vgl. Fünfer 2013: 20). In der Forschung der SLT wurde vor allem in den letzten zwei Jahrzehnten die steigende Spontaneität der Sprache unter die Lupe genommen und die damit einhergehende Begleiterscheinung der Komplexität betrachtet. VERBMOBIL und NESPOLE! verwendeten fachbereichsspezifische Grammatik in der Spracherkennung und interlinguale Zugänge für die Übersetzung. Die Schwierigkeit der Limitierung auf bestimmte Bereiche lässt Spontaneität tendenziell nicht zu (vgl. Kumar et al. 2014: 3231).

Eine weitere Grundsatzfrage ist, ob man bei maschinellem Dolmetschen überhaupt von einer Dolmetschung sprechen kann. Nach der Definition von Kade (1968: 35) besteht der wesentliche Unterschied zwischen einer Übersetzung und einer Dolmetschung darin, dass der Ausgangstext einmalig vorliegt und nicht wie die Übersetzung permanent und beliebig oft wiederholt werden kann. Strenggenommen wird beim maschinellen Dolmetschen das Gesprochene sofort in den Computer aufgenommen und gespeichert. Ob diese Daten für längere Zeit gespeichert werden und ein zweites Mal aufrufbar wären oder nach der Umsetzung gleich wieder gelöscht werden, kann hier nicht beantwortet werden. Im Endeffekt und für diese Arbeit ist lediglich relevant, dass der Prozess der Spracherkennung und -verarbeitung, der Übersetzung sowie der Sprachsynthese hier als Dolmetschung bezeichnet wird (vgl. Härtel 2016: 68-69).

1.2 Neuronale Netzwerke und *Big Data*

Maschinelle Dolmetschsysteme und die damit verbundenen Prozesse basieren auf der Sprachtechnologie der angewandten Computerlinguistik und berühren somit auch das Feld der Künstlichen Intelligenz. Um adäquate Dolmetschungen erstellen zu können, müssen die Systeme über Datenbanken auf eine enorme Menge von Informationen und Weltwissen zurückgreifen können. Dieser Datenfluss nimmt heutzutage rasant zu, zumal das Interesse an online zugänglichen Informationen im öffentlichen, privaten und kommerziellen Bereich stetig wächst. Diese Tatsache zeigt sich unter anderem an der steigenden Frequenz bei der Nutzung sozialer Netzwerke und diverser Smart-Technologien und führt dazu, dass derzeit alle 18 Monate von einer Verdopplung der nutzbaren Datenmengen gesprochen werden kann (vgl. Bughin et al. 2010). Dieser Informationstrend ist allgemein unter dem Begriff „Big Data“ bekannt. Technologie wird eingesetzt, um Informationen festzuhalten, zu analysieren und zugänglich und preiswert zu machen.

Die Verarbeitung der *Big Data* erfolgt über so genannte Artificial Neural Networks (ANNs). Diese künstlichen Netzwerke sind mathematische Modelle von niederschwelligen Strömungen im menschlichen Gehirn und sind schon seit den 1950er Jahren ein Begriff. Dabei dienen künstliche Neuronen, so genannte Perzeptoren, welche in den 1960er Jahren von Frank Rosenblatt entwickelt wurden und als künstliches neuronales Netz fungieren, dazu, Eingabevektoren (Input) in Ausgabevektoren (Output)

umzuwandeln. Mehrere Inputs mit unterschiedlichen Gewichtungen werden im Zuge dessen zu einem Output kanalisiert, dann mit anderen, schon bestehenden Daten verglichen, ausgewertet und schließlich zur Erkennung von Phonemen herangezogen. Als erweiterte Version von Neural Networks gibt es auch noch Deep Neural Networks, die vielschichtiger aufgebaut sind und somit noch bessere Ergebnisse aufweisen (vgl. Nielsen 2015).

Neural Networks werden in vielen unterschiedlichen Disziplinen angewandt, so zum Beispiel auch bei der Bilderkennung. Für die wahrscheinlichkeitsstatistische Zuordnung von Prozessen in der Spracherkennung, beziehungsweise die Auswertung auditiver Daten, wird zumeist das bereits erwähnte Hidden Markov Model (ANN-HMM) verwendet. Das von Skype genannte Modell ist unter Context-Dependent Deep Neural Network Hidden Markov Model (CD-DNN-HMM) bekannt und gilt als Basis von *Skype Translator*. Das Team von *Microsoft* Research machte nach eigenen Angaben 2013 einen gewaltigen Fortschritt, indem es die Vielschichtigkeit seines Systems noch vertiefte und dieses somit die Leistung konventioneller Systeme überragte, was zu sehr guten Ergebnissen der Word Error Rates (WER-Wert) führte (vgl. Dahl et al. 2012: 30).

Das Konzept der Word Error Rate basiert auf der Differenz zwischen dem, was erkannt werden sollte, und dem tatsächlichen erkannten Output. Mithilfe eines Transkripts wird somit evaluiert, inwieweit sich diese beiden Darstellungen überschneiden. Die sogenannte Distanz ergibt sich aus den Wörtern, die ersetzt, hinzugefügt oder herausgelöscht wurden. Für gewöhnlich werden hierbei alle drei Komponenten gleichwertig betrachtet (vgl. WER in 2.1). Die Methode dieser Evaluierung der Spracherkennung wird in der Literatur viel diskutiert, da sich die WER-Messung hauptsächlich auf die Worterkennungsgenauigkeit, nicht aber auf das Verständnis bezieht (vgl. Wang et al. 2003: 577f).

1.3 Automatisiertes Lernen

Automatisiertes Lernen ist essenziell für die Weiterentwicklung von Dolmetschsystemen wie dem von *Skype Translator*. Mithilfe der richtigen Software und großen Datensätzen als Grundbausteine der Trainingsdaten kann die Qualität weiterentwickelt werden. Je größer die Datenmengen und Sprachmaterialien, desto

adäquater können die einzelnen Prozesse durchgeführt werden. Darüber hinaus basiert die Grundlage dieses Systems nicht auf konventioneller Programmierung, sondern darauf, dass ProgrammiererInnen dem Computer kleine Aufgaben geben, die dieser leicht und klar definiert lösen kann und so dazulernt. Neural Networks erhalten nämlich keine Anweisungen, sondern lernen das Lösen von Problemen anhand von Datenmengen. Automatisiertes Lernen klingt vielversprechend, war aber für lange Zeit eine schwer überwindbare Hürde, da man nicht wusste, wie man die Netzwerke dazu programmieren konnte, über konventionelle Ansätze hinauszugehen und sich selbst weiterzuentwickeln (vgl. Microsoft 2014: 7).

Automatisiertes Lernen, welches auf den DNNs basiert, erfolgt in jedem der Teilbereiche der Maschinellen Dolmetschung, also sowohl in der Spracherkennung (ASR), der automatischen Übersetzung (MÜ) und der Sprachsynthese (STT) als auch generell im Rahmen der Verbesserung im Hinblick auf die Benutzerfreundlichkeit des Programms. Durch das Lernen in der Trainingsphase und mithilfe so genannter Lernprotokolle kann die Software vermeintlich auch verschiedene Themen, Akzente und Sprachvariationen der *Skype Translator*-NutzerInnen besser zuteilen sowie Störelemente wie Füllwörter und Unterbrechungen isolieren (vgl. Skype 2014).

Die Quellen der Trainingsdaten kommen von übersetzten Websites, Videos mit Untertitelung und zuvor übersetzten und transkribierten Konversationen. *Skype Translator* nimmt Gespräche auf, um Transkripte anzufertigen und damit das System zu trainieren, zu verbessern, zu ergänzen (vgl. Microsoft 2014: 21).

Hierbei ist es als besondere Herausforderung anzusehen, umgangssprachliche und mündliche Sprache dolmetschen zu können, welche von *Skype Translator* durch Übersetzungen von kolloquialen Ausdrücken und Schreibwendungen aus sozialen Netzwerken wie zum Beispiel *Facebook* ermöglicht wird. Diese Form der Sprache wird als eine Art Mischung aus mündlicher und schriftlicher Sprache gesehen und ist damit als Basis zum Übersetzen und Dolmetschen mündlicher Texte geeignet. Viele Menschen spendeten auch Daten in Form von Sprachproben, indem sie beispielsweise dem Aufnehmen ihrer Skype-Konversationen zustimmten, damit diese für statistische Modelle des Trainingsmaterials verwendet werden konnten. Damit wurden Spracherkennung und maschinelle Übersetzung modelliert beziehungsweise für die jeweils andere Sprache ausgebaut. Die NutzerInnen von *Skype Translator* in der

Preview Phase wurden ganz explizit zu Beginn der Konversation darüber benachrichtigt, dass ihr Gespräch aufgenommen und dazu verwendet würde, die Qualität von *Microsofts* Systemen zu verbessern (vgl. Microsoft 2014: 22).

Sowohl in der Spracherkennung als auch in der maschinellen Übersetzung werden so genannte Hidden-Markov-Modelle (HMM) und Markov-Ketten genutzt, um mittels stochastischer Prozesse das Wissen, das von vorangegangenen oder aktuellen Zuständen erlangt wurde, zur Berechnung wahrscheinlicher späterer Zustände zu verarbeiten. Mit diesem System kann also die Verkettung von den zu erwartenden Abfolgen von z.B. Phonemen zu einer Silbe, zu einem Wort und diese wiederum zu einem Satz aufgebaut werden (vgl. Härtel 2016: 69-70).

1.4 Zusammenfassung

Viele kleine Teilprozesse sind notwendig, um *Skype Translator* funktionstüchtig zu machen. Die maschinelle Dolmetschung funktioniert mithilfe vier großer Phasen: der Spracherkennung, der Analyse der Spracherkennung und Verschriftlichung dieser, der maschinellen Übersetzung dieser Verschriftlichung und zuletzt der künstlichen vokalen Wiedergabe dieser Übersetzung. Die Daten für die Spracherkennung und die maschinelle Übersetzung stammen aus einem riesigen Datenpool, der *Big Data* genannt wird und vom Software-Giganten *Microsoft* durch NutzerInnen aufgezeichnet wurde und wird. Mithilfe dieser Daten kann das System automatisiert lernen, wie wahrscheinlich es ist, ob eine Übersetzung gut in den durch diverse Schlüsselwörter bestimmten Kontext passt. Je mehr Daten gespeichert und analysiert werden, desto wahrscheinlicher ist es, dass Skype nach der Evaluierungsmethode der Word Error Rate passendere Ergebnisse erreicht.

2 Hörverständnis und Spracherkennung beim Dolmetschen

Wenn davon ausgegangen wird, dass maschinelle Dolmetschungen ähnlich wie bei Humandolmetschungen (siehe Abschnitt 1.1) aus ebendiesen Prozessen des Hörverständnisses und der Analyse sowie der Dolmetschausgabe (vgl. Gile 1995) und der Gedächtnisleistung bestehen, dann beschäftigt sich diese Arbeit ausschließlich mit der ersten Phase, der Spracherkennung und des Hörverständnisses.

Verständnis erfordert nicht allein das Erkennen von Wörtern, sondern bedarf auch einer situationalen und kontextuellen Analyse der Informationen (vgl. Pöchhacker 2004: 118f). Die Literatur betrachtet es von vielerlei Perspektiven. In der Kognitionswissenschaft sowie der Computerlinguistik und ähnlichen Forschungsbereichen beginnt der erste Schritt von Verständnis auf der Ebene der Spracherkennung, also der Wahrnehmung von Schallwellen, die als Phoneme erkannt und anschließend zu Wörtern zusammengefügt werden, um dann eine eindeutige lexikalische Bedeutung zu erhalten, sowie in Verbindung miteinander einen Satzbau ergeben (vgl. Härtel 2016: 69-70)

Der zweite Schritt, auf den vor allem auch in der Dolmetschwissenschaft Augenmerk gelegt wird, erforscht das „Sinnverständnis“, also die analysierten Informationen, die den Diskurs bestimmen, welcher sich aus dem kontextuellen, situationellen und enzyklopädischen Wissen ergibt (vgl. Pöchhacker 2004: 119). Dies ist unbestreitbar für die Dolmetschwissenschaft relevant, da sich Dolmetschleistungen auch oft in fachlichen Kontexten abspielen, wo die Expertise, das Hintergrundwissen und die Analysefähigkeit der DolmetscherInnen unabdingbar für eine gute Leistung sind. Nachvollziehbar wird dies angesichts der häufig vorliegenden Ausgangslage, dass DolmetscherInnen nicht zwangsläufig über dieselbe Expertise wie die RednerInnen verfügen, (vgl. Gile 2004: 86) weswegen sie sich deswegen mithilfe des Kontextes und ihrem Hintergrundwissen zu helfen wissen müssen. Nach Gile muss dieses Sinnverständnis so weit gehen, dass die Intention der SprecherInnen antizipiert werden kann, bevor die gesamte Sinneinheit vorgetragen wurde.

Gleichzeitig sollte angemerkt werden, dass Verständnis niemals als absolut anzusehen ist, sondern stets ein subjektives, sehr komplexes Konstrukt ist. Gemäß Chernov (1979: 104f) ist Verständnis im Zusammenhang mit Humandolmetschung eine dynamische Analyse von Sinnstrukturen. Mithilfe dieses Prozesses können

Informationen einerseits in wichtig und unwichtig separiert und andererseits vorhergesehen werden. Diesen letzten Bereich des Verständnisses nennt man in der Dolmetschwissenschaft auch Antizipationsfähigkeit (vgl. Chernov 1979).

Gleichzeitig kann nicht unmittelbar festgestellt werden, ob der „verstandene" Sinn der DolmetscherInnen sich mit dem „beabsichtigten Sinn" der RednerInnen deckt (vgl. Gile 2009: 162). Nach Seleskovitch und Lederer (1989) ist dies die „interpretative Natur" des Dolmetschens, bei der das Verständnis und die Analyse des Ausgangstextes vor allem von der persönlichen Interpretation der DolmetscherInnen in der unmittelbaren Situation abhängen. Dies bestätigt auch den gewaltigen Unterschied zwischen HumandolmetscherInnen, ÜbersetzerInnen und maschinellen Dolmetschprogrammen, welche rein auf die linguistische Analyse aufgebaut sind.

Gile postuliert (2009: 162), dass maschinelle Dolmetschprogramme einen Mangel an Fähigkeit haben, linguistische Zeichen miteinander sowie diese mit Weltwissen zu verknüpfen, wie auch Ambiguität zu entkräften. Des Weiteren könnten auch andere Probleme wie inhaltliche Schwierigkeiten, Abweichungen der Standardsprache und Logik, oder auch senderbezogene Schwierigkeiten auftreten. Beispielsweise können bei Betonungen von Wörtern, welche nicht nur innerhalb von unterschiedlichen SprecherInnen, sondern bei SprecherInnen individuell unterschiedlich ausfallen, Probleme auftreten. Es ist nachvollziehbar, dass für Maschinen demnach die Erkennung von Wörtern in Sonogrammen, durch welche graphische Repräsentationen von Lauten durchgeführt werden, besonders schwierig ist (vgl. Guibert 1979).

Aus diesen Gründen ist es nicht verwunderlich, dass Maschinen nicht immer die lexikalischen und syntaktischen Informationen richtig auswerten und deswegen inhaltliche Schwierigkeiten sowie Abweichungen im Bereich von Standardsprache und Logik häufig Schwierigkeiten bereiten. Demzufolge reicht zumindest derzeit die Leistung maschineller Dolmetschsysteme nicht an die von HumandolmetscherInnen heran (vgl. Gile 2009: 162).

2.1 Maschinelle Spracherkennung und *TrueText*

Kommerziell erwerbbare Spracherkennungstechnologie steckt so gut wie hinter jeder Applikation, die mit Speech-to-Text Software oder automatischem *Telephone Service*, also Spracherkennungssystemen, zu tun hat. Ganz oben auf der Prioritätenliste steht

zumeist Genauigkeit. Für automatische Verdolmetschungen ist eine kontinuierliche Spracheingabe notwendig. So genannte diskrete Spracherkennungssysteme sind nicht geeignet, da sie von den SprecherInnen nach jedem Wort eine Pause verlangen und so als Befehle nur einzelne Wörter zulassen (vgl. Chang 2011).

Da jeder Mensch individuelle Züge in seiner Vortragsweise hat und sprecher-abhängige Systeme für jede/n einzelne/n SprecherIn neu trainiert werden müssten, ist das ultimative Ziel, sprecherInnenunabhängige Spracherkennungsdienstleistungen durchführen zu können. Diese sollten dann ohne erforderliches Benutzertraining für alle NutzerInnen unter jeglichen Konditionen gute Leistung erbringen (vgl. Chang 2011).

Um diesem Ideal in irgendeiner Weise näher zu kommen, ist es notwendig, auch das Szenario zu erfassen, in dem die maschinelle Dolmetschung durchgeführt werden soll. In diesem Zusammenhang liegt ein großer Unterschied darin, ob eine einzelne Person in monologischem Modus eine Rede hält oder ob es sich um eine Situation handelt, in der das permanente Wechseln zwischen Sprachen stattfindet, wofür *Skype Translator* entwickelt wurde. Das Spracherkennungssystem sollte in diesem Fall auch erkennen können, um was für eine Sprache es sich handelt. Im Mittelpunkt eines Spracherkennungssystems steht allerdings der Inhalt des Gesagten, womit die Frage nach dem, was gesagt wurde, zentral ist (vgl. Bahl et al. 1989: 1001). Wie im Kapitel der *Big Data* beschrieben, sind die Datenpools, die für die jeweiligen Phasen von *Skype Translator* genutzt werden, immens. Nichtsdestotrotz ist die derzeit verwendete Spracherkennungstechnologie von *Microsoft* lediglich auf Wort- und Buchstabenebene ausgelegt. Dadurch, dass Sprache jedoch nicht nur auf verbaler Ebene stattfindet, sondern auch die nonverbale Kommunikation eine ebenso große Rolle spielt, ist anzunehmen, dass die Entwicklung der Spracherkennungstechnologie noch lange nicht ihr Ende erreicht hat (vgl. Forsberg 2003: 2).

Tatsächlich ist nonverbale Kommunikation ein integraler Bestandteil von menschlicher Kommunikation und auch für HumandolmetscherInnen als Interpretationshilfe von sprachlichen Äußerungen essenziell. Gestik und nonverbale Komponenten enthalten Botschaften, deren Abwesenheit bei maschinellen Dolmetschungen schnell zu Missverständnissen führen kann. Die Forschung in diesem Gebiet ist noch in ihrer Anfangsphase, jedoch werden bereits Studien durchgeführt, die sich mit „visual computing“ auseinandersetzen, wobei der Spracherkennung durch

visuelle Inputs zusätzliche nonverbale Informationen zukommen sollen (vgl. Spilker et al. 2000: 135).

Gleichzeitig gibt es eine ganze Bandbreite an zusätzlichen Herausforderungen: Geräusche im Hintergrund, Verzögerungslaute, Wortschatzumfang, Perplexität der Sprache (unerwartete Wortreihungen/Satzkonstruktionen etc.), Speicherauslastung, Geschwindigkeit, Anforderungen an den Prozessor und selbst die Position des Mikrofons können allesamt zu Erkennungsfehlern führen (vgl. Pieraccini & Rabiner 2012: 139). Diese Thematik wird in Abschnitt 2.2.1 über Problemauslöser genauer ausgeführt.

2.1.1 Funktionsweise

In der Spracherkennung werden akustische Modell-Inputs in Phoneme übertragen, indem die SprecherInnen in ein Mikrophon sprechen und Schallwellen als Ton im Gerät eingefangen werden. Zur gleichen Zeit wird versucht, Hintergrundgeräusche auszublenden beziehungsweise zu reduzieren und die Lautstärke zu normalisieren. Diese gefilterten Schallwellen werden in winzige Einheiten, also Phoneme, aufgespalten. Von einem Phonem gibt es zwischen 15 und 50 Betonungssymbole in einem Wörterbuch einer Sprache. Diese Töne werden gebraucht, um ein Wort zu bilden, und jedes Phonem ist wie eine Verbindung in einer Kette, da jede Sprache stark verbreitete sequenzielle Muster und übliche Verbindungen hat und somit in unterschiedliche Gruppen eingeteilt werden kann. Phoneme kann man auch noch weiter in sogenannte Senone aufspalten, von welchen es wohl tausende in jeder Sprache gibt (vgl. Li et al. 2013: 1136). Anschließend kombiniert ein Sprachmodell diese phonetischen Informationen mithilfe von statistischen Modellen und Trainingsdaten zu Wörtern und Sätzen, die als eine Art Paralleltext[1] verwendet werden (vgl. Li et al. 2013: 1138).

Anschließend müssen noch Satzsegmente erfasst werden, die für die nächste Phase der maschinellen Übersetzung unbedingt erforderlich sind (vgl. Waibel & Fügen 2008: 71-72). Oft passiert es auch, dass die SprecherInnen sehr lange Sequenzen von sich geben, bei denen das Programm gefordert ist, mittendrin Segmentierungen

[1] Angelehnt an Göpferich (1999: 184f) werden in dieser Arbeit unter Paralleltexten verschiedensprachige Texte verstanden, die von MuttersprachlerInnen erstellt wurden und ein Thema behandeln, das dem ähnelt, welches von den TranslatorInnen bearbeitet wird.

durchzuführen. Segmentierungsalgorithmen verwenden statistische natürliche Pausen, Sprachmodell-Statistiken und prosodische Stichworte, um solche Segmente einzuteilen (vgl. Fünfer 2013: 23). Wie bereits in Kapitel 1.2 ausgeführt wurde, werden in der Spracherkennungsforschung bislang die Fehlerquoten häufig in WER (Word Error Rate) gemessen, die für Speech-to-Speech Technologie unbedingt unter 10% liegen sollte, um sinnvolle Übersetzungen durchführen zu können (vgl. Waibel & Fügen 2008: 71-72).

Gleichzeitig wird diese Methode dafür kritisiert, dass sie Wortfehler nur auf Wortebene misst, und nicht auf Ebene von Kontext und Syntax. Während es für den Bereich der maschinellen Übersetzung tendenziell viele Evaluierungsmethoden bezüglich Qualität gibt (siehe Wonisch 2017), ist für die Spracherkennung als Analysemethode tendenziell ausschließlich WER weitgehend akzeptiert. Aus diesem Grund wurden beispielsweise von He et al. (2011: 5632) Versuche gestartet, den Ansatz von BLEU zu etablieren, einem System, welches sich auf die WER-Methode stützt, jedoch nicht auf Wortbasis arbeitet sondern zusätzlich auch die Satz- beziehungsweise Segmentbasis evaluiert, wodurch der syntaktische Aspekt miteinbezogen wäre.

2.1.2 TrueText

TrueText ist ein für *Skype Translator* entwickelter Prozess, der nach der phonetischen Spracherkennungsphase eingesetzt wird, um Sprachkorrekturen durchzuführen. Nach *Microsoft* ist es ein eigenes Translationssystem „in sich selbst“ und verwandelt gesprochene Sprache in etwas, das der schriftlichen Sprache näher kommen soll, da die tatsächlichen Dolmetschsysteme in der geschriebenen Sprache aufgrund der verfügbaren höheren Datenmenge weitaus exakter und flüssiger funktionieren (vgl. Wendt 2016: 45). Das Programm basiert auf Paralleltexten mit vielen Ungereimtheiten, wodurch das System lernt, Zielsprachendaten zu reinigen. Die größte Schwierigkeit ist es, Paralleltexte zu generieren, die aus genügend Ungereimtheiten bestehen (vgl. Kumar et al. 2014: 3232). Diese Ungereimtheiten werden *Sprachdisfluenzen* genannt und sind oft Filterwörter, die verwendet werden, um nächste Gedanken zu fassen. Sie präsentieren sich beispielsweise in Form von Grunzen, Husten, Füllwörtern wie *ähm* oder *uhm*, Fehlstarts und Wiederholungen und sollen schlussendlich von *TrueText* eliminiert werden. *TrueText* wird durch Transkripte, die noch fehlerhaft sind, und das

selbige in korrigierter Form trainiert. *Saubere Transkripte* nennt man in diesem Zusammenhang jene Transkripte, die bereits verbessert wurden, während *Originaltranskripte* jene sind, die aus der originalen Audiodatei generiert wurden (vgl. Wendt 2016a: 47). Auch für die ErfinderInnen des Programms hinter Skype ist klar, dass mündliche Sprache sich sehr von schriftlicher Sprache unterscheidet und Menschen in den meisten Fällen keine klaren, perfekten Sätze formulieren.

Die Phase der Sprachkorrektur wird in dieser Arbeit als Teil der Spracherkennung betrachtet und transformiert die mündliche Redensart durch automatische Eliminierung der aufkommenden Disfluenzen in Richtung schriftlicher Sprache. Denn neben der Eliminierung werden auch Satzzeichen hinzugefügt und durch automatische Satzpausen, Satzzeichen und Groß- und Kleinschreibung wird versucht, den Text lesbarer und strukturierter zu machen (vgl. Wendt 2016a: 47f). Die nachstehende Abbildung ist von Skype publiziert worden und stellt das Phasenmodell von *TrueText* dar (vgl. Wendt 2016: 8).

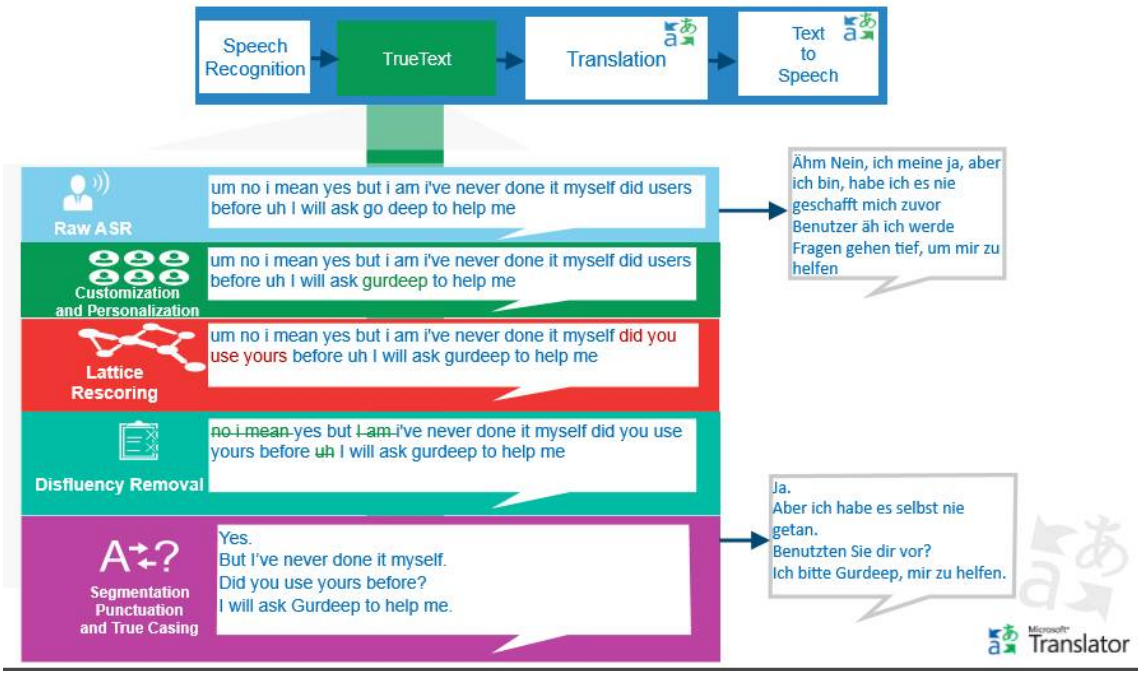

Abbildung 1: TrueText Phasenmodell

Nachdem die Rohdateien aufgenommen und phonetisch analysiert wurden und die Struktur des Textes aufgenommen ist, wird im nächsten Schritt eine Textanalyse durchgeführt, bei der evaluiert wird, ob es sich um eine positive oder eine negative Aussage handelt. Anschließend werden die Schlüsselphrasen extrahiert und einem

möglichen Thema zugeordnet. Ähnlich wie bei HumandolmetscherInnen erfolgt auch eine Art Antizipationsverfahren, bei dem die Wörter als solche und wahrscheinliche linke und rechte Nachbarn beziehungsweise Kollokationen antizipiert werden. Die meisten dieser Beispiele werden durch das *Bing*-System evaluiert, welches gleichzeitig auch als Rechtschreibprüfung genutzt wird. Die dabei analysierten Segmente bestehen teils aus einzelnen Wortkombinationen, Namen, Markennamen, idiomatischen Ausdrücken und Umgangssprache oder auch Homonymen. Schlüsselwörter werden also nach der phonetischen Analyse einer weiteren unterzogen, die diese miteinander in Zusammenhang setzen und so die wahrscheinlich richtige Lösung auswählen soll (vgl. Jefford 2017: 13ff).

2.2 Problemauslösende Faktoren

Der Dolmetschprozess wird in der (Human-)Dolmetschwissenschaft in mehrere Unterprozesse aufgeteilt: in „Listening and Analysis Effort“, also den Hör- und Analyseprozess, den „Production Effort“, also den Produktionsprozess sowie den „Short-Term Memory Effect“, also den Kurzzeitgedächtnisprozess (vgl. Gile 2009: 171). Alle drei beanspruchen ein bestimmtes Volumen der limitierten Kognitiven Kapazität. Probleme kommen dann auf, wenn die Anforderungen an die Prozesskapazität zu hoch werden und diese überlasten. Daraus resultieren so genannte *Problem Trigger,* also Problemauslöser, denen DolmetscherInnen mit bestimmten Strategien begegnen (2009: 190). Elemente, die speziell für Dolmetschungen und deren einzelne Prozesse Probleme auslösen können, gibt es in sämtlichen Phasen. In der Parallelarbeit von Wonisch (2017) wird auf Problemauslöser eingegangen, die vor allem in anderen Phasen der (maschinellen) Dolmetschleistung auftreten.

Die Problemauslöser finden sich beispielsweise in Parametern wie Ausgangssprache, Vortragsgeschwindigkeit, Stil, Fachlichkeit der Ausgangsrede, Betonung, Umweltgeräusche oder auch der Temperatur in der Dolmetschkabine. In nahezu jeder Dolmetschsituation gibt es potenzielle Störfaktoren. Zur besseren Übersicht wurden Kategorien entwickelt, die sämtliche problemauslösende Faktoren zusammenfassen.

So brachte Mankauskienė (2016) beispielsweise in Anlehnung an Giles Klassifikation folgende vier Unterteilungsgruppen vor: senderbezogene Problemquellen (Akzent,

Vortragsgeschwindigkeit etc.), redenbezogene Problemquellen (Syntax, Terminologie etc.), auf DolmetscherInnen bezogene Problemquellen (Erfahrung, Kompetenz, Erschöpfung etc.) sowie technik- und umweltbezogene Problemquellen (Versagen der Dolmetschtechnik, externe Geräuschquellen etc.).

Für diese Arbeit sind vor allem die redenbezogenen Problemauslöser relevant, die sich in der Spracherkennung bemerkbar machen. Dazu muss gesagt werden, dass Problemauslöser nicht zwangsläufig zu tatsächlichen Problemen führen müssen, sondern immer nur „mögliche Problemauslöser" sind (vgl. Gile 2009: 171).

2.2.1 Redenbezogene Problemauslöser

Im folgenden Unterkapitel werden unterschiedliche Faktoren angeführt, die im Prozess der Spracherkennung als mögliche redenbezogene Problemauslöser identifiziert und im nachfolgenden Experiment evaluiert werden. Dafür wurden Ausgangstextkomplexität, Fachtermini, (Jahres-)Zahlen, Eigennamen, Akronyme, Syntax und Sprachregister als für diese Arbeit relevante Bereiche ausgewählt. Im empirischen Teil der Arbeit werden zudem auch allgemeinsprachliche Elemente als Kontrolle zu Fachtermini benutzt, welche hier jedoch nicht extra erklärt sind.

2.2.1.1 Ausgangstextkomplexität

Sinnverständnis des Ausgangstextes ist für DolmetscherInnen essenziell, um Dolmetschleistungen erfolgreich durchführen zu können (vgl. Pöchhacker 2004: 131). Es ist schwierig, einen Ausgangstext bezüglich seiner Komplexität zu messen. So könnte ein Ausgangstext beispielsweise auf Basis seiner lexikalischen Varietät, Wortfrequenz oder seiner redundanten oder nicht-redundanten Einschübe wie auch der Häufigkeit von Eigennamen oder Zahlen evaluiert werden (vgl. Gile 1995: 170ff). Informationskomplexität spielt sich also nicht zwangsläufig auf der lexikalischen Ebene ab.

Aus diesem Grund nutzten Liu & Chiu (2001: 248) zur Komplexitätsmessung ihres Ausgangsmaterials für Dolmetschleistungen die Lesbarkeitsformel „Flesch-Kincaid", welche mithilfe einer mathematischen Formel die Satzlänge in Relation zur Wortlänge berechnet. Gleichzeitig sollte aber beachtet werden, dass die Flesch-Kincaid-Formel „Lesbarkeit" evaluiert und nicht die Verständlichkeit beim Hören.

In der Übersetzungswissenschaft entwickelte Christiane Nord eine „translationsorientierte Ausgangstextanalyse“ (Nord 2004: 243), die sich allgemein auf Texte ummünzen lässt, während sie spezifisch genug ist, um für die Übersetzung mögliche Problemfelder aufzuzeigen. Auch wenn dieser Zugang für die Evaluierung von Dolmetschausgangstexten für gewöhnlich nicht üblich ist, soll er hier einen besseren Einblick in den situativen Zugang geben (vgl. Nord 1999: 270).

2.2.1.2 Fachtermini

Für TranslatorInnen ist es wichtig, die funktionale und logische Infrastruktur der Sätze ausreichend zu verstehen, um diese in der Zielsprache zu reproduzieren. Oft ist es nämlich nicht ausreichend, die gleiche linguistische Struktur einer Sprache in die andere zu übertragen, da diese in der Zielsprache sonst holprig beziehungsweise inakzeptabel klingt. Dies wird vor allem in Wort-für-Wort Translationen sichtbar (vgl. Gile 2009).

Wenn dieses Verständnis gegeben ist, kann es für die TranslatorInnen ausreichend sein, eine adäquate terminologische Äquivalenz von Nomen oder Nomengruppen und teilweise auch von Verben zu nutzen, um die Translation erfolgreich durchzuführen, obwohl ihre Expertise eventuell nicht mit der der ExpertInnen gleichzustellen ist. Dies sagt auch aus, dass zukünftige DolmetscherInnen nicht zusätzlich die jeweiligen Fachgebiete studieren müssen, um erfolgreich zu dolmetschen, sondern dass es in den meisten Fällen ausreichend ist, sich mit einzelnen Teilbereichen der Gebiete zu beschäftigen, um äquivalente Fachbegriffe und deren Bedeutungen in der Fremdsprache zu erlernen (vgl. Gile 2009: 86). Gleichzeitig sind Glossare und Wörterbücher niemals ausreichend, niemals absolut vertrauenswürdig und selten präzise genug, um nicht-spezialisierte TranslatorInnen mit definitiven Lösungen in terminologischen Problemsituationen zu versorgen (vgl. Gile 2009: 95).

Sowohl DolmetscherInnen als auch ÜbersetzerInnen nutzen den gegebenen Kontext, der von den Texten und Reden ausgeht, um mehr Wissen über das Thema zur Analyse für eventuelle Hinweise zu erlangen oder um ihr Verständnis der daraus erfolgenden Zieltextsegmente zu verbessern. Während die Reden und Texte verarbeitet werden, machen sie sich mit dem Thema vertraut, um ein besseres Verständnis zu erlangen (vgl. Gile 2009: 96).

Durch die Weiterentwicklungen in Wissenschaft und Technologie entstehen jedes Jahr viele neue technische und wissenschaftliche Ausdrücke. Je nachdem in welchem Land

beziehungsweise in welcher Sprache diese Entwicklungen entstehen, sind sich andere Nationen dieser neuen Wörter nicht immer bewusst. Um verschiedene kommunikative Bedürfnisse der NutzerInnen einer Sprache zu erfüllen, werden diese neuen Begriffe eingeführt. Haugen war 1959 der Erste, der in der Literatur von so genanntem „Language Planning" sprach, da das Kreieren von Sprache sorgfältig geplant sein sollte (vgl. Haugen 1966: 50f). Diese neuen Wortkreationen werden „Neologismen" genannt. Gleichzeitig sind Begriffe, die aus anderen Sprachen geborgt werden, unter „Lehnwörtern" bekannt.

2.2.1.3 Zahlen, Eigennamen und Akronyme

Die Schwierigkeit, Eigennamen korrekt auszusprechen, bezieht sich nicht nur auf eingliedrige, sondern auch auf mehrgliedrige Ausdrücke und ist problematisch, da diese meist durch ähnlich klingende Konsonanten, Silben beziehungsweise Vokale leicht verwechselt werden können (vgl. Gile 2009: 194). Dasselbe trifft auch auf die Kategorie der Akronyme oder Abkürzungen zu.
(Jahres-)Zahlen werden als Problemauslöser gesehen, da DolmetscherInnen unterscheiden können müssen, ob es sich bei diesen um ungefähre Zahleneinheiten oder bestimmte Messwerte handelt. Die Strategie allgemein gehaltenen Zahlen wie zum Beispiel „52,3%" gibt DolmetscherInnen die Möglichkeit, auf eine andere Ebene zu gehen und diese mit „ungefähr die Hälfte" auszudrücken. Gleichzeitig wäre diese Strategie jedoch bei einem Wert wie „873,5" Milligramm nicht immer ratsam, da manche Kontexte es erfordern, die Daten exakt auszudrücken. In manchen Fällen könnten genau derartige Verallgemeinerungen verheerende Auswirkungen haben (vgl. Nolan 2012: 288).

Zahlen, Eigennamen und Akronyme sind vor allem deswegen auch zu den Problemauslösern zu zählen, da sie eine geringe Redundanz aufweisen und somit oft mehr Energie als andere Einheiten benötigen (vgl. Gile 1995: 108).

2.2.1.4 Syntax

Syntax kann großen Einfluss auf die Verständlichkeit von Texten haben. Texte, die mit vielen Einschüben und Relativsätzen gespickt sind, gelten tendenziell als schwer verständlich. Dies hat Einfluss darauf, ob neue Informationen als solche wahrgenommen werden und anschließend mit den bereits gegebenen Informationen

verknüpft werden können (vgl. Nolan 2012: 25ff). Zusätzlich beanspruchen kompliziertere Formulierungen die kognitive Prozesskapazität stärker, was wiederum nach Giles *Effort Model* schneller zur Überlastung führt.

Gesprochene und geschriebene Sprache haben einige syntaktische Unterschiede. So sind zum Beispiel verschiedene morphosyntaktische Fehler in der gesprochenen Sprache durchaus zulässig, während sie es in der geschriebenen Sprache nicht sind. Auch Abbrüche und Neuformulierungen kommen häufig vor und können darauf zurückgeführt werden, dass SprecherInnen einer ersten Intention folgend den Satz beginnen, dann aber ändern und spontan eine neue Aussage formulieren. In derartigen Fällen kann es zu lexikalischen oder syntaktischen Mischformen kommen.

Gesprochene Sprache ist demnach oft mit Disfluenzen und Redundanz übersät, während geschriebene Sprache meist eine höhere Informationsdichte aufweist (vgl. Crevatin 1989: 22). Selbstkorrekturen wie Wiederholungen, Satzabbrüche, perplexe Sprache, ebenso wie Verzögerungsphänomene sind in gesprochener Sprache häufig mögliche Fehlerquellen für maschinelle Dolmetschsysteme (vgl. Pieraccini & Rabiner 2012: 139).

2.2.1.5 Idiomatik und Sprachregister

In formalen und vor allem diplomatischen Szenarien sollten sich DolmetscherInnen der Konventionen und der politisch korrekten Umgangsformen bewusst sein, die von Kontext und Gegebenheiten bestimmt werden. In Fällen formeller Umgebung werden Kraftausdrücke zum Beispiel tendenziell vermieden, was aber nicht heißt, dass Ausdrücke moralischer Verurteilung oder Ablehnung nicht ausgedrückt werden. Meist ist es von den DolmetscherInnen abhängig zu entscheiden, mit welcher Gewichtung sie die Ebene der Wortwahl wiedergeben. Dies ist auch vom Ton und der Intention der SprecherInnen abhängig und die präzise Wiedergabe der Schattierung der Aussagen kann sehr wichtig sein. Da die DolmetscherInnen durchaus dafür verantwortlich gemacht werden können, wenn Eskalationssituationen im Nachhinein auf die Dolmetschung zurückgeführt werden, sind viele DolmetscherInnen dazu geneigt, in Fällen, in denen es hart auf hart kommt, eher neutrale Ausdrucksweisen zu verwenden (vgl. Nolan 2012: 127f). Im Fall von *Skype Translator* hat sich durch Eigenerprobung herausgestellt, dass bei Erkennung von Kraftausdrücken ein Filter verwendet wird,

welcher im Transkript besagte Ausdrücke mithilfe der Symbole „***“ darstellt und in der Dolmetschung diese durch einen Piep-Ton ersetzt.

Gleichzeitig ist jedoch auch die Meinung unter DolmetscherInnen vertreten, dass Auseinandersetzungen auch in derselben Intensivität wiedergegeben werden sollten wie sie dargebracht wurden, da diese einen Zweck verfolgen. So bewegen sich DolmetscherInnen oft zwischen diplomatischem Zugang und strenger Debatte auf Messers Schneide. Nolan (2012: 173f) nennt Rhetoriker wie Cicero, die finden, dass Register und Wortwahl drei Kriterien entsprechen sollten: der Adäquatheit des Themas, der Angepasstheit an die Situation und der Anpassung an die ZuhörerInnen. Aus diesem Grund ist es wichtig, dass DolmetscherInnen sich der Intention der SprecherInnen bewusst sind, um dadurch so exakt und vollständig wie möglich arbeiten zu können. Die Intonation der SprecherInnen ist in vielen Fällen genauso wichtig wie auch der Inhalt der Rede.

Darüber hinaus ist es nicht unwesentlich, dass unterschiedliche Sprachen unterschiedliche Stile haben. So gilt die spanische Sprache als *blumiger* als die Englische. DolmetscherInnen, die vom Spanischen oder Französischen ins Englische arbeiten, haben tendenziell öfter Probleme mit formellen Ausdrücken, da für das Spanische oder Französische Eloquenz, Komplexität und reichhaltige Wortwahl als kultiviert gelten, während diese Art der Ausdrucksweise für das Englische eher als veraltet oder barock wahrgenommen wird (vgl. Nolan 2012: 174).

Wenn DolmetscherInnen in solchen Situationen also zu nah am Ausgangstext dolmetschen, kann die Dolmetschung schnell lächerlich klingen und vom Zeitgeist entfernt sein. Darüber hinaus gilt Humor unter DolmetscherInnen als schwierig zu dolmetschen, da ZuhörerInnen, die lachen, Aufmerksamkeit erregen und gleichzeitig jene ZuhörerInnen, die von den Dolmetschungen abhängig sind, begierig den Auslöser der Heiterkeit erfahren wollen. DolmetscherInnen sollten sich daher über den Zweck des Humors bewusst sein, um eine entspannte Beziehung zu ihren ZuhörerInnen aufbauen zu können (vgl. Nolan 2012: 258).

2.2.2 Weitere Problemauslöser

Die in diesem Kapitel angeführten Problemauslöser, die besonders in der Spracherkennung relevant sind, werden für den empirischen Teil dieser Arbeit nicht explizit benötigt. Nichtsdestotrotz werden sie als Ergänzung für einen möglichst umfassenden Überblick angeführt. Senderbezogene Problemauslöser findet man bei Wort- und Satzakzenten (also auch bei Dialekten und Akzenten von Personen), Tonverläufen einer Silbe, Melodien von Wort, Satz oder Äußerungen sowie bei der Quantität aller sprachlicher Einheiten und Pausen, außerdem noch bei der Geschwindigkeit und der Lautstärke einer Aussage. Oftmals treten diese in Kombination und nicht getrennt voneinander auf. In vielen Sprachen ist die Akzentuierung essenziell zur Unterscheidung von gleich geschriebenen Wörtern mit unterschiedlicher Bedeutung. Auch gesundheitliche oder emotionale Zustände können durch Prosodie dargestellt werden. Durch Satzmelodie und Pausen werden zudem auch oft Informationen zu Satzgrenzen, Fragen, Ausrufen und Aussagen verdeutlicht (vgl. Ahrens 2005: 1f).

2.2.2.1 Redetempo

SprecherInnen, die kurze, klare Sätze in einem mäßigen Tempo bilden, können DolmetscherInnen ihre Arbeit erleichtern. Komplizierte Syntax sollte von professionellen DolmetscherInnen auch ohne größere Anstrengung gedolmetscht werden können. Nichtsdestotrotz bedeutet es zusätzliche Belastung, wenn komplexe und lange Satzkonstruktionen mit schnellem Tempo kombiniert werden (vgl. Nolan 2012: 25).

Eine sehr große Herausforderung für DolmetscherInnen ist es, wenn der Text so schnell vorgetragen wird, dass die Output-Geschwindigkeit nicht mehr erhöht werden kann (vgl. Li 2010: 23). In Fällen, in denen DolmetscherInnen ihre eigene Sprechgeschwindigkeit nicht mehr erhöhen können, kann eine Strategie angewandt werden, bei der der Sinn des Ausgangstextes in der Zielsprache zusammengefasst wird. Aufgrund redundanter Informationen, die ausgelassen werden, können somit die Texte kürzer und prägnanter gestaltet werden.

Wenn Reden langsam vorgetragen werden, kann dies zwar die Kapazitätsanforderungen an die DolmetscherInnen minimieren, impliziert jedoch die

Schwierigkeit, dass die Informationen länger im Kurzzeitgedächtnis behalten werden müssen, was wiederum kognitiv herausfordernd ist (vgl. Gile 2009: 200). Demnach können langsame Reden den Dolmetschprozess ebenso erschweren wie schnelle.
Nach AIIC[2] (2002) gilt beim Konferenzdolmetschen eine Geschwindigkeit zwischen 100 und 120 Wörtern pro Minute als angenehm. Über oder unter der Geschwindigkeit zwischen 95 oder 120 Wörtern pro Minute wurde in einer Studie von Gerver (1969: 66) eine Abnahme der korrekt gedolmetschten Textstellen sowie eine Erhöhung des *Time Lag* und der Pausen beobachtet.

Zur Diskussion steht hierbei, inwieweit die Redegeschwindigkeit in unterschiedlichen Sprachen als von der Norm abweichend wahrgenommen wird und auch, ob das Redetempo in der Ausgangssprache auch der Geschwindigkeit entspricht, wie diese wahrgenommen wird (vgl. Déjean le Féal 1982: 222f). Hierbei kommt auch wieder die Diskussion des Unterschiedes zwischen gesprochener oder geschriebener Sprache zutage (vgl. Tannen 1982).

2.2.2.2 Prosodische Elemente wie Akzent, Stimmhöhe, nonverbale Kommunikation

In der Arbeitswelt von DolmetscherInnen gibt es oft Situationen, in denen die Vorträge in einer Sprache gehalten werden, die nicht die Muttersprache der Vortragenden ist. Aus diesem Grund ist es für DolmetscherInnen wichtig, mit den unterschiedlichen Sprachvarietäten und Akzenten ihrer Arbeitssprachen vertraut zu sein. Nach Kurz (2005: 61) bietet ein unbekannter Akzent eine besonders große Herausforderung für DolmetscherInnen. Der Akzent impliziert nicht nur soziale oder regionale Spezifika sondern bei nichtmuttersprachlichen RednerInnen oft auch das Unvermögen, die Zielsprache phonetisch korrekt zu betonen. Besonders in der Arbeitssprache Englisch kommen DolmetscherInnen häufiger in die Situation, mit unterschiedlichen Akzenten zu arbeiten, da Englisch mittlerweile als *Lingua franca* verwendet wird (2005: 60).

Das Verstehen von Sprachlauten hängt jeweils wieder von der Vertrautheit, dem Vorwissen und der Erfahrung mit jenen Akzenten ab (vgl. Cooper et al. 1982: 104, AIIC 2002: 25). Mazzetti (1999: 125) brachte auf den Punkt, dass ein „ausländischer Akzent“ sich nicht nur in der ungewöhnlichen Betonung von Phonemen und auf

[2] AIIC steht für „Association Internationale des Interprètes de Conférence“ und ist der internationale Verband der Konferenzdolmetscher

syntaktischer und lexikalischer Ebene äußert, sondern dass Intonation und andere Komponenten der Prosodie, also auch Tempo oder Rhythmus, einen Einfluss auf die Verständlichkeit des Akzentes und somit auf den Dolmetschprozess haben.
In diesem Zusammenhang haben Emotionen in der gesprochenen Sprache eine immense Wichtigkeit. So verliert eine ursprünglich aggressive politische Wahlkampfrede ohne Emotionen ihren Charakter und zeigt im Zieltext eine vollkommen andere Wirkung. Emotionen spiegeln sich demnach nicht nur in der Wortwahl, sondern zusätzlich in der Sprachmelodie wider. So können zum Beispiel Schimpfwörter Frustration oder Aggression zum Ausdruck bringen. Bisher haben Spracherkennungssysteme große Schwierigkeiten, Emotionen in die Analyse einzubinden, da der Trainingsaufwand dafür sehr hoch ist und Emotionskategorien festgelegt werden müssen (vgl. Akagi et al. 2014: 2). Es wurden aber bereits Versuche gestartet, diese Kategorisierungen in maschinelle Dolmetschsysteme einzubauen (siehe Abschnitt 2.1).

Emotionen können auch Einfluss auf die Stimmhöhe nehmen. Davon abgesehen gibt es große Unterschiede in der Stimmfrequenz bei Menschen unterschiedlichen Geschlechtes und Alters. Dies wurde sprachübergreifend und auch diskursübergreifend bestätigt. Die Durchschnittsfrequenz bei Männern liegt bei ungefähr 120 Hz, während jene von Frauen bei etwa 210 Hz liegt. Diese Durchschnittswerte verändern sich ein wenig im Alter (vgl. Traunmüller & Eriksson 1994). Die Stimmen „verdunkeln“ sich um etwa 30 Hz in Form eines Stimmbruchs meist um die Zeit der Menopause. Männer erleben eine immer weiter leicht abfallende Kurve nach dem Stimmbruch in der Pubertät, welche sich bis zum 35. Lebensjahr leicht fortsetzt und dann wieder stark zunimmt, um ab dem 50. Lebensjahr bis zu 30 Hz anzusteigen (vgl. Linville 2001: 170ff).

Im Alter verändert sich auch die Artikulationsfähigkeit von älteren Erwachsenen. Zusätzlich reduziert sich auch ihre Sprechgeschwindigkeit, sodass die natürliche Redegeschwindigkeit deutlich langsamer wird als die von jüngeren Erwachsenen. Auch die Sprachqualität und die Wortwahl verändern sich. So scheinen ältere Männer eine Tendenz dazu zu haben, die Vokale besonders zu betonen, da sie weniger Kontrolle über ihre Zungenmuskulatur haben (vgl. Vipperla 2011: 8). Zusammengefasst tendieren sie im Vergleich zu jüngeren Erwachsenen stärker zu Wortfindungsproblemen und verwenden weniger komplexe Sätze, während sie ihre

Wahrnehmungsschwierigkeiten durch eine stärkere Betonung von prosodischen Elementen überkompensieren.

2.2.2.3 Technik- und umweltbedingte Problemauslöser

Geräusche im Hintergrund, Verzögerungslaute, Wortschatzumfang, Speicherauslastung, Anforderungen an den Prozessor sowie die Position des Mikrofons sind allesamt Aspekte, die in die Kategorie der technischen und umweltbedingten Problemauslöser fallen (vgl. Waibel & Fügen 2008: 71-72).

Für DolmetscherInnen können Faktoren wie Tonqualität oder Hintergrundgeräusche erheblich sein, wenn es darum geht, die Konzentration aufrecht zu erhalten.

Spontansprachliche wie auch außerlinguistische Schallereignisse wie Hüsteln, Räuspern und Schlucken sind Geräusche, die in einem Szenario durchaus zulässig und üblich sind und somit vom Spracherkennungssystem einberechnet werden sollten.

Bis zu einem gewissen Grad werden diese Geräusche als Störungen erkannt und in Abschnitt 2.1.2, im Fall von *Skype Translator* durch *TrueText*, entsprechend verworfen. Nicht alle Geräusche gehen zwangsweise von den RednerInnen aus und können als äußerliche Faktoren auch schnell zu Informationsverlust zwischen SprecherInnen und Mikrofon führen, wobei erwähnt werden muss, dass Systeme inzwischen üblicherweise mit akustischen Funktionalitäten wie Echokompensation, Enthallung etc. ausgestattet sind (vgl. Gile 2009: 193).

Neben den Störgeräuschen zählt auch die Tonqualität selbst zu den wesentlichen Faktoren für die Machbarkeit der Dolmetschungen. Hochwertige Dolmetschtechnikanlagen zur idealen Übertragungsqualität wie auch gut isolierte Kabinen sind somit essenziell, um die passende Lautstärke und die notwenige Signalqualität gewährleisten zu können. Auch das Sichtfeld der DolmetscherInnen kann hierbei von Bedeutung für ihr Verständnis sein (vgl. Pöchhacker 2004: 127).

Microsoft empfiehlt für die Nutzung von *Skype Translator,* nicht das computerintegrierte Mikrofon zu nutzen, sondern ein separates Headset zu verwenden, da die Qualität der Spracherkennung sonst in Mitleidenschaft gezogen werden könnte (vgl. Microsoft 2014: 7).

So wird beispielsweise in der Spracherkennungstechnologie (siehe Kapitel 2.1), vor allem mit den Werten des WER gearbeitet, um festzustellen, wie viele Wörter exakt

erkannt, ausgelassen, ersetzt oder hinzugefügt wurden. Auch in der Beurteilung von HumandolmetscherInnen ließe sich ein derartiges Verfahren anwenden (vgl. Barik 1971) und ist auch für den empirischen Teil dieser Arbeit relevant.

2.3 Zusammenfassung

Spracherkennung und das damit zusammenhängende Verständnis ist ein essenzieller Teil jeder Dolmetschleistung. Maschinelle Spracherkennungstechnologie nutzt *Big Data* zur Lokalisierung von Paralleltexten, welche durch Schlüsselbegriffe und damit entstehende Kontextbereiche erkannt werden, um anschließend Segment für Segment korrekt in Schrift umzuwandeln. *SkypeTranslator* nutzt als zusätzliches Element ein Sprachkorrekturprogramm namens *TrueText*, durch welches Ungereimtheiten in dem bis dahin vorhandenen Material eliminiert werden sollen.

HumandolmetscherInnen arbeiten häufig am Rande ihrer Gedächtnisleistungskapazität und können tendenziell leicht durch zusätzlich auftretende Faktoren aus dem Gleichgewicht gebracht werden. Diese Faktoren werden Problemauslöser genannt. Unter redenbezogenen Problemauslösern versteht man beispielsweise Elemente wie die Komplexität des Ausgangstextes, unbekannte Fachtermini, Zahlen, Eigennamen, Akronyme oder Registerwahl.

Ansonsten können in der Phase der Spracherkennung auch Problemauslöser wie schnelle Redegeschwindigkeit, unbekannte Akzente oder fehlende nonverbale Kommunikation auftauchen. Auch die Tonqualität der Aufnahme oder Störgeräusche von außen können erheblichen Einfluss auf die Qualität der Dolmetschleistung haben.

3 Methodik

Diese Arbeit beschäftigt sich mit der Spracherkennungskomponente beim maschinellen Dolmetschsystem *Skype Translator*. Ziel der Arbeit ist es herauszufinden, wie *Skype Translator* im Spracherkennungsprozess mit redenbezogenen problemauslösenden Faktoren umgeht. Dies soll erreicht werden, indem mehrere textuell idente Sprachaufnahmen von *Skype Translator* gedolmetscht und diese dann sequenziell miteinander verglichen werden. Dadurch, dass die Dolmetschleistung von *Skype Translator* durch Transkripte aufgezeichnet wird, können anhand dieser die einzelnen problemauslösenden Sequenzen überprüft werden.

Der Vorgang wurde initiiert, indem ein rezeptives Interview[3] über ein wissenschaftliches Thema in Monologform in englischer Sprache aufgenommen und anschließend so zusammengeschnitten wurde, dass es statt 12 Minuten nur noch ein Interview von 6 Minuten war. Davon wurden dann zwei Transkripte (siehe Anhang 02) angefertigt: ein *Originaltranskript*, in dem Disfluenzen aufgezeichnet und ein *sauberes Transkript*, aus dem selbige bereits entfernt wurden. Die englische Sprache wurde gewählt, weil diese die Ursprungssprache von Skype ist, weswegen davon auszugehen ist, dass das Programm in dieser die besten Ergebnisse aufweist. Anschließend wurde das Transkript mit den Disfluenzen erstmals auf mögliche Problemauslöser überprüft.

Hinterher sprachen 31 SprecherInnen mithilfe von „Shadowing“ (nachstehend in Kapitel 3.2 erklärt) den vorgegebenen Text auf ein Tonbandgerät nach. Die insgesamt 32 im Vorfeld aufgenommenen Audiodateien wurden daraufhin direkt als Input-Quelle in *Skype Translator* eingespielt (siehe Kapitel 3.3).

Schließlich standen 64 von *Skype Translator* gedolmetschte Transkripte vom Englischen ins Englische für die Analyse zur Verfügung. Die bereits im Vorfeld ausgewählten redenbezogenen Problemauslöser-Sequenzen wurden aus sämtlichen gedolmetschten Transkripten herausgesucht. Die Kategorien der redenbezogenen problemauslösenden Faktoren in der Spracherkennung wurden aus der Literatur abgeleitet (siehe Abschnitt 2.2.1) und in folgende Bereiche unterteilt: „Fachtermini“, „Focus on Form“ als spezieller Fachterminus, „Jahreszahlen“, „Eigennamen“, „Syntax“,

[3] Nach Lamnek (2006: 383) sind rezeptive Interviews „die offenste und am wenigsten prädeterminierende Form aller qualitativen Interviews“. ForscherInnen sind „interviewende BeobachterInnen und beobachtende InterviewerInnen“ und halten sich zurück beziehungsweise stellen keine antwortproduzierenden Fragen.

„Betonung“, „Intonation determinierende Syntax“ und „Registerwechsel“. Zuletzt wurde auch noch eine weitere Kategorie namens „Kontrollwörter“ hinzugefügt, die aus allgemeinsprachlichen Wörtern besteht und als Gegenüberstellung für Fachtermini gedacht war.

Für die Bewertung dieser Kategorien wurden vier Skalen entwickelt, mithilfe derer „semantische“, „phonetische“, „typologische“ und „spezielle“ Bewertungen durchgeführt wurden und welche jeweils aus vier Bewertungskriterien bestanden. Diese Kriterien bestanden aus JA, NEIN, JEIN und LEER. Allgemein gesprochen wurden diese in etwa so eingestuft, dass JA als „voll zutreffend“, JEIN als „halb zutreffend“, NEIN als „nicht zutreffend“ und LEER als „Daten nicht evaluierbar“ eingeteilt wurde.

Eine genauere Darstellung ist einerseits in Anhang 01, einige Beispiele sind in Kapitel 3.4 zu finden. Die Daten wurden anschließend anhand der ihnen zugeteilten Skalen quantitativ mit den jeweilig für sie zutreffenden Bewertungsmethoden ausgewertet. Zur besseren Übersicht über die Ergebnisse wurden in einem letzten Schritt die Bewertungskriterien nach einem dafür entwickelten Score aufgeschlüsselt.

3.1 Der Ausgangstext/Redenbezogene Faktoren

Das rezeptive Interview, das in monologischer Form dargestellt wurde, beschäftigt sich mit dem Themenbereich der Fremdsprachendidaktik und wurde mit einer Expertin in diesem Gebiet eigens für diese und die damit einhergehende Parallelarbeit von Wonisch (2017) durchgeführt. Aufgegriffen wurde das von *Skype Translator* genannte Szenario der/des PhD-StudentIn, der/die sich mithilfe von *Skype Translator* Informationen von ExpertInnen aus dem Ausland (in anderer Sprache) holen möchte. In diesem Sinne ist der empirische Teil dieser Arbeit in Form einer Simulation durchgeführt worden.

Das Interview der Ausgangssituation wurde demnach von jemandem durchgeführt, der/die sich über neue Konzepte in diesem Fachbereich mithilfe des *Skype Translator*-Modus informieren wollte und dafür die Form eines rezeptiven Interviews wählte. Die interviewte Person berichtete dann über das Konzept von *Focus on Form* (FOF). Das Experiment ging davon aus, dass der/die Interviewende anschließend viele andere Interviews über dasselbe Thema durchführte und somit der Verlauf und die persönliche Historie des genutzten Skype-Kontos gespeichert wurden, wodurch *Skype*

Translator registrierte, in welchem Fachbereich sich diese/r NutzerIn besonders bewegte.

Der Ausgangstext wurde in einer sehr schnellen Geschwindigkeit vorgetragen (etwa 162 WPM). Bereits hier wird deutlich, dass hinter dieser Arbeit die Absicht steht, die Grenzen des Programms auszutesten, mit anderen Worten, einem „Stresstest" zu unterziehen. Die nächsten beiden Unterkapitel dienen dazu, den Inhalt und auch die Form des Ausgangstextes näher zu beschreiben. Der Schwierigkeitsgrad wird mithilfe der Lesbarkeitsformel „Flesch-Kincaid" (vgl. Flesch 1948) gemessen, die auch schon von Liu & Chiu (2009: 248) genutzt wurde, um Ausgangstexte für Dolmetschungen zu evaluieren. Darüber hinaus wird eine Textsortenanalyse durchgeführt, die den Ausgangstext in Bezug auf textexterne wie auch textinterne Faktoren evaluiert.

3.1.1 Textsorte, Fachbereich und Textkomplexitätsmessung

Um sich der Intention der Autorin des Ausgangstextes bewusst zu werden, wird dieser anhand der Textanalyse von Nord (1988: 270) untersucht. Diese ist eigens für die Analyse von Ausgangstexten für Übersetzungsleistungen gedacht und liefert textinterne sowie textexterne Faktoren, die in Tabelle 1 angeführt werden. Die textexternen Faktoren wurden erhoben, um sämtliche W-Fragen zu beantworten. Konkret, um zu benennen, wer an wen über welches Medium wo, wann, zu welchem Anlass, wozu und mit welcher Funktion diesen Text produziert hat.

Tabelle 1: Textexterne Faktoren

WER (TextproduzentIn)	WissenschafterIn mit Doktorat-Abschluss für Didaktik in der Zweitsprache wird interviewt (monologischer Vortrag) von einem/einer PhD-StudentIn
WOZU	Diese Expertin wurde zurate gezogen, um interdisziplinär mehr Informationen über bestimmte Konzepte für die Forschung der Spracherwerbswissenschaft zu erhalten.
WEM (EmpfängerIn/ EmpfängerIn-erwartungen)	WissenschafterIn im Bereich des Zweitspracherwerb, der/die seine Doktorarbeit schreibt und sich erhofft, neue Informationen/Zugänge für seine Arbeit von ExpertInnen aus dem Ausland zu holen.

WELCHES MEDIUM (Medium/Kanal)	Medium ist das Internet via Skype-Gespräch in englischer Sprache. Es wird ein maschinelles Dolmetschsystem hinzugezogen.
WO	Internet
WANN	Jänner 2017
WARUM	Anlass ist eine kommunikative Handlung: Ein/Eine PhD-StudentIn erkundigt sich nach neuen Konzepten, die er/sie für seine/ihre Forschungsarbeit brauchen kann.
FUNKTION	Informieren über wissenschaftliche Zugänge

Die textinternen Faktoren beschreiben die nicht ganz so offensichtlichen Elemente, wie zum Beispiel die Wortwahl und die damit einhergehende Stimmung, die Wirkung, das Register und den Einsatz von nonverbalen Elementen beziehungsweise den Aufbau des Textes.

Tabelle 2: Textinterne Faktoren

WORÜBER (Fachbereich/Thema)	Zweitsprachen/ Kognitionswissenschaften/ Spracherwerbswissenschaften/Sprachwissenschaften/Didaktik
WAS	Focus on Form/Focus on Meaning/Focus on FormS – was man unter diesen Zugängen versteht, wie sie angewandt werden und wie sie sich äußern
WAS/WAS NICHT (Präsuppositionen)	Erklärt wird, wie sich dieser Wissenschaftsbereich entwickelte/ wie dieser Bereich im Klassenzimmer angewandt wird und wie LehrerInnen und SchülerInnen bestmöglich damit umgehen sollten.
WELCHE REIHENFOLGE (Textaufbau)	Textaufbau – Entwicklung von FOF/ Was man darunter versteht/ Was man versucht zu tun/ Wer Forschung darüber betrieben hat/ Welche Methoden für FOF angewandt werden (Feedback vs. Input Flooding)/ Wie FOF kognitiv verarbeitet wird
WELCHE NON-VERBALEN ELEMENTE	legerer Umgangston/ erklärend/ eher ein entspanntes Gespräch/ kollegial und professionell

WELCHE WORTE (Lexik)	viele Fremdwörter/Fachbegriffe, dabei auch einige kolloquiale Ausdrücke
WAS FÜR SÄTZE	Syntax – sehr lange – verschachtelte Sätze. Frei gesprochene Sprache/ Disfluenzen/ Wiederholungen/ Aposiopesen, sehr schnelles Redetempo (ca. 162 WPM)
WELCHER TON	Suprasegmentale Merkmale – wissenschaftlich - informativ

3.1.2 Ausgangstextkomplexitätsmessung nach Flesch-Kincaid

Neben Textsortenanalyse und Fachbereichsevaluierung wird hier für ein ganzheitliches Bild auch der Schwierigkeitsgrad des Ausgangstextes mithilfe der Flesch-Kincaid-Lesbarkeitsformel gemessen. Liu & Chiu (2009: 248) bestätigten, dass Ausgangstexte für Dolmetschungen schwierig zu messen sind, da vergleichbare intrinsische Elemente nicht leicht quantifizierbar sind. In ihrer Arbeit im Jahre 2009 versuchten sie sich daran, ein Modell zu entwickeln, durch das eine solche Evaluierung durchgeführt werden könnte. Dazu maßen sie unterschiedliche Elemente wie Informationsdichte und Lesbarkeitsindextoken und bezogen zudem eine ExpertInnenevaluierung mit ein. Der Lesbarkeitstoken ist ein potenzieller Indikator für Wort- und Satzlänge, welcher wesentlich zum Leseverständnis eines Textes beiträgt. Da hierbei jedoch von mündlich vorgetragenen Ausgangstexten ausgegangen wird, stellte sich die Frage, ob Lesbarkeit auch ein Indiz für Hörbarkeit ist.

Die Frage, ob Lesbarkeitsformeln auch für die Messung von gesprochenen Texten verwendbar sind, wurde in der Literatur viel diskutiert. William H. DuBay (2007: 12) kam zu dem Schluss, dass Rudolf Fleschs Lesbarkeitsformel von 1943 ein Instrument ist, welches vor allem dann hilfreich für die Spracherkennungsevaluierung ist, wenn die speziellen Merkmale des Hörverständnisses in Betracht gezogen werden.

Aus diesem Grund wurde der Ausgangstext für diese Arbeit nach dem *Flesch Reading Ease Score* sowie dem *Flesch-Kincaid Grade Level* evaluiert. Dafür wurden beide angefertigten Transkripte (sowohl die Version mit als auch die ohne Disfluenzen) herangezogen. Nach der Analyse wurden die Ergebnisse miteinander verglichen, um die tatsächliche Schwierigkeit des Textes einschätzen zu können.

Der *Flesch Reading Ease Score* (o.J.) geht von einem System aus, bei dem die Leseleichtigkeit (RE – readability ease), die durchschnittliche Satzlänge (ASL – average

sentence length) sowie die durchschnittliche Anzahl an Silben pro Wort (ASW – average number of syllables per word) miteinander in Beziehung gesetzt werden.

Die Skala der Schwierigkeitsgrade geht von „sehr leicht“ über „leicht“, „relativ leicht“, „Standard“, „relativ schwer“ und „schwer“ zu „sehr verwirrend“. Während „sehr leicht“ mit 10 Punkten zwischen 90 und 100 angelegt ist, liegt der Großteil ab „leicht“ bis „relativ schwer“ mit jeweils 9 Punkten zwischen 89 und 59. „Schwer“ ist zwischen 30 und 49 angesiedelt und „sehr verwirrend“ zwischen 0 und 29.

Die zweite Formel, der Flesch-Kincaid Grade Level (o.J.), gibt an, welcher Bildungsstufe der Text angepasst ist. Hierbei wird in vier Schritten analysiert. Im ersten Schritt wird die Durchschnittsanzahl der Wörter pro Satz errechnet, im zweiten die der Silben pro Wort. Im dritten Schritt werden beide Werte miteinander in Bezug gesetzt und im vierten Schritt in einer mathematischen Formel errechnet. Diese Formel evaluiert das ungefähre Lesealter und die Schulstufe, in der dieser Text eingesetzt werden könnte. Mit anderen Worten wird evaluiert, wie die durchschnittliche Satzlänge ist und wie viele Silben durchschnittlich ein Wort besitzt.

Transkript A (ohne Disfluenzen) wurde mit einem Ergebnis von 57,1 als „relativ schwer zu lesen“ eingestuft. In Bezug auf das Flesch-Kincaid Grade Level wurde ein Wert von 12,9 erreicht, wodurch der Text auf das Niveau von „College“ einstuft werden kann. Transkript B (mit Disfluenzen) wurde mit einem Wert von 56,9 ebenso in die Kategorie „relativ schwer zu lesen“ eingestuft und im Flesch-Kincaid Grade Level bei einem Wert von 13,2 ebenso für das Niveau „College“ eingestuft. Daraus geht hervor, dass das *saubere Transkript* ohne Disfluenzen offenbar geringfügig leichter zu verstehen ist, als das *Originaltranskript* mit Disfluenzen.

3.1.3 Problemauslöser im Ausgangstext

Während der maschinellen Dolmetschung von *Skype Translator* konnte teilweise beobachtet werden, wie bestimmte Wörter, die erkannt wurden, auf dem Bildschirm aufschienen, kurz bevor sie sich in eine neue Wortkombination transformierten. Insgesamt gab es 29 eingeplante Disfluenzen im Ausgangstext. Wie erwartet schienen keine davon im Transkript auf, sondern wurden in der Phase der Sprachkorrektur (*TrueText* – siehe Abschnitt 2.1) eliminiert.

In diesem Kapitel werden Sequenzen dargestellt, die aufgrund ihrer problemauslösenden Charakteristika zur Evaluierung ausgewählt wurden. In der linken Spalte befinden sich die Abkürzungen der Fachtermini, in der rechten sind die Fachtermini mit Zusatzerklärungen versehen.

In Tabelle 3 werden die Fachtermini des Bereichs der Fremdsprachendidaktik aufgelistet. F2 und F3 sind doppelt vorkommende Wörter. Mehrmals bestehen die Ausdrücke auch aus mehrgliedrigen Elementen. In der Phase der maschinellen Übersetzung zeigt sich, ob das Dolmetschsystem die Fachtermini als solche erkennt und sie in ihrer Originalform belässt, oder ob sämtliche Elemente übersetzt werden. Dies tangiert jedoch nicht die Spracherkennung, weswegen hier vor allem darauf geachtet wird, ob die Ausdrücke in ihrer Kombination richtig erkannt werden.

Tabelle 3: Fachtermini

Abk.	**F - Fachtermini**	**Anmerkungen**
F_1	[…] we call that „the **noticing the gap**“ between […]	mehrgliedrig
F_2	[…] the **interlanguage** and the intake […]	2x
F_3	**interlanguage**	siehe F_2
F_4	[…] so we use implicit techniques such as **input flooding** of texts or […]	mehrgliedrig
F_5	[…] **input flooding** of the language of the teacher […]	mehrgliedrig
F_6	[…] you are a **second language acquisition researcher** […]	mehrgliedrig

„Focus on Form“ ist der am häufigsten“ vorkommende Ausdruck im Ausgangstext. In dem für diese Arbeit ausgewählten Textabschnitt kommt er insgesamt zehn Mal in seiner reinen Form und zwei weitere Male abgewandelt vor. Im Fall von F_9 wurde die Mehrzahl von *Focus on Form* und gleichzeitig die Erklärung, dass es sich dabei um ein großgeschriebenes S am Ende des Ausdruckes handelt, mit dem Code (F_10) hinzugefügt (Tabelle 4). Damit wurde vermittelt, dass *Focus on FormS* ein eigener Zugang ist, der sich vom Verständnis des zuvor behandelten *Focus on Form* abgrenzt.

Tabelle 4: Focus on Form als Fachterminus

Abk.	**F – Fachtermini, A - Abkürzung**	**Anmerkungen**
F_7	Focus on Form	Insgesamt 10x
A_1	FOF	
F_8	Focus on Meaning	mehrgliedrig
A_2	FOM	
F_9	Focus on FormS […]	
F_10	[…] with a capital S at the end	Zusatz F_9

Tabelle 5 zeigt zehn Wörter und eine Jahreszahl, die für die Evaluierung der Kategorien „Zahlen, Eigennamen und Kontrollwörter" ausgewählt wurden. Als Kontrollwörter wurden zu diesem Behelfe einzelne Wörter ausgewählt, die nicht als Fachtermini sondern allgemeinsprachlich verstanden werden sollten. Nichtsdestotrotz wird davon ausgegangen, dass die hier gewählten Kontrollwörter vor allem durch ihre phonetische Ähnlichkeit zu anderen Wörtern für das Spracherkennungssystem schwer zu verstehen sind. Tabelle 5 enthält zwei Mal den Eigennamen „Ellis", eine Jahreszahl, eine Nationalbezeichnung sowie sieben der eben genannten Kontrollwörter.

Tabelle 5: Eigennamen, Kontrollwörter und Jahreszahlen

Abk.	**E - Eigennamen, J - Jahreszahlen, K - Kontrollwörter**	**Anmerkungen**
E_1	[…] by **Ellis**	2x
E_2	[…] recent publication by **Ellis** […]	siehe E_1
E_3	[…] many **French** people […]	Nationalität
J_1	[…] **2016**	
K_1	**communicative turn**	mehrgliedrig
K_2	[…] we also try to **vary** the feedback […]	Verb
K_3	[…] the **intake** that comes from the outside […]	Substantiv
K_4	[…] **either** because we think […]	Adjektiv
K_5	[…] that it does **arise** as a teacher or […]	Verb
K_6	[…] because a student makes an **error** […]	Substantiv

Tabelle 6 zeigt die zwei Passagen, durch die die „Intonation determinierende Syntax" gemessen werden soll. Hierbei wurde eine Fragestellung hinzugefügt, welche lediglich aus einem einzelnen Wort besteht. Die darauffolgende kurze Pause wird als Verneinung

der Frage verstanden, wobei es auch möglich ist, dass in besagter kurzer Pause ein verneinendes Kopfschütteln erfolgt sein könnte. Die zweite Passage besteht aus einem Teil, bei dem die interviewte Person das Formulierte humorvoll betonte. Anzunehmen ist, dass die interviewte Person hier einen kleinen persönlichen Scherz einfügt und diese Aussage somit zu einer Textstelle macht, die sich vor allem durch ihre Intonation auszeichnet. Dadurch, dass *Skype Translator* in der Sprachsynthesephase nicht auf die Intonationserkennung ausgelegt ist, sollte hierbei beobachtet werden, ob *Skype Translator* den Satz, der belustigend gemeint war, von den anderen Textteilen abspaltete und durch Zeichensetzung markierte.

Tabelle 6: Intonation determinierende Syntax

Abk.	**I – Intonation determinierende Syntax**	**Anmerkungen**
I_1	I don't know whether you have heard of Focus on Form**, no? Ok, well,** I'll give you a quick explanation then.	Frage
I_2	[…] because your working memory is limited **...even if you are a very smart person** – your working memory is still limited […]	Lachend

Wie in der Textsortenanalyse besprochen, handelt es sich vor allem um einen wissenschaftlich-formellen Text. Nichtsdestotrotz kommt es im Interview bei bestimmten Passagen vor, dass eine etwas informellere Formulierung gewählt wurde, als von der erklärenden Kommunikationsweise zum persönlichen Ansprechen des Gegenübers gewechselt wurde. Der sprachübergreifende Registerunterschied zwischen Englisch und Deutsch (Sie-Form) wird von Wonisch (2017) in der maschinellen Übersetzung bearbeitet. Gleichzeitig können bereits einige Schwierigkeiten in der Spracherkennung beobachtet werden. Drei der behandelten Ausdrücke sind „yeah“ - ein ambivalentes Wort, das im Sprachgebrauch auch als Disfluenz oder als Bejahung wahrgenommen werden könnte. Dadurch, dass es nicht wie die anderen Füllwörter durch *TrueText* entfernt wurde, konnte dieses Element genauer unter die Lupe genommen werden.

Im Fall von „wanna“ sollte beobachtet werden, ob *Skype Translator* diesen Ausdruck in seiner umgangssprachlichen Form erkannte, oder ob es dies zu einem „want to“ umwandelte, da die schriftliche Sprache der mündlichen in *Skype Translator* aus dem Grund vorgezogen wird, dass mehr Paralleldaten diesbezüglich in der Datenbank vorhanden sind. Wie auch von einem Mitarbeiter von *Skype* beschrieben

(vgl. Wendt 2016a), ist es essenziell, die Elemente der mündlichen Sprache auch zu erkennen und einzubringen, ohne jedoch den Sprachfluss zu stören. Aus diesem Grund sind die restlichen Ausdrücke wie „thing“, „stuff“ und „guy“ vor allem solche, die in der mündlichen, umgangssprachlichen Rede verwendet, jedoch in der schriftlichen Sprache tendenziell vermieden werden.

Tabelle 7: Register

Abk.	**R - Register**	**Anmerkungen**
R_1	[…] just give me your email or send me your email in a **text editor thing** […]	
R_2	[…] so if you **wanna** read about it I would suggest starting with this […]	Umgangs-sprachlich
R_3	[…] so the **guy,** […]	
R_4	[…] who wrote all the important **stuff** on Focus on Form […]	Zusatz R_3
R_5	[…] a second language acquisition researcher because in...**yeah** and the role of the working memory […]	Füllwort (3x)
R_6	[…] **yeah**, and that has a lot to do with […]	Siehe R_5
R_7	**Yeah** haha	Siehe R_5

„Intonation determinierende Syntax“ und „Syntax“ sind sich in der hier ausgeführten Analysemethode sehr ähnlich. Dennoch werden die beiden Kategorien getrennt behandelt, da die Vortragsform sich voneinander unterscheidet. Im Fall der Syntax und den damit einhergehenden Problemen geht es vor allem um Sätze, die abgebrochen beziehungsweise neu angefangen wurden oder sich durch Änderung der Äußerung auszeichnen.

Tabelle 8: Syntax

Abk.	**S - Syntax**	**Anmerkungen**
S_1	[…] **because in...yeah** and the role of the working memory […]	Satzabbruch
S_2	[…] because Focus on Form **it has...well** I can send you the publication […]	Satzabbruch
S_3	[…] because **the prob the problem** we get with that is ...attention.	Selbstkorrektur
S_4 S_5	[…] **that's what I what I think** we see in many French people **for example.**	Wiederholung
S_6	[…] and you artificially have to alter your language **to contern contain a certain form** […]	Selbstkorrektur

3.2 Die SprecherInnen

Um den Fokus der geplanten Arbeit aufrecht zu erhalten sowie aus Gründen der Übersichtlichkeit, wurde das ursprünglich zwölfminütige Interview auf ungefähr sechs Minuten gekürzt, ohne jedoch an Kohärenz einzubüßen. Wichtig war es, spezifische Elemente wie Namen, eine Jahreszahl und Satzabbrüche im Text zu behalten.
Für die Aufnahmen der Texte der SprecherInnen wurde eine Methode genutzt, welche „Shadowing" genannt wird und besonders für Dolmetsch-AnfängerInnen eine gute Übung ist. Dabei wird durch aktives Zuhören ein Text in derselben Sprache unmittelbar und so exakt wie möglich wiedergegeben, wodurch die Sprachkompetenz in der jeweiligen Sprache sowie die Prozesskapazität trainiert wird (vgl. Yo 2015).

Für die hier durchgeführte empirische Arbeit wurde das „Shadowing" genutzt, um Texte mit den gleichen Worten und möglichst auch der gleichen Betonung des Ausgangstextes, jedoch mit unterschiedlichen Akzenten, Geschwindigkeiten, und Stimmfrequenzen zusammenzutragen.
Aufgrund der Komplexität des Inhalts wurde zusätzlich zu dem Vorspielen des Ausgangstextes über Kopfhörer den SprecherInnen ein Transkript vorgelegt, während sie das „Shadowing" durchführten.

Das Transkript sollte den SprecherInnen die Möglichkeit geben, so zu lesen, dass Verzögerungslaute und gewisse Pausen beziehungsweise Lachen als solche gelesen werden konnten (siehe Anhang 02). Als Maßnahme zur besseren Lesbarkeit des Transkripts wurde eine Schriftgröße von 16-Punkt gewählt und ein Zeilenabstand von 1,5 eingefügt. Außerdem wurde das dann 4-seitige Transkript ausgedruckt, das weiße Papier auf einen größeren, bunten Hintergrund geklebt und anschließend laminiert. Dadurch wurde das Transkript leichter „greifbar", während gleichzeitig Hintergrundgeräusche wie Papierrascheln bei den Aufnahmen auf ein Minimum reduziert werden konnten (siehe Abbildung 21 im Anhang).

Um allen SprecherInnen dieselben Informationen bezüglich ihrer Aufgaben zur Verfügung zu stellen, wurden zusätzlich zur mündlichen Instruktion drei Auftragskarten im A5-Format mit den Anleitungen (siehe Abbildung 22 in Anhang 03) dazugegeben. Die erste Auftragskarte informierte darüber, dass es sich um das Transkript eines Monologes handelte, das vorzulesen ist, während der Originaltext ertönt. Dabei wurde das Tempo der Originalaufnahme dem Lesetempo der SprecherInnen angepasst, sodass

diese den Text ohne großen Zeitdruck vorlesen konnten. Des Weiteren wurden sie auf der ersten Karte dazu aufgefordert, sich mit dem Text und der Aufnahme vertraut zu machen, bevor es zur tatsächlichen Sprachaufnahme kam. Mithilfe der Auftragskarten wurden alle SprecherInnen über sämtliche Schritte informiert. Die zweite Auftragskarte gab Informationen darüber, wie der Text vorgelesen werden sollte. Es wurde darüber informiert, dass es bei dieser Aufgabe vor allem darum ging, verschiedene Akzente miteinander vergleichen zu können. Aus diesem Grund wurden die SprecherInnen dazu aufgefordert, sich auf eine ihnen möglichst natürliche Aussprache zu konzentrieren und sich nicht an der Betonung der Aufnahme zu orientieren. Außerdem wurden sie instruiert, alle Füllwörter und die im Transkript angegebenen Informationen bezüglich „humorvollem Ausdruck“ oder Wortabbrüchen ebenfalls mitzusprechen. Zusätzlich sollten sie versuchen, mit der Geschwindigkeit der Originalaufnahme (welche an ihre jeweiligen Tempi angepasst wurden) mitzuhalten und möglichst auch die Satzmelodie der Originalaufnahme nachzuahmen. Es wurde ihnen freigestellt, den Text in mehreren Abschnitten vorzulesen beziehungsweise nach bestimmten Absätzen zu pausieren oder den Text nur einmal oder mehrmals vorzusprechen. Das mehrmalige Vorsprechen hatte den Vorteil, dass die SprecherInnen dadurch mit dem Text vertrauter wurden und diesen fehlerfrei lesen konnten.

Schlussendlich wurden die Aufnahmen zusammengeschnitten oder jene Aufnahmen zur Auswertung herangezogen, die inhaltlich vollständig waren. Auf der dritten Auftragskarte wurden die SprecherInnen gebeten, Informationen bezüglich ihrer Geburtsorte, Geburtsdaten, ihrer derzeitigen und ehemaligen Wohnorte sowie ihre Einschätzung darüber, wo sie ihre persönlichen Akzente im Englischen zuordnen würden, preiszugeben.

3.2.1 Rahmenbedingungen

Die Aufnahmen der SprecherInnen wurden an sehr unterschiedlichen Orten durchgeführt, was ungleiche Ausgangssituationen bezüglich der Störgeräusche im Hintergrund darstellte. Die Aufnahmen wurden mit dem externen Aufnahmegerät ZOOMh1 durchgeführt, welches über die zusätzliche Funktion der Hintergrundgeräuscheliminierung verfügt und als semiprofessionelles Aufnahmegerät vertrieben wird. Die Tonaufnahmen wurden jeweils in einem .wav-Format aufgenommen und mit dem Programm *audacity* bearbeitet.

Unter den Aufnahmeorten befanden sich Innenräume von Wohngebäuden (Küche, Wohnzimmer, Schlafzimmer – ohne merkbarem Hall bei der Aufnahme), ein Frisörsalon (deutlich hörbare Hintergrundgeräusche, kein Hall), ein Gemeinschaftsaufenthaltsraum (ohne zusätzliche Personen im Raum mit starkem Hall) mit Spielplatz davor (leise Hintergrundgeräusche) Seminarräume (ohne akustische Absonderlichkeiten), der Billardraum eines SeniorInnenzentrums (leise Hintergrundgeräusche, kein Hall) und noch viele andere. Die Aufnahmen wurden mit geschlossenen DJ-Kopfhörern der Firma Stagg vorgespielt. Von den 32 Aufnahmen wurden insgesamt 8 in Wien und Wien-Umgebung durchgeführt, 2 in Ennis, Irland, 8 in Las Vegas, Nevada, USA und 14 in Boulder, Colorado, USA.

3.2.2 Samplingbeschreibung

Die SprecherInnen wurden mithilfe des Schneeballprinzips akquiriert. Ausschlusskriterium für die Sprachaufnahme war ein unzureichendes Englischniveau. Mit anderen Worten mussten alle SprecherInnen über sehr gute Englischkenntnisse verfügen und sich in der englischen Sprache gut ausdrücken können. Das Englischniveau wurde jedoch nicht extra getestet, sondern ergab sich im mündlichen Gespräch und aus der Tatsache heraus, dass die SprecherInnen das Transkript und den Ausgangstext in Audioform verstehen und ihn nachvollziehen konnten.

Die SprecherInnen hatten sehr unterschiedliche professionelle Hintergründe. Unter ihnen waren (Physik-, Dolmetsch-, Wirtschaft-) StudentInnen, ein Frisör, PensionistInnen, PhysikerInnen, eine Unternehmerin, eine Lehrerin, ein Berater, ForscherInnen beziehungsweise WissenschafterInnen.

3.2.2.1 Herkunftsorte

Abbildung 2 zeigt, dass der Großteil der SprecherInnen vor allem aus dem nordamerikanischen und europäischen Raum stammt. Genauer gesagt, stammen rund 42%, nämlich 13 der 32 SprecherInnen, aus den USA oder Kanada und sprachen amerikanisches Englisch. 5 SprecherInnen, also ca. 16%, stammen aus englischsprachigen Ländern in Europa, also Großbritannien, Schottland sowie Irland. 14 SprecherInnen, also rund 45 %, gaben an, dass Englisch nicht ihre Muttersprache ist. Diese SprecherInnen kommen aus den Ländern Österreich, Brasilien, Deutschland, Griechenland, Taiwan, Bulgarien, Spanien, Italien, Schweden, Frankreich und Uruguay.

Abbildung 2: Herkunftsländer

Wie auf dieser Grafik ebenfalls zu sehen ist, sind die SprecherInnen eine relativ heterogene Gruppe. Die Ortszuschreibungen sind darauf bezogen, wo die SprecherInnen die längste Zeit ihrer Kindheit verbracht haben und nach eigenen Angaben aufgewachsen sind. In vier von den 32 Fällen ist dies nicht der Geburtsort. In Abbildung 2 können bei genauem Hinsehen weniger als 32 Pfeile ausgemacht werden, was darauf zurückzuführen ist, dass einige der Personen in derselben Stadt aufgewachsen sind (beispielsweise London oder Wien).

3.2.2.2 Stimmfrequenz

Um etwaige Verständnisprobleme seitens *Skype Translator* aufgrund von untypischen Stimmfrequenzlagen bei dieser empirischen Untersuchung auszuschließen, wurde die Durchschnittsfrequenz aller SprecherInnen mithilfe des Phonetikprogrammes PRAAT (vgl. Boersma & Weenink 2006) evaluiert. Obwohl es in der Praxis bei der Untersuchung von Stimmfrequenzen üblich ist, nur Sequenzmessungen vorzunehmen, wurde in dieser Arbeit der Durchschnitt über den gesamten gelesenen Text gemessen, da es hier – im Gegensatz zu den üblichen Untersuchungen im Zusammenhang mit Stimmfrequenzmessungen – nicht um das Ausmachen unterschiedlicher Emotionen anhand von Stimmhöhenunterschieden ging, sondern um das Vermeiden von verfälschten Ergebnissen aufgrund anormaler Stimmlagen.

Die durchschnittliche Stimmfrequenzlage von Männerstimmen befindet sich,wie im theoretischen Teil beschrieben (siehe Abschnitt 2.2.2.2) zwischen 100 und 150 Hz, die der Frauen zwischen 190 und 250 Hz und die der Kinder zwischen 350 und 500 Hz (vgl. Pétursson & Neppert 2002: 125f). Die Grafik zeigt, dass bei Männern und Frauen jeweils relativ tiefe Stimmen vertreten waren. Dadurch, dass alle SprecherInnen über 19 Jahre alt sind, ist nachvollziehbar, warum hier keine „Kinderstimmfrequenzen“ vertreten sind.

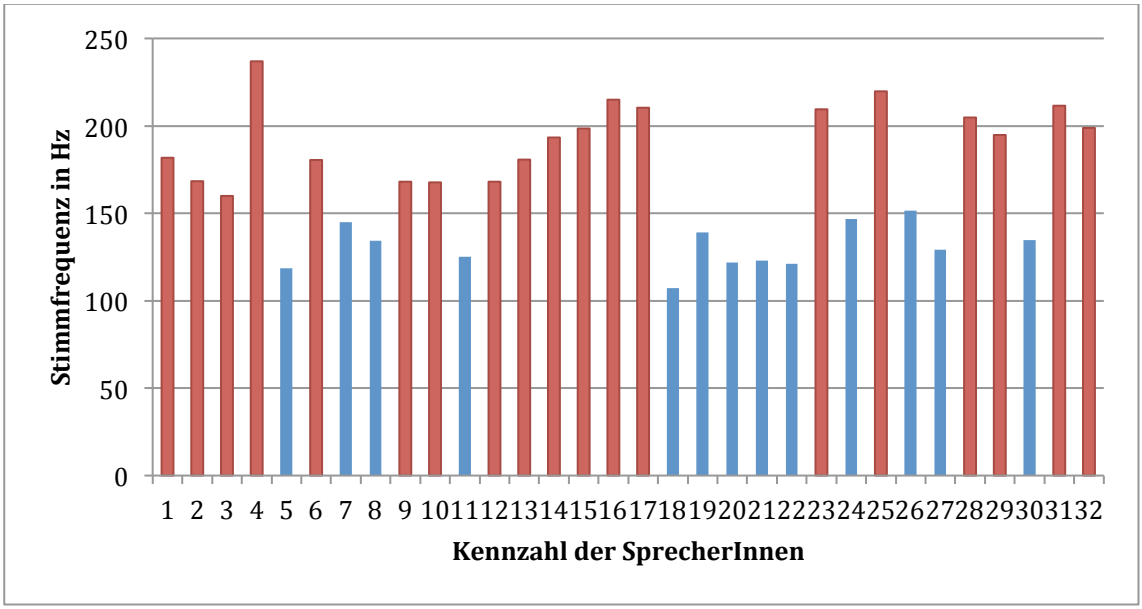

Abbildung 3: Übersicht zur durchschnittlichen Stimmfrequenz der SprecherInnen in Hz. Blaue Balken: Männer; Rote Balken: Frauen.

3.2.2.3 Alter

In der heterogenen Gruppe der SprecherInnen war die jüngste Person 20, während die zwei Ältesten, beide im Alter von 93, schon als hochaltrig bezeichnet werden können. Das Durchschnittsalter der SprecherInnen liegt bei etwa 43 Jahren, die meisten SprecherInnen sind aber zwischen 20 und 30 Jahre alt. Da sich alle Stimmfrequenzen im Normalbereich befinden, ist anzunehmen, dass die Sprechproben von der Datenbank abgedeckt werden und die Stimmfrequenz demzufolge als potenzieller Problemauslöser für *Skype Translator* auszuschließen ist. Einzig die Stimmfrequenz des Sprechers mit der Nummer 18 ist auffallend niedrig, verlässt aber nicht den Normbereich.

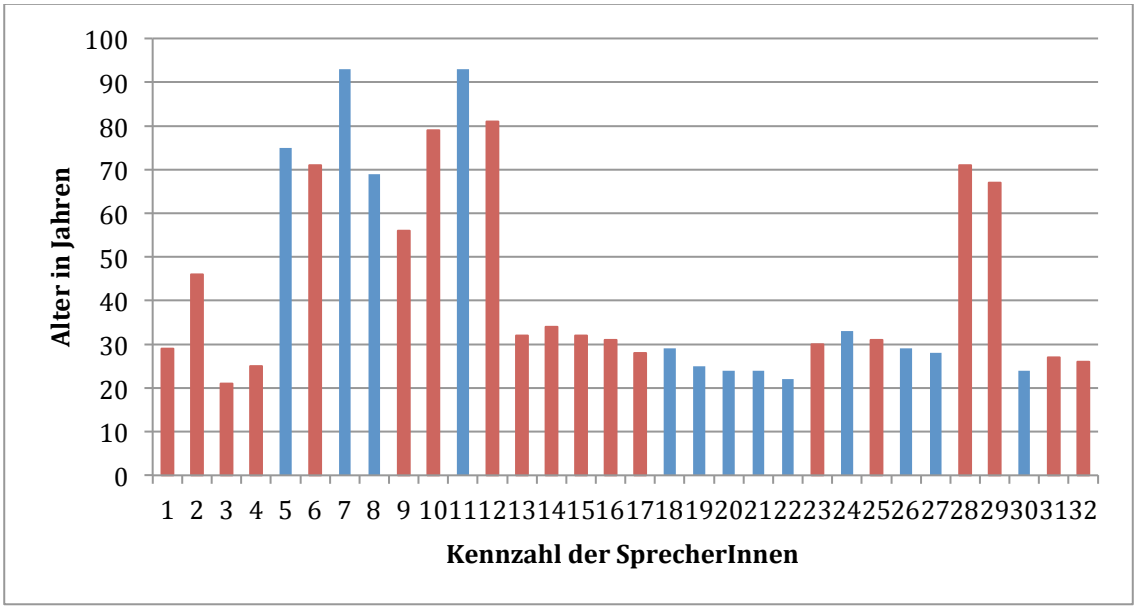

Abbildung 4: Übersicht über das Alter der teilnehmenden SprecherInnen.
Blaue Balken: Männer; Rote Balken: Frauen

3.2.2.4 Redegeschwindigkeit

Ein wichtiges Kriterium für die Dolmetschbarkeit eines Textes ist auch die Redegeschwindigkeit. Da die Idealgeschwindigkeit zum Dolmetschen (siehe Abschnitt 2.2.2.1) bei etwa 120 WPM liegt, ist auf den ersten Blick zu sehen, dass die SprecherInnen in dieser Untersuchung vergleichsweise sehr schnell sprachen. Die violetten Balken in Abbildung 5 geben die Reden an, die mit den Transkripten gemessen wurden, welche sämtliche Disfluenzen beinhalten, während die orangen Balken jene darstellen, aus denen die Disfluenzen entfernt wurden. Das erste Transkript weist 1078 Wörter auf, das zweite 1031 Wörter. Die Originalrede (Nummer 1) wurde

mit dem ersten Transkript in einem Tempo von 169 WPM vorgetragen. Sechs weitere SprecherInnen absolvierten das „Shadowing“ in einem vergleichbar schnellen Tempo (168-169 WPM). Bei den Messungen mit dem zweiten Transkript verlangsamten sich sämtliche Reden um rund 5 bis 8 Wörter pro Minute.

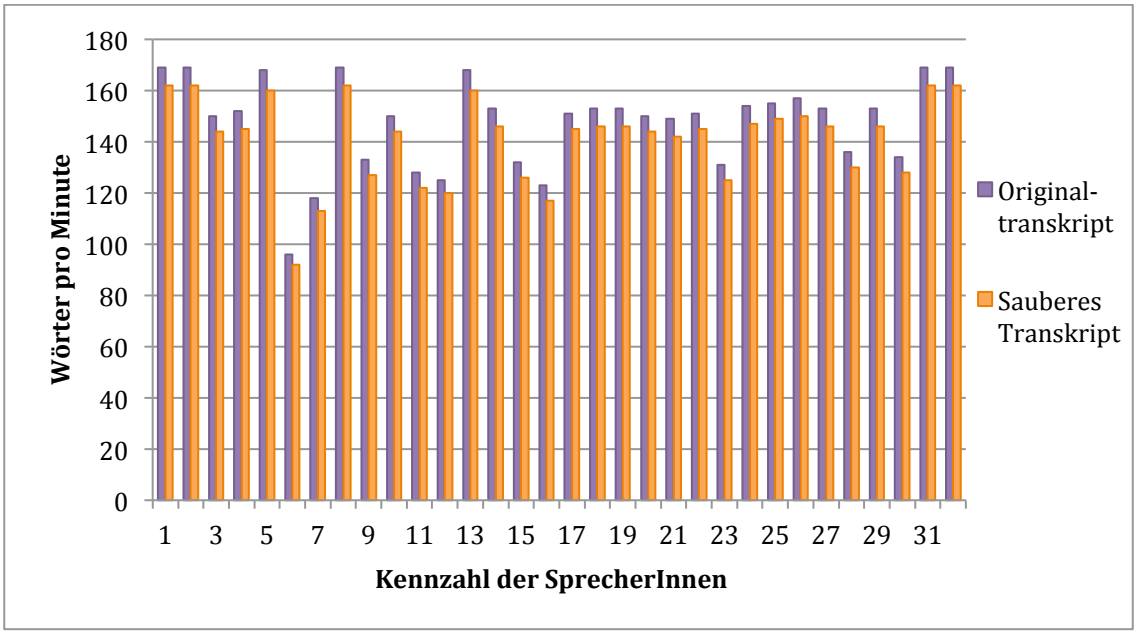

Abbildung 5: Übersicht über die Sprechgeschwindigkeit in WPM (gesprochenen Wörtern pro Minute)

3.2.2.5 Aufnahmeintensität

Es ist anzunehmen, dass die Intensität, also die Sprechlautstärke der SprecherInnen, von den örtlichen Gegebenheiten und Hintergrundgeräuschen bei der Aufnahme beeinflusst wurde. Für sämtliche Aufnahmen wurde immer das gleiche Aufnahmegerät verwendet. Ebenso wurde darauf geachtet, dass der Abstand zwischen Mikrofon und SprecherInnen ähnlich weit gehalten wurde und das Aufnahmegerät stets auf einer festen Oberfläche auflag und somit keine Positionsveränderung erfolgte.

Obwohl das verwendete Mikrofon auch eine Funktion zur Dämmung der Hintergrundgeräusche hat, die auch für die Aufnahmen aktiviert wurde, muss man davon ausgehen, dass sich dieser Umstand in der Tonqualität der Aufnahmen wiederspiegelt.

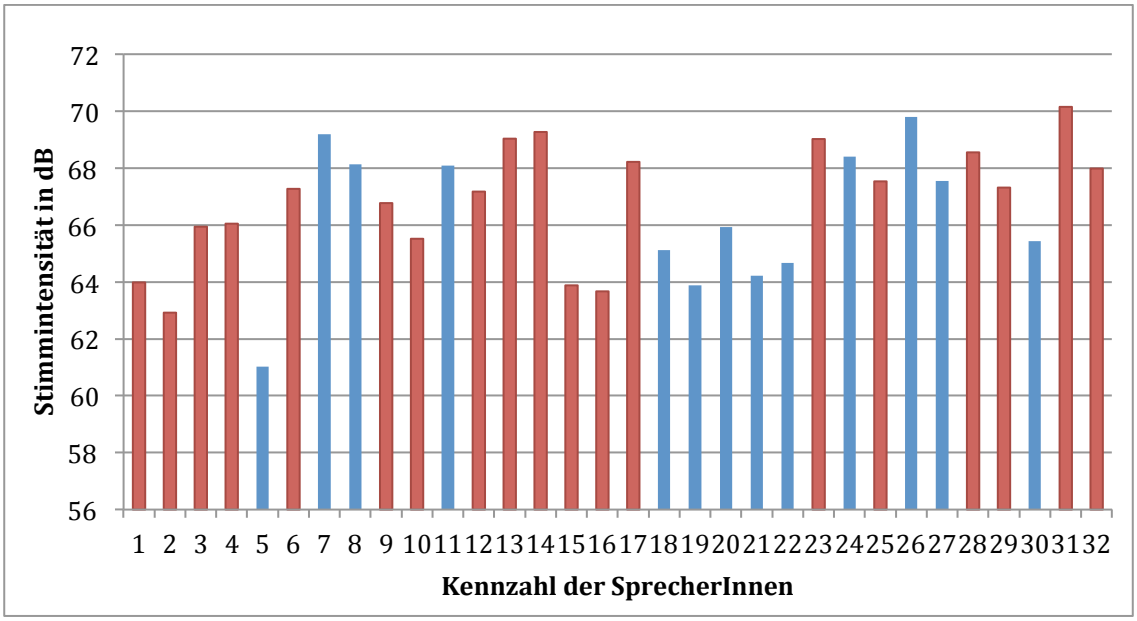

Abbildung 6: Intensität der Aufnahmen in dB. Blaue Balken: Männer; Rote Balken: Frauen.

3.3 Ablauf/Aufzeichnung der Ergebnisse

Die Transkripte der Skype-Dolmetschungen wurden generiert, als die Sprachaufnahmen der SprecherInnen als direkte Input-Quelle während des Skype-Gesprächs in den Computer gespielt wurden. Dafür wurden zwei Skype-Konten auf unterschiedlichen Geräten genutzt, durch die eine Telefonverbindung mit aktiviertem *Skype Translator* hergestellt wurde. Anschließend wurden die Audiodateien der SprecherInnen als Eingabequelle für die Unterhaltung von einem der Konten eingespielt. Dieser Vorgang wurde wiederholt, da die Dolmetschung des *Skype Translator* trotz identer Audiodatei nicht immer die gleichen Ergebnisse lieferte. Auf diese Weise konnte der Aspekt des automatisierten Lernens in der Analyse berücksichtigt werden. Der erste Durchlauf wurde an zwei aufeinanderfolgenden Tagen Ende Mai durchgeführt. Für jede Audiodatei wurde eine eigene Telefonverbindung via Skype hergestellt, nach etwa 5 Minuten abgebrochen und anschließend neu verbunden. Das war erforderlich, da sich herausstellte, dass *Skype Translator* nach etwa 6 Minuten nicht mehr „dolmetschte“. Der zweite Durchgang erfolgte Mitte Juni ebenso an zwei aufeinanderfolgenden Tagen. Dieselbe Methode wie für Durchgang 1 wurde auch hier mit einem Unterschied angewandt: der Abbruch und die Wiederherstellung der Verbindung erfolgte bereits nach etwa 3 Minuten. Der Grundgedanke war, dass, sollte dieser Verbindungsabbruch einen wesentlichen Einfluss auf die Leistung von Skype Translator haben, die

beeinflussten Sequenzen in beiden Durchläufen nicht übereinstimmen und dadurch nicht zu fehlgeleiteten Schlussfolgerungen führen. Wie in der Theorie beschrieben, siehe Abschnitt 2.1.1, wird für die Evaluierung der Spracherkennung der WER-Wert genutzt. Diese WER-Evaluierungsmethode wurde nicht für diese Arbeit genutzt, weil einerseits der Aufwand der Originaltransskripterstellung sämtlicher Sprachaufnahmen in diesem Fall zu groß wäre, andererseits der WER-Wert in der Kritik steht, syntaktische Elemente nicht zu berücksichtigen. Vielmehr versucht diese Arbeit bewusst, die Spracherkennung des Skype Translator von einem anderen, in diesem Fall dolmetschwissenschaftlichen Blickwinkel aus zu bewerten.

3.4 Das Analysevorgehen

Wie beschrieben, wurden die 32 Aufnahmen der SprecherInnen dem Programm *Skype Translator* zwei Mal vorgespielt, um das maschinelle Lernen ebenso in die Ergebnisse miteinzubeziehen.

Die „gedolmetschten“ Ergebnisse von *Skype* wurden dann über Transkripte ausgegeben, welche hinsichtlich ausgewählter Sequenzen und Wörter analysiert und verglichen wurden (siehe Kapitel 4). Für die quantitative Datenanalyse wurden vier mögliche Bewertungskriterien gewählt, denen sämtliche problemauslösende Textstellen zugeteilt wurden. Wie in der Einleitung des Methodenkapitels beschrieben, zählen dazu JA, JEIN, NEIN und LEER. Diese Bewertungskriterien wurden wiederum in Skalen eingebettet, die angeben, nach welchem Schema die Elemente bewertet wurden. Die vier Skalen richten sich nach den Kriterien der Semantik, der Phonetik sowie der Typologie und dienten auch zur Analyse der Reaktionen von *Skype Translator* in Sonderfällen.

Demnach wird bei Skala 01 das Bewertungskriterium JA als „semantisch exakt übereinstimmend/nachvollziehbar”, JEIN als „semantisch übereinstimmend/ nachvollziehbar“, NEIN als „semantisch nicht übereinstimmend/nachvollziehbar“ und LEER als „Daten nicht identifizierbar/nicht vorhanden“ gewertet. Dementsprechend wurde beispielsweise der Begriff „Focus on Form“ dann als JA bewertet, wenn der Begriff exakt wiedergegeben wurde. Groß- und Kleinschreibung wurde in dieser Skala nicht beachtet, da sie keinen direkten Einfluss auf das mündliche Ergebnis hatte. Unter JEIN wurden Ergebnisse wie „focus on the form“ eingegliedert, da es dem gewünschten

Ergebnis nahe kommt, aber nicht dem exakten Wortlaut entspricht. NEIN wurde beispielsweise im Fall von „forks on phone“ zugeteilt, da die Bedeutung vollkommen verändert wurde. LEER waren sämtliche Daten, die nicht als bewertbar ausgemacht werden konnten oder einfach nicht vorhanden waren. Dies ist darauf zurückzuführen, dass entweder *Skype Translator* die Daten nicht als solche erkennen konnte, oder auch darauf, dass bereits bei den Aufnahmen und dem „Shadowing“ vereinzelt Schwierigkeiten auftraten. Zur besseren Nachvollziehbarkeit der Bewertungsmethode zeigen die Tabellen 9, 10 und 11 Beispiele der jeweils erklärten Skalen. Sämtliche Elemente mit ihren jeweiligen Bewertungen sind in Anhang 01 angeführt.

Tabelle 9: Skala 01- Beispiele der Bewertungsmethode

SKALA 01* - Semantisch		**JA**	**JAIN**	**NEIN (Beispiele)**
F_1	„noticing the gap“	noticing the gap	noticing that gap; noticing gap	notion that gap, nothing they can, notion the gap, thing that gap, notice in the gap, noticing Jack, nursing the gap, nazi that ***, nursing gap, nursing yet, not dissing the gap, not a single gap,
E_1 E_2	„Ellis“	Ellis		Alice, hours, others, allah’s, enlist, ends, list, vieilleus, hours, Harris, at least, endless, palace, Diana’s
K_2	„vary“	vary		very, Barry, bury, pray, bari, buried

Bei Skala 02 werden die Bewertungskriterien folgendermaßen zugeordnet: JA wird als „phonetisch exakt nachvollziehbar“, JEIN als „phonetisch nachvollziehbar“ und NEIN als „phonetisch nicht nachvollziehbar“ interpretiert. LEER impliziert, dass die Daten „nicht identifizierbar oder nicht vorhanden“ waren. Beispielsweise wurde der Fall von F_10 „with a capital S at the end“ als Sequenz analysiert. Dadurch, dass *Skype Translator* einzelne Buchstaben nicht ausmachen kann, wurden Wörter erkannt, die eine phonetische Ähnlichkeit mit den Buchstaben haben. Aus diesem Grund wurde Skala 02 entwickelt, wodurch JA jenen Fällen zugeordnet wurde, in denen sämtliche Wörter außer dem einzelnen „S“ richtig erkannt wurden und eine phonetische Äquivalenz zu diesem erkannt wurde, wie zum Beispiel „with a capital as at the end“. Die Ergebnisse wurden JEIN zugeordnet, wenn mehrere Silben als nur das „S“ nicht richtig erkannt wurden. Ein Beispiel dazu ist „with a capital asset the end“. NEIN wurde zugeordnet, wenn das Original nicht mehr erkennbar war, wie zum Beispiel bei „with the captured at

that the end“. Und LEER kam wie bei Skala 01 dann zum Einsatz, wenn keine Daten abgeleitet werden konnten.

Tabelle 10: Skala 02 - Beispiele der Bewertungsmethode

SKALA 02 – Phonetisch		**JA**	**JEIN**	**NEIN**
A_1	FOF	FOF, SOS, FAO,	affo effe, Sos,	more Sos, more ffff, for every half, core more Fomc, more effort of length, our death forever, forever,

Mit Skala 03 wurden typologische Bewertungen durchgeführt. Das heißt, es wurde versucht zu eruieren, ob *Skype Translator* beispielsweise Satzabbrüche erkannte und die jeweiligen Elemente, die kurz nach dem Abbruch passierten, typologisch von dem Abbruch abgrenzte. In Fällen wie diesen wurden die Ergebnisse dann JA zugeordnet, wenn zum Beispiel im Fall von „because in…yeah. And…“ das „yeah“ rechts und links mit Satzzeichen versehen wurde, während JEIN dann ausgewählt wurde, wenn „yeah“ nur mit einem Satzzeichen links oder rechts versehen war. NEIN wurden sämtliche Daten zugeordnet, die zwar erkannt, aber nicht als sinnvolle Zusammenhänge dargestellt werden konnten, und LEER waren jene Fälle, die nicht als Daten ausgemacht werden konnten.

Tabelle 11: Skala 03 - Beispiele der Bewertungsmethode

SKALA 03* - Satzzeichen		**JA**	**JEIN**	**NEIN**
S_2	it has ... well, I	,well, ; .well, ;	Has well, ; ,well I;	has what I can send you to, as well. And; it hurts well
S_5	for example.	,for example	. For example	

Die vierte und letzte Skala bezieht sich auf Spezialfälle wie die „Selbstkorrektur“, bei denen auch zugelassen wurde, dass einzelne Elemente nicht dargestellt wurden, wie beispielsweise im Fall von „the prob- the problem“. Die Ergebnisse wurden JA zugeordnet, wenn entweder „the problem“ als solches erkannt wurde oder „the prob- the problem“ als gesamte Sequenz erkannt wurden. Der Gedanke dahinter war, dass die Eliminierung des „prob-“ den Sinn nicht stört. Die Daten wurden dann JEIN zugeordnet, wenn beispielsweise „the pro- the problem“ erkannt wurde, da „pro“ gegebenenfalls als die Abkürzung von „professional“ erkannt werden könnte und dies eine inhaltliche Veränderung mit sich bringen würde. NEIN kam wie in den anderen

Fällen dann zum Einsatz, wenn sich das Ergebnis vollkommen vom Original unterschied und LEER, wenn die Daten als solche nicht erkannt werden konnten.

Tabelle 12: Skala 04 - Beispiele der Bewertungsmethode

SKALA 04* Spezielle Fälle		**JA**	**JEIN**	**NEIN**
S_6	the prob- the problem	„the prob- the problem“	The prob problem, the problem, the prob- problem, the problem the problem	The pro- the problem; the prop; the problem
	„what I what I think“	That’s what I think; that’s what I what I think		what I would I think, when I think I said, That’s why I don’t think, with the cops to I said, yes, that’s what I am what I, that’s what I want. I think we see. what I would I think, when I think I said, That’s why I don’t think, with the cops to I said, yes, that’s what I am what I, that’s what I want. I think we see.

Zusammenfassend kann man sagen, dass sämtliche Skalen vor allem danach ausgerichtet waren, den „Sinn“ der Aussage in die Phase der maschinellen Übersetzung übertragen zu können. Dadurch, dass dies wie zum Beispiel bei den Akronymen nicht in allen Fällen möglich war, wurden die Skalen an ein Niveau angepasst, das die vorgegebenen Inhalte am ehesten nachvollziehen lässt. Zur besseren Visualisierung der beschriebenen Ergebnisse wurde zusätzlich ein Score herangenommen, für den die Auswertungen der Kategorien zusammengezogen wurden. Dabei wird JA durch einen Wert von +1 repräsentiert, JEIN mit +0,5, NEIN entspricht Wert -1 und LEER einen Wert von 0. Der hier genutzte Score geht davon aus, dass die Erkennung der Informationen zur Weiterbearbeitung das höchste Gut sind und falsch erkannte Informationen einen negativeren Einfluss auf die Dolmetschung haben können als nicht erkannte Informationen, wobei dies natürlich zur Diskussion steht. Die Berechnungen der Gesamt-Scores der Kategorien erfolgte nach folgender Formel: JA+ JEIN – NEIN. Die Ergebnisse der Formel in den jeweilig angeführten Tabellen dienen einerseits zur Visualisierung der Gegenüberstellung von Transkript 1 und Transkript 2 und dadurch dem automatisierten Lernen, andererseits soll damit auf einen Blick dargestellt werden, ob und in welchem Ausmaß die exakt oder fast richtig erkannten Daten überwiegen oder ob es sich genau andersherum verhält.

4 Ergebnisse

Wie in Abschnitt 3.1 bereits angeführt, wurden die Ergebnisse nach den Analysekriterien „Fachtermini“, „Focus on Form“ als spezieller Fachterminus, „Akronyme“, „Eigennamen“, „Kontrollwörter“, „Jahreszahlen“, „Intonation determinierende Syntax“, „Registerwechsel“ und „Syntax“ dargestellt. Die Werte in den angegebenen Abbildungen sind in natürlichen Zahlen angegeben. Die angegebenen Prozentwerte im Text wurden gerundet. Im Anschluss an die jeweiligen Ergebnisse pro Kategorie wurde der in Kapitel 3.4 beschriebene Score errechnet.

4.1 Fachtermini

In der Kategorie „Fachtermini“ wurden sechs Elemente ausgewählt, von denen zwei im Ausgangstext doppelt vorkamen. Sämtliche Problemauslöser in Abbildung 7 und 8 wurden mithilfe der Skala 01 ausgewertet. Wie man in Abbildung 7 auf den ersten Blick erkennen kann, ist vor allem der NEIN-Anteil herausragend. Insgesamt wurden im ersten Durchlauf also bei 32 Transkripten und jeweils sechs evaluierten Termini eine Anzahl von 54 JA, 91 NEIN und 15 JEIN erreicht. 32 der Begriffe konnten nicht als Daten erkannt werden. Auffallend ist, dass F_6, „second language acquisition researcher“, als einziger Fall mit 17 JA, 6 NEIN und 7 JEIN ausgezählt wurde und damit der JA- und der JEIN-Anteil stärker vertreten sind als der NEIN-Anteil. Es ist anzunehmen, dass die Erkennung aufgrund der höheren Anzahl an Schlüsselwörtern, die miteinander in Verbindung gesetzt werden können, höhere Ergebnisse erzielt hat.

In Abbildung 8, die die Auswertung des zweiten Transkripts zeigt, ist eine starke Veränderung zu den Ergebnissen von Transkript 1 zu erkennen. Die Daten, die in Transkript 2 aufscheinen, sind dieselben Sequenzen und Audiodateien wie in Transkript 1, mit dem Unterschied, dass *Skype Translator* die Audiodateien schon 32 Mal analysieren konnte und somit das automatisierte Lernen (siehe Abschnitt 1.3) seine Wirkung zeigen müsste.

Wie in Abbildung 8 deutlich zu erkennen ist, haben sich die JA-Werte drastisch erhöht. Insgesamt wurden 98 JA, 65 NEIN und 14 JEIN verzeichnet; 15 konnten nicht identifiziert werden. Während die Summe der JEIN-Werte also sehr ähnlich blieb (wobei die „JEIN-Fälle“ nun anders verteilt sind als zuvor), haben sich die nicht

identifizierbaren Daten zwischen Transkript 1 und 2 um nahezu 50% verringert. Während beim Transkript 1 noch bei drei verschiedenen Messungen über fünf (bei F_4, „input flooding“ sogar zehn) Elemente nicht verarbeitet werden konnten, gab es beim Transkript 2 keine Rubrik, in der fünf Aufnahmen nicht ausgewertet werden konnten. F_5 ist in Transkript 2 das einzige Element, dessen JA-Wert weiterhin vom NEIN-Wert überragt wird. Im Vergleich zu Transkript 1 erhöhte sich der JA-Wert jedoch signifikant, der NEIN-Wert hingegen nur marginal.

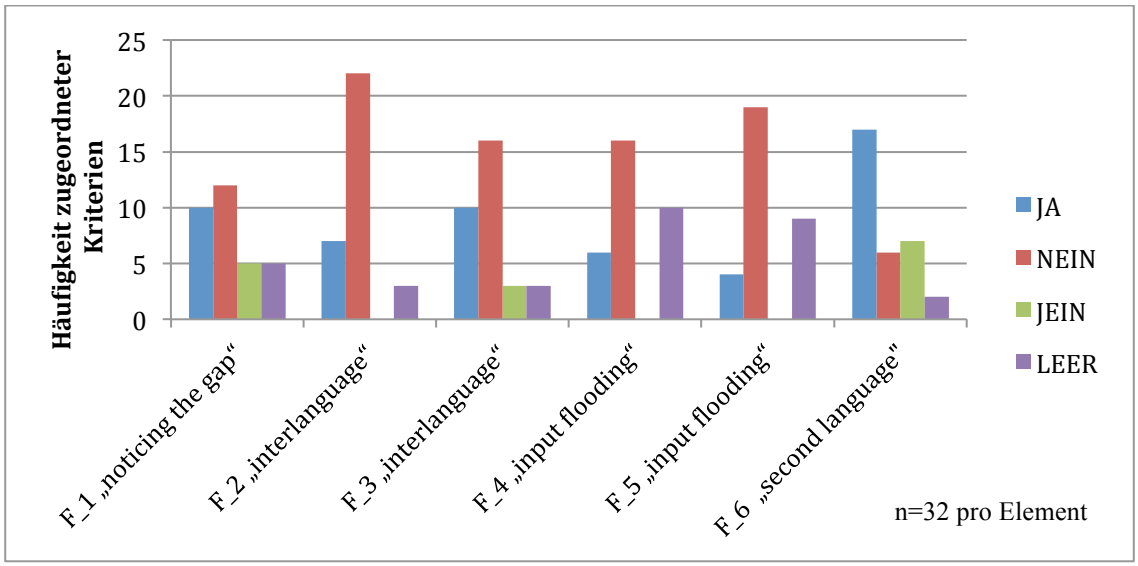

Abbildung 7: Auswertung der Problemfälle, die zu den „Fachtermini“ im Transkript 1 überprüft wurden

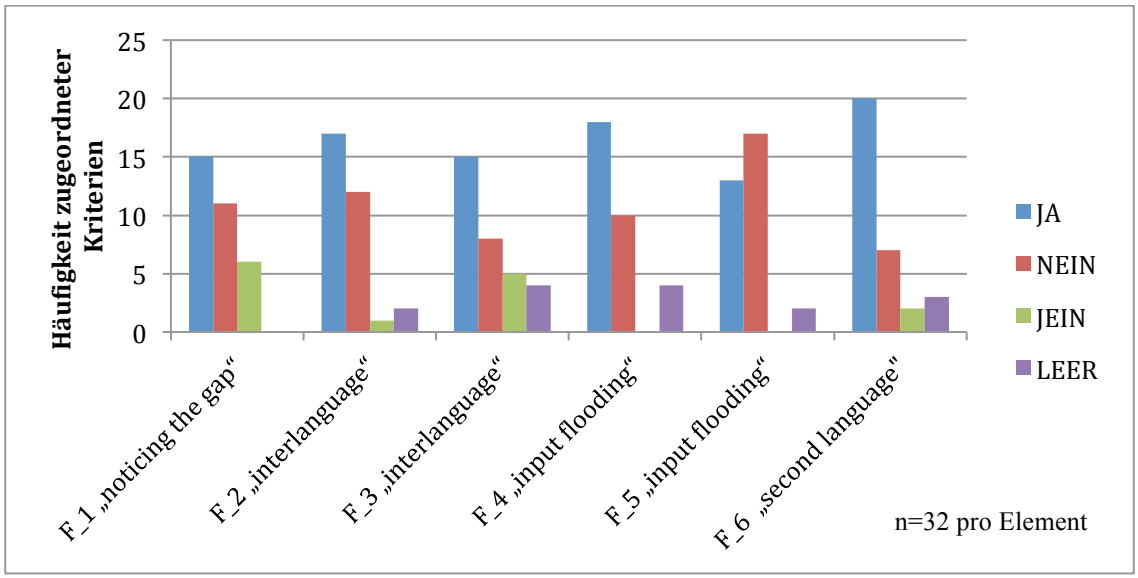

Abbildung 8: Auswertung der Problemfälle, die zu den „Fachtermini“ im Transkript 2 überprüft wurden

Nach dem im Kapitel 3.4 beschriebenen Score (JA +1; JEIN +0,5; NEIN -1, LEER 0) fielen die Ergebnisse der Kategorie „Fachtermini“ bei einer Grundgesamtheit von 192 Einträgen in 6 Elementen so aus, dass in Transkript 1 ein deutlicher NEIN-Anteil zu vermessen ist. Dazu verbesserten sich die Werte im Gegenzug um fast 70 Transkripteinträge in Transkript 2, was auf eine deutliche Lernkurve hinweist.

Tabelle 13: Score der Elemente in der Kategorie „Fachtermini“

	SCORE Transkript 1	**SCORE Transkript 2**
F_1	0,5	7
F_2	-15	5,5
F_3	-4,5	9,5
F_4	-10	8
F_5	-15	-4
F_6	14,5	14
GESAMTSUMME (max. Score pro Transkript: 192)	**-29,5**	**+40**

4.2 Focus on Form als Fachterminus

„Focus on Form“ zählt zur Kategorie der Fachtermini, wurde jedoch in dieser Arbeit separat behandelt. Es wurde ebenso wie die zuvor analysierten Elemente mithilfe der Skala 01 analysiert. *Focus on Form* ist ein häufiger Begriff im Ausgangstext und kam insgesamt 10 Mal in der ausgewählten Textstelle vor. Insgesamt ergibt das pro Transkript also 320 Erwähnungen und letztlich 640 analysierte Elemente. Auf den ersten Blick kann in den untenstehenden Kreisdiagrammen sofort erkannt werden, dass sowohl in Transkript 1 wie auch in Transkript 2 der NEIN-Wert dominant ist (Transkript 1 mit 58% und Transkript 2 mit 46% der analysierten Elemente). Die JA-Werte erhöhten sich zwischen Transkript 1 und Transkript 2 um 7%. Der JEIN-Wert erhöhte sich zwischen Transkript 1 und 2 um 3% während sich der Wert an nicht identifizierbaren Daten ebenfalls um 3% erhöhte. Das führt zu dem überraschenden Ergebnis, dass der Ausdruck „Focus on Form“ oder abgewandelte Versionen davon beim zweiten Durchlauf seltener identifiziert werden konnten als beim ersten.

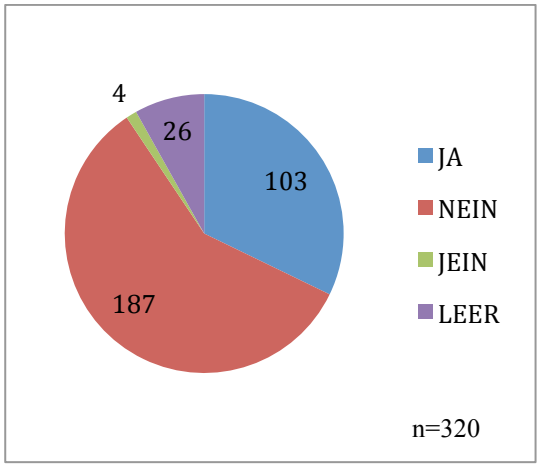

Abbildung 9: Auswertung der Problemfälle, die zum Fachterminus „Focus on Form“ im Transkript 1 überprüft wurden

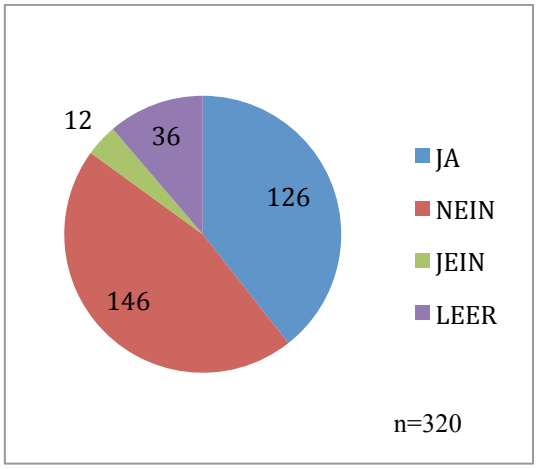

Abbildung 10: Auswertung der Problemfälle, die zum Fachterminus „Focus on Form“ im Transkript 2 überprüft wurden

Auch die Kategorie des speziellen Fachterminus „Focus on Form“ zeigte bei einer Grundgesamtheit von 320 Einträgen bei nur einem Element eine deutliche Leistungsverbesserung, obwohl in beiden Transkripten der NEIN-Anteil die erkannten Daten überragte.

Tabelle 14: Score der Elemente in der Kategorie „Focus on Form als Fachterminus“

	SCORE Transkript 1	**SCORE Transkript 2**
F_7 GESAMT (max. Score pro Transkript: 320)	**-82**	**-14**

4.3 Akronyme und spezielle Fachtermini

Die folgende Analyse bezieht sich auf Elemente, die unterschiedlichen Analysekategorien zugeordnet werden könnten, jedoch aufgrund ihrer inhaltlichen Übereinstimmung gemeinsam in einer Darstellung visualisiert wurden. F_8 und F_9 wurden beide mithilfe von Skala 01, also hinsichtlich silbengetreuer Darstellung, bewertet. A_1 „FOF“ und A_2 „FOM“ wurden mithilfe von Skala 02 bewertet, da die Akronyme der hier behandelten Wörter in keinem der Fälle von *Skype Translator* erkannt worden waren. Es scheint, dass *Skype Translator* nicht auf die Erkennung von Akronymen oder einzelnen Buchstaben ausgelegt ist. Demnach wurden im Fall von A_1 und A_2 die Werte dann mit JA bewertet, wenn *Skype Translator* zumindest einen Teil des Akronyms als Großbuchstaben erkannte. JEIN wurde dann zugeordnet, wenn

das Wort fast richtig erkannt wurde, und NEIN, wenn die Transkripteinträge so eingeschätzt wurden, dass sie unbrauchbar für die Weiterverarbeitung in Phase 2, der maschinellen Übersetzung, waren.

F_10 „with a capital S at the end“ ist ein spezieller Fall und wurde mit Skala 01 gemessen. Hier verhielt es sich ähnlich wie bei der Auswertung der Akronyme, da es *Skype Translator* nicht möglich war, den einzelnen Buchstaben „S“ als solchen zu erkennen. Aus diesem Grund wurde der Bewertungsvorgang so angepasst, dass beobachtet werden konnte, welche Werte von *Skype Translator* stattdessen wahrgenommen wurden. Die Sequenz „with a capital S at the end“ wurde in ihrer Gesamtheit gemessen. Wenn der Buchstabe „S“ mit einem phonetisch äquivalenten Element ersetzt wurde, wurde das Element trotzdem mit JA bewertet. Auch „with a capital at the end“ wurde mit JA bewertet. Diese Entscheidung wurde getroffen, da die gesamte Sequenz dann „Focus on FormS with a capital at the end“ lauten würde und die Erkennung sich dementsprechend nicht sinnverändernd auswirkt. Die Akronyme wiesen keine JA, 59 NEIN, ein JEIN und vier nicht identifizierbare Bewertungen im ersten Transkript auf.

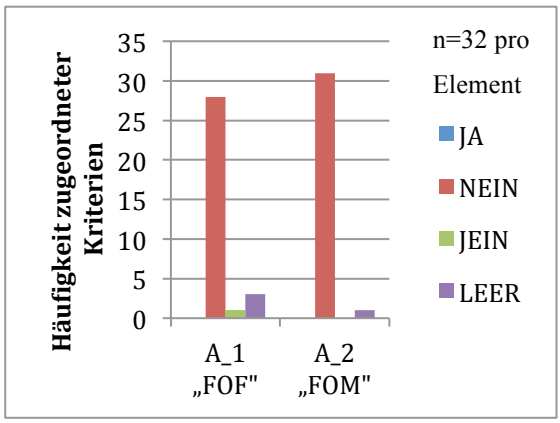

Abbildung 11: Auswertung der Problemfälle, die zu „Akronymen“ im Transkript 1 überprüft wurden

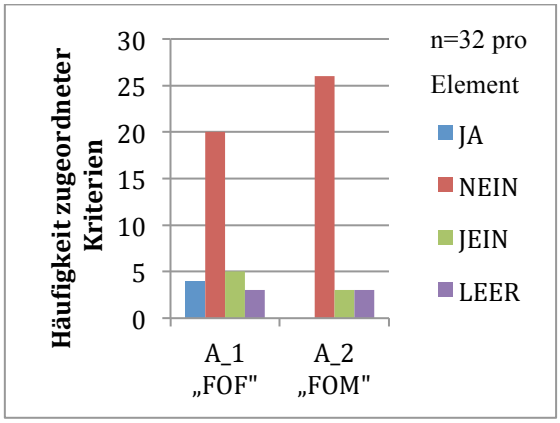

Abbildung 12: Auswertung der Problemfälle, die zu „Akronymen“ im Transkript 2 überprüft wurden

Abbildung 13 zeigt klar, dass auch in diesem Fall die NEIN-Werte am häufigsten waren, wobei auch die JA-Werte relativ hoch ausfielen. Die Fachbegriffe und Ausdrücke (F_8, F_9 und F_10) in Abbildung 13 weisen insgesamt eine Anzahl von 34 JA, 49 NEIN, 10 JEIN sowie 3 nicht identifizierbare Datensätze auf. In weiten Teilen zeigten die Ergebnisse des zweiten Transkripts eine bessere Spracherkennung als zuvor.

So wurde bei sämtlichen Fällen die Anzahl der mit NEIN beurteilten Spracherkennungen verringert, während der Anteil der mit JA beurteilten merklich anstieg – mit einer Ausnahme: Das Ergebnis von F_10, „with a capital S at the end“, zeigte allein eine Steigerung der NEIN-Werte, während alle anderen abnahmen. Auch hier zeigt sich also wieder eine Ausnahme, bei der das automatische Lernen von *Skype Translator* die Fehlerquote geringfügig erhöhte, anstatt sie zu senken. Allgemein lässt sich jedoch sagen, dass die nicht auswertbaren Daten bei der Mehrheit der überprüften Kriterien abgenommen haben.

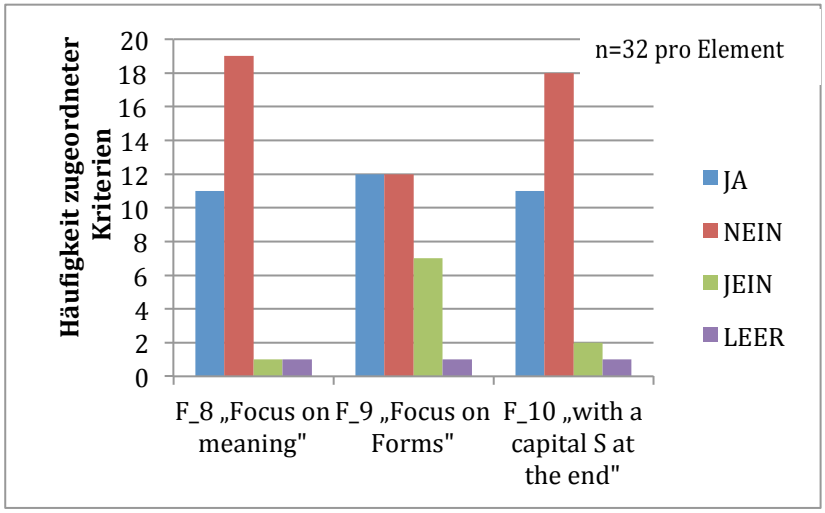

Abbildung 13: Auswertung der Problemfälle, die zu „einzelne Fachtermini“ im Transkript 1 überprüft wurden

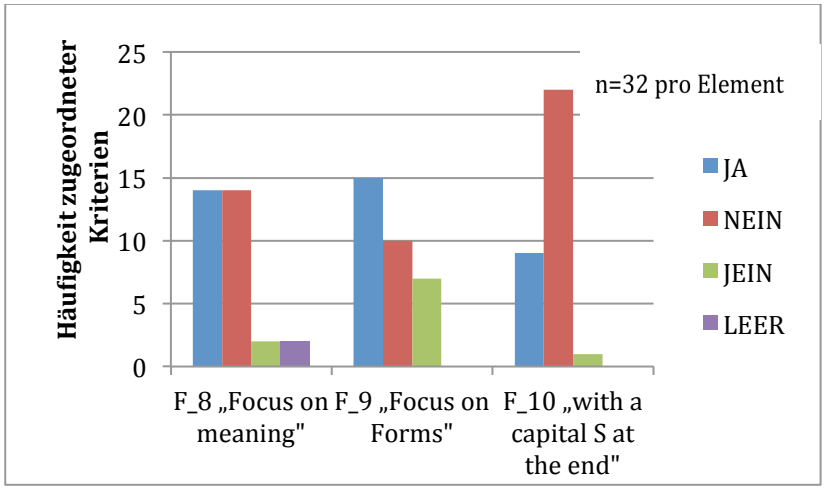

Abbildung 14: Auswertung der Problemfälle, die zu „einzelne Fachtermin“ im Transkript 2 überprüft wurden

Auch in der Kategorie der „Akronyme und einzelnen Fachtermini“ konnte weniger als die Hälfte der Datensätze erkannt werden. Die 5 Elemente erreichten bei einer Grundgesamtheit von 161 Einträgen folgende Scores:

Tabelle 15: Score der Elemente in der Kategorie „Akronyme und einzelne Fachtermini“

	SCORE Transkript 1	**SCORE Transkript 2**
A_1	-27,5	-13,5
A_2	-31	-24,5
F_8	-7,5	1
F_9	3,5	8,5
F_10	-6	-12,5
GESAMTSUMME (max. Score pro Transkript: 161)	**-68,5**	**-41**

4.4 Eigennamen, Kontrollwörter und Jahreszahlen

Dieses Unterkapitel umfasst die Auswertung bezüglich der im Interview vorhandenen Eigennamen, der Kontrollwörter und der Jahreszahl. Ein Eigenname tritt zweimal im Ausgangstext auf und gibt den Namen eines Wissenschafters wieder, der andere Eigenname gilt als Nationalbegriff für die „Franzosen“, also „French people“. Darüber hinaus ist die Publikation, die von eben diesem Wissenschafter genannt wurde, aus dem Jahr 2016 – welche die hier verwendete Jahreszahl zur Analyse ist. Die ansonsten angeführten sechs Ausdrücke gehören nicht direkt zur Kategorie der Problemauslöser, sondern stellen eine Repräsentation von Allgemeinsprache dar. In diesem Fall wurden sämtliche Werte nach Skala 01 bewertet, also nach exakter Darstellung des Gehörten.

Abbildung 13 und Abbildung 15 zeigen, dass die meisten überprüften Elemente öfter korrekt als falsch beurteilt wurden. Von der Grundgesamtheit von 320 wurden insgesamt 151 JA (47%), 94 NEIN (29%) und 12 JEIN (3,7%) eruiert. 63 Werte (19%) waren nicht als brauchbare Daten identifizierbar. Insgesamt überwiegen die JA-Werte. Bei E_1, „Ellis“ und K_1, „communicative turn“, waren die meisten Ergebnisse inkorrekt. Beim zweiten Auftreten des Eigennamens wurde dieser überraschenderweise geringfügig häufiger korrekt als falsch erkannt, doch zeigt sich auch ein wesentlich größerer Anteil an nicht identifizierbaren Daten als zuvor. Die Kriterien J_1, „2016“, K_2, „vary“, K_3, „intake“ und K_4, „either“, zeigen einen klar erkennbaren Unterschied zwischen korrekt und inkorrekt verarbeiteten Daten, während die Kriterien

E_3, „intake“, und K_5, „arise“, mehr als dreimal mehr JA- als NEIN-Werte erkennen lassen.

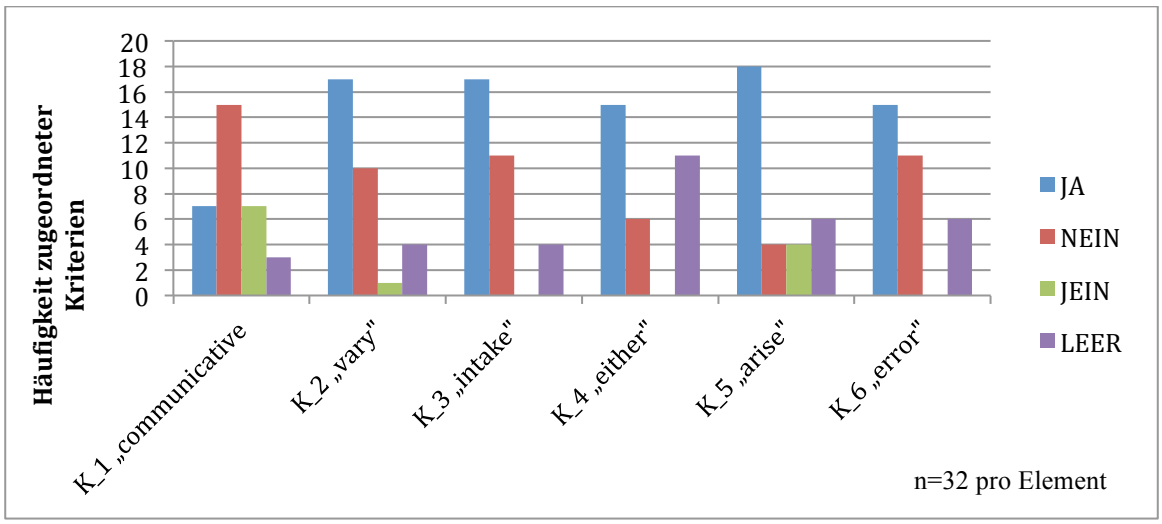

Abbildung 15: Auswertung der Problemfälle, die zu „Kontrollwörtern“ im Transkript 1 überprüft wurden

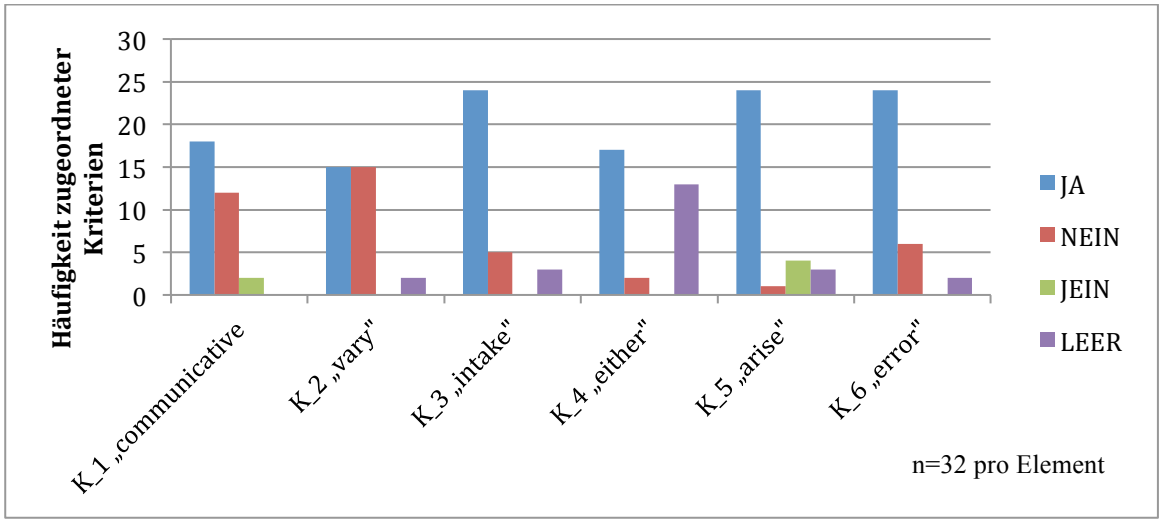

Abbildung 16: Auswertung der Problemfälle, die zu „Kontrollwörtern“ im Transkript 2 überprüft wurden

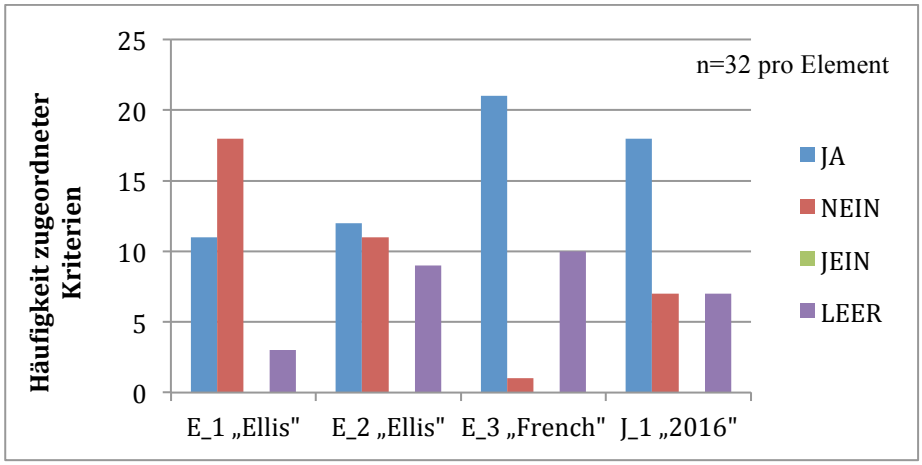

Abbildung 17: Auswertung der Problemfälle, die zu „Eigennamen und Jahreszahlen" im Transkript 1 überprüft wurden

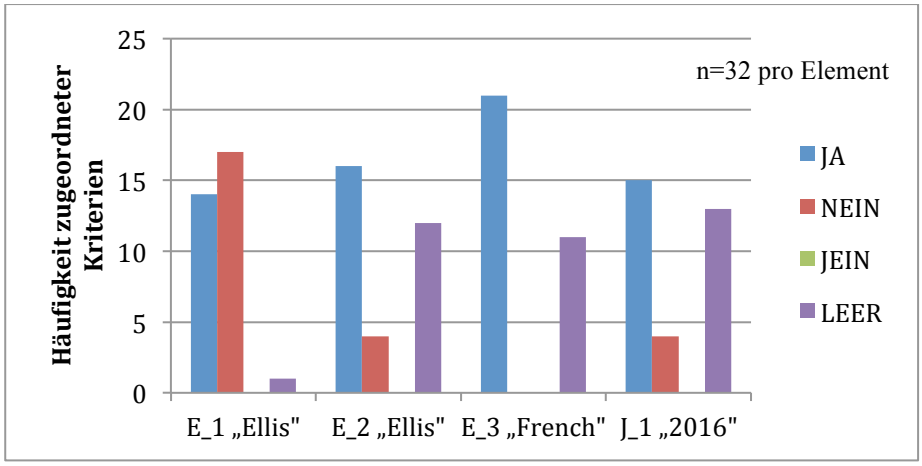

Abbildung 18: Auswertung der Problemfälle, die zu „Eigennamen und Jahreszahlen" im Transkript 2 überprüft wurden

Ausgehend von der Grundgesamtheit wurden in Transkript 2 insgesamt 188 JA (+11%), 66 NEIN (-8%), 6 „JEIN" (-1,9%) und 60 (-0,75%) nicht identifizierbare Daten vermessen. Im Unterschied zum ersten Transkript haben die Kriterien E_1 „Ellis" , E_2 „Ellis" , K_1 „communicative turn", K_3 „intake", K_5 „arise" und K_6 „error" an JA-Werten bei gleichzeitiger Abnahme von NEIN-Werten auffällig zugenommen, wobei beim Kriterium E_1 weiterhin die NEIN-Werte überwiegen. Bei K_1 wurde der auffälligste Zuwachs an JA-Werten verzeichnet: Obwohl zuvor die JA- und JEIN-Werte jeweils um mehr als das Doppelte von den NEIN-Werten überragt wurden, übertreffen nun die korrekten Messungen die inkorrekten um beinahe ein Drittel. Auch der starke Rückgang an NEIN-Werten ist bei diversen Kriterien frappant: Die zweitmalige Verwendung des Eigennamens im Text (E_2) zeigt nun mehr als dreifach so viele

korrekte als falsche Verarbeitungen, obwohl beim ersten Durchlauf die besagten Werte nahezu gleichauf waren. Ebenda sowie bei K_3 „intake“ verringert sich der NEIN-Wert um mehr als die Hälfte.

Trotzdem zeigt sich in mehreren Fällen ein überraschender Rückgang der korrekten Datenaufnahme: Die Jahreszahl (J_1) verliert einige korrekte Ergebnisse an nicht zu verarbeitende Daten, während bei K_2 sogar ein Rückgang der korrekten bei gleichzeitiger Zunahme von inkorrekten Daten zu verzeichnen ist, sodass letztlich trotz des vormalig passablen Ergebnisses nun die JA- und NEIN-Werte zahlenmäßig gleich stark vertreten sind. Während insgesamt beim ersten Transkript der Großteil der JA-Werte zwischen 15 und 20 Entsprechungen aufwies (ein Wert darüber, zwei darunter und weitere zwei genau auf 15), liegt beim zweiten Transkript nur noch ein Wert (E_1 „Ellis“) unter 15, während zwei genau auf 15 und der Rest darüber liegt und sogar vier Kriterien über 20 JA-Werte verzeichnen.

Der Score, welcher von einer Grundgesamtheit von 320 ausging, zeigte in seiner Gesamtheit, dass die Kategorie „Eigennamen, Kontrollwörter und Jahreszahlen“ mehr als die Hälfte der Transkripteinträge als richtig beziehungsweise halbrichtig verzeichnete. Demgegenüber fiel das erste Mal, in dem der Eigenname „Ellis“ vorkam, schlechter aus als das zweite Mal. Die Nationalbezeichnung „French people“ sowie die Jahreszahl erhielten gleich von Beginn an einen sehr hohen Score und die Veränderung zwischen Transkript 1 zu 2 fiel minimal bis gar nicht aus. Insgesamt konnte eine deutliche Besserung hinsichtlich der Worterkennung zwischen Transkript 1 und 2 vermessen werden.

Tabelle 16: Score der Elemente in der Kategorie „Eigennamen, Kontrollwörter und Jahreszahlen“

	SCORE Transkript 1	**SCORE Transkript 2**
K_1	-4,5	7
K_2	7,5	0
K_3	6	19
K_4	9	15
K_5	16	25
K_6	4	18
E_1	-7	-3
E_2	1	12
E_3	20	21
J_1	11	11
GESAMTSUMME (max. Score pro Transkript: 320)	**63**	**125**

4.5 Intonation determinierende Syntax

Dieses Unterkapitel behandelt die Elemente der Intonation determinierenden Syntax. *Skype Translator* ist vor allem auf die Wortebene programmiert, nicht aber auf die Intonation. Gleichzeitig ist mithilfe des *TrueText* durchaus auch korrekte Zeichensetzung zu erwarten. Mit den beiden hier exerzierten Fällen sollte analysiert werden, ob es sich syntaktisch nachvollziehen lässt, wann eine Veränderung der Intonation vollzogen wurde. I_1 „No? okay,...“ ist eine Frage der/des Interviewten, welche von der/dem Interviewenden im Ausgangstextinterview durch ein Kopfnicken oder nonverbale Kommunikation beantwortet wurde. Die gestellte Frage enthält nicht den üblichen syntaktischen Aufbau einer Frage mit Fragewort, Prädikat und Subjekt, sondern wird mit nur einem Wort durch Intonation und Zusammenhang als solche erkannt. Dadurch, dass die Fragezeichen in keinem der Fälle an dieser Stelle durch *TrueText* eingefügt wurden (da *Skype Translator* nicht auf Intonation ausgerichtet ist), wurde auch hierfür die Skala, mit der gemessen wurde, leicht adaptiert. In diesem Fall handelte es sich um Skala 04 für spezielle Fälle.

I_1 wurde mit JA bewertet, wenn sowohl „No“ als auch „Okay“ davor und danach mit einem Satzzeichen vom Rest des Satzes abgegrenzt wurde. JEIN wurde zugeteilt, wenn zumindest vor oder nach „No“ beziehungsweise „okay“ durch Satzzeichen abgegrenzt wurde oder die Wörter nicht richtig erkannt wurden. Alles, was nicht klar abgegrenzt war und somit in den Inhalt des Satzes davor oder danach eingeflossen ist, konnte dementsprechend für die nächste Phase der maschinellen Übersetzung nicht erfolgreich verarbeitet werden und wurde mit NEIN bewertet.

Ein ähnliches Verfahren wurde auch für I_2 angewandt. Hierbei war fraglich, ob die belustigt eingebrachte Parenthese „your working memory is limited – *even if you are a very smart person* – your working memory is still limited“ von *TrueText* als solche erkannt und gekennzeichnet wurde.

Auffällig ist, dass für I_1 und I_2 sehr ähnliche Resultate sowohl in Transkript 1 als auch in Transkript 2 registriert wurden. Insgesamt wurden in Transkript 1 von den 64 Fällen 38 JA (59%), 12 NEIN (19%), 13 JEIN (20%) und 1 LEER verzeichnet.
Transkript 2 blieb mit 36 JA-Beurteilungen, also einem Unterschied von 2, dem ersten Transkript sehr ähnlich. Genauso wurde auch der NEIN-Wert mit einem Unterschied

von 2 mit 10 vermerkt und JEIN mit 14, was demnach keinen signifikanten Unterschied aufweist. Einzig die nicht identifizierbaren Elemente verringerten sich minimal.

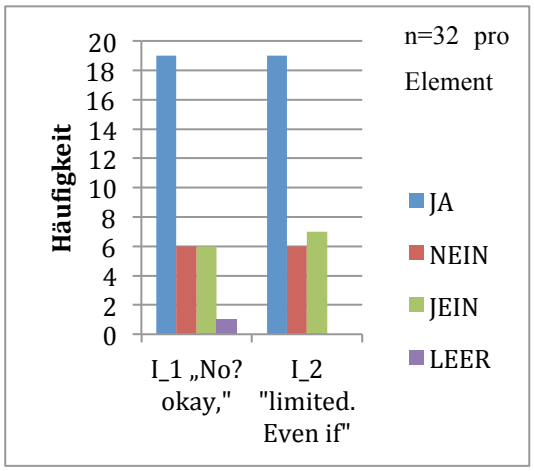

Abbildung 19: Auswertung der Problemfälle, die zur „Intonation determinierender Syntax" im Transkript 1 überprüft wurden

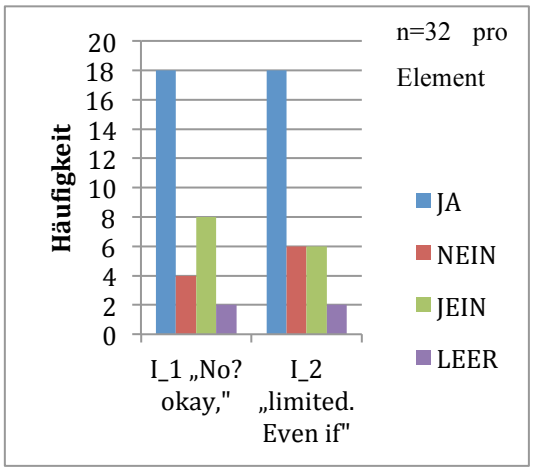

Abbildung 20: Auswertung der Problemfälle, die zur „Intonation determinierender Syntax" im Transkript 2 überprüft wurden

Beim errechneten Score der „Intonation determinierenden Syntax" mit einer Grundgesamtheit von 64 bei 2 Elementen fiel auf, dass sich die Ergebnisse beim Score zwischen Transkript 1 und 2 kaum veränderten.

Tabelle 17: Score der Elemente in der Kategorie „Intonation determinierende Syntax"

	SCORE Transkript 1	**SCORE Transkript 2**
K_1	16	18
K_2	16,5	15
GESAMTSUMME (max. Score: 64)	**32,5**	**33**

4.6 Registerwechsel

Für die Kategorie Registerwechsel wurden vornehmlich Ausdrücke aus dem Ausgangstext ausgewählt, die in der mündlichen Sprache für gewöhnlich bei informellen Anlässen verwendet werden, aber, wie hier zu sehen ist, auch in einem Interview mit WissenschafterInnen Anwendung finden. Der Ausdruck „yeah", welcher für R_5, R_6 und R_7 ausgewählt wurde, ist insofern speziell, da dieses Wort auch als eine Art Partikel interpretiert werden könnte. Auch der Ausdruck, der bei R_1 überprüft wurde, birgt eine Mischung informeller und allgemeiner Sprache. Sämtliche Begriffe wurden nach Skala 01 analysiert.

R_1 „text editor thing“ wurde dann mit JA evaluiert, wenn alle drei Wörter exakt erkannt wurden. Die Fälle, die entweder als „text editor“ oder „editor thing“ erkannt wurden, wurden als JEIN eingestuft. Insgesamt wurden nach dieser Bewertungsmethode also 11 JA (34%), 8 NEIN (25%) und 9 JEIN (28%) gemessen, während vier Daten (13%) nicht identifiziert werden konnten.

Für R_2 ist auffällig, dass der Anteil an JA-Werten mit 7 dem gleichen Wert entspricht wie JEIN. Hierbei ist zu beachten, dass JEIN dann bewertet wurde, wenn *Skype Translator* „wanna“ zu „want to“ veränderte. Blieb „wanna“ unverändert, wurde es mit JA beurteilt. NEIN fielen mit 4 (12%) etwas geringer aus, während 14 (44%) Transkripteinträge, also ein sehr hoher Anteil, nicht identifiziert werden konnten.

Bei R_1 ist für Transkript 2 im Vergleich zu Transkript 1 mit 17 JA ein Zuwachs von 19% ersichtlich. Die anderen Bewertungskriterien von R_1 nahmen geringfügig ab. Die Ergebnisse von R_2 veränderten sich im Vergleich zu Transkript 1 immens, was im Vergleich zu der im Transkript 1 erhobenen JA-Werte eine Abnahme von 16% der gesamten Datenmenge ausmacht. So wurden nach der zweiten Evaluierung insgesamt nur 2 JA-Werte vermerkt, was im Vergleich zu den bei Transkript 1 erhobenen Daten einen Rückgang von 16% der gesamten Datenmenge ausmacht. 16 Fälle wurden JEIN zugeordnet, also zu „want to” ausgebessert. Trotz der 14 nicht identifizierbaren Werte wurde kein einziges NEIN verzeichnet.

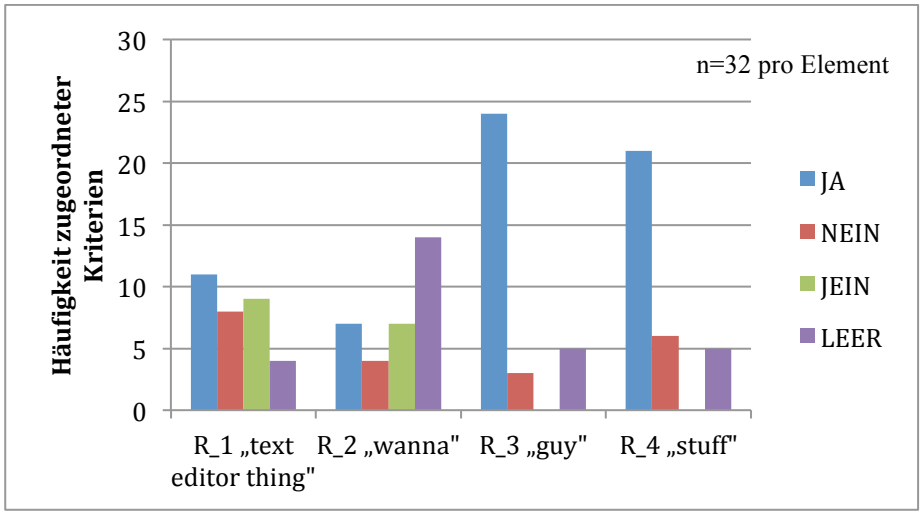

Abbildung 21: Auswertung der Problemfälle, die zum „Registerwechsel" im Transkript 1 überprüft wurden

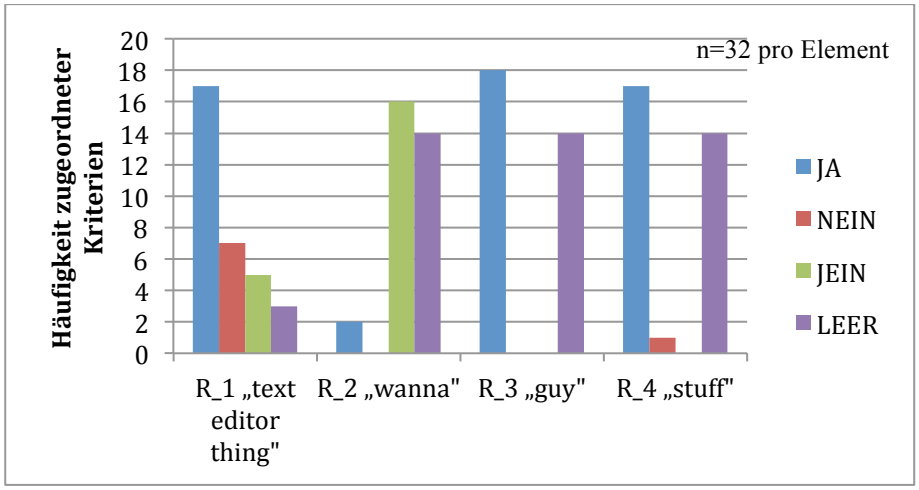

Abbildung 22: Auswertung der Problemfälle, die zum „Registerwechsel" im Transkript 2 überprüft wurden

Auffallend ist, dass die „JA-Komponente von R_5 bis R_7 „yeah" im Transkript 1 mit 67 (69%) von 96 stark dominant ist. Die Werte von R_5 bis R_7 wiesen insgesamt 83 JA 6 NEIN, 7 JEIN und 64 LEER auf. Besonders auffällig ist hierbei der Zuwachs des LEER-Wertes bei R_5, der von unter 5 bei Transkript 1 auf über 15 bei Transkript 2 anstieg. Interessanterweise steigen beinah genauso stark die JEIN-Werte bei R_3 „guy" und R_4 „stuff" an, die jedoch beide nicht das vielfältig verwendbare „yeah" verkörpern. Außerdem ist zu bemerken, dass die anderen Beurteilungen von „yeah" bei R_5 und R_6 an JA-Werten ab- und gleichzeitig bei JEIN-Werten zunahmen, während es sich bei „yeah" von R_7 genau andersherum verhält. Die tendenzielle Abnahme der

mit JA bewerteten Kriterien lässt vermuten, dass *Skype Translator* eher darauf ausgerichtet ist, schriftsprachliche Translationen abzuliefern.

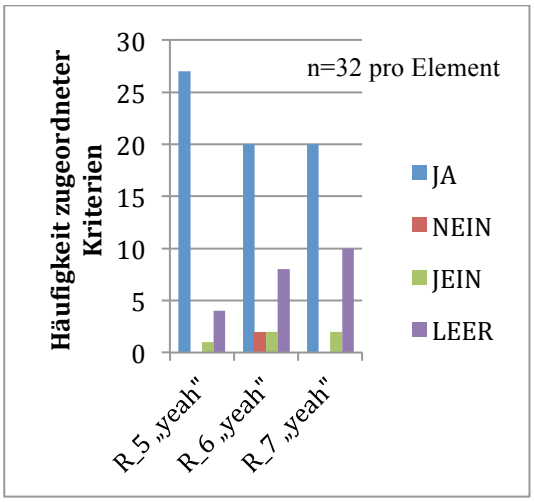

Abbildung 23: Auswertung der Problemfälle, die zum „Registerwechsel - yeah" im Transkript 1 überprüft wurden.

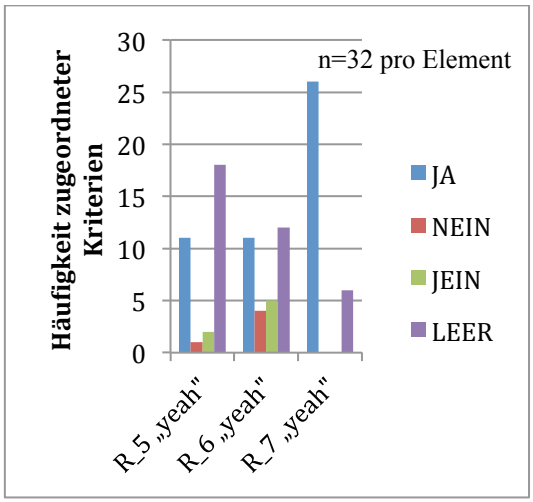

Abbildung 24: Auswertung der Problemfälle, die zum „Registerwechsel - yeah" im Transkript 2 überprüft wurden.

Die Kategorie „Registerwechsel" wies bei einer Grundgesamtheit von 224 bei 7 Elementen von Anfang an eine deutliche Mehrheit an exakt erkannten beziehungsweise erkannten Daten auf. Überraschend ist jedoch, dass sich die Daten im Gesamten von Transkript 1 auf 2 nicht verbesserten sondern laut Score verschlechterten, was darauf zurückzuführen ist, dass das automatisierte Lernen hierbei nicht in Aktion trat.

Tabelle 18: Score der Elemente in der Kategorie „Registerwechsel"

	SCORE Transkript 1	**SCORE Transkript 2**
R_1	7,5	12,5
R_2	6,5	10
R_3	21	18
R_4	15	16
R_5	27,5	11
R_6	19	9,5
R_7	21	26
GESAMTSUMME (max. Score: 224)	**117,5**	**103**

4.7 Syntax

Die Evaluierung der Kategorie Syntax erfolgteähnlich wie bei der „Intonation determinierenden Syntax". Hierbei wurden S_1 „because in,...“, S_2 „it has ... well, I“ sowie S_5 „„for example“ nach Skala 03 evaluiert, welche sich vor allem bezüglich Typologie und Abgrenzung der Sätze und Wörter richtet. In diesen Fällen wurde JA dann als Kriterium ausgewählt, wenn im Fall von S_1 „yeah“, im Fall von S_2 „well“ mit Satzzeichen eingegrenzt wurden. Im Fall von S_1 wurde dann JEIN als Kriterium gewählt, wenn statt „yeah“ ein „yes“ erkannt wurde oder das „yeah“ nur mit einem einzelnen Satzzeichen abgegrenzt wurde. Im Fall von S_2 wurde ein ähnliches Vorgehen gewählt, wobei der Wert JEIN dann vorkam, wenn „well“ nur mit einem einzelnen Satzzeichen abgegrenzt wurde. In diesen beiden Fällen befindet sich im Ausgangstext ein Satzabbruch, weswegen interessant ist, wie *Skype Translator* diesen ins Transkript einarbeitet.

S_3 „to contern contain a certain form“ ist als Spezialfall anzusehen, da hier ein Versprecher dazu führte, dass ein Wort gesagt wurde, welches in der englischen Sprache nicht existiert, und es dementsprechend dem Programm auch nicht möglich sein dürfte, dieses zu erkennen. Die letzten drei aufgezählten Fälle, S_3, S_4 und S_6 wurden nach Skala 04 evaluiert, welche für derartige spezielle Fälle angelegt wurde. Im diesem Fall wurde das Kriterium JA dann ausgewählt, wenn alle Wörter bis auf das Wort „contern“ richtig erkannt wurden und ein phonetisch nachvollziehbares Äquivalent als Lösung für ebenjenes Wort gewählt wurde. JEIN wurde dann vergeben, wenn zusätzlich einzelne andere Elemente kleine Fehler aufwiesen.

In den Fällen von S_4 „that’s what I what I think“ und S_6 „the prob- the problem“ geht es vor allem darum, wie *Skype Translator* mit den Selbstkorrekturen der SprecherInnen umgegangen ist. Von der Annahme ausgehend, dass auch HumandolmetscherInnen diesen abgebrochenen Satz nicht wiederholen, sondern einen flüssigen Satz bilden würden, wurden die Daten auch dann unter JA eingestuft, wenn die Wiederholung der Wörter eliminiert wurde oder wenn sämtliche Elemente dargestellt wurden und diese inhaltlich keine Veränderungen implizierten. JEIN wurde vergeben, wenn die Lösungen dem Original ähnlich waren, jedoch kleine Ungereimtheiten beinhalteten.

Beim Kriterium S_5 wurde ein Satz mit „for example“ beendet. Mit JA wurden jene Sätze beurteilt, an deren Ende „for example“ mit anschließendem Schlusspunkt gefunden wurde. Jene Fälle, in denen *Skype Translator* „For example“ mit dem Großbuchstaben an den Anfang des nächsten Satzes stellte, wurden unter die Bewertungskategorie JEIN gestellt. Genauere Details zur Beurteilung sind in Anhang 01 oder in Kapitel 3.3 zu finden.

S_1 und S_2 wiesen mit einer Grundgesamtheit von 64 Elementen in Transkript 1 insgesamt 36 JA, 1 NEIN, 8 JEIN und 19 nicht identifizierbare Elemente auf. In Transkript 2 verliert S_1 im Vergleich zu Transkript 1 viele richtige Spracherkennungen, während bei S_2 nun zwar mehr Daten ausgewertet werden können, diese aber nur minimal die JA-Werte bereichern. Stattdessen sind nun die falschen Spracherkennungen gleichauf mit den beinahe richtigen.

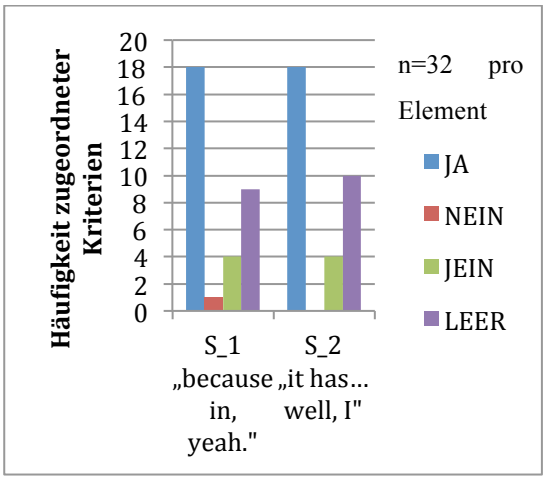

Abbildung 25: Auswertung der Problemfälle, die zur „Syntax - Satzabbruch“ im Transkript 1 überprüft wurden

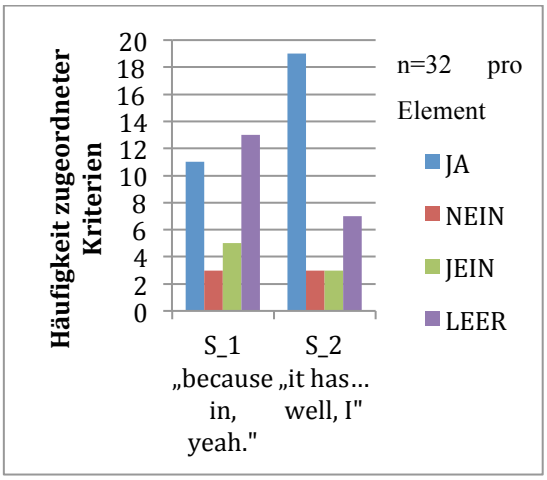

Abbildung 26: Auswertung der Problemfälle, die zur „Syntax - Satzabbruch“ im Transkript 2 überprüft wurden

Das Element S_5 zeigt hingegen durch den hohen Wert von 21 JEIN, dass *Skype Translator* „For example“ in den meisten Fällen an den Anfang des nachfolgenden Satzes stellte. JA- und NEIN-Werte sind jeweils nur einmal vertreten, 9 Datensätze wurden nicht identifiziert.

Die beiden Spezialfälle der Selbstkorrekturen S_4 und S_6 wiesen insgesamt 36 JA, 10 NEIN, 9 JEIN und 9 nicht identifizierbare Daten auf. Dabei ist schnell ersichtlich, dass die beiden Ergebnisse sehr unterschiedlich verteilt sind und S_6 relativ und absolut einen sehr viel höheren JA-Anteil hat. Der verbleibende Sonderfall S_3 ist

mit insgesamt 7 JA, 9 NEIN, 11 JEIN und 4 JEIN auch in seiner Auswertung besonders, da alle vier Werte relativ eng beieinander liegen, was bisher kaum vorgekommen ist (Ausnahme Registerwechsel R_1).

Den nahezu optimalen Erfolg durch das automatische Lernen weist das Element S_6 auf, da im Transkript 2 alle falschen Aufzeichnungen richtiggestellt werden konnten. Bei S_3 sind deutliche Verbesserungen zu vermerken, nun überragt der JA-Wert die anderen, lag er doch zuvor hinter den identifizierbaren Daten zurück. Auf den ersten Blick erweckt auch S_4 den Eindruck, als hätten sich die Werte verbessert, doch stellt sich dies als Trugschluss heraus: Obwohl die JA-Beurteilungen minimal zunehmen, verschwinden die beinahe richtigen Spracherkennungen fast zur Gänze, während sich die falschen mehr als verdoppeln.

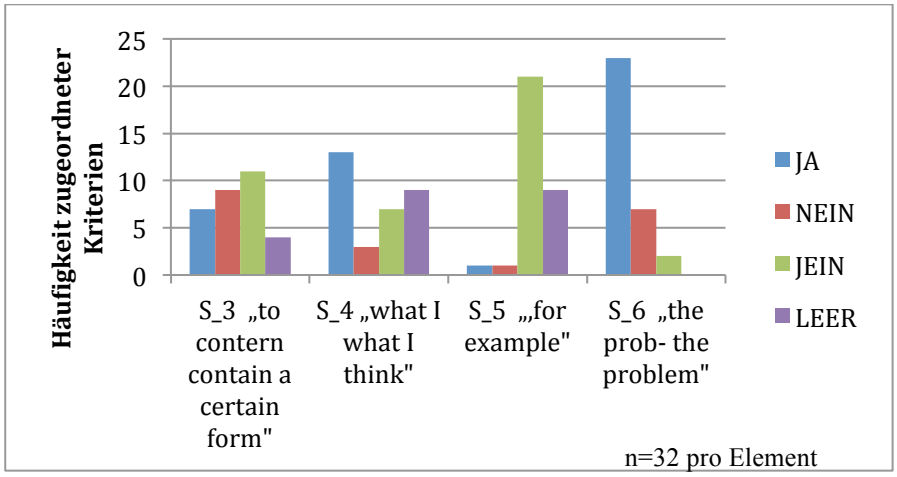

Abbildung 27: Auswertung der Problemfälle, die zur „Syntax" im Transkript 1 überprüft wurden.

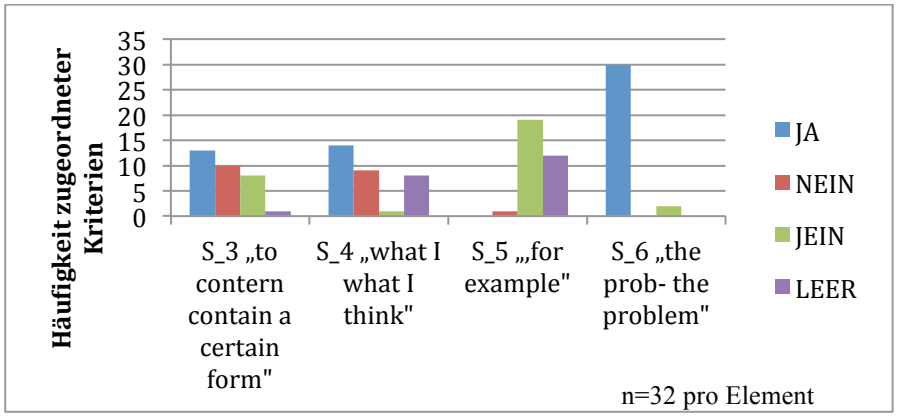

Abbildung 28: Auswertung der Problemfälle, die zur „Syntax" im Transkript 2 überprüft wurden.

Ebenso wie bei der Kategorie „Registerwechsel“ fiel der Score bei der Kategorie „Syntax“, welcher von einer Grundgesamtheit von 192 bei 6 Einträgen ausgeht, so aus, dass das automatisierte Lernen nicht nachvollzogen werden kann, da die Daten sich nicht verbesserten.

Tabelle 19: Score der Elemente in der Kategorie „Syntax“

	SCORE Transkript 1	**SCORE Transkript 2**
S_1	21	13
S_2	22	19
S_3	9	11
S_4	17	6
S_5	21	18
S_6	18	32
GESAMTSUMME (max. Score: 192)	**108**	**99**

5 Diskussion und Schlussfolgerungen

Ziel dieser Arbeit war es, herauszufinden, wie das innovative und weltweit zugängliche maschinelle Dolmetschsystem *Skype Translator* auf ausgewählte problemauslösende Elemente des Ausgangstextes reagiert.

Dank des in *Skype Translator* integrierten automatisierten Lernsystems ist künftig generell von einer steten Lernkurve und großen Fortschritten in der Erkennung der Ausgangsrede auszugehen. Mit der in dieser Arbeit angewandten Methode konnte vor allem beleuchtet werden, bei welchen dieser unterschiedlichen problemauslösenden Faktoren *Skype Translator* tatsächlich Fortschritte aufzeigte und bei welchen dieser Lernprozess nicht eintrat.

Für die Auswertung dieser Elemente in insgesamt 64 Dolmetschungen von *Skype Translator* spendeten 32 SprecherInnen ihre Stimmen zur Generierung dieser textuell übereinstimmenden Ausgangstexte. Durch das zweimalige Einspielen dieser Ausgangstexte in Skype wurde *Skype Translator* darauf getestet, ob das Programm einerseits durch die Vielzahl der Texte mit übereinstimmender Wortwahl ein bestimmtes Muster vernahm, und ob sich andererseits dadurch die Erkennung der identen Tonaufnahme beim zweiten Durchlauf veränderte.

Die quantitative Studie hebt die starken Unterschiede hervor, die der automatisierte Lernprozess von *Skype Translator* im Laufe der Datensammlung von Transkript 1 und 2 ausmacht. Auffallend ist, dass vor allem im Bereich der exakten semantischen Worterkennung besonders große Fortschritte in der Spracherkennung gemacht wurden. So wurden bei Fachtermini, die aus zwei bis drei Wörtern gebildet wurden, beim zweiten Durchgang wesentlich bessere Ergebnisse verzeichnet als bei Transkript 1. Jene Fachtermini, die aus noch mehr Wörtern gebildet wurden, wie zum Beispiel „second language acquisition researcher", zeigten sogar bereits beim ersten Durchgang erstaunlich hohe JA-Werte. Dieser Umstand deckt sich mit der Angabe von *Skype Translator*, nämlich, dass sich mehrere Wörter, die mehr Kontextumfang implizieren, leichter in einen gewissen Kontext bringen lassen.

Außerdem fiel auf, dass vor allem Wörter, die beim ersten Durchgang nicht erkannt wurden, teilweise beim zweiten Durchgang in die Kategorie JEIN oder NEIN eingestuft wurden, wie es beispielsweise im Fall von „interlanguage" geschah (siehe

Abschnitt 4.1). Dies eröffnet Spielraum für weitere Verbesserungen und kann mit weiteren Datensätzen in einer korrekten Spracherkennung gipfeln.

Das Beispiel von F_7 „Focus on Form“ ist ein Spezialfall, da insgesamt von einer wesentlich größeren Grundgesamtheit ausgegangen werden konnte als bei den anderen Begriffen. Von insgesamt 640 Fällen (32 SprecherInnen, die bei jeweils zehn Gelegenheiten „Focus on Form“ vortrugen, dies wiederum zwei Mal *Skype Translator* vorgespielt wurde) wurden insgesamt 32% korrekt erkannt, 9% als „der richtigen Version ähnlich“ eingestuft, 52% als falsch klassifiziert, während 7% nicht als brauchbare Daten identifiziert werden konnten. Auch hierbei ist anzumerken, dass sich von Transkript 1 zu Transkript 2 eine Verbesserung ausmachen ließ, welche nochmals durch die Vergleichsfälle von „Focus on Meaning“ und „Focus on FormS“ bestätigt werden konnte.

Auch der Bereich der Kontrollwörter, also jener Ausdrücke, die der Allgemeinsprache zugeordnet werden können, weist auf die Annahme hin, dass die Erkennung der Wörter durch mehrmalige Verwendung im System tendenziell besser wird. Einzig und allein der Begriff K_2 „vary“ wies diese Verbesserung nicht auf, was sich damit erklären lassen könnte, dass es sich hierbei um ein Wort handelt, welches in erster Linie nicht durch phonetische, sondern vielmehr durch syntaktische beziehungsweise kontextuelle Zusammenhänge erkannt wird.

Wie bereits im Unterkapitel der quantitativen Analyse über Akronyme eruiert wurde (siehe Abschnitt 4.3), wurden diese von *Skype Translator* fast gar nicht erkannt. Im zweiten Durchlauf wurden einzelne allgemeine Akronyme wie SOS (Internationales Notsignal) oder FOMC (Federal Open Market Committee) als Akronyme erkannt. Ähnlichkeiten von Buchstabierungen wurden vereinzelt vor allem in Form von Lautketten (eff o eff) dargestellt. Davon abgesehen wurden diese jedoch nicht als einzelne Buchstaben wahrgenommen. Demnach ist davon auszugehen, dass die vorhandene Datenbank, welche *Skype Translator* als Basis für Paralleltexte und Kontexte für die Spracherkennung und die maschinelle Übersetzung dient, keine große Menge an Akronymen als Parallelquellen parat hat oder *Skype Translator* nicht auf Akronyme programmiert worden ist.

Dadurch, dass *Skype Translator* auf Wort-für-Wort und Schlüsselwortbasis arbeitet und keine Analyse auf semantischer Ebene (siehe 1.1) macht, wurde in den

Ergebnissen wie erwartet analysiert, dass Zusatzerklärungen wie im Fall von „with a capital S at the end“ keinen Einfluss auf das beschriebene Wort haben, was für HumandolmetscherInnen, selbst wenn sie mit der Materie nicht vertraut wären, ein großes Hilfsmittel für das Verständnis des großen Ganzen bedeuten könnte.

Die Ergebnisse der „Intonation determinierenden Syntax“ und den dabei ersichtlichen typologischen Abgrenzungen ergaben, dass die zwischen Transkript 1 und Transkript 2 angefallenen Unterschiede insignifikant sind. Daraus kann wiederum geschlossen werden, dass sich der „Lernprozess“ nicht direkt auf die Abgrenzung von Satzzeichen bezieht, da diese von *TrueText* vermutlich aufgrund der syntaktischen Aufstellung der Sätze erfolgen. Es wäre ebenfalls möglich, dass die Sprechpausen vor und nach der Artikulation Einfluss auf das Ergebnis haben. Bei längeren Sprechpausen werden nämlich neue Segmentierungen initiiert. Dies könnte auch in der Kategorie „Redeflussunterbrechungen“ evaluiert werden, wie sie in der Parallelarbeit von Wonisch (2017) bearbeitet wurden.

Auch der Bereich des „Registerwechsels“ legt einige Schlüsse nahe. Die Auswertung der Formulierung „wanna“ (7 in Transkript 1, 16 in Transkript 2) bestätigt, dass Wörter, die umgangssprachlich ausgedrückt werden, tendenziell eher in Schriftsprache umgewandelt werden („wanna“ zu „want to“). Dies könnte auch grundlegend als Dolmetschstrategie verstanden werden. Hierbei bestätigt sich, wie in der Literatur beschrieben, dass der Grund dieses Registerwechsels vor allem darauf fußt, dass formell-schriftlicher Sprache eine höhere Datenmenge an Paralleltexten für die maschinelle Übersetzung zur Verfügung steht.

Diese Schlussfolgerungen decken sich auch mit denen der syntaktischen Elemente, welche nach einem ähnlichen Prinzip evaluiert wurden. In sämtlichen Fällen, in denen es vor allem um Satzzeichen als Grenzziehung zwischen zwei Sätzen ging, konnten keine starken Veränderungen zwischen Transkript 1 und Transkript 2 nachgewiesen werden. Das lässt vermuten, dass es durchaus Bereiche gibt, in denen das automatisierte Lernen hinter *Skype Translator* noch gewisse Mängel bezüglich seiner Adaptationsfähigkeit aufweist.

In jenen Fällen, in denen es nicht um Satzzeichen und Abgrenzungen ging, sondern wiederum um den Umgang mit den Problemauslösern im Zusammenhang mit „Satzabbruch“, waren durchaus Verbesserungen zu vermerken. S_6, also „prob-

problem“, wurde im zweiten Transkript in beinahe allen Sprechproben fast vollständig korrekt erkannt. Daraus kann geschlossen werden, dass *Skype Translator,* oder besser gesagt *TrueText,* ein sich wiederholendes Muster der Satzkorrektur erkannte und eine entsprechende Lösung dafür fand.

Es hat den Anschein, dass für jene Audiodateien, die von *Skype Translator* auf Anhieb dem richtigen Kontext und Fachbereich zugeordnet werden können, tendenziell eine relativ verständliche Darstellung der Ausgangstexte erfolgt. Diesem Gedanken folgend hieße das, dass jene von Skype auszuwertenden Daten, die von Anfang an von Skype Translator keinem passenden Kontext zugeordnet werden konnten, weniger korrekte Zuordnungen erhielten. Hieraus kann man schlussfolgern, dass *Skype Translator* einen Kontext nicht nur mithilfe einzelner Segmentierungen zuordnet, sondern sämtliche Textsegmentierungen miteinander in Bezug setzt, um den Kontext auszumachen. Dies wäre auch eine mögliche Erklärung dafür, dass *Skype Translator* eine begrenzte Anzahl an Minuten zur maschinellen Dolmetschung zur Verfügung stellt. Schließlich müssten dafür sämtliche Segmente und Schlüsselwörter miteinander in Beziehung gesetzt werden, was an einem bestimmten Punkt zur Maximalauslastung des Dolmetschsystems führen würde.

Es wurde rasch klar, dass das Themengebiet dieser Arbeit ein sehr komplexes und interdisziplinäres ist, bei dem jeder einzelne Punkt ganze Bücher, wenn nicht sogar Bibliotheken füllen könnte. Spannend war hierbei, ein Gebiet, das zum Großteil aus Algorithmen und Programmierung besteht, aus den Augen einer Translatorin beziehungsweise angehenden Dolmetscherin zu betrachten, die gewiss einen anderen Zugang zu dem Thema pflegt als das *Microsoft*-Team aus dem Bereich der Künstlichen Intelligenz. Wer sich den Aufbau und die Programmierung einer möglichst effizienten, künstlichen Imitation der Humandolmetschung ansieht, macht sich zwangsläufig bewusst, wie verflochten und spannend Sprache, Kommunikation und vor allem der Prozess des Dolmetschens sind.

5.1 Kritische Reflexion und Limitationen der Studie

Das durchgeführte Interview und der Ausgangstext für diese Arbeit inszenieren nur eines der vielen mit *Skype Translator* möglichen Nutzungs-Szenarien. Wie in der Theorie beschrieben (siehe Kapitel 1) geht *Microsoft* von vier verschiedenen

Kommunikationsmethoden aus. Gleichzeitig ist davon auszugehen, dass das Szenario an sich keinen Unterschied auf die Ergebnisse der hier durchgeführten Analyse hatte, da nur einzelne Aspekte sequenziell gemessen wurden.

Darüber hinaus sind die hier ausgewählten Sequenzen der Problemauslöser und Kategorien erst ausgewählt worden, nachdem das Interview durchgeführt war, da zuerst der Text auf mögliche Problemauslöser gesichtet werden musste. Man könnte kritisieren, dass die Kategorien und die Beispiele hinsichtlich ihrer Anzahl nicht ausgeglichen waren und nicht stichprobenartig, sondern subjektiv gewählt wurden.

Wesentlich bei der Aufnahme des Ausgangstextes war zudem, dass dieser sehr schnell gesprochen wurde und vermutlich in einer realen Gesprächssituation mit dem *Skype Translator* das Gesprächstempo gedrosselt worden wäre. Eventuell hätte bei einer anderen Geschwindigkeit auch der Syntaxverlauf anders ausgesehen. Die Originalgeschwindigkeit nahm vermutlich Einfluss auf die SprecherInnen. Insgesamt wurde das Durchschnittsredetempo bei 148 WPM gemessen, was für eine Humandolmetschleistung nicht als die Idealgeschwindigkeit eingestuft wird.

Die Durchführung des „Shadowing“ mit vorbereitetem Transkript zur Unterstützung war für die SprecherInnen allgemein eine ungewöhnlich herausfordernde Situation. Mit Hinblick auf Giles *Effort Model* wurden die SprecherInnen dazu aufgefordert, drei Tätigkeiten zur gleichen Zeit durchzuführen. Zum Einen hörten sie der Audiodatei und so der interviewten Person über Kopfhörer zu, zum Zweiten lasen sie den Text mit, während sie zum Dritten den Text auch wiederholten und versuchten, dem Tempo der Rednerin aus der Audiodatei zu folgen. Auffällig war, dass die Personen, die mit Mehrsprachigkeit und ganz besonders Dolmetscherfahrung im Vorfeld zu tun gehabt hatten, sich wesentlich leichter taten als jene, die aus einem Umfeld kamen, in dem derartige „Multitasking-Prozesse“ nicht üblich waren. Der Grundtenor der Rückmeldungen durch die SprecherInnen war, dass die Aufgabe als sehr herausfordernd, zugleich aber als sehr spannend empfunden wurde und beim Durchführen Spaß bereitet habe. Es sollte in Betracht gezogen werden, diese Art von Simulation ausschließlich mit Menschen zu wiederholen, die über Dolmetscherfahrung verfügen.

Einige der SprecherInnen entschieden sich dafür, den Text mehrmals durchzusprechen, während andere den Text im Voraus lediglich still durchlasen und

gleichzeitig der Audiodatei lauschten. Die Anpassung der Geschwindigkeit erfolgte nach den ersten paar Sätzen des Ausgangstextes. Die Audiodatei wurde ihnen einige Sekunden lang vorgespielt, woraufhin sie entschieden, ob sie die Geschwindigkeit zum Vorlesen reduziert haben wollten und um wie viel. Das Bemühen vom Großteil der SprecherInnen, die Originalgeschwindigkeit möglichst einzuhalten, lässt darauf schließen, dass die TeilnehmerInnen entschlossen waren, eine gute Leistung zu erbringen. Es wäre deswegen durchaus denkbar, dass das Vorspielen der Originalrede in der Originalgeschwindigkeit den SprecherInnen suggerierte, dass die vorgegebene Geschwindigkeit erstrebenswert sei. Womöglich wurde in manchen Fällen deswegen nicht die Geschwindigkeit gewählt, die sie für sich persönlich in einer entsprechenden Situation gewählt hätten. Gleichzeitig wurde ihnen allerdings vermittelt, dass es notwendig ist, jene Geschwindigkeit für das „Shadowing" zu wählen, bei der sie sich wohl fühlen.

Rückblickend wäre es wohl die besser geeignete Methode gewesen, zuerst die SprecherInnen den Text mit dem Transkript in einem Probedurchlauf ohne den Einfluss der Audiodatei vorlesen zu lassen. Dann wäre es einfacher gewesen, die Geschwindigkeit der Audiodatei an die Sprechgeschwindigkeit der SprecherInnen anzupassen und nicht umgekehrt. Dadurch, dass manche der Aufnahmen länger als 6 Minuten dauerten und die Dolmetschungen genau nach dieser Zeitspanne abbrachen und neu angefangen werden mussten, könnte es sein, dass auch dieser Aspekt Einfluss auf die Ergebnisse hatte.

Mit dieser Ausgangssituation muss damit gerechnet werden, dass nicht alle ausgewerteten Audiodateien fehlerfrei waren. Mit anderen Worten: Es wurden einige Sequenzen bereits bei der Aufnahme der SprecherInnen nicht vollständig formuliert, wodurch wiederum das System keine Möglichkeit hatte, jene Elemente zu erkennen.

Dies berücksichtigend wurde zudem eine Auszählung von jenen Elementen gemacht, die sowohl im ersten als auch im zweiten Transkript mit dem Kriterium JEIN bewertet wurden. Von den einzelnen Analysekategorien ausgehend wurden im Fall des „Focus on Form" von den 640 analysierten Elementen insgesamt 14 in beiden Transkripten nicht erkannt. Für sämtliche Analysekategorien konnten solche Fälle gefunden werden, welche sich im Endeffekt bei einer Anzahl von gesamt 2944 ausgewerteten Datensätzen auf insgesamt 192 Elemente belaufen, die von *Skype*

Translator in beiden Transkripten nicht erkannt wurden. Deswegen kann darauf geschlossen werden, dass diese bereits im Prozess der Ausgangstextproduktion verloren gingen.

5.2 Ausblick

Wie es sich so oft mit umfangreichen Themen verhält, könnte man auch diese Arbeit nahezu grenzenlos ausweiten. So wäre eine Komponente, die sich mit den bereits ausgewerteten Daten direkt weiterführen ließe, die der senderbezogenen Problemauslöser. Daher könnte beispielsweise evaluiert werden, welche der SprecherInnen im Speziellen die meisten richtig erkannten Werte aufwies. Anschließend wäre es möglich, Schlüsse daraus zu ziehen, welche Geschwindigkeit, welcher Akzent, welche Stimmfrequenz oder allgemein welche Gegebenheiten momentan die besten Voraussetzungen für erfolgreiche Spracherkennung sind.

Dadurch, dass in beiden Transkripten durchaus Satzzeichen – nicht nur Beistriche, Punkte sondern auch Fragezeichen und fett geschriebene Elemente, welche hier möglicherweise als Ausrufe interpretiert werden könnten, in der Erkennung von *Skype* auftraten, wäre in einer weiteren Analyse eine Evaluierung eben dieser vorkommenden Elemente hinsichtlich genauerer Ergebnisse sinnvoll.

Von *Skype* wurde hervorgehoben (vgl. Wendt 2016a: 46), dass die Variable der Akzente einen besonders großen Einfluss auf die Ergebnisse der Spracherkennung hat. In diesem Zusammenhang ist beispielsweise aufgefallen, dass in den für diese Arbeit behandelten Transkripten eine relativ hohe Anzahl an Schimpfwörtern fälschlicherweise „erkannt“ wurde, obwohl sie im vorgelegten Text nicht eingefügt waren. Dies stach vor allem bei den zwei SprecherInnen mit italienischem Hintergrund innerhalb des Ausdruckes *Focus on Form* hervor. *Skype* vermerkte diese stets mit ***. Es ist anzunehmen, dass ähnliche Interpretationsfehler auch bei anderen Akzenten aufzufinden sind, was einen weiten Forschungsraum für die Zukunft eröffnen könnte.

Auffallend war auch, dass ältere SprecherInnen tendenziell langsamere Sprechgeschwindigkeiten wählten. Man könnte diesbezüglich nachforschen, ob eher die langsamere Sprechgeschwindigkeit und somit die vermeintlich „genauere“ Aussprache zu besseren Ergebnissen in der Spracherkennung führt, oder ob eventuell durch die

Veränderung der Muskulatur bei SeniorInnen die Artikulationsfähigkeit derart abgeschwächt ist, dass wiederum junge Erwachsene bessere Ergebnisse aufweisen.

Dass sowohl zu langsames wie auch zu schnelles Sprechen Schwierigkeiten für maschinelle und menschliche Dolmetschungen bedeuten, wurde bereits in der Dolmetschwissenschaft belegt (siehe Abschnitt 2.2.2.1). Dies und die Frage, ob es für *Skype Translator* eine Idealgeschwindigkeit gibt, könnten auch für maschinelle DolmetscherInnen erhoben werden. Es ist davon auszugehen, dass die sich im Normalbereich befindlichen unterschiedlichen Stimmfrequenzen der Datenbank, von der Skype seine Lösungen bezieht, abgedeckt sind und es keine Probleme diesbezüglich bei der Erkennung gibt. Nichtsdestotrotz wäre es auch nützlich herauszufinden, ob besonders hohe und besonders tiefe Frequenzen zu unterschiedlichen Ergebnissen führen.

Zusammenfassend ergeben sich aus den für diese Arbeit generierten Daten neue Fragestellungen und Forschungsideen, beispielsweise eine Evaluierung der SprecherInnen:

- Wann konnten bei wem welche Ergebnisse verzeichnet werden und warum?
- Welchen Einfluss auf das Ergebnis der Spracherkennung haben die verschiedenen Variablen wie Stimmhöhe, Akzent, Aussprachegenauigkeit, mögliche Sprachfehler Füllwörter oder Sprechgeschwindigkeit? Diese Variablen könnten einzeln betrachtet werden, um verschieden Muster herauszuarbeiten, wie sie beispielsweise im Fall der Akzente und der beschriebenen Kraftausdrücke auftraten. Darüber hinaus könnte ebenso analysiert werden, ob es Variablen gibt, die größeren oder kleineren Einfluss auf die Erkennung haben als andere, wie zum Beispiel langsamer aber ungenauer Aussprache (ältere Erwachsene), gegenüber schneller, aber exakter Aussprache (jüngere Erwachsene).
- Welche Rolle spielen Drittvariablen, wie Hintergrundgeräusche, lautes Atmen, Hall?
- Gibt es für *Skype Translator* eine Idealgeschwindigkeit der Ausgangsrede?
- Welche Beachtung verdient die Satzzeichensetzung von *TrueText* in den Transkripten von *Skype Translator*?

In sämtlichen Forschungsfeldern sollte natürlich immer mitbedacht werden, dass Drittvariablen mit statistischen Methoden wissenschaftlich ausgeschlossen werden sollten, wenn sie möglicherweise Einfluss auf die gemessenen Outputs haben können.

Quellenverzeichnis

Ahrens, Barbara (2005) Analysing Prosody in Simultaneous Interpreting: Difficulties and possible Solutions. *The Interpreters' Newsletter* 13, 1-14.

AIIC (2002) „Workload Study." http://.aiic.net/page/657/interpreter-workload-study-full-report/lang/ (20.08.2017).

Akagi, Masato, Han, Xiao, Elbarougy, Reda, Hamada Yasuhiro & Li, Junfeng (2014) Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages. In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, Siem Reap, Cambodia, 09-12 Dezember 2014*, 1-10.

Bahl, Lalit R., Brown, Peter F., de Souza Peter V. & Mercer, Robert L. (1989) A tree based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustic Speech and Signal Processing* 37 (7), 1001-1008.

Black, Alan W. & Lenzo, Kevin A. (2004) Multilingual text-to-speech synthesis. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP) Montreal, Canada, 17–21 Mai 2004* (3), 761–764.

Boersma, Paul & Weenink, David (n.d.) „Praat: doing phonetics by computer." University of Amsterdam, Niederlande. http://www.praat.org/ (20.08.2017).

Bughin, Jacques, Chui, Michael & Manyika, James (2010) Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly* 56 (1), 75-86.

Chang, Janie (2011) „Speech Recognition Leaps Forward." http://research.microsoft.com/en-us/news/features/speechrecognition-082911.aspx (20.03.2016).

Chernov, Ghelly (1979) Semantic Aspects of Psycholinguistic Research in Simultaneous Interpretation. In: Franz Pöchhacker & Miriam Shlesinger (Eds.) (2002) *The Interpreting Studies Reader*. London/New York: Routledge, 99-109.

Chernov, Ghelly (1994) Message Redundancy and Message Anticipation in Simultaneous Interpretation. In: Sylvie Lambert & Barbara Moser-Mercer (Eds.) (1994) *Bridging the gap: Empirical research in simultaneous interpretation.* Amsterdam/Philadelphia: John Benjamins Publishing, 139-53.[3]

Cooper, Cary L., Davies, Rachel & Tung, Rosalie L. (1982) Interpreting Stress: Sources of Job Stress among Conference Interpreters, *Multilingua* 1 (2), 97-107.

Crevatin, Franco (1989) Directions in Research Towards a Theory of Interpretation. In: Laura Gran & John Dodds (Eds.) *The Theoretical and Practical Aspects of Teaching Conference Interpretation*. Udine: Campanotto, 21-22.

Dahl, George E., Yu, Dong, Deng, Li & Acero, Alex (2012) Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing* 20 (1), 30-42.

Déjean le Féal, Karla (1982) Why Impromptu Speech is Easy to Understand. In: Nils E. Enkvist (Ed.) (1982) *Impromptu Speech: A Symposium. November 20-22, Finland: Åbo Åkademi,* 78, 221-239.

Flesch Formula (o.J.) „Flesch Reading Ease Readability Formula." http://www.readabilityformulas.com/flesch-reading-ease-readability-formula.php (20.08.2017).

Flesch Formula (o.J.) „Flesch-Kincaid Grade Level Readabiliy Formula." http://www.readabiliyformulas.com/flesch-grade-level-readability-formula.php (20.08.2017).

Flesch, Rudolf (1943) *Marks of a readable style*. Teachers College Contributions to Education¸ Columbia University.

Flesch, Rudolf (1948) A new readability yardstick. *Journal of Applied Psychology* 32 (3), 221-233.

Forsberg, Markus (2003) *Why is Speech Recognition Difficult?* Göteborg, Sweden: Chalmers University of Technology.

Fünfer, Sarah ([2]2013) *Mensch oder Maschine? Dolmetscher und maschinelles Dolmetschsystem im Vergleich*. Berlin: Frank & Timme.

Gerver, David (1969) The Effects of Source Language Presentation Rate on the Performance of Simultaneous Conference Interpreters. In: Franz Pöchhacker & Miriam Shlesinger (Eds.) (2002) *The Interpreting Studies Reader*. London/ New York: Routledge, 53-66.

Gile, Daniel (1995) Basic Concepts and Models for Interpreter and Translator Training. Amsterdam/Philadelphia: John Benjamins.

Gile, Daniel ([2]2009) Basic Concepts and Models for Interpreter and Translator Training. Amsterdam/Philadelphia: John Benjamins.

Göpferich, Susanne (1998) Paralleltexte. In: Mary Snell-Hornby, Hans G. Hönig, Paul Kußmaul & Peter A. Schmitt (Hgg.) ([2]1999) *Handbuch Translation*. Tübingen: Stauffenberg, 184-186.

Grbić, Nadja (2008) Constructing interpreting quality. *Interpreting* 10 (2), 232–257.

Guibert, Jean (1979) *La parole. Compréhension et synthèse de la parole par les ordinateurs*. Paris: PUF.

Härtel, Johannes (2016) Vollautomatisches Dolmetschen – Möglichkeiten und Grenzen. *Lebende Sprachen* 61 (1), 67-116.

Haugen, Einar (1966) Linguistics and language planning. In: William Bright (Ed.) *Sociolinguistics: Proceedings of the UCLA Sociolinguistics Conference 1964*, The Hague/Paris: Mouton, 50-71.

He, Xiaodong & Acero, Alex (2011) Why Word Error Rate is not a good metric for speech recognizer training for the speech translation task? In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICSSAP), Prague, Czech Republic, 22-27 Mai 2011,* 5632-5635.

Jefford, Darren (2017) „How Microsoft Cognitive Services can help your apps communicate with people.“ https://channel9.msdn.com/Events/Build/2017/B8090 (20.08.2017).

Kade, Otto (1968) *Zufall und Gesetzmäßigkeit in der Übersetzung.* Leipzig: VEB Verlag Enzyklopädie.

Kitano, Hiroaki (1994) *Speech-to-speech Translation: A Massively Parallel Memory-based Approach*. Boston: Kluwer.

Kumar, Gaurav, Post, Matt, Povey, Daniel & Khudanpur, Sanjeev (2014) Some insights from translating conversational telephone speech. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICSSAP), Florence, Italy, 4–9 Mai 2014*, 3231-3235.

Kurz, Ingrid (2005) Akzent und Dolmetschen – Informationsverlust bei einem nichtmuttersprachlichen Redner. *Bulletin VALS-ASLA* 81, 57-71.

Lamnek, Siegfried ([4]2006) *Qualitative Sozialforschung: Lehrbuch.* Weinheim, Basel: Beltz.

Li, Changshuan (2010) Coping strategies for fast delivery in simultaneous interpretation. *The Journal of Specialized Translation* 13, 19-25.

Li, Haizhou, Bin Ma, & Kong Aik Lee (2013) Spoken Language Recognition: From Fundamentals to Practice. *Proceedings of the IEEE* 101 (5), 1136-1159.

Linville, Sue E. (2001) *Vocal Aging*. San Diego: Singular Thomson Learning.

Liu, Minhua & Chiu, Yu-Hsien (2009) Assessing source material difficulty for consecutive interpreting: Quantifiable measures and holistic judgment. *Interpreting* 11 (2), 244-266.

Mankauskienė, Dalia (2016) Problem Trigger classification and its applications for empirical research. In: Marina Platonova & Larisa Ilinska (Eds.) *Procedia-Social and Behavioral Sciences. International Conference; Meaning in Translation: Illusion of Precision, Riga Latvia, 11-13 May 2016,* 231, 43-148.

Mazzetti Andrea (1999) The Influence of Segmental and Prosodic Deviations on Source-text Comprehension in Simultaneous Interpretation. *The Interpreters' Newsletter* 9, 125-147.

Microsoft (2014) „Skype Translator Reviewers Guide." https://news.microsoft.com/download/presskits/skype/docs/SkypeTranslatorRG.pdf (20.08.2017).

Microsoft (2016) „Microsoft Translator brings end-to-end speech translation to everyone with the world's first Speech Translation API." https://blogs.msdn.microsoft.com/translation/2016/03/30/microsoft-translator-brings-end-to-end-speech-translation-to-everyone-with-the-worlds-first-speech-translation-api/ (20.08.2017).

Nielsen, Michael A. (2015) „Neural Networks and Deep Learning." http://neuralnetworksanddeeplearning.com (20.08.2017).

Nolan, James ([2]2005) *Interpretation: Techniques and exercises*. Bristol/Buffalo: Multilingual Matters.

Nord, Christiane (1988) *Textanalyse und Übersetzen. Theoretische Grundlage, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse*. Heidelberg: Julius Groos Verlag.

Nord, Christiane (2004) Loyalität als ethisches Verhalten im Translationsprozess. In: Ina Müller (Hg.) (2004) *Und sie bewegt sich doch. Translationswissenschaft in*

Ost und West. Festschrift für Heidemarie Salevsky zum 60. Geburtstag. Frankfurt am Main: Lang, 235-245.

Och, Franz J. & Ney, Hermann (2004) The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics* 30 (4), 417-449.

Pétursson, Magnús & Neppert, Joachim (2002) *Elementarbuch der Phonetik*. Hamburg: Buske Verlag.

Pieraccini, Roberto & Rabiner, Lawrence (2012) *The Voice in the Machine: Building Computers that understand Speech.* Cambridge, Mass: MIT Press.

Pöchhacker, Franz (2004) *Introducing Interpreting Studies*. London/New York: Routledge.

Seleskovitch, Danica & Lederer, Marianne (1989) *Pédagogie raisonnée de l'interprétation*. Paris: Didier erudition.

Skype (2014) „Skype Translator – How it Works." http://blogs.skype.com/2014/12/15/skype-translator-how-it-works/ (20.08.2017).

Skype (2014a) „Now you're speaking my language (literally)." https://news.microsoft.com/download/presskits/skype/docs/SkypeTranslatorInfo.pdf (20.08.2017).

Skype (2015) „你好! Ciao! Skype Translator Now Speaks Italian and Mandarin." http://blogs.skype.com/2015/04/08/%E4%BD%A0%E5%A5%BD-ciao-skype-translator-now-speaks-italian-and-mandarin/ (20.08.2017).

Skype (2015a) „Skype Translator Preview App Brings Our World a Little Closer by Introducing French and German." http://blogs.skype.com/2015/06/18/skypetranslator-preview-app-brings-our-world-a-little-closer-by-introducing-french-and-german/ (20.08.2017).

Skype (2015b) „Olá! Skype Translator welcomes Portuguese." http://blogs.skype.com/2015/12/10/ola-skype-translator-welcomes-portuguese/ (20.08.2017).

Skype (2015c) „Skype Translator Preview Is Coming to the Skype for Windows Desktop App." http://blogs.skype.com/2015/06/08/skype-translator-available-for-skype-for-windows-desktop-by-end-of-summer/ (20.08.2017).

Skype (2015d) „Skype Translator Preview Access Just Got Easier!“ http://blogs.skype.com/2015/05/12/skype-translator-preview-access-just-got-easier/ (20.08.2017).

Skype (2015e) „Skype Translator unveils the magic to more people around the world.“ http://blogs.skype.com/2015/10/01/skype-translator-unveils-the-magic-to-more-people-around-the-world/ (20.08.2017).

Skype (2016) „About Skype.“ http://www.skype.com/en/about/ (20.08.2017).

Skype (2016a) „Salam, Skype Translator now speaks Arabic.“ http://blogs.skype.com/2016/03/08/salam-skype-translator-now-speaks-arabic/ (20.08.2017).

Skype (2016b) „Привет! Skype Translator says Hello to Russian.“ https://blogs.skype.com/news/2016/10/11/привет-skype-translator-says-hello-to-russian/ (20.08.2017).

Skype (2016c) „Skype Translator empowers more people.“ http://blogs.skype.com/2016/01/13/skype-translator-empowers-more-people/ (20.08.2017).

Skype (2017) „Skype Translator offers Japanese as its 10th real-time spoken language.“ https://blogs.skype.com/news/2017/04/06/skype-translator-offers-japanese-10th-real-time-spoken-language/ (20.08.2017).

Spilker, Jörg, Klarner, Martin & Görz, Günther (2000) Processing of Self-Corrections in a Speech-to-Speech System. In: Wolfgang Wahlster (Ed.) (2000) *Verbmobil. Foundations of speech-to-speech translation*. Berlin/New York: Springer, 131-140.

Tannen, Deborah (1980) Spoken/Written Language and the Oral/Literate Continuum in Discourse. In: *6th Annual Meeting of the Berkeley Linguistics Society. California, Berkeley, 16-18 Februar 1980.* California, Berkeley: University of Berkeley, 207-218.

Traunmüller, Hartmut & Anders, Eriksson (1994) *The frequency range of the voice fundamental in the speech of male and female adults*. Manuscript, Department of Linguistics, University of Stockholm.

Vipperla, Ravichander (2011) *Automatic Speech Recognition for Ageing Voices.* Dissertation, University of Edinburgh.

Wahlster, Wolfgang (2000) Verbmobil. Foundations of speech-to-speech translation. Berlin/New York: Springer.

Waibel, Alex & Fügen, Christian (2008) Spoken Language Translation. *IEEE Signal Processing Magazine*, 71–73.

Wang, Ye-Yi, Acero, Alex & Chelba, Ciprian (2003) Is Word Error rate a good indicator for spoken language understanding accuracy? In: *IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, US Virgin Islands, 30 November-3 December 2003*, 577-582.

Wendt, Chris (2010) „Better translations with user collaboration – Integrated MT at Microsoft." http://amta2010.amtaweb.org/AMTA/papers/4-10-Wendt.pdf (20.08.2017).

Wendt, Chris (2016a) „Speech translation in human-to-human interaction: Skype Translator." http://www.interact25.org/downloads/WENDT_InterAct25_ ChrisWendt_160714.pdf (20.08.2017).

Wendt, Chris (2016b) Behind Skype's machine interpreting. *Multilingual* 27 (6), 46f.

Wonisch, Alexander (2017) Skype Translator: Funktionsweise und Analyse der Dolmetschleistung in der Sprachrichtung Englisch – Deutsch. Masterarbeit, Universität Wien.

Yo, Hamada (2015) Shadowing: Who benefits and how? Uncovering a booming ELF teaching technique for listening comprehension. *Language Teaching Research* 20 (1), 35-52.

Anhang

Anhang 01 - Skalen mit Evaluierungsbeispielen

Tabelle 20: Skala 01 - Semantik

SKALA 01* - Semantisch		JA	JAIN	NEIN
F_1	„noticing the gap“	noticing the gap	noticing that gap; noticing gap	Bsp. notion that gap, nothing they can, notion the gap, thing that gap, notice in the gap, noticing Jack, nursing the gap, nazi that ***, nursing gap, nursing yet, not dissing the gap, not a single gap,
F_2 F_3	„inter-language“	interlanguage; inter language	internal languague,	into language, interim language, internet english, ages language, in true language, internal world, anuar language, tearing internal, the language,
F_4 F_5	„input flooding“	input flooding		in perfecting, in perverse floating out, from float, in put flooding, imports lighting, import flooding, oil put flooding, input for putting up, him put flooding, sporting for flighting, input for the text, input for them, implementing, input slighting, in Sex, input footing, improved floating, input florine, input flat, simplifying, providing, important for, input for them, input fighting
F_6	„second language acquisition researcher “	second language acquisition researcher	second language acquisition research; second language acquisition researchers	second language courses acquisition, second pushing researcher, second language acquisition reception, second language acquisition deception, Language Acquisition Reese,second language exercise search, second acquisition researcher
F_7	„Focus on Form“	Focus on Form	focused on form; focus in form; focus on the form; focus of form	focus o four, focus on pharma, focus on phone, focusing form, focus on former, focus on for him. free *** ISM form, for some foreign, focus on farm, focused on phone, focuses form, *** is in form,access low the fox from, *** so far, focus on farming, form sexually, I’m form, focus on former, boss Francie. Anna Christine., probably some form,focus on four minutes, it will farm and has weapons, I’m form it says, focus on four, forks here’s some form to
F_8	„Focus on Meaning“	Focus on Meaning	Focus on the meaning, focusing on meaning	focus on meeting, focusing meaning, focus on Manning, focus on me now,
F_9	„Focus on Forms“	Focus on Forms	focus on form as, focus on form	books on forms, focus on farms, focus on Ford’s, focus on form as, focus on problems, focus on forums, focus on. focus on their forms, focus on informants
R_1	„text editor thing“	text editor thing	Editor thing; text editor	tech sector, text ed is the same, textured, male text her thing, texted ge of the thing, take Zander thing,
R_2	„wanna“	wanna	Want to	

R_3	„guy“	Guy		
R_4	„stuff“	stuff		Staff
R_5 R_6 R_7	„yeah“	yeah	Yes	Year, air,
E_1 E_2	„Ellis“	Ellis		Alice, hours, others, allah’s, enlist, ends, list, vieilleus, hours, Harris, at least, endless, palace, Diana’s etc.
E_3	„French“	French		fridge
J_1	„2016“	2016		Sending 16, 2 6 16, 16, 2006, 8,
K_1	„comm-unicative turn“	comm-unicative turn	communicati ve turning, communique turn, communicati on turn,	communication journal. communicator and, completely turn, communicative term, communicative time turning, communicative china, communicative charm, communicate turn, communicated to turn,communicative kerning
K_2	„vary“	vary		very, Barry, bury, pray, bari, buried
K_3	„intake“	intake		been sick, inside, incentives well, antigens, Inti,
K_4	„either“	either		oughta, raises, mice, or eyes. I did,
K_5	„arise“	arise	rise, rises,	rice,
K_6	„error“	error		narrow, next cinara, era, air, are, rode

Tabelle 21: Skala 02 - Phonetik

SKALA 02 – Phonetisch		JA	JEIN	NEIN
A_1	FOF	FOF, SOS, FAO,	affo effe, Sos,	more Sos, more ffff, for every half, core more Fomc, more effort of length, our death forever, forever,
A_2	FOM	FOM, FOMC	ephoren b, Fomc	every whim, ever,, ham, eco I am, ever lamb, at following, as I am, ask, 4. A.M.
F_10	With a capital S at the end	with a capital as at the end, with a capital ** at the end, with a capital at the end	with a capital is at the end, with a capital asset the end, with a capital as in the end, with a capitals *** at the end,	with the capital ists at the end, with the capture death at the end, they capital *** at the end, capital less, in the capital mess at the end, with the cops to I said, Well at the end, with a capitalism band, with a capital idea, with a capital *** and the end with a capital less at the end, with a capital sin, camera less at the end, with a capital assets tea end, capitalist at the end, but it kept as the end

Tabelle 22: Skala 03 - Satzzeichen

SKALA 03* - Satzzeichen		JA	JEIN	NEIN
S_1	because in, yeah. And	.yeah,; ,yeah, .yeah.; ,yeah,	,yes,; .yes.; ,yeah; yeah,; yeah.; .yeah	
S_2	it has ... well, I	,well, ; . well, ;	Has well,; ,well I;	has what I can send you to, as well. And; it hurts well
S_5	for example.	,for example	.For example	

Tabelle 23: Skala 04 - Spezielle Fälle

SKALA 04* Spezielle Fälle	JA	JEIN	NEIN

S_3	to contern contain a certain form	To contain a certain form, to „*Wort unbekannt*“ contain a certain form	To can turn contain a certain phone	
S_4	„what I what I think“	That’s what I think; that’s what I what I think		what I would I think, when I think I said, That’s why I don’t think, with the cops to I said, yes, that’s what I am what I, that’s what I want. I think we see. what I would I think, when I think I said, That’s why I don’t think, with the cops to I said, yes, that’s what I am what I, that’s what I want. I think we see.
S_6	the prob- the problem	„the prob- the problem“	The prob problem, the problem, the prob- problem, the problem the problem	The pro- the problem; the prop; the problem
I_1	„No? Okay.“	.No, okay.; ,No, okay,; .no, okay,; ,no, okay.	No. Well, ; no.; now, okay.	
I_2	„Limited. Even if“	Even if; even if		

*Groß- und Kleinschreibung blieb für die Auswertung unbeachtet

Anhang 02 - Originaltranskript des Ausgangstextes

Well, first of all I’d like to thank you for contacting me in this matter. Our approach uh what we are doing is we are using focus on form. I don’t know whether you have heard of focus on form, no? Ok, ah well, I’ll give you a quick explanation then. So, focus on form or FOF as it is abbreviated it’s not a very new concept but newer than all the other concepts to language teaching and you’re teaching a subject but you’re teaching a language at the same time. uh

So in the beginning you know we had it was only used for language teaching but in the beginning of language teaching the focus was merely on explicit grammar teaching so probably at school you also had explicit grammar teaching when you learned a foreign language or you learned as how does the past tense work and how does the present tense work. And then there was a communicative turn in language teaching um (harrumph) and they started with what we now call focus on meaning or FOM um because they saw that explicit language teaching or focus on formS with a capital s at the end you got a very high accuracy but you got a very low fluency because

you know people would know everything about language explicitly but they couldn't talk. Um That's what I what I think we see in many French people for example. They have very explicit language teaching so they know all the grammar rules but they're unable to say one straight English sentence most of the time. Ah And then you had the communicative turn so you know of course people said this explicit language teaching doesn't work so let's go to the exact opposite and they started doing um (harrumph) this focus on meaning were they say: well, let's not talk about language at all let's just interact the way it is for example you know when you teach your baby or child a foreign language you don't explicitly talk about past tense but you just interact in meaningful communication. And with that you get a very high fluency but the accuracy is lower. And focus on form combines both of these approaches so we're trying to have language learning in meaningful interactions and that's why it is so suitable for subject language teaching um because the prob the problem we get with that is attention. 'Cause you know your working memory is limited – *even if you are a very smart person* – your working memory is still limited. So what we're doing is we are doing subject language teaching and then um when the possibility arises either because we think that it does arise as a teacher or because a student makes an error we shift the focus on the language and then we have these very short episodes were language is discussed and then we go back to the subject language teaching basically. And this way we try to promote both fluency and accuracy and subject teaching, what's subject learning if we talk from the student's perspective as well.

Focus on form it's actually…but if I say approach I'm using the wrong word ah because focus on form it has …well, I can send you the publication if you want. There has been a very recent publication by Ellis. Just give me your E-Mail or send me your E-Mail in a text editor thing. Then I can just copy it and send you the publication. Because it's coming out in 2016 by Ellis. So, the guy who wrote all the important stuff about focus on form. So if you wanna um read about it I would I would suggest starting with this because it's the newest thing uh where he calls focus on form more a set of various techniques than an approach. So if I'm saying approach it's actually wrong. Ya So we use implicit techniques such as input flooding of texts or input flooding of the language of the teacher, which is very hard of course if you talk to people and you artificially have to alter your language to contern contain a certain form. Well, we train

them on doing that and then we have more explicit techniques like um feedback. Um And we also try to vary the feedback types ah because well in research we have seen that if you have a very explicit feedback such as saying, "well, that's just plain wrong what you are saying and this and this is the correct form" um that tends to have an immediate effect usually, because you know students notice if you tell them that they are wrong um and they probably correct themselves. But if you use implicit techniques such as just uh rephrasing the utterance the student has said, it's possible that they don't react immediately. But studies have shown that that is has a longer effect, so in the long run they would kind of profit more from the implicit feedback. So we try to combine kind of both, 'cause we want them to be accurate immediately and on their own run.

Ah Well, I think it's very interesting that we are talking about this as you are a second language acquisition researcher because in … yeah and the role of the working memory and the attention I was talking about earlier, ah because you know second language researchers we have this kind of input um so all that kind of comes at you, all the language that comes at you and then in focus on form we have the intake as well, so that is kind of the tiny part your working memory actually processes. And that's why it's very important in focus on form to play with the attention. So we are consciously trying to um manipulate the intake of our students, which is very hard of course 'cause of their working memory, 'cause they have to constantly um kind of reflect what is their their interlanguage, so what is the language they already know and what is the intake they're getting kind of comparing those to each other. Yeah And that has a lot to do with noticing and then we call that the "noticing the gap" um between the interlanguage and the intake that comes from the outside. So I think there is a lot of very interesting research possibilities for second language acquisition there.

Anhang 03 - Transkript des Ausgangstextes für SprecherInnen

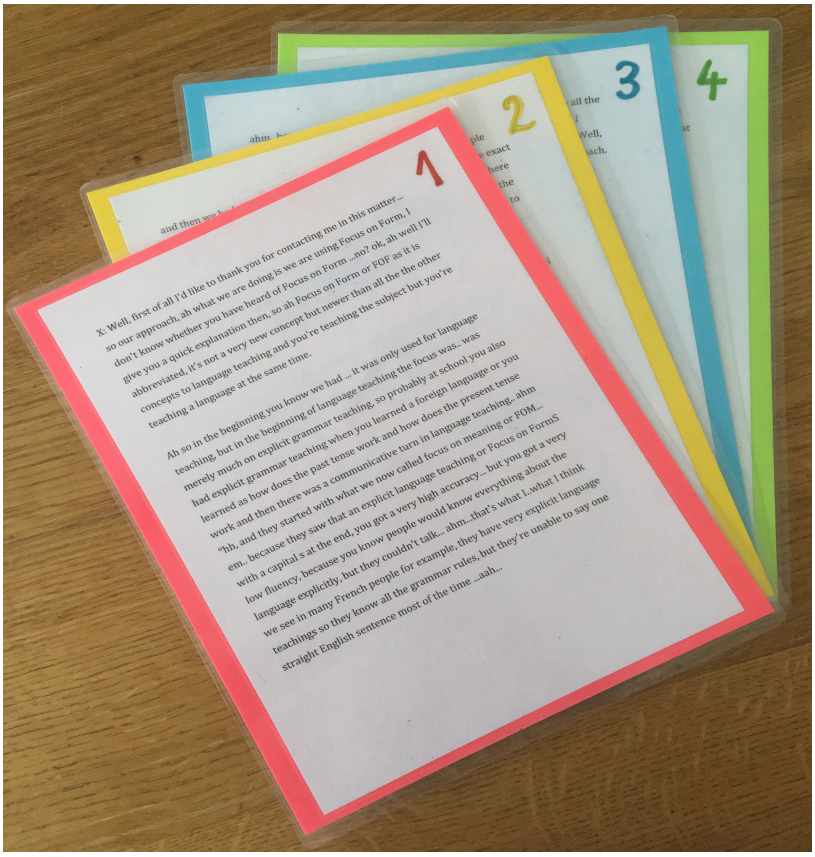

Abbildung 29: Transkript des Ausgangstextes für SprecherInnen

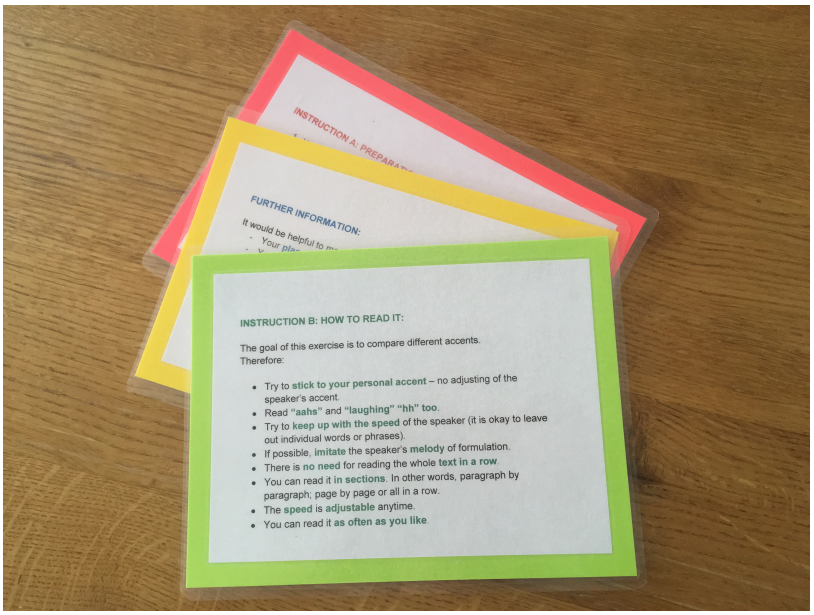

Abbildung 30: Instruktionskarten für SprecherInnen

Anhang 04 - Abstract

Abstract (English)

Machine translation systems inclusive of automatic speech recognition software are no longer science-fiction but form part of our everyday life. Skype Translator is one of those free speech-to-speech systems of which speech recognition software is the first of four phases generating this system. The aim of this study was to find out how skype translator's speech recognition reacts on problem triggers in the source text. For this study Skype Translator performed 64 machine translations altogether, for which 32 speakers presented the same source text. Sequences from its results were analysed and compared to each other. Results showed that in particular automatic learning had significant influence on some of the categories, while others were not influenced at all.

Abstract (Deutsch)

Maschinelle Dolmetschsysteme und damit einhergehende Spracherkennungssoftware sind keine futuristische Phantasie mehr, sondern werden bereits in vielen Bereichen unseres täglichen Lebens eingesetzt. Skype Translator ist eines der frei verfügbaren Dolmetschsysteme, in denen Spracherkennungssoftware als erste von vier Phasen den ersten Grundpfeiler darstellt. Ziel dieser Arbeit war es herauszufinden, durch welche Faktoren es in der Spracherkennung von gesprochenen Texten für die Software zu Problemen kam und wie diese gelöst wurden. Insgesamt führte Skype Translator 64 maschinelle Dolmetschungen durch, für die derselbe Ausgangstext von 32 SprecherInnen Skype Translator vorgespielt wurde. Daraus wurden einzelne problemauslösende Textsegmente miteinander verglichen und ausgewertet. Auffallend war, dass vor allem automatisiertes Lernen einige Kategorien stark beeinflusste, während dies bei einzelnen anderen wenig Einfluss zeigte.