Karlsruher Institut für Technologie

Automatische Eingabeselektion in der Maschinellen Übersetzung

Bachelorarbeit
von

Bettina Weller

An der Fakultät für Informatik
Institut für Anthropomatik und Robotik (IAR)

Erstgutachter:	**Prof. Dr. Alexander Waibel**
Zweitgutachter:	**Dr. Sebastian Stüker**
Betreuender Mitarbeiter:	**Dr. Jan Niehues**

Bearbeitungszeit: 8. Juli 2015 – 7. November 2015

Erklärung:

Ich versichere hiermit, dass ich die Arbeit selbstständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe und die Satzung des Karlsruher Instituts für Technologie zur Sicherung guter wissentschaftlicher Praxis beachtet habe.

Karlsruhe, den 7. November 2015

Bettina Weller

Kurzzusammenfassung:

Im Folgenden werden Untersuchungen durchgeführt mit dem Ziel maschinelle Übersetzung zu verbessern. Dabei wird der Ansatz betrachtet, dass Daten besser zu übersetzen sind, wenn sie einfacher sind. Ebenso sollte eine einfachere Übersetzung eine bessere Qualität aufweisen.
Um festzustellen, wann Daten einfacher sind, werden Kriterien benötigt, die dies festlegen. In dieser Arbeit werden solche mögliche Kriterien auf ihren Einfluss untersucht und daraus entstehende Ergebnisse evaluiert und zusammengefasst. Dafür werden verschiedenen Experimente vorgenommen und automatisch evaluiert.
Die Untersuchungen finden dabei auf mehreren Eingaben und deren Übersetzungen statt. Diese sind Paraphrasierungen voneinander. Es ist nicht garantiert, dass die Eingaben eine unterschiedliche Einfachheit aufweisen. Anhand einer manuellen Evaluation werden sie daher qualitativ bewertet und eingeordnet.
Insgesamt bildet diese Arbeit einen Ansatz für weitere Untersuchungen in derselben Richtung.

Inhaltsverzeichnis

1 Einleitung

Maschinelle Übersetzung ist inzwischen weitverbreitet und wird oft genutzt. Die einfachste Version ist ein digitales Wörterbuch, das nur einzelne Worte sucht und deren Äquivalent in der entsprechenden Sprache ausgibt. Die Übersetzung ganzer Sätze bleibt dabei dem Menschen selbst überlassen. Weitergehende Software kann Sprachen erkennen und zusätzlich komplette Sätze in die gewünschte Sprache übersetzen. Noch weitergehende Software reagiert auf Spracheingabe und bietet Sprachausgabe an.

Dabei gibt es qualitative Unterschiede zwischen verschiedenen Übersetzungen. Diese können sowohl grammatikalischer wie auch stilistischer Art sein oder lediglich auf der Wortwahl beruhen. Trotz mangelhafter Übersetzung kann diese verständlich sein, insbesondere bei kürzeren, unkomplizierteren Sätzen. Teilweise kann die Verständlichkeit oder auch der Stil durch andere Wortwahl erhöht werden. So kann statt dem Wort *Metapher* beispielsweise auch *Wortmalerei* oder eine Umschreibung genutzt werden, was die Verständlichkeit erhöht. Wird dagegen *Metapher* statt einer Umschreibung genutzt, verbessert dies den Stil. Somit gibt es nicht nur eine einzige richtige Übersetzung.

Verschiedene Übersetzungen durch maschinelle Übersetzung können durch Nutzung unterschiedlicher Übersetzungssysteme oder durch Bearbeitung der Eingabe- bzw. Ausgabedaten entstehen. Wie in obigem Beispiel kann ein Wort sowohl bei Vor- als auch Nachbearbeitung durch eine Synoym ersetzt werden, wobei das genutzte Übersetzungssystem dieses nun anders übersetzen würde bzw. es durch die Nachbearbeitung verändert wurde.

1.1 Motivation

Ziel dieser Arbeit ist es die Qualität einer Übersetzung als Ergebnis von zwei Eingaben zu erhöhen, insbesondere deren Verständlichkeit. Bisher sind die Ergebnisse von maschineller Übersetzung noch nicht perfekt, weisen jedoch unterschiedliche Erfolge vor, abhängig von Ziel- und Quellsprache, und machen weiterhin Fortschritte.

Abhängig vom Anwendungsgebiet liegt eher die inhaltliche Verständlichkeit bzw. die sprachliche Qualität im Vordergrund einer Übersetzung. In manchen Fällen ist allerdings auch beides von Bedeutung, beispielsweise bei simultanen Übersetzungen öffentlicher Reden.

Es gibt verschiedene Ansätze um die Qualität zu verbessern, wie eine Vereinfachung der Eingabe. Ist die Eingabe einfacher aufgebaut, also weniger komplex, so sollte auch der Vorgang des Übersetzens einfacher werden. Grund dafür ist, dass z.B. bei einer einfacheren Satzstruktur weniger zu beachten ist, und die Häufigkeit eines ähnlichen Satzes steigt. Das Gleiche gilt für eine Anpassung des Vokabulars an das Übersetzungssystem. Häufigere Wörter kann es passender übersetzen. Dies sind Beispiele für Kriterien, welche Einfachheit beeinflussen. Mit Einfachheit ist gemeint, dass die Eingabe für das Übersetzungssystem einfacher zu übersetzen ist.

Um eine einfachere Eingabe zu erhalten, muss zunächst bekannt sein, wie ein System automatisch erkennen soll, ob eine Eingabe einfacher ist als eine andere. In dieser Arbeit werden dafür zwei Eingaben betrachtet und auf unterschiedliche Kriterien untersucht, um entscheiden zu können warum eine Eingabe einfacher ist. Für die Qualitätsbewertung wird hierbei die Evaluationsmetrik BLEU verwendet und satzweise betrachtet.

1.2 Gliederung

Zunächst werden einige Grundlagen im Bezug auf maschinelle Übersetzung vorgestellt. Anschließend werden Arbeiten in ähnlichen Forschungsrichtungen vorgestellt (Kapitel 3) und ihr Bezug zu dieser Arbeit erläutert. Weiterhin wird in Kapitel 4 die Ausgangslage vor jeglichen Experimenten erläutert, sowie das Thema dieser Arbeit näher erklärt. Die Ausgangslage beinhaltet die verwendeten Daten, sowie einen Überblick zu deren Erstellung. Anschließend folgt eine Beschreibung des experimentellen Vorgehens innerhalb dieser Arbeit, um das dort näher erläuterte Ziel zu erreichen. Auf der Ausgangslage der Daten werden verschiedene Experimente vorgenommen und in Kapitel 5 beschrieben, ausgewertet und interpretiert. Als Abschluss folgt eine Zusammenfassung der Ergebnisse (Kapitel 6) mit Ausblick auf zukünftige Fortführungsmöglichkeiten dieser Arbeit (Kapitel 7).

2 Grundlagen

In der maschinellen Übersetzung geht es um das Transformieren von einem Text in einer Quellsprache in einen Text in eine oder mehrere Zielsprachen. Hierbei gelten sowohl Verständlichkeit wie auch Qualität der Übersetzung als Ziel. Für die Transformation existieren verschiedene Ansätze, die im Folgenden kurz vorgestellt werden. Anschließend wir auf weitere Merkmale der maschinellen Übersetzung sowie verschiedene Bewertungsmaße, sogenannte Evaluationsmetriken, eingegangen, mit deren Hilfe die Qualität einer Übersetzung eingeschätzt werden kann.

2.1 Maschinelle Übersetzung

Prinzipiell existieren 3 Basisansätze der maschinellen Übersetzung:

- Transfer auf eine Interlingua: eine sprachenunabhängige Darstellung
- regelbasierte Übersetzung auf verschiedenen Abstraktionsebenen: Transfer von Quell- in Zielsprache auf syntaktischer und semantischer Ebene
- direkte Übersetzung: Transfer von Quell- in Zielsprache anhand bisher bekannter passender Übersetzungen

Zusätzlich sind Kombination der Grundtechniken und sowohl Vor- als auch Nachbearbeitungen möglich. Die einzelnen Stufen sind in Abb. 2.1, dem Vauquois-Dreieck, aufgezeigt. Hierbei ist ganz unten der Ansatz der direkten Übersetzung zu sehen, darüber jener der verschiedenen Ebenen des regelbasierten Übersetzens und ganz oben der Ansatz der Interlingua. Für den regelbasierten Ansatz werden syntaktischer und semantischer Transfer als Zwischenebenen angesehen, welche durch Aufstellung von Regeln und deren Anwendung zu erreichen sind. Dies wird in der Regel in Form von Grammatiken realisiert.

Im Folgenden wird auf den Ansatz der direkten Übersetzung eingegangen. Hierbei existieren verschiedene Ausprägungen. Wie bei einem Wörterbuch kann man Übersetzungen speichern und von diesen die

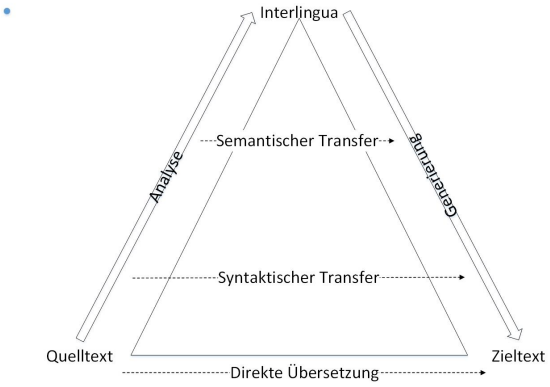

Abbildung 2.1: Vauquois-Dreieck: Darstellung der Basisansätze für maschinelle Übersetzung (Interlingua, Transfer und direkte Übersetzung); Abbildung in Anlehnung an Koehn (2009)

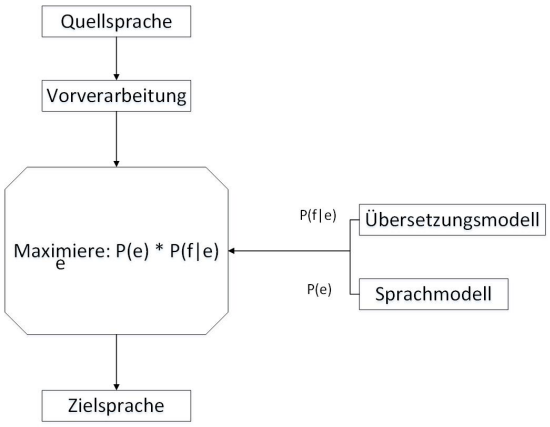

Abbildung 2.2: Übersetzungsystem und seine Komponenten (Vorverarbeitung und eigentlicher Übersetzungsprozess); Abbildung in Anlehnung an Koehn (2009)

passendste auswählen und modifizieren. Modifizierte Übersetzungen kann man im Falle einer Wiederholung erneut verwenden. Da normalerweise mehrere verschiedene Übersetzungen möglich sind, können diesen Wahrscheinlichkeiten zugewiesen werden, und somit kann daraus die wahrscheinlichste Übersetzung zusammengebaut werden. Dieser Ansatz wird *statistical machine translation* (SMT) genannt. Dafür wird das System auf *parallelen Korpora*, zweisprachigen Daten, trainiert.

Bei SMT wird das Ziel verfolgt, eine Übersetzung zu finden, welche die Wahrscheinlichkeit nach Gleichung 2.1 maximiert.

$$\hat{e} = \underset{e}{argmax} P(e) * P(e|f) \tag{2.1}$$

Dabei ist f ein Satz aus der Quellsprache und e ein Satz aus der Zielsprache, welcher die höchste Wahrscheinlichkeit aufweist, die zugeörige Übersetzung zu f zu sein. *P(e)* gibt die Wahrscheinlichkeit für den Zielsatz an, während *P(e|f)* die bedingte Wahrscheinlichkeit angibt, dass e in der Übersetzung genutzt wird und f der zugehörige Quellsatz ist. Somit ist $\hat{e}$ die maximale Wahrscheinlichkeit für die Übersetzung von f aus der Quellsprache zu e in die Zielsprache.

In Abbildung 2.2 werden nochmals die für eine Übersetzung wichtigen Komponenten aufgezeigt. Auf den in der Quellsprache vorhandenen Daten können vor Beginn der Übersetzung Vorverarbeitungen vorgenommen werden, beispielsweise Fehlerkorrekturen. Um während der Übersetzung Gleichung 2.1 zu maximieren, kommen das *Translation Model* und das *Language Model* zum Einsatz. Das *Translation Model* sorgt für die Übersetzung an sich. Die Rolle des *Language Model* dagegen ist die Anpassung an die Regeln der Zielsprache.

2.1.1 Translation Model

Beim *Translation Model* gibt es einen wortbasierten Brown u. a. (1993) und einen phrasenbasierten Koehn u. a. (2003) Ansatz. Also eine Übersetzung Wort für Wort einzeln zu generieren oder für eine Gruppe von Wörtern. Bei beiden Ansätzen wird ein *Alignment* benötigt. Dieses gibt an, welche Wörter aus dem Quellsatz auf welche Wörter aus dem Zielsatz abgebildet werden und somit Übersetzungen voneinander sind. Für den phrasenbasierten Ansatz wird des weiteren eine *Phrase Table* erzeugt.

Quellsprache	Zielsprache	Wahrscheinlichkeit
home	Haus	0.658
home	Haus,	0.101
home	Zuhause	0.241

Tabelle 2.1: Beispiel für den Ausschnitt einer *Phrase Table* mit Wahrscheinlichkeiten

Phrase Table

Die *Phrase Table* enthält eine Liste von Phrasen, also eine Liste mit nichtleeren Folgen von Wörtern, welche linguistisch zusammenhängen. Hierbei stehen Phrasen aus Quell- und Zielsprache, welche gegenseitige Übersetzungen sind, gemeinsam als Phrasenpaare. Eine Phrase aus der Quellsprache kann mehr als eine Übersetzung haben, daher werden für diese Wahrscheinlichkeiten berechnet.

Um eine *Phrase Table* zu erzeugen, wird zunächst ein *Alignment* berechnet. Anschließend folgt die Extraktion von Phrasen aus diesem *Alignment*. Währenddessen ist darauf zu achten, dass alle Wörter einer Phrase auf Quellseite nur zu Wörtern aus der Phrase auf Zielseite oder zu keinem Wort aliniert sind. Dasselbe gilt auch für Phrasen auf der Zielseite. Hinzu kommt die Berechnung der Wahrscheinlichkeit eines Phrasenpaares. Diese hängt von dem während der Phrasenextraktion verwendeten Algorithmus ab, ebenso wie die Phrasenpaare selbst.

In Tabelle 2.1 wird der Ausschnitt einer *Phrase Table* beispielhaft dargestellt. In der linken Spalte steht die Quellphrase, daneben die Zielphrase und ganz rechts die zugehörige Wahrscheinlichkeit. Es können noch weitere Details in einer *Phrase Table* enthalten sein, wie etwa die zugehörigen Wortarten oder andere Werte.

2.1.2 Language Model

Wie bereits erwähnt soll das *Language Model* für flüssige, natürliche Sprache sorgen. Ähnlich wie bei dem *Translation Model* gibt es auch hier einen stochastischen Ansatz. In Trainingsdaten wird das Vorkommen von Wörtern gezählt und darauf basierend werden diesen Wahrscheinlichkeiten zugeteilt. Die Wahrscheinlichkeiten hängen von den vorherigen Wörtern innerhalb eines Satzes ab, sodass bewertet wird, welches Wort wie wahrscheinlich als nächstes folgt. Dafür werden n-Gramme genutzt, welche sich wie folgend berechnen lassen:

$$p(w_N|w_{N-n+1},...,w_{N-1}) = \frac{count(w_{N-n+1},...,w_N)}{\sum_w count(w_{N-n+1},...w_{N-1},w)} \tag{2.2}$$

Wie in Gleichung 2.2 dargestellt werden n - 1 voranstehende Worte betrachtet. Daraus wird der Anteil berechnet, wie oft diese in Kombination mit dem n-ten Wort w_N vorkommen, im Vergleich zu deren Vorkommen mit einem beliebigen Nachfolgewort w. N bezeichnet dabei die Wortposition innerhalb eines Satzes. Neben diesem Ansatz existieren noch weitere, wie z.B. das auf kontextfreien Grammatiken beruhende *Language Model*.

In den zu übersetzenden Daten tauchen immer wieder unbekannte Wörter auf, sogenannte *out of vocabulary* (*OOV*) Wörter. In den Trainingsdaten sind sie nicht aufgetaucht und erhalten somit für das *Language Model* eine n-Gramm-Wahrscheinlichkeit von null. Für den Umgang mit *OOV*-Wörtern existieren verschiedene Ansätze. Man kann sie direkt übernehmen oder *smoothing*-Techniken (Chen und Goodman (1998)) verwenden.

Perplexität

Die Perplexität wird genutzt, um ein *Language Model* bewerten und mit anderen vergleichen zu können. Für ihre Berechnung wird die Kreuz-Entropie (Gleichung 2.3) genutzt.

$$H(p_{LM}) = -\frac{1}{n} * \sum_{i=1}^{n} log(p_{LM}(w_i|w_1,...,w_{i-1})) \tag{2.3}$$

Dabei wird für jedes Wort die n-Gramm-Wahrscheinlichkeit in Abhängigkeit des *Language Model* berechnet und deren Logarithmus aufsummiert. Bei der n-Gramm-Wahrscheinlichkeit werden alle voranstehenden Worte betrachtet. Das Ergebnis wird anschließend gemittelt. Für die Berechnung werden Testdaten benötigt und für einen sinnvollen Vergleich der Ergebnisse müssen dieselben Daten verwendet werden.

Aus der Kreuzentropie lässt sich die Perplexität wie folgend berechnen:

$$PP = 2^{H(p_{LM})} \tag{2.4}$$

Die Perplexität gibt an, wie gut die durch ein *Language Model* gegebene Wahrscheinlichkeitsverteilung auf die Daten zutrifft, welche zuvor zur Berechnung der Kreuzentropie genutzt wurden. Je niedriger sie ist, desto besser modelliert ein *Language Model* deren Wahrscheinlichkeiten.

SRILM

SRI Language Modelling Toolkit (SRILM) ist ein Toolkit, mit welchem statistische *Language Model* erzeugt und bewertet werden können. Es wurde 1995 vom *SRI Speech Technology and Research Laboratory* entwickelt und wird seitdem weiterhin modifiziert.

In Stolcke (2002) werden einige Basisbefehle vorgestellt: Mit dem Befehl *ngram-count* kann ein n-Gramm basiertes *Language Model* erzeugt und mit dem Befehl *ngram* kann es bewertet werden. Bei der Bewertung wird die Perplexität im Bezug auf die angegeben Daten berechnet, sowie OOV gezählt. Durch weitere Parameter lassen sich weitere Dinge festlegen, wie etwa die n-Gramm-Art, welche genutzt werden soll. Ebenso existieren noch einige weitere Befehle zur Manipulation für *Language Model*, wie auch die Möglichkeit andere Arten von *Language Model* zu erzeugen. Diese und noch weitere Details zu SRILM werden in Stolcke (2002) genauer aufgezeigt.

2.1.3 Pre-editing

Pre-editing bezeichnet die Vorverarbeitung der zu übersetzenden Daten. Dabei kann unter Vorverarbeitung diverses zu verstehen sein, wie u. a. Fehlerkorrekturen (Grammatikfehler ebenso wie Rechtschreibfehler) oder größere Umstrukturierungen, um die Textstruktur zu vereinfachen. Ziel dieser Anpassungen ist es, eine bessere Qualität bei der anschließenden Übersetzung zu erzielen.

Somit dient *Pre-editing* einer Anpassung an das anschließend genutzte Übersetzungssystem. Dazu gehört auch eine Anpassung des Vokabulars, um unbekannte oder seltene Wörter zu vermeiden. Für einige Vorverarbeitungen existieren Programme, wie etwa eine Rechtschreibkorrektur, sodass die Verbesserungen automatisch vorgenommen werden können (vgl. Densmer).

2.1.4 Post-editing

Gemeinsam mit *Pre-editing* bildet *Post-editing* ein Forschungsgebiet. Bei *Post-editing* werden die Ergebnisdaten nach der Übersetzung überarbeitet und Fehler korrigiert oder die Daten anderweitig aufbereitet, wie etwa durch Normalisierung. Beide Bearbeitungen haben das Ziel die Qualität der Ergebnisübersetzung zu verbessern, dabei findet eine in der Quellsprache statt, während die andere in der Zielsprache statt findet.

Ob sowohl *Pre-editing* als auch *Post-editing* eingesetzt werden, hängt von der Qualität ab, die man erreichen möchte und auch von jener, welche man ohne Bearbeitung erreichen kann. Davon abhängig kann auch der Bearbeitungsgrad variieren. Abhängig vom Verwendungszweck sind beispielsweise sprachliche Fehler akzeptabel, wenn der Inhalt verständlich ist.

2.2 Evaluationsmetriken

Für die Qualitätsbeurteilung von Übersetzungen wird ein schneller Mechanismus benötigt. Somit können viele Daten mit relativ wenig Aufwand verglichen und evaluiert werden. Im Allgemeinen werden dafür automatische Evaluationsmetriken genutzt. Insbesondere in der Forschung findet dies Einsatz. Nachfolgend wird die Evaluationsmetrik BLEU erläutert, weitere sind beispielsweise TER (Snover u. a. (2006)) und METEOR (Denkowski und Lavie (2014)).

2.2.1 BLEU

BLEU ist eine Evaluationsmetrik, die eine Hypothese und eine Referenz benötigt. Die Hypothese ist die zu bewertende Übersetzung und bei der Referenz wird davon ausgegangen, dass sie eine korrekte Übersetzung ist, mit welcher die Hypothese verglichen wird. Dieser Ansatz basiert auf der Idee, dass eine Übersetzung gut ist, wenn sie ähnlich zu einer manuellen Übersetzung ist (vgl. (Papineni u. a. (2002))).

Gleichung 2.5 gibt die Berechnung des BLEU-Wertes an. Diese beruht auf einer modifizierten n-Gramm-Präzision. Dabei wird für die n-Gramme der Hypothese ihr Vorkommen in der Hypothese aufsummiert und durch ihr Vorkommen in der Referenz geteilt. In dem Fall, dass mehr als eine Referenz genutzt wird, wird das maximal mögliche Vorkommen innerhalb einer Referenz zur Berechnung genutzt. Für 1-, 2-, 3- und 4-Gramme wird die Berechnung satzweise durchgeführt. Der BLEU-Wert ergibt sich aus dem gewichteten geometrischen Mittel und liegt zwischen Null und Eins.

$$BLEU = BP * exp(\sum_{n=1}^{4} w_n * log(p_n)) \tag{2.5}$$

Um den Fall abzudecken, dass die Hypothese kürzer ist als die Referenz, sich also eindeutig unterscheidet, wird eine Strafgewichtung *BP* (brevity panelty) genutzt. Durch diese bekommen die Ergebnisse kürzerer Hypothesen einen niedrigeren BLEU-Wert.

3 Weiterführende Literatur

Wie bereits erwähnt sind *Pre-editing* und *Post-editing* eigene Forschungsgebiete in der maschinellen Übersetzung. Beide beschäftigen sich damit, die Qualität einer Übersetzung zu verbessern, betrachten diese allerdings zu unterschiedlichen Verarbeitungsmomenten. In dieser Arbeit betrachten wir hauptsächlich die Daten nach der Übersetzung. Daher werden in diesem Kapitel Ansätze vorgestellt, die ähnliche Ziele oder Vorgehensweisen verfolgen.

Während wir nach Merkmalen suchen, durch welche sich Eingaben qualitätsmäßig unterscheiden, verfolgt Barancikova und Tamchyna (2014) einen anderen Ansatz. Hier wird ein Übersetzungssystem erstellt, welches für eine Eingabe eine Paraphrase in derselben Sprache als Ausgabe erzeugt, welche evaluiert wird. Hierfür wird mithilfe einer Pivotsprache eine *Phrase Table* erstellt, also durch übersetzen in eine andere Sprache und anschließendes Zurückübersetzen. Diese *Phrase Table* wird weiterbearbeitet, um eventuelle Fehler herauszufiltern.
Die Evaluation ergibt, dass die originale Referenz nicht übertroffen wird, sowie dass die Grammatik des Ergebnisses bei zu geringer Gewichtung des *Language Model* sehr schlecht ist. Dagegen unterscheidet sich der Inhalt bei höherer Gewichtung stark von der Referenz. Anhand der Ergebnisse dieser Arbeit und Fortführung in dieser Richtung ist es möglich die Bewertung solcher Paraphrasen zu erleichtern. Der allgemeine Fall, dass Kriterien mehrerer Eingaben untersucht werden, entspricht in etwa dem Fall, dass Paraphrasen untersucht werden.

Auch in Zhao u. a. (2007) wird das Thema Paraphrasierung behandelt. Anders als in dieser Arbeit gehen sie dabei nicht satzweise, sondern wortweise bzw. phrasenweise vor und legen das Augenmerk auf den passenden Kontext. Ziel dahinter ist eine bessere Verständnisextraktion, sodass nicht nur die Anwendung im Bereich maschineller Übersetzung verbessert wird, sondern auch z.B. die Antwortfindung[1] bei Dialogen.
Im ersten Schritt werden Phrasenkandidaten aus ähnlichen Sätzen extrahiert und in einem weiteren Schritt auf Verwendbarkeit überprüft. Die Phrasenextraktion basiert auf zwei Prinzipien: Jeder Mensch tendiert zur Nutzung eines anderen persönlichen Vokabulars und lexikalische Paraphrasen haben gleichzeitig eine ähnliche syntaktische Funktion.
Das Vorgehen zeigt, dass auch Untersuchungen unterhalb der Satzebene sinnvoll sind. Ebenso zeigt es, dass Paraphrasen nicht nur statische Eigenschaften besitzen, wie z. B. ihre Länge. Zusätzlich besitzen sie auch dynamische Eigenschaften, sodass sich die Eigenschaften derselben Paraphrase abhängig von ihrem Vorkommen unterscheiden können.

Paraphrasen sind u. a. ein Ansatz, um mit unbekannten Wörtern umzugehen. Die Übersetzung solcher Wörter, *OOV*-Wörter, stellen ein besonderes Problem dar. In Marton u. a. (2009) wird ein Verfahren vorgestellt, wie Phrasenextraktion für Paraphrasen auf monolingualen Daten ermöglicht wird. Dafür werden Profile für *OOV*-Wörter angelegt. In diesen Profilen wird festgehalten, wie oft sie mit welchen Wörtern gemeinsam auftreten. Dabei werden nur Wörter betrachtet, die innerhalb eines maximalen Abstandes zu diesem Wort liegen. Über zwei Profile kann eine Ähnlichkeit berechnet und somit über mehrere Profile eine Bestenliste aufgestellt werden. Dies garantiert nicht, dass die gefundenen Kandidaten Paraphrasen sind.
In beiden Fällen ließen sich Ansätze wie unsere Arbeit nutzen, um die Kandidaten zu bewerten. Sie stellen somit eine Alternative zu Marton u. a. (2009) dar, welche jedoch Kontext und konkreten Inhalt vernachlässigt.

[1] Suche nach einer passenden Antwort innerhalb eines Dialogs

Dabei handelt es sich bisher nur um Paraphrasen, also inhaltlich gleichbedeutende Teile. Anstelle von ausschließlicher Paraphrasensuche wird in Mirkin u. a. (2009) nach einfacheren Wörtern und Phrasen gesucht, welche gleichbedeutend oder verallgemeinernd sein können. Es werden dabei Regeln aufgestellt, welches Wort zu welchem Wort spezialisierend bzw. in Gegenrichtung verallgemeinernd ist. Paraphrasen sind demnach beides in beide Richtungen.
Abhängig vom Kontext sind nicht alle Verallgemeinerungen korrekt. In der Anwendung sollen daher auf die Kandidaten, welche durch die Regeln herausgesucht werden, Kontextmodelle angewendet werden, um festzustellen, wie passend sie sind. Aus diesen Kandidaten und ihren Bewertungen wird eine Bestenliste zur weiteren Übersetzung generiert.
Diese Überlegungen bezüglich Verallgemeinerungen und Spezialisierungen verschiedener Eingaben stellt einen weiteren Schritt dar, welcher nach dieser Arbeit zu untersuchen ist. Wir untersuchen lediglich Merkmale der Eingabe und vernachlässigen dabei deren Bedeutungen. Hierbei ist zu beachten, dass bekannt ist, dass die für diese Arbeit verwendeten Daten inhaltlich gleich sind. Weiterhin ist es trotzdem möglich, dass der Informationsgehalt gewissen Schwankungen aufgrund der Übersetzungsprozesse unterliegt.
Das Aufstellen von Paraphrasen ist sowohl in Ziel- als auch in Quellsprache möglich. Dies würde dann unter die Kategorien *Pre-editing* und *Post-editing* fallen. Wie viel Nutzen derartige Bearbeitungen ergeben, ist abhängig von dem Ausmaß der Bearbeitung sowie der vorherigen Qualität der Daten.

In einem weiteren Artikel wird der Nutzen von *Pre-editing* genauer betrachtet. Anhand einer Studie, in welcher Nutzer eines Online-Forums Artikel anderer Nutzer nach vorgegeben Regeln bearbeiten müssen, wird in Bouillon u. a. (2014) untersucht, wie stark sich dies auf eine anschließende Übersetzung auswirkt.
Die Nutzer bekommen dabei Unterstützung durch ein Programm, welches nach vorher definierten Regeln, arbeitet und den Prozess des *Pre-editing* für den Nutzer erleichtern soll. Die Regeln lassen sich in manuell auszuführende und automatische unterteilen. Zu den manuellen Regeln gehören beispielsweise Zeit- oder Satzzeichenfehlerkorrekturen und auch Vorschläge, z.B. den Stil betreffend, welche das Programm liefert. Die automatischen Regeln werden von dem Hilfsprogramm direkt ausgeführt und betreffen die Wortreihenfolge oder das Ersetzen von Wörtern.
Desweiteren wurden die Übersetzungen der zuvor von den Nutzern editierten Daten, die Übersetzungen der unbearbeiteten Daten und Übersetzungen, welche zuvor von Experten bearbeitet wurden, verglichen. Durch *Pre-editing* wurde die Qualität der Übersetzung deutlich gesteigert. Es ist jedoch nicht eindeutig erkennbar, wie stark die Übersetzung nach Bearbeitung von Experten sich von jener von den Nutzern bearbeiteten unterscheidet. Über das Vornehmen von *Pre-editing* muss von der Übersetzung entschieden werden. Anders ist es bei *Post-editing*. Hierfür können die Übersetzungen betrachtet werden und anschließend kann entschieden werden, ob *Post-editing* angewendet werden soll.

Entscheidungen bezüglich *Pre-editing* lassen sich wie in der gerade beschriebenen Studie ebenfalls manuell treffen. Dabei kann der manuelle Einfluss über die Auswahl aus mehreren Vorschlägen hinausgehen und die vorherige Übersetzung als Unterstützung von Dolmetschern genutzt werden. Eine Studie in dieser Richtung wurde für das Europäische Parlament durchgeführt, um den möglichen Nutzen von maschineller Übersetzung zu prüfen (Poulis und Kolovratnik (2012)).
Für die Evaluation wurden Daten in Quellsprache sowie deren Übersetzungen in Zielsprache zur Verfügung gestellt. Diese sollen in Kategorien einsortiert werden, aus welchen sich die noch nötige Nachbearbeitung einschätzen lässt. Als Bewertungskriterien wurden inhaltliche Verständlichkeit und sprachliche Qualität betrachtet. Zusätzlich sollen die Probanden *Post-editing* vornehmen, wenn sie dies für erforderlich halten. Die dafür benötigte Zeit wird mit der Zeit verglichen, die ein Proband selbst zum Übersetzen benötigt. Daraus lässt sich der Nutzen von *Post-editing* abschätzen.
Basierend auf Poulis und Kolovratnik (2012) kann Eingabeselektion ebenso genutzt und getestet werden. Kombiniert man diese mit Verfahren, um Paraphrasen aufzustellen, wird der Aufwand für *Post-editing* verringert. Dies führt zu einer besseren Möglichkeit der praktischen Anwendung.

4 Eingabeselektion

Hat man mehrere Eingaben mit demselben Inhalt, kann man alle übersetzen und von diesen die beste auswählen. Dafür muss festgelegt werden, was die beste Übersetzung ausmacht. Kriterien können Evalutationsmetriken, aber auch andere Merkmale der Übersetzungen sein. Nicht immer sind alle für die Evaluationsmetriken benötigten Daten vorhanden, weshalb die Merkmale der Übersetzungen selbst betrachtet werden müssen.

Den Aufwand und die Zeit, welche für das Übersetzen aller Eingaben und der Evaluierung der Übersetzungen benötigt werden, kann man reduzieren, indem man nur einen Teil der Eingaben übersetzt. Dabei kann es passieren, dass die Eingabe, welche die beste Übersetzung liefert, ausgelassen wird. Um weiterhin die beste Übersetzung zu erhalten, die mit den gegeben Eingaben möglich ist, müssen die Eingaben sinnvoll gefiltert werden. Hierbei stehen im Normalfall keine Referenzen zum Vergleich zur Verfügung und es müssen andere Kriterien zur Auswahl betrachtet werden.
In Abbildung 4.1 wird die Situation abgebildet. Zunächst existieren *n* Eingaben aus welchen so ausgewählt wird, dass man die bestmögliche Übersetzung daraus generieren kann.

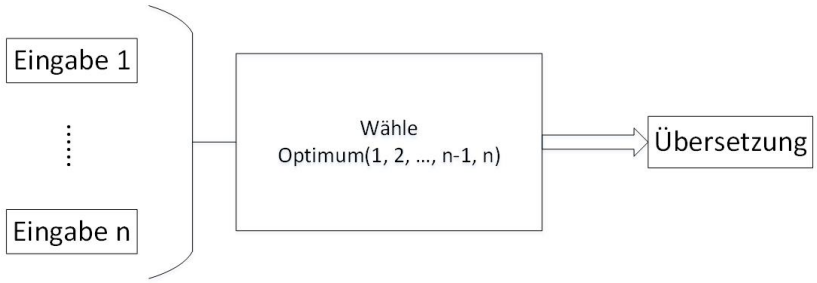

Abbildung 4.1: Eingabeselektion: Auswahl der besten Eingabe aus n Eingaben und Ausgabe einer Übersetzung

4.1 Ziel

Zu Beginn sind mehrere Eingaben vorhanden, welche übersetzt werden. Die beste und die schlechteste Übersetzung bilden die Begrenzungen der Qualität dieser Eingaben ohne weitere Bearbeitungen. Wir möchten die Qualität der Übersetzung weiterhin steigern. Dafür nutzen wir Kombinationen aus den Eingaben.

Aus den Eingaben wollen wir die einfachste auswählen, welche für das Übersetzungssystem besser zu übersetzen ist und somit ein besseres Ergebnis liefert. Dafür müssen zunächst Kriterien gefunden werden, welche die Einfachheit einer Eingabe beeinflussen.

4.2 Daten

Die in dieser Arbeit verwendeten Daten waren aufgrund eines vorherigen Projektes bereits vorhanden. Im Folgenden wird erläutert wie sie zusammengestellt wurden.

Aus verschiedenen Zeitungsartikeln wurden ein deutscher und ein französischer Text erstellt. Diese wurden von Hand übersetzt (Deutsch nach Französisch und Französisch nach Deutsch). Hierbei wurde jeder der Texte zweimal übersetzt, sodass zu dem französischen zwei deutsche und zu dem deutschen zwei französische existieren. Für diese gilt, dass die beiden Übersetzungen dieselbe Satzanzahl haben und die Sätze inhaltlich gleich sind.

Diese manuellen Übersetzungen wurden mittels maschineller Übersetzung wieder in ihre ursprüngliche Sprache zurück übersetzt. Damit existieren zwei Hypothesen mit einer Referenz je Sprache. Die manuellen Übersetzungen werden im Folgenden als *Quellseite Deutsch 0*, *Quellseite Deutsch 1* (manuelle Übersetzung ins Deutsche) und *Quellseite Französisch 0*, *Quellseite Französisch 1* (manuelle Übersetzung ins Französische) bezeichnet. Entsprechend werden die maschinellen Übersetzungen *Zielseite Französisch 0*, *Zielseite Französisch 1* (Deutsch nach Französisch) und *Zielseite Deutsch 0*, *Zielseite Deutsch 1* (Französisch nach Deutsch) genannt. Der Prozess des Übersetzens und Rückübersetzens der Daten wird in Abbildung 4.2 veranschaulicht. Auf den Rückübersetzungen werden Untersuchungen vorgenommen, um qualitative Unterschiede zu bemerken und mögliche Gründe für jene unterschiedlichen Qualitäten zu finden, um ein besseres Ergebnis zu erzielen.

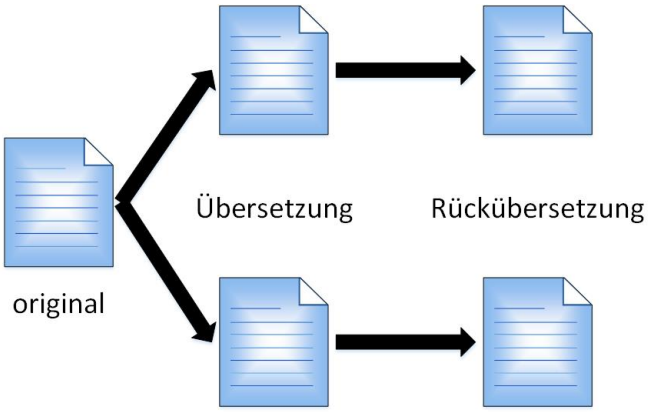

Abbildung 4.2: Entstehung der genutzten Daten

Für derartige Bewertungen gibt es verschiedene Ansätze. Einerseits kann eine Evaluation manuell von Menschen durchgeführt werden, andererseits können verschiedene Evaluationsmetriken (vgl. Kapitel 2.2) automatisch berechnet werden. Ersteres ist relativ subjektiv, jedoch sind auch die Evaluationsmetriken nicht immer aussagekräftig und können von menschlichen Meinungen abweichen.

4.3 Übersetzungssystem

Für die Übersetzungen der französischen und deutschen Texte wurde ein von *Quaero*[1] erzeugtes System genutzt. Es wird als *QuaeroP6* bezeichnet und wurde für eine interne Evaluation erzeugt. *Quaero* ist ein Zusammenschluss mehrerer Organisationen. Darunter fallen sowohl Unternehmen wie auch staatliche Einrichtungen. Ihr gemeinsames Ziel ist eine automatische Verarbeitung von multilingualen Daten verschiedener Form.

[1] www.quaero.org

Von Partnerorganisationen wurden *Korpora* zur Verfügung gestellt, welche während des Trainings genutzt wurden. Dazu gehören Übersetzungsdaten des Europäischen Parlaments, sowie Nachrichtendaten und Daten aus dem Internet. Aufgrund der unterschiedlichen Qualität wurden sie zunächst vorverarbeitet und gefiltert. Zu den verwendeten *Korpora* zählen sowohl bi- wie auch monolinguale *Korpora*. Beides wurde für monolinguales Training verwendet. Zusätzliche wurden Daten aus *Quaero P3* und *Quaero P4* genutzt.

Das Training fand unter Verwendung des GIZA++ Toolkits statt. Mit diesem wurde das *Alignment* unter Verwendung der *grow-diag-final-and*-Heuristik erzeugt. Weiterhin wurde die *Phrase Table* mit Moses Toolkit unter Nutzung von Kneser-Ney (Foster u. a. (2006)) generiert.

Die *Language Model* wurde mittels SRILM (Stolcke (2002)) als 4-Gramm *Language Model* erstellt. Dabei wurden Kneser-Ney oder Witten-Bell als *Smoothing* verwendet. Weitere verwendete *Language Model* sind *Gigaword Language Model*[2], *bilinguale Language Model*, *POS Lanugage Model* (basierend auf Schmid und Laws (2008)) und *Cluster Language Model* (basierend auf Och (1999)).

4.4 Kriterien

Erhält man mehrere Eingaben und möchte aus diesen die einfachste wählen, um die beste Übersetzung zu generieren, benötigt man zunächst Eigenschaften nach denen man die Auswahl trifft. In dieser Arbeit betrachten wir verschiedene Kriterien, welche die Einfachheit einer Eingabe beeinflussen können. Dazu erörtern wir zunächst, warum ein Kriterium eine Auswirkung haben kann, und führen in Kapitel 5 Experimente dazu durch.

4.4.1 Perplexität

Die Perplexität (Kapitel 2.1.2) gibt an, wie passend die Eingabe unter Betrachtung des verwendeten *Language Model* ist. Somit ist eine schlechte Perplexität, also eine hoher Wert, ein Anzeichen für schlechtere Eignung und damit schlechterer Wortwahl. Es werden also mehr Wörter in der Eingabe genutzt, welche das Übersetzungssystem nicht kennt oder weniger häufig gesehen hat.

In dieser Arbeit betrachten wir die Perplexität der Übersetzungsergebnisse anhand von zwei *Language Model*, welche mittels SRILM (Kapitel 2.1.2) erstellt wurden. Eines der *Language Model* hat einen großen Wortschatz (*big lm*), dagegen hat das andere einen kleineren dafür themenbezogeneren Wortschatz (*indomain lm*). Für die Berechnung der Perplexität wird ebenfalls SRILM benutzt. Auch wenn hier die Ausgabe betrachtet wird, gibt dies zugleich Aufschluss über die Eingabe, da die Ausgabe von der Eignung der Eingabe für das Übersetzungssystem abhängig ist.

Mithilfe der Perplexität wird aus zwei Texten satzweise ein neuer Text erstellt. Die Auswahl erfolgt wie folgend:

$$\text{Satz} = \begin{cases} 1 & , \quad \text{Perplexität(S1)} > \text{Perplexität(S0)} \\ 0 & , \quad \text{sonst} \end{cases} \tag{4.1}$$

Es wird der Satz mit geringerer Perplexität ausgewählt. Sollten beide Sätze dieselbe Perplexität aufweisen, so wird der Satz aus Text 0 gewählt. Dies wird paarweise auf die Texte 0 und 1 der Übersetzungen ausgeführt. Die zugehörigen Ergebnisse (*Ziel Deutsch* und *Ziel Französisch*) und deren Evaluation sind in Kapitel 5.2 zu finden.

[2] unter Verwendung des Gigaword-Korpus erstellt

4.4.2 OOV

OOV-Wörter beeinflussen die Perplexität, welche bereits in dem Abschnitt zuvor betrachtet wurde. Während dem Training sind diese Wörter nicht aufgetaucht und somit für das Übersetzungssystem unbekannt. Daher werden sie durch die *Language Model* nicht modelliert. Zusätzlich liegen keine Informationen vor, wie das Übersetzungssystem diese Wörter anpassen oder in Relation mit anderen Wörtern setzen soll. Entsprechend ist mit einem besseren Ergebnis zu rechnen, wenn weniger *OOV*-Wörter vorkommen.

Hier verfolgen wir diesen Ansatz, aus den Eingaben eine Ausgabe mit der minimalen Anzahl an *OOV*-Wörtern zu erzeugen. Dabei werden keine anderen Parameter betrachtet, sondern ausschließlich der Einfluss der Anzahl der *OOV*-Wörter.

Nun wird aus zwei Eingaben jeweils der Satz mit weniger *OOV*-Wörtern ausgewählt und diese Sätze werden erneut zu einem neuen Text zusammengefügt.

$$\text{Satz} = \begin{cases} 1 & , \quad \text{OOV(S0)} > \text{OOV(S1)} \\ 0 & , \quad \text{sonst} \end{cases} \tag{4.2}$$

Hat Text 1 weniger *OOV*-Wörter als Text 0, so wird der Satz aus Text 1 gewählt, ansonsten der Satz aus Text 0. Dies wird paarweise mit den Übersetzungen ausgeführt.

4.4.3 Phrasen

Phrasen beeinflussen die Natürlichkeit einer Übersetzung. Besonders bei längeren Phrasen ist es eindeutiger, wie sie zu übersetzen bzw. in Kontext des Satzes zu stellen sind. Besteht ein Satz vorwiegend aus langen Phrasen, steigt seine Natürlichkeit und er wirkt flüssiger. Besteht er dagegen vorwiegend aus kurzen Phrasen, sinkt die Wahrscheinlichkeit, da das Übersetzungssystem mehr Phrasen verbinden muss und dabei mehr Fehler entstehen können.

Wir verfolgen das Ziel eine Ausgabe zu erzeugen, welche pro Satz eine möglichst hohe durchschnittliche Phrasenlänge hat. Je höher diese ist, desto geringer ist die Anzahl an verwendeten Phrasen pro Satz, wodurch es flüssiger erscheint.

Im folgenden wird für beide Eingaben die durchschnittliche Phrasenlänge pro Satz, der verwendeten Phrasen berechnet. Von den Eingaben wird diejenige ausgewählt, welche die höhere durchschnittliche Phrasenlänge aufweist.

$$\text{Satz} = \begin{cases} 1 & , \quad \varnothing \text{ Phrasenlänge(S1)} > \varnothing \text{ Phrasenlänge(S0)} \\ 0 & , \quad \text{sonst} \end{cases} \tag{4.3}$$

Weisen ein Satz aus Text 0 und der zugehörige Satz aus Text 1 dieselbe durschnittliche Phrasenlänge auf, so wird der Satz aus Text 0 gewählt. Die Auswahl findet paarweise für die Übersetzungen statt.

4.5 Vorgehen

Zuvor wurde aufgezeigt, welche Eigenschaften wir aus welchen Gründen untersuchen. Ebenso wurde das prinzipielle Vorgehen bei den verschiedenen Experimenten erläutert. In diesem Abschnitt betrachten wir das gesamte Vorgehen.

Die Übersetzungen der Eingaben unterscheiden sich in ihrer Qualität. Diese Qualitäten geben den Rahmen der ohne Bearbeitung möglichen Ergebnisse an und sind somit Vergleichswerte für die Ergebnisse der Experimente. Daher werden die Eingabetexte im ersten Schritt sowohl manuell als auch automatisch im Hinblick auf ihre Qualität evaluiert. Für die manuelle Evalution (vgl. Kapitel 5.1) wird eine Umfrage zur Bewertung ausgewählter Sätze genutzt. In Kapitel 5.1.1 wird dies genauer erläutert.

Die automatische Evaluation findet mithilfe der Evaluationsmetrik BLEU statt, welche in Kapitel 2.2.1 vorgestellt wurde. Somit wird die Ausgangsqualität der Eingabedaten untersucht. Gleichzeitig findet eine Einschätzung statt, welche der Eingaben das insgesamt beste Ergebnis erzeugt und somit auch in Anbetracht der Einfachheit besser abschneiden sollte. Das bedeutet, dass bei den Experimenten aus dieser Eingabe ein höherer Anteil an Sätzen ausgewählt werden sollte.

Im nächsten Schritt werden, wie bereits beschrieben, die Eingabetexte satzweise verglichen und aus den Sätzen, welche je nach der untersuchten Eigenschaft die besseren sind, wird eine neuer Ergebnistext generiert. Die Auswahl wird mehrfach aufgrund verschiedener Kriterien durchgeführt (vgl. Kapitel 4.4) und anschließend wird erneut das Ergebnis auf seine Qualität mittels BLEU untersucht. Anhand dieser Qualität wollen wir Aussagen über den Einfluss der betrachteten Kriterien machen. Bei einem deutlich besseren Ergebnis kann ein positiver Zusammenhang zwischen dem untersuchten Merkmal und der Qualität der Übersetzung bestehen.

Als nächster Schritt werden die Ergebnisse aus den vorherigen Experimenten miteinander verglichen. Dabei wird eingeschätzt wie stark der Einfluss der untersuchten Merkmale im Vergleich zu den anderen ist und wie stark sich die Qualität der Ergebnisse der Experimente unterscheiden.

5 Experimente und Evalutation

In diesem Kapitel werden die verschiedenen Experimente betrachtet, welche Aufschluss über die Ursachen unterschiedlicher Qualität geben sollen. Dafür wird zunächst durch eine manuelle Evaluation die Qualität der Übersetzungen der Eingaben evaluiert, sowie mit den zugehörigen Ergebnissen einer automatischen Evaluation verglichen. Anschließend werden verschiedene Experimente erläutert und evaluliert.

5.1 Manuelle Evaluation

Die für diese Arbeit genutzten Daten wurden mit der Evaluationsmetrik BLEU untersucht, um einschätzen zu können, welche der Eingaben zu einer besseren Übersetzungsqualität führt. Um zu überprüfen, wie geeignet BLEU als Bewertung für die gegeben Daten ist, führen wir eine manuelle Evaluation durch. Resultieren die Ergebnisse einer manuellen Evaluation und die BLEU-Werte der Daten in ähnlichen Vergleichsbewertungen, deutet dies darauf hin, dass BLEU für eine automatische Evaluation innerhalb dieser Arbeit geeignet ist. Wir benutzen BLEU, um Qualitätsänderungen und Qualitätsunterschiede zu bemerken.

Hierfür wurde eine Online-Umfrage erstellt, deren Ergebnisse im Folgenden ausgewertet werden. Die Umfrage ist auf die ursprünglichen maschinellen Übersetzungen ohne jegliche Veränderungen (Übersetzung 0 und 1 nach Französisch, sowie Übersetzung 0 und 1 nach Deutsch) bezogen. Dabei wurde einerseits nach der sprachlichen Qualität andererseits nach inhaltlicher Verständlichkeit gefragt. Zusätzlich wurde nach der bevorzugten Übersetzung seitens des Teilnehmers gefragt. Dies ist zwar sehr subjektiv, gleichzeitig erhält man die Auskunft, welche Übersetzung im Alltag eher genutzt wird bzw. natürlicher klingt.

5.1.1 Umfrage

Für die Erstellung der Umfrage und die Erhebung zugehöriger Ergebnisse wurde *limesurvey*[1], ein Online-Tool, genutzt. Zunächst wird der Aufbau der Umfrage beschrieben, anschließend folgt die Auswertung der Ergebnisse.

Aufbau

Der Aufbau einer Frage ist wie folgend und in Abbildung 5.1 zu sehen. Zunächst kommt ein Satz und dessen Übersetzung aus Text 0. Diese Übersetzung soll anhand der beiden Kriterien Verständlichkeit und sprachlicher Richtigkeit bewertet werden. Anschließend findet dasselbe noch einmal mit den zugehörigen Sätzen aus Text 1 statt. Zum Abschluss einer Frage sollen diese beiden Übersetzungen verglichen werden, es soll also angegeben werden, welche nach subjektiver Meinung die bessere ist. Eine solche aus drei Teilen bestehende Frage wird jeweils für 10 ausgewählte Sätze in beide Übersetzungsrichtungen durchgeführt. Zusätzlich wird nach dem Sprachniveau (Französisch und Deutsch) des Teilnehmers gefragt und von diesem eingeschätzt.

[1] www.limesurvey.org

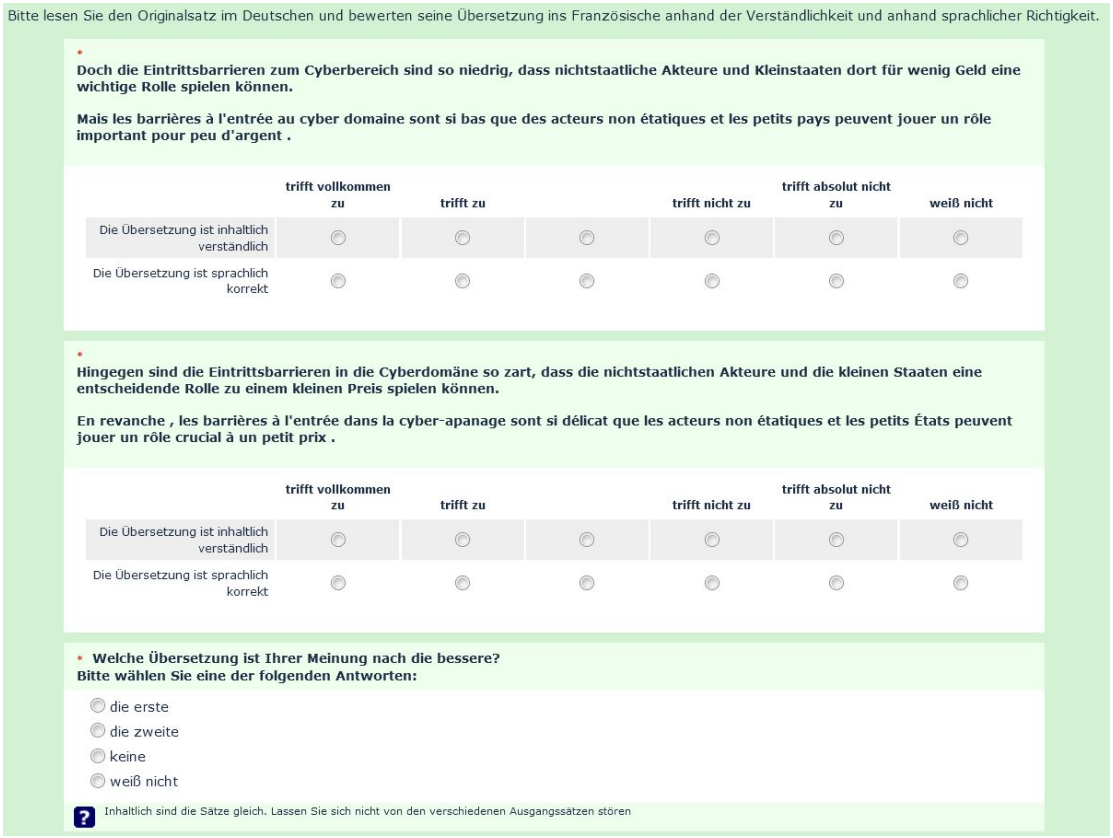

Abbildung 5.1: Umfrage

5.1.2 Auswertung

8 Personen haben die Umfrage komplett durchgeführt. Des weiteren haben Personen die Umfrage lediglich teilweise durchgeführt. Ein Teilnehmer gab das Feedback, dass die Nuancen zwischen den Sätzen für Sprachkenntnisse auf Niveau B2 (Französisch) nicht auf Korrektheit hin unterscheidbar sind. Aus diesem Grund wurde die Teilnahme abgebrochen. Ähnliches gilt vermutlich auch für weitere Teilnehmer. Aufgrund der bereits geringen Anzahl an Probanden wurden auch die Bewertungen aus angefangenen Teilnahmen in die Evaluation miteinbezogen. Es wurden 90 Übersetzungen von Französisch nach Deutsch und 128 Übersetzungen von Deutsch nach Französisch verglichen. Aufgrund von Enthaltungen können die Werte leicht schwanken.

Die Ergebnisse der Umfrage werden im Folgenden ausgewertet und verglichen. Dabei werden die Übersetzungen von Französisch nach Deutsch und die Übersetzungen von Deutsch nach Französisch einzeln betrachtet. Wir erhoffen uns davon, dass die BLEU-Werte der Daten bestärkt werden und somit die später betrachteten Kriterien einen Einfluss haben.

Französisch-Deutsch

Um einen Eindruck davon zu gewinnen, ob eine der Übersetzungen gegenüber der anderen bevorzugt wird, werden zu Beginn die Ergebnisse der Vergleiche betrachtet (vgl. Tabelle 5.2). Anhand der Antworten ist zu erkennen, dass die Sätze aus Text 1 besser erscheinen als die Sätze aus Text 0. Etwa die Hälfte der Probanden ist dieser Meinung. Dagegen ist ungefähr ein Viertel der Meinung, dass die ausgewählten Sätze beider Texte gleich gut sind. Diese Anzahl an Stimmen für Text 0 beträgt ebenfalls in etwa ein

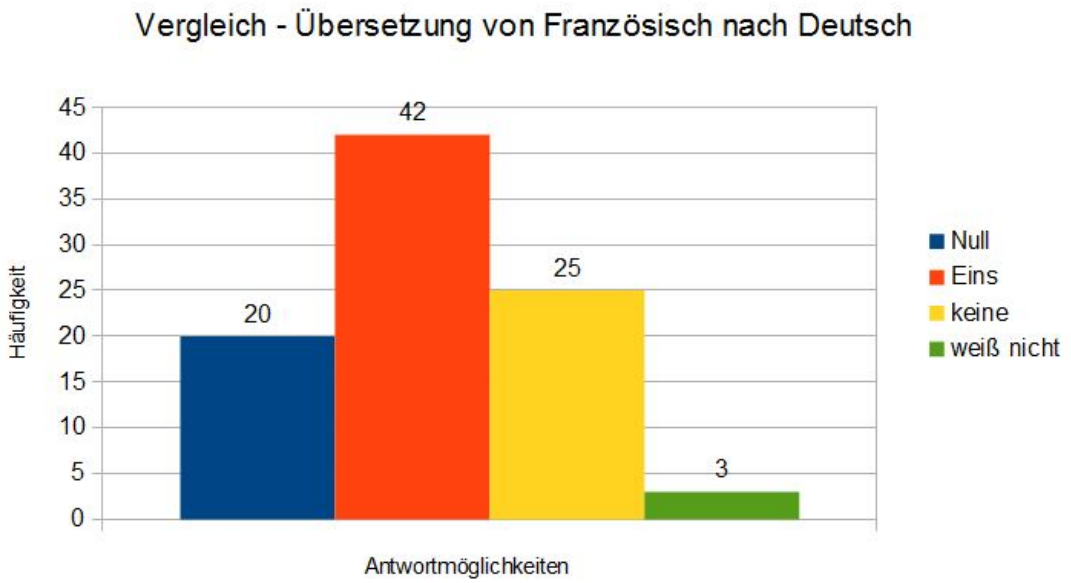

Abbildung 5.2: Vergleich: Ergebnisse der manuellen Evaluation im Bezug darauf, welche der Übersetzungen besser ist

Viertel, ist jedoch geringer. Daraus lässt sich folgern, dass Text 1 insgesamt eine deutlich bessere Qualität hat. Dabei gilt, dass einige Sätze auf ähnlichem Niveau einzuschätzen sind, wie aus Text 0. Über die eigentliche Qualität der beiden Texte lassen sich dadurch keine Schlüsse ziehen. Sie kann sowohl sehr schlecht wie auch sehr gut sein.

Die Qualität lässt sich besonders anhand von zwei Kriterien einschätzen. Einmal die sprachliche Qualität und einmal die inhaltliche Verständlichkeit. Entsprechend waren diese beiden Kriterien in der Umfrage zu bewerten.

In der Praxis kann eine Übersetzung, welche zu zu geringer inhaltlicher Verständlichkeit führt nicht verwendet werden. Daher betrachten wir die zugehörigen Ergebnisse. In Tabelle 5.3 sind die Umfrageergebnisse zu Text 0 und in Tabelle 5.4 jene zu Text 1 dargestellt. Für Übersetzung 0 ergibt sich, dass sie eher schlecht verständlich ist. Die meisten Meinungen verteilen sich auf dem negativen Bewertungsraum. Aufgrund der groben Einstufungsmöglichkeiten lässt sich der genaue Grad nicht bestimmen.

Text 1 hingegen scheint deutlich verständlicher zu sein. Auch hier existieren negative Meinung, welche von den positiven überwogen werden. Somit wird Text 1 als inhaltlich verständlicher gegeüber Text 0 bewertet. Dieses Ergebnis stimmt mit dem vorherigen aus den Vergleichen der Sätze beider Texte überein und stellt einen möglichen Grund dafür dar.

Neben der Verständlichkeit des Inhalts sollte auch die sprachliche Qualität nicht zu schlecht sein. Ähnlich wie die inhaltliche Verständlichkeit wird die sprachliche Qualität von Text 0 negativ eingeschätzt. Hierbei ist das Ergebnis eindeutiger und die Meinungen sind weniger verteil, was dafür spricht, dass diese tatsächlich sehr gering ist.

Für Text 1 wird die sprachliche Qualität sehr ähnlich, geringfügig besser, eingeschätzt. Da dieser Unterschied sehr klein ist, gehen wir davon aus, dass er ignoriert werden kann und bei einer größeren Probandenmenge verschwunden wäre.

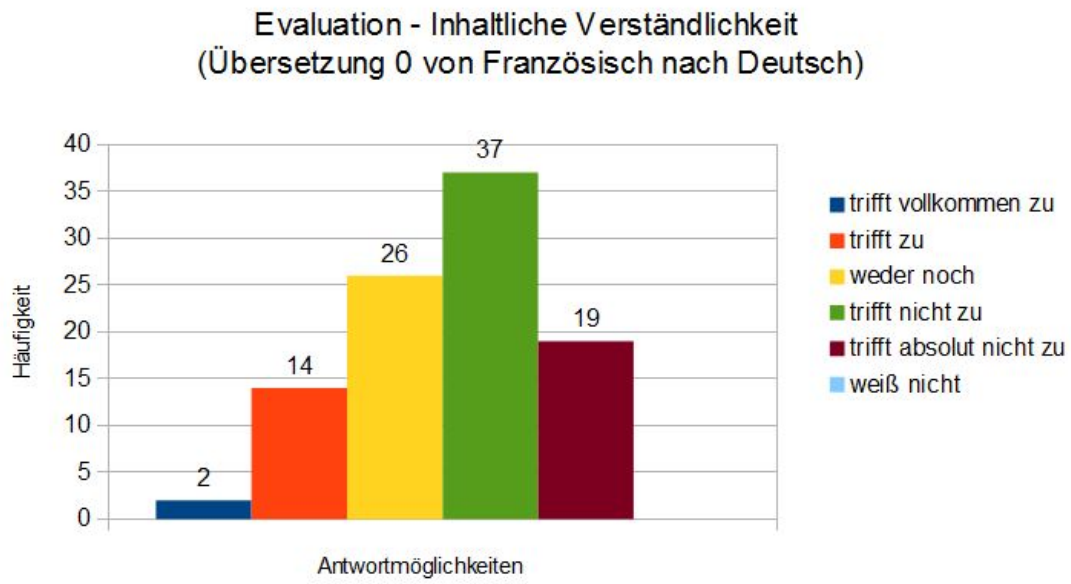

Abbildung 5.3: Umfrage: Inhaltliche Verständlichkeit von Übersetzung 0 von Französisch nach Deutsch

Aus den Ergebnissen der Umfrage lässt sich für die Übersetzungen von Französisch nach Deutsch Folgendes schließen:
Text 0 und Text 1 weisen unterschiedliche Qualitäten auf, wobei Text 1 als besser angesehen wird. Grund dafür ist vor allem die inhaltliche Verständlichkeit, während die sprachliche Qualität für beide Texte eher mangelhaft ist.

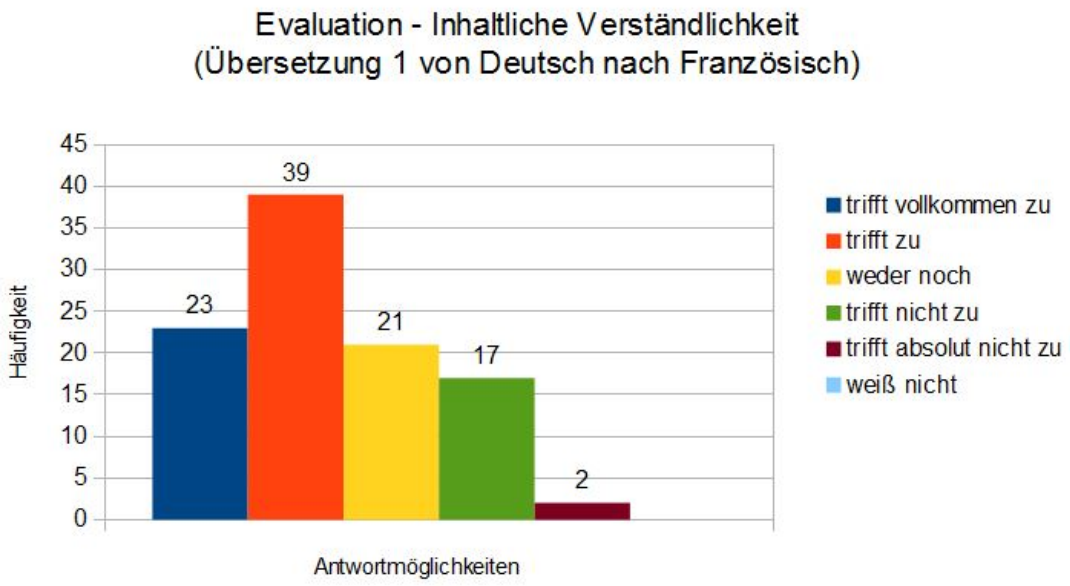

Abbildung 5.4: Umfrage: Inhaltliche Verständlichkeit von Übersetzung 1 von Französisch nach Deutsch

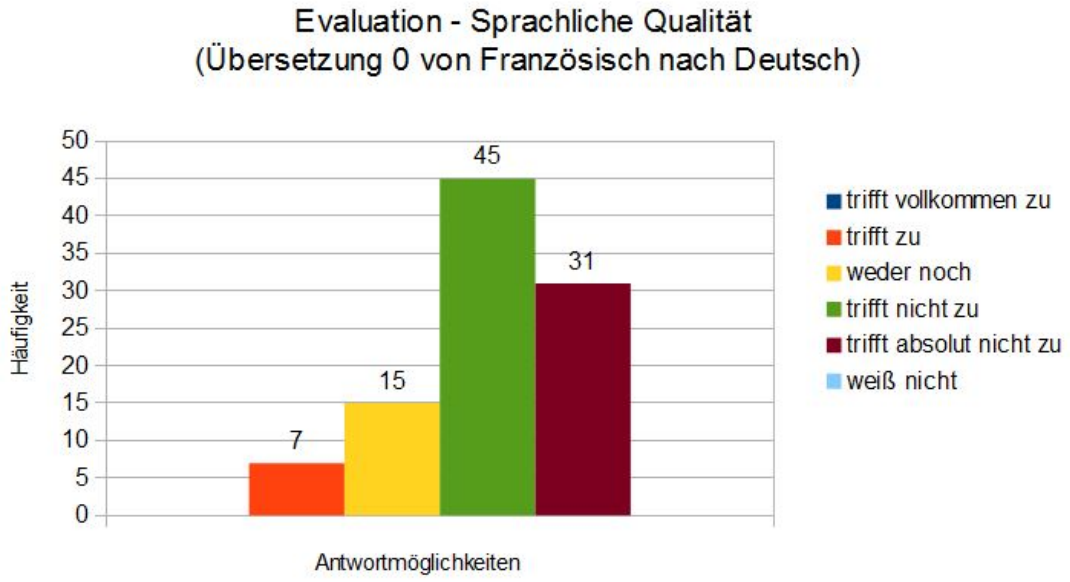

Abbildung 5.5: Umfrage: Sprachliche Qualität von Übersetzung 0 von Französisch nach Deutsch

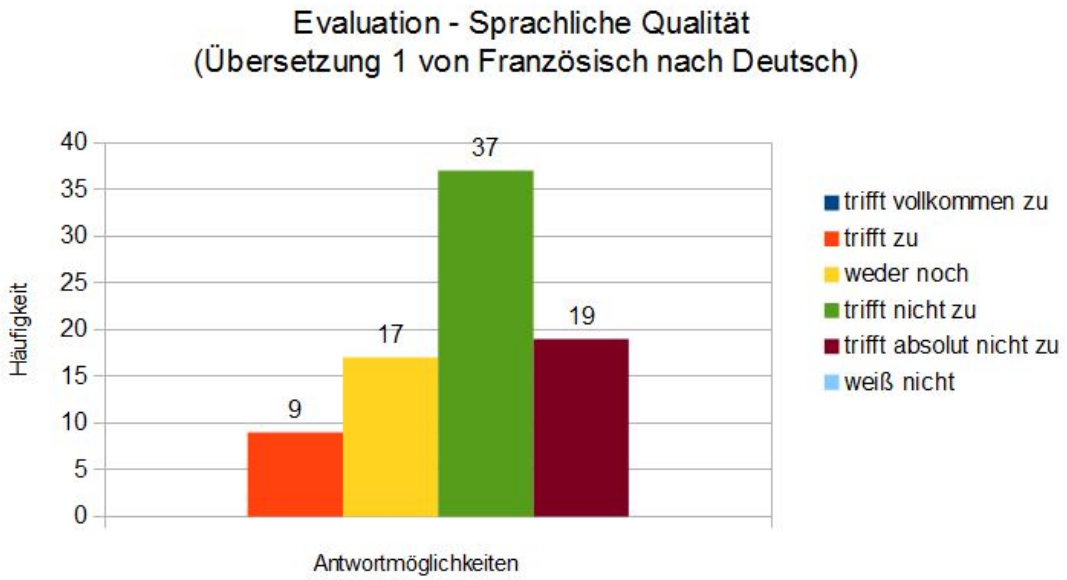

Abbildung 5.6: Umfrage: Sprachliche Qualität von Übersetzung 1 von Französisch nach Deutsch

Deutsch-Französisch

In dem vorherigen Unterkapitel wurden die Übersetzungen von Französisch nach Deutsch untersucht. Nun betrachten wir die Übersetzungen der Gegenrichtung, also Deutsch nach Französisch. Für diese gilt es ebenfalls herauszufinden, wie sie zueinander und wie sie qualitativ einzuordnen sind.

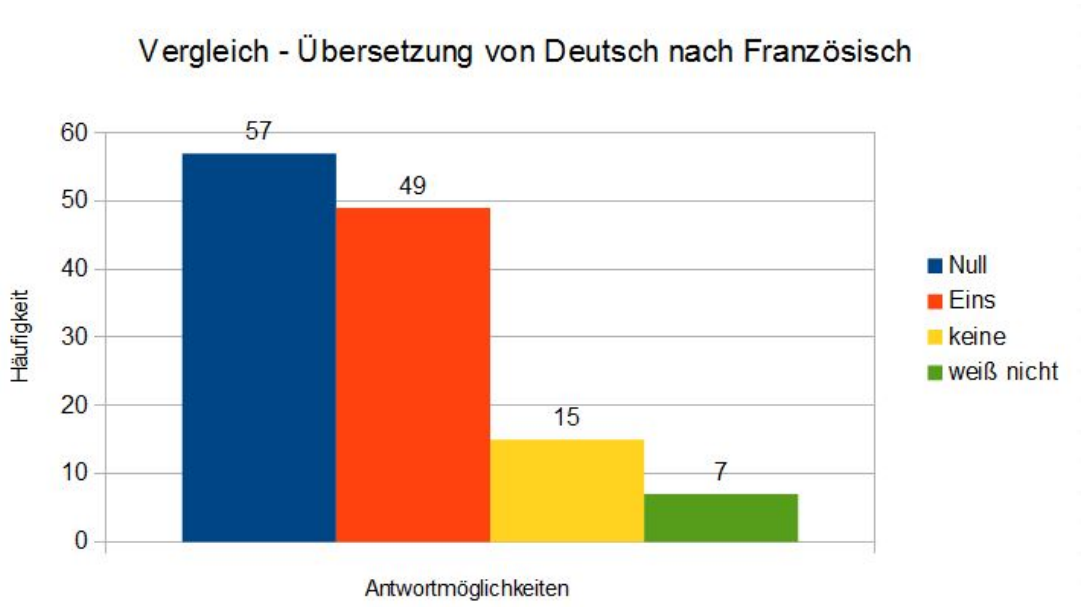

Abbildung 5.7: Vergleich: Ergebnisse der manuellen Evaluation im Bezug darauf, welche der Übersetzungen besser ist

Zunächst werden die beiden Übersetzungen verglichen. Dabei existieren wenige Meinungen, welche sich nicht für einen der Texte entscheiden. Text 0 wird hierbei Text 1 vorgezogen. Der Unterschied zwischen der Anzahl an Meinungen entspricht 6,25% der Gesamtzahl und ca. 7,55% nur auf die Stimmen für Text 0 und Text 1 bezogen. Somit lassen sich die beiden Übersetzungen bei in etwas gleicher Gesamtqualität einordnen. Dabei ist zu beachten, dass soweit bekannt lediglich einer der Probanden als Muttersprache Französisch hatte. Es besteht die Möglichkeit, dass das Ergebnis anders ausgefallen wäre, wenn mehr französische Muttersprachler beteiligt worden wären.

Da sich aufgrund der bisherigen Ergebnisse keiner der Texte als besser hat einstufen lassen, suchen wir nach Unterschieden zwischen deren inhaltlicher Verständlichkeit und sprachlichen Qualität. Diese können sich trotzdem unterscheiden. Des weiteren geben sie Aufschluss über die qualitative Einordnung der beiden Übersetzungen.

Ohne ausreichende inhaltliche oder sprachliche Qualität ist eine Übersetzung einer anderen nicht vorzuziehen. Wir betrachten zunächst die Bewertungen der inhaltlichen Qualität. Sowohl für Text 0 als auch für Text 1 ergibt die Umfrage, dass beide eher inhaltlich verständlich sind. Dabei existieren auch abweichende Meinungen, welche eine deutlich geringere Anzahl erreichen. Beide Texte werden in etwa gleich anhand ihrer inhaltlichen Qualität eingeordnet. Dies könnte ein möglicher Grund dafür sein, dass es unter einem direkten Vergleich zwischen Text 0 und Text 1 zu keinem eindeutigen Ergebnis kam.

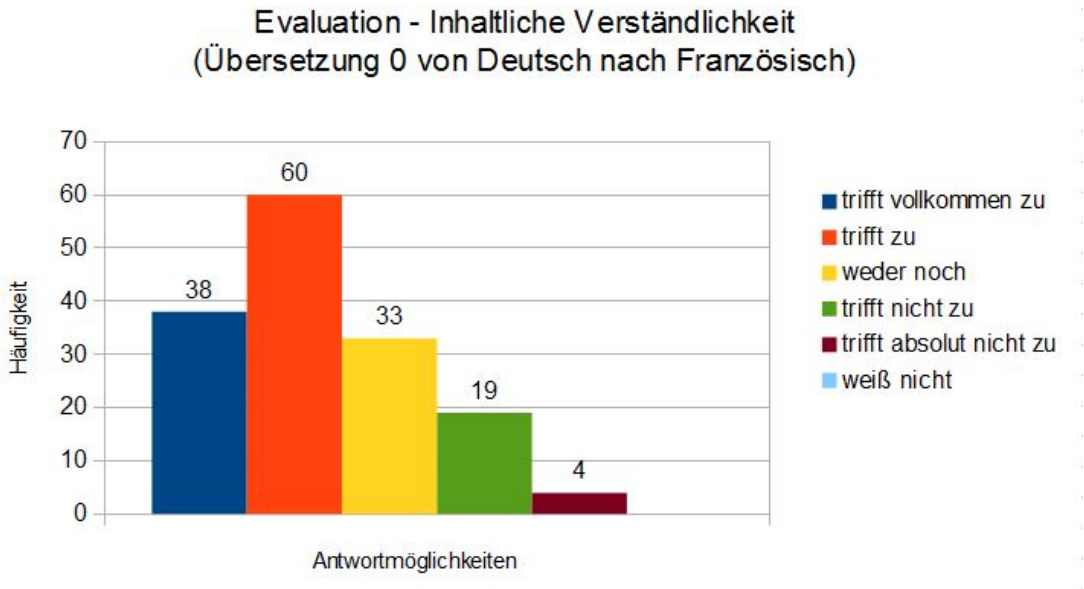

Abbildung 5.8: Umfrage: Inhaltliche Verständlichkeit von Übersetzung 0 von Deutsch nach Französisch

Unter Beachtung der bisherigen Ergebnisse der Umfrage würden wir erwarten, dass sich die Texte auch in Hinsicht auf sprachliche Qualität nicht unterscheiden. Das Ergebnis sieht wir folgend aus. Beide Texte lassen sich nach sprachlicher Qualität weder sonderlich herausragend noch sonderlich schlecht einordnen. Text 0 weist dabei eine leichte Tendenz in die positivere Richtung auf, welche bei größerer Probandenmenge verschwinden könnte.

Somit lässt sich schließen, dass Text 0 und Text 1 in etwa ein ähnliches Niveau besitzen. Weiterhin weisen beide eine eher mittelmäßige sprachliche Qualität und eher positive inhaltliche Verständlichkeit auf. Insgesamt ist somit keiner dem anderen deutlich vorzuziehen.

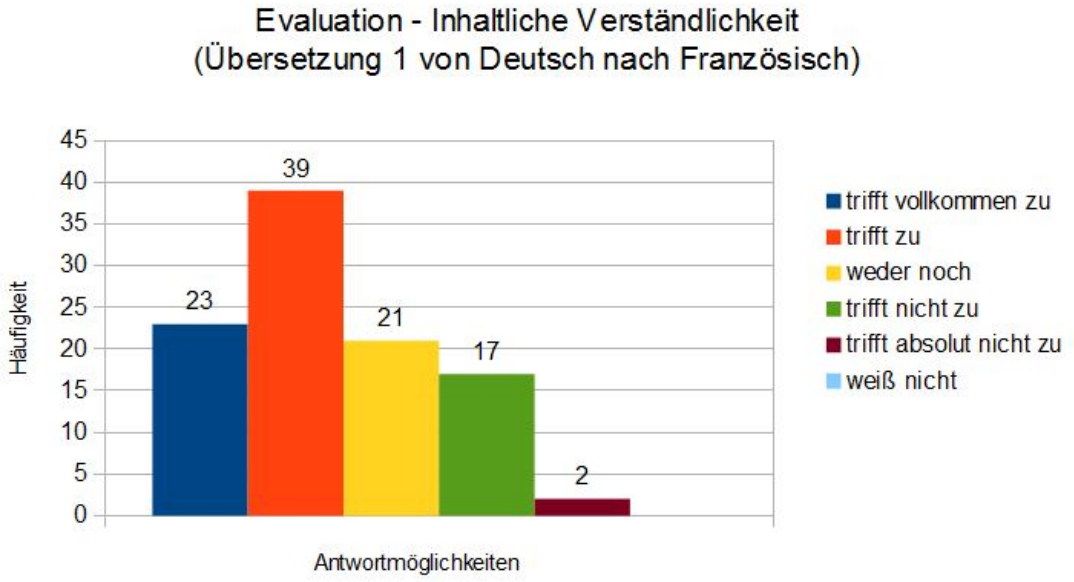

Abbildung 5.9: Umfrage: Inhaltliche Verständlichkeit von Übersetzung 1 von Deutsch nach Französisch

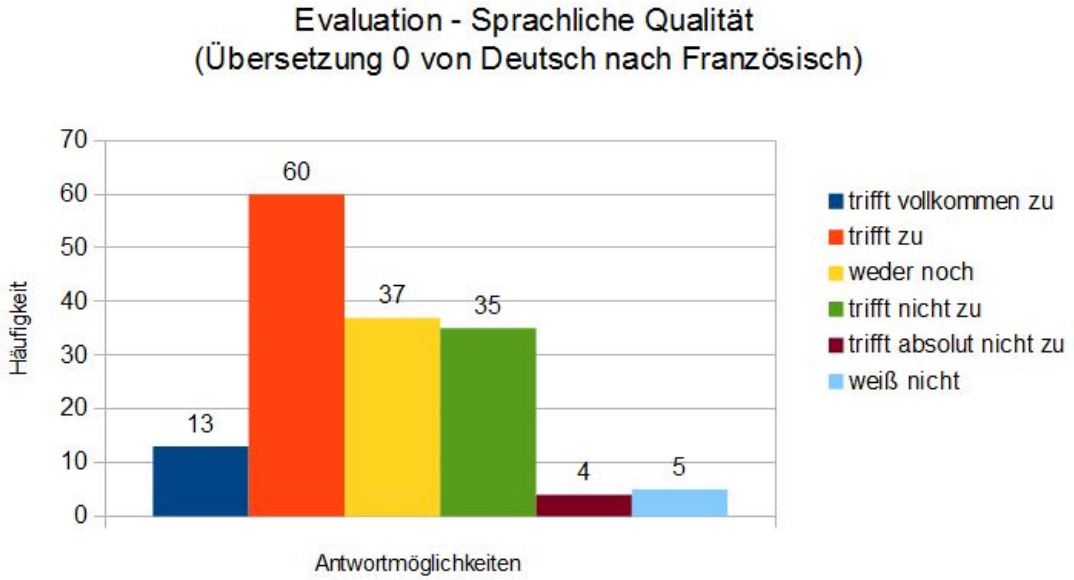

Abbildung 5.10: Umfrage: Sprachliche Qualität von Übersetzung 0 von Deutsch nach Französisch

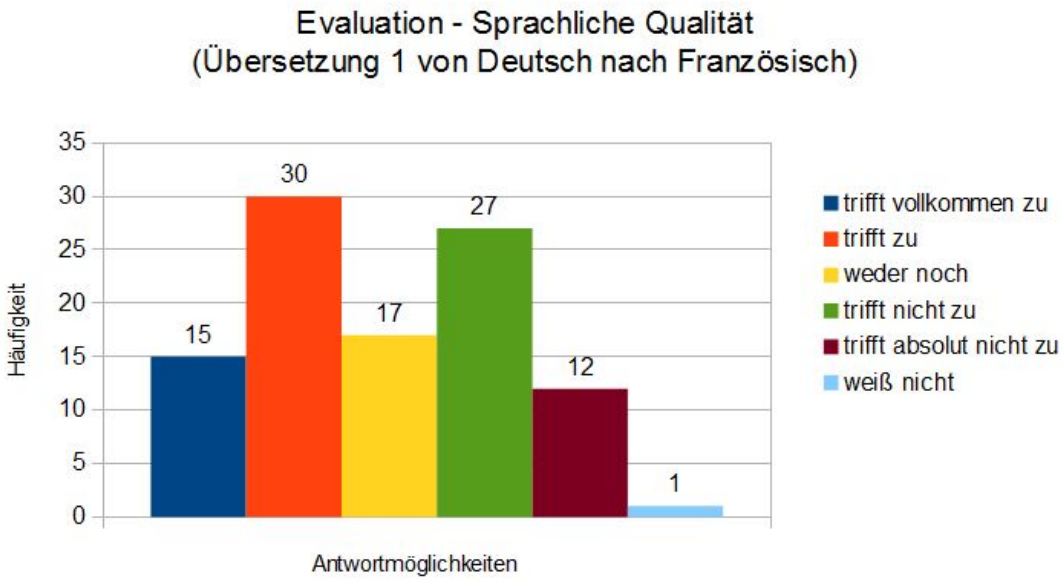

Abbildung 5.11: Umfrage: Sprachliche Qualität von Übersetzung 1 von Deutsch nach Französisch

5.2 Automatische Evaluation

Dieser Teil der Arbeit befasst sich mit der Auswertung der Experimente und den Originaldaten sowie Vergleichen zwischen deren Ergebnissen. Für die Auswertung wird die Evaluationsmetrik BLEU (vgl. Kapitel 2.2.1) verwendet.

5.2.1 Eingabedaten und unbearbeitete Übersetzungen

Zunächst werden die Eingabedaten selbst auf ihre Qualitäten untersucht, sowie auf ihre Werte der für die Experimente genutzten Merkmale.

Daten	BLEU	∅Phrasenlänge(best)	min∅(best)	max∅(best)
Ziel Deutsch 0	17.55%	1.8180	1.0	5.0
Ziel Deutsch 1	24.40%	1.8453	1.0	5.0
Ziel Franz. 0	23.14%	2.1037	1.0	7.0
Ziel Franz. 1	36.80%	2.0500	1.0	5.0

Tabelle 5.1: Für die automatischen Übersetzungen der französischen und deutschen Texte: BLEU-Wert, durchschnittlich Phrasenlänge pro Satz (unter Betrachtung der besten Hypothese), sowie minimale und maximale durchschnittliche Phrasenlänge pro Satz (unter Betrachtung der besten Hypothese)

Sowohl bei der Übersetzung von Französisch nach Deutsch als auch bei der Übersetzung von Deutsch nach Französisch weist Text 1 einen höheren BLEU-Wert auf als Text 0 und unterstützt somit teilweise das Ergebnis der Umfrage in Kapitel 5.1.1. Die genauen BLEU-Werte sind in Tabelle 5.1 zu sehen. Bei der Übersetzung von Französisch nach Deutsch stimmen das Ergebnis der manuellen Umfrage und der BLEU-Wert insofern überein, dass Text 1 eine bessere Gesamtqualität als Text 0 hat, beide jedoch keine gute Gesamtqualität aufweisen. Für die Übersetzungen von Deutsch nach Französisch stimmen die Ergebnisse nicht überein. Aus der manuellen Evaluation lassen sich keine großen Qualitätsunterschiede schließen, während der BLEU-Wert Text 1 als besser bewertet.

Der BLEU-Wert ist somit für eine Bewertung einigermaßen geeignet, da er lediglich zum Vergleich der Experimentalergebnisse verwendet wird und nicht für das Bewerten einer Gesamtqualität. Wir gehen also davon aus, dass eine einfachere Übersetzung eine bessere Qualität aufweist als eine schwierigere. Entsprechend verbessert oder verschlechtert sich der BLEU-Wert.

Betrachtet man des weiteren das Merkmal "durchschnittliche Phrasenlänge", so kann man keine klare Aussage über deren Einfluss auf die Einfachheit einer Eingabe machen. Im Fall einer Übersetzung von Französisch nach Deutsch weist Text 1 eine höhere durchschnittliche Phrasenlänge auf, im Fall der Übersetzung von Deutsch nach Französisch hingegen gilt, dass Text 0 eine höhere durchschnittliche Phrasenlänge besitzt. Ersteres stützt die These, dass die Nutzung von mehr längeren Phrasen zu einer besseren Übersetzung führt, während letzteres dieser widerspricht. Wodurch gemeinsam keine eindeutige Aussage darüber gemacht werden kann.

Dabei ist der Unterschied der durchschnittlichen Phrasenlänge in beiden Fällen äußerst gering und die Eingaben somit in Hinsicht auf Phrasenlänge recht ähnlich. Beide nutzen als kürzeste Phrasen Phrasen der Länge null, unterscheiden sich jedoch in der längsten genutzten Phrase. Für die Übersetzungen von Französisch nach Deutsch gilt, dass sie denselben Wert besitzen (vgl. Tabelle 5.1). In Gegenrichtung weist Text 0 einen leicht höheren Wert auf. Dies lässt uns ebenfalls keinen Zusammenhang zwischen einer einfacheren Eingabe und einer besseren Übersetzungsqualität erkennen.

In Tabelle 5.2 werden die Merkmal Perplexität und *OOV*-Wörter in Abhängigkeit von dem *Language Model big lm* und in Tabelle 5.3 in Abhängigkeit von dem *Language Model indomain lm* aufgezeigt.

Daten	Perplexität	OOV
Quelle Franz. 0	109.287	14
Quelle Franz. 1	101.636	15
Ziel Deutsch 0	170.652	73
Ziel Deutsch 1	158.786	67
Quelle Deutsch 0	344.27	10
Quelle Deutsch 1	349.975	34
Ziel Franz. 0	96.2486	60
Ziel Franz. 1	95.494	119

Tabelle 5.2: Für die übersetzten französischen und deutschen Texte sowie ihre Originaltexte: Perplexität und Anzahl an *OOV*-Wörtern in Abhängigkeit des *Language Model big lm*

Zunächst betrachten wir die Perplexität für die französischen Quelltexte nach *big lm*. Hier weist Text 0 einen höheren Wert als Text 1 auf und wird somit schlechter von *big lm* modelliert. Im Gegensatz dazu besitzt Text 1 ein *OOV*-Wort mehr, was nicht sehr ins Gewicht fallen sollte. Passend dazu besitzt die Übersetzung von Text 1 einen besseren BLEU-Wert.

Untersucht man die Übersetzungen ins Deutsche auf ihre Perplexität, so zeigt sich, dass auch hier Text 0 einen höheren Wert als Text 1 aufweist. Dasselbe gilt nun auch für die Anzahl an *OOV*-Wörtern. Diese Werte sprechen für einen Zusammenhang zwischen einer niedrigeren Perplexität und einer höheren Einfachheit einer Eingabe.

Für die deutschen Quelltexte gilt, dass Text 1 eine höhere Perplexität als Text 0 besitzt. Ebenso besitzt Text 1 eine höhere Anzahl an *OOV*-Wörtern. Im Gegensatz dazu weist die Übersetzung von Text 1 einen besseren BLEU-Wert auf.

Anders als zuvor gilt für die übersetzten Texte von Deutsch nach Französisch, dass Text 0 eine höhere Perplexität als Text 1 besitzt. Text 1 hat weiterhin eine größere Zahl an *OOV*-Wörtern. Daraus könnte man schließen, dass die Perplexität einen größeren Einfluss als die Zahl an *OOV*-Wörtern besitzt.

Daten	Perplexität	OOV
Quelle Franz. 0	103.744	249
Quelle Franz. 1	96.6908	217
Ziel Deutsch 0	153.631	346
Ziel Deutsch 1	145.785	292
Quelle Deutsch 0	217.22	191
Quelle Deutsch 1	229.124	391
Ziel Franz. 0	81.2772	186
Ziel Franz. 1	80.2922	338

Tabelle 5.3: Für die übersetzten französischen und deutschen Texte sowie ihre Originaltexte: Perplexität und Anzahl an *OOV*-Wörtern in Abhängigkeit des *Language Model indomain lm*

Unter Nutzung von dem *Language Model indomain lm* sind die zugehörigen Zahlen in Tabelle 5.3 abgebildet. Hierbei ergeben Vergleiche der Texte 0 und 1 fast dieselben Ergebnisse wie unter Verwendung von *Language Model big lm*. Die einzige Ausnahme bildet der französische Quelltext. Anders als zuvor besitzt nun Text 1 die höhere Zahl an *OOV*-Wörtern.

Bei einem direkten Vergleich der Texte 0 und 1 werden somit von beiden *Lanugage Model* dieselben Texte besser modelliert. Bei einer satzweisen Auswahl anhand der Perplexität würden wir daher ähnliche Ergebnisse erwarten, wenn dies mit beiden *Language Model* einzeln ausgeführt wird.

Insgesamt werden die nach *big lm* gezogenen Schlüsse durch *indomain lm* weiter bestätigt. Ein möglicher Zusammenhang zwischen Perplexität und Einfachheit existiert, ebenso wie ein evenuteller Einfluss durch die Anzahl an *OOV*-Wörtern.

5.2.2 Perplexität

Anhand der beiden *Language Model big lm* und *indomain lm* wurde für jeden übersetzen Satz der Eingaben die Perplexität berechnet. Anhand dieser wurden neue Texte aus den Sätzen mit besserer, also geringerer, Perplexität erstellt. Die Eigenschaften dieser Texte werden im Folgenden betrachtet.

Tabelle 5.4 zeigt die Eigenschaften Perplexität, Anzahl an *OOV*-Wörtern und BLEU-Wert in Abhängigkeit der *Language Model big lm* und *indomain lm* an. Für *Ziel Franz. (best)* wurde aus den Übersetzungen 0 und 1 ins Französische jeweils der Satz mit geringerer Perplexität ausgewählt und zu einem Text zusammengefügt. Dabei wurde einmal die Perplexität genutzt, welche durch *big lm* berechnet wurde und einmal die durch *indomain lm* berechnete. Um Auswirkungen zu erkennen, werden die Ergebnisse jeweils mit dem *Language Model* betrachtet, mit dem sie erzeugt wurden.

Daten	LM	Perplexität	OOV	BLEU
Ziel Deutsch (best)	big lm	138.531	68	22.19%
Ziel Deutsch (best)	indomain lm	124.66	330	21.71%

Tabelle 5.4: Für *Ziel Deutsch*(aus den Sätzen mit besserer Perplexität aus den Texten 0 und 1 erzeugt): Perplexität und Anzahl an *OOV*-Wörtern in Abhängigkeit zweier *Language Model* (*big lm* und *indomain lm*), sowie BLEU-Werte

Der Erwartung entsprechend weisen die Daten eine deutlich bessere Perplexität auf als beide Eingaben. Somit wurde der durch die Eingaben vorgegeben Rahmen übertroffen und ein sichtbarer Einfluss der Perplexität sollte erkennbar sein. Dies gilt für beide *Lanugage Model indomain lm* und *big lm*. Daraus lässt sich schließen, dass sich Eingaben trotz selbem Inhalt sehr stark unterscheiden können. Sonst wäre es nicht möglich gewesen mittels einer Selektion nach der Perplexität diese so stark zu verbessern.

Für die Anzahl an *OOV*-Wörtern gilt dies nicht. Nur für das *Language Model big lm* wurde der vorherige Rahmen unterschritten. Nach *indomain lm* ist die Zahl an *OOV*-Wörtern geringer als die gegebene Obergrenze, erreicht die Untergrenze jedoch nicht. Daher gehen wir davon aus, dass der Einfluss von *OOV*-Wörtern für diesen Fall nicht maßgebend ist.

Betrachtet man die BLEU-Werte der neuen Daten auf der Zielseite, so haben diese sich verbessert. Weiterhin liegen sie dabei zwischen den Werten von Text 0 und Text 1, sind jedoch deutlich besser als der BLEU-Wert von Text 0. Daraus lässt sich ein Zusammenhang zwischen geringerer Perplexität und höherer Qualität für eine Übersetzung von Französisch nach Deutsch schließen.

Dasselbe Szenario betrachten wir nun für die Übersetzungen von Deutsch nach Französisch. Die Daten wurden einmal nach geringerer Perplexität unter *big domain* ausgewählt und erneut unter diesem betrachtet, sowie einmal unter *indomain lm*. Die zugehörigen Zahlen sind in Tabelle 5.5 dargestellt.

Dabei erhalten wir ähnlich Ergebnisse wie bei den Daten aus der Übersetzung von Französisch nach Deutsch. Auch hier unterschreitet die Perplexität für beide *Language Model* die vorgegebenen unteren

Daten	LM	Perplexität	OOV	BLEU
Ziel Franz.(best)	big lm	81.1009	90	31.59%
Ziel Franz.(best)	indomain lm	67.8204	245	31.18%

Tabelle 5.5: Für *Ziel Franz.(best)* (aus den Sätzen mit besserer Perplexität aus den Texten 0 und 1 erzeugt): Perplexität und Anzahl an *OOV*-Wörtern in Abhängigkeit zweier *Language Model* (*big lm* und *indomain lm*), sowie BLEU-Werte

Grenzen. Gleichzeitig verringert sich die Anzahl an *OOV*-Wörtern nicht soweit, dass sie die vorherige Grenze unterschreitet. Damit lässt sich erneut ein Zusammenhang zwischen der Perplexität und der Qualität folgern. Dabei ist zu beachten, dass auch anderen Kriterien, welche hier nicht beachtet wurden, dies mitbeeinflusst haben können.

5.2.3 OOV

Im vorherigen Abschnitt wurde ein möglicher Einfluss der Anzahl an *OOV*-Wörter bereits berücksichtigt, da diese auch die Perplexität mitbeeinflusst. Im Folgenden betrachten wir vor allem den Einfluss der *OOV*-Wörter, da eine geringere Anzahl einen Text einfacher und vor allem verständlicher machen sollte.

Mittels der *Language Model big lm* und *indomain lm* wurde jeweils einmal die Anzahl an *OOV*-Wörter pro Satz bestimmt und aus den Sätzen mit der geringeren Anzahl an *OOV*-wörtern wurden zwei Texte *Ziel Deutsch (best)* und zwei Texte *Ziel Französisch (best)* erzeugt. Tabelle 5.6 zeigt Perplexität, Anzahl an *OOV*-Wörtern sowie den BLEU-Wert zu *Ziel Deutsch (best)* an.

Daten	LM	Perplexität	OOV	BLEU
Ziel Deutsch(best)	big lm	170.671	35	17.84%
Ziel Deutsch(best)	indomain lm	154.996	208	18.49%

Tabelle 5.6: Für *Ziel Deutsch (best)* (aus den Sätzen mit weniger OOV aus den Texten 0 und 1 erzeugt): Perplexität und Anzahl an *OOV*-Wörtern in Abhängigkeit der *Language Model big lm* und *indomain lm*, sowie BLEU-Werte

Wie erwartet ist die Anzahl an *OOV*-Wörter gesunken und ist sowohl nach *big lm* als auch nach *indomain lm* deutlich geringer als in *Ziel Deutsch 0* und *Ziel Deutsch 1*. Die neuen Werte bilden nach den genutzten *Langugage Model* eine untere Grenze, welche ohne weiteres Bearbeiten der Daten nicht unterschritten werden kann.

Im Gegenteil zur Anzahl an *OOV*-Wörter ist die Perplexität gestiegen und liegt in beiden Fällen etwas über dem zuvor höheren Wert. Somit sind aus diesem Experiment mögliche zu erkennende Zusammenhänge nicht von einer sinkenden Perplexität verursacht.

Vergleicht man die BLEU-Werte mit denen der Übersetzungen ins Deutsch, so stellt man fest, dass sie erneut zwischen ihnen liegen. Die neuen Werte sind nur leicht besser als der schlechtere Wert, jener von Text 0. Es liegt nahe, dass eine geringere Anzahl an *OOV*-Wörtern für eine bessere Qualität sorgt und sie somit die Einfachheit der Daten beeinflussen. Wie stark dieser Einfluss ist, ist nicht klar erkenntlich.

Tabelle 5.7 zeigt die Ergebnisse für eine satzweise Selektion von Text 0 und Text 1 (Übersetzungen von Deutsch nach Französisch). Die Selektion basiert auf der geringsten Anzahl an *OOV*-Wörtern. Dies erfolgt einmal nach dem *Language Model big lm* und einmal nach *indomain lm*.

Entsprechend dem Auswahlkriterium sinkt die Anzahl an *OOV*-Wörtern. Hierbei unterschreitet sie in

Daten	LM	Perplexität	OOV	BLEU
Ziel Franz.(best)	big lm	96.1128	24	35.93%
Ziel Franz.(best)	indomain lm	79.5162	129	35.27%

Tabelle 5.7: Für *Ziel Deutsch (best)* (aus den Sätzen mit weniger OOV aus den Texten 0 und 1 erzeugt): Perplexität und Anzahl an *OOV*-Wörtern in Abhängigkeit der *Language Model big lm* und *indomain lm*, sowie BLEU-Werte

beiden Fällen den geringeren Wert der vorherigen Texte. Ohne weitere Bearbeitung der Daten können diese Werte nicht unterschritten werden.

Wir betrachten die neuen Werte der Perplexität, da die Zahl an *OOV*-Wörtern diese mitbeeinflusst. Anders als bei derselben Untersuchung auf *Ziel Deutsch (best)* ist die Perplexität hierbei in beiden Fällen im Vergleich zum vorherigen Maximum gesunken. Für *big lm* unterschreitet der Wert der Perplexität das vorherige Minimum nicht. Für *indomain lm* dagegen wird das Minimum unterschritten.

Ebenfalls verbessert haben sich die BLEU-Werte. Sowohl für *big lm* als auch *indomain lm* übertreffen sie die Werte von Text 0 und Text 1 deutlich. Ein möglicher Grund ist, dass beide *Language Model* das Selektionsergebnis besser modellieren und wir zuvor einen möglichen Zusammenhang zwischen der Perplexität und dem BLEU-Wert festgestellt haben. Ob die gesamte Verbesserung der BLEU-Werte lediglich durch die Selektion nach Anzahl an *OOV*-Wörtern verursacht wird, ist nicht schließbar. Wir betrachten die Verbesserung an sich als Indiz für eine gestiegene Einfachheit, bewerten ihren Grad dabei jedoch nicht.

5.2.4 Phrasen

Wir untersuchen die durchschnittliche Phrasenlänge, da diese Aufschluss darüber gibt, ob eher kurze oder eher lange Phrasen für eine Übersetzung genutzt wurden. Wobei längere Phrasen auf eine einfachere Eingabe hinweisen. Erwartet wird, dass mit Länge der Phrasen die Ergebnisqualität steigt.

Aus den Übersetzungen ins Deutsche wird jeweils der Satz mit einer höheren durchschnittlichen Phrasenlänge gewählt und zu einem neuen Text, *Ziel Deutsch (best)* zusammengefügt. Die zugehörigen Werte der durchschnittlichen Phrasenlänge pro Satz sowie der BLEU-Wert werden in Tabelle 5.8 dargestellt.

Daten	∅Phrasenlänge(best)	BLEU
Ziel Deutsch (best)	1.9488	21.20%

Tabelle 5.8: Für *Ziel Deutsch (best)* (aus den Sätzen mit größter durchschnittlicher Phrasenlänge der besten Hypothese aus Text 0 und Text 1 erzeugt): durchschnittliche Phrasenlänge und BLEU-Wert

Wie nach Vorgabe der Selektion erwartet ist die durchschnittliche Phrasenlänge gestiegen. Im Vergleich zu Text 0 und Text 1 ist sie nur geringfügig größer. Ein möglicher Grund dafür wäre, dass sich viele der genutzten Phrasen von Text 0 und 1 ähneln, wodurch eine große Verbesserung nicht möglich wäre.

Für den BLEU-Wert von *Ziel Deutsch (best)* gilt, dass er sich im Vergleich zu Text 0 verbessert hat. Weiterhin liegt er unterhalb des BLEU-Wertes von Text 1. Daraus folgern wir, dass eine längere durchschnittliche Phrasenlänge für einen besseren BLEU-Wert und damit auch einen einfacheren Text sorgen kann.

Denselben Fall wie zuvor betrachten wir nun für die Übersetzungen von Deutsch nach Französisch.

Dabei werden die Sätze mit der größten durchschnittlichen Phrasenlänge ausgewählt und auf diese untersucht. Das zugehörige Ergebnis sowie der BLEU-Wert sind in Tabelle 5.9 abgebildet.

Daten	∅Phrasenlänge(best)	BLEU
Ziel Franz.(best)	2.2097	29.88%

Tabelle 5.9: Für *Ziel Französisch (best)* (aus den Sätzen mit größter durchschnittlicher Phrasenlänge der besten Hypothese aus Text 0 und Text 1 erzeugt): durschnittliche Phrasenlänge und BLEU-Wert

Auch hier ergibt sich eine etwas größere durchschnittliche Phrasenlänge als für die Texte 0 und 1. Zusätzlich verbessert sich der BLEU-Wert gegenüber dem vorherigen Minimum. Damit erzielen wir eine ähnliches Ergebnis wie bei der Übersetzung von Französisch nach Deutsch. Gemeinsam lässt sich in beider Übersetzungsrichtungen ein möglicher Zusammenhang zwischen größerer Phrasenlänge und Qualität schließen.

Da die bisherigen Untersuchungen darauf hinweisen, dass die Phrasenlänge einen Einfluss auf die Einfachheit der Daten hat, betrachten wir denselben Fall wie zuvor erneut mit einem Unterschied. Er unterscheidet sich dadurch, dass nicht nur die Texte 0 und 1 für die Selektion in Frage kommen, sondern alle n-besten Hypothesen, welche während der Übersetzung erzeugt wurden. Aus diesen werden die Sätze mit der höchsten durchschnittlichen Phrasenlänge ausgewählt und betrachtet. In Tabelle 5.10 werden die zugehörigen Ergebnisse dargestellt.

Daten	∅Phrasenlänge(n)	BLEU
Ziel Deutsch (nbest)	2.4471	19.37%
Ziel Franz.(nbest)	2.7769	27.43%

Tabelle 5.10: Für *Ziel Deutsch (nbest)* und *Ziel Französisch (nbest)* (jeweils aus den Sätzen mit größter durchschnittlicher Phrasenlänge der n-besten Hypothesen von Text 0 und Text 1 erzeugt): durchschnittliche Phrasenlänge und BLEU-Wert

Für beide Übersetzungsrichtungen gilt, dass die durchschnittliche Phrasenlänge deutlich steigt. Sie ist sowohl höher als jene der Texte 0 und 1 als auch der nach Phrasenlänge selektierten Ergebnisse dieser. Gleichzeitig sind auch die BLEU-Werte höher als in Text 0 und 1. Sie sind jedoch geringer als die Werte des zuvor betrachteten Experiments. Dies war zu erwarten. Da die Selektion auf der n-Bestenliste durchgeführt wird, wird eine gewisse Endqualität garantiert. Während der Selektion werden die Bewertungen der Hypothesen ignoriert. Sie wird ausschließlich nach der durchschnittlichen Phrasenlänge durchgeführt.

5.2.5 Oracle

Durch die bisherigen Experimente wurde einige mögliche Zusammenhänge gefunden. Um diese zu validieren wurde ein *Oracle* genutzt. D.h. mithilfe eines Tools wurde der bestmögliche BLEU-Wert berechnet, welcher mit den gegebenen Eingaben möglich ist. Die Daten dazu stehen in Tabelle 5.11.

Durch die *Oracle*-Werte ist erkennbar, dass die Eingaben zwar eine unterschiedliche Qualität haben, jedoch beide nicht die best- bzw. schlechtmöglichste Qualität aufweisen. Somit ist es durch satzweise Selektion möglich ein besseres Ergebnis zu erreichen, wodurch die vorherigen Experimente gerechtfertigt werden.

Daten	topbest score:	oracle score:	worst-oracle score:
Ziel Deutsch	24.40%	25.55%	16.26%
Ziel Deutsch (rand)	21.06%	23.10%	18.84%
Ziel Franz.	36.80%	37.83%	21.93%
Ziel Franz.(rand)	30.25%	34.07%	26.37%

Tabelle 5.11: *Oracle*-BLEU-Wert für je zwei Eingaben (*Ziel Deutsch 0* mit *Ziel Deutsch 1*; *Ziel Französisch 0* mit *Ziel Französisch 1*) sowie für eine zufällige Eingabe

Betrachtet man die BLEU-Werte der zufälligen Eingaben, so liegen diese zwischen den besten und schlechtesten Werten im Bezug auf die *Oracle*-Werte. Sie bilden eine gute Vergleichsbasis für die Ergebnisse der Experimente, da zumindest zufällige Werte überschritten werden sollten, um zu zeigen, dass die untersuchten Merkmale einen gewissen Einfluss haben.

5.2.6 Überblick

In Tabelle 5.12 sind die BLEU-Werte aus den verschiedenen Experimenten zusammengetragen. Durch die unterschiedlichen Werte lässt sich ein unterschiedlicher Einfluss der untersuchten Merkmale schätzen.

Daten	Ziel Franz.	Ziel Deutsch
0	23.14%	17.55%
1	36.80%	24.40%
Perplexität (big lm)	31.59%	22.19%
Perplexität (indomain lm)	31.18%	21.71%
OOV (big lm)	35.93%	17.84%
OOV (indomain lm)	35.27%	18.49%
best durchschnittliche Phrasenlänge	29.88%	21.20%
nbest durchschnittliche Phrasenlänge	27.43%	19.37%
oracle	37.83%	25.55%
rand	30.25%	21.06%

Tabelle 5.12: Vergleich der BLEU-Werte aus den Experimenten sowie der Originaldaten(Spalten). In den Zeilen werden *Ziel Französisch* und *Ziel Deutsch* nebeneinander dargestellt.

Insgesamt ist erkennbar, dass der BLEU-Wert nach jedem Experiment besser als der schlechtere Wert der Eingaben ist. Die Werte unterscheiden sich teilweise stark. Somit ist ein möglicher Einfluss der untersuchten Kriterien sichtbar. Dabei ist nicht auszuschließen, dass die BLEU-Werte auch von anderen Kriterien beeinflusst werden.

Des weiteren übertreffen die BLEU-Werte der Experimente nach Perplexität und Anzahl an *OOV*-Wörtern den Wert der Zufallseingabe für die Übersetzung ins Französische. Daraus können wir schließen, dass sowohl Perplexität als auch Anzahl an *OOV*-Wörtern die Einfachheit eines Textes beeinflussen. Für die Phrasenlänge gilt dies nicht. Sie liegt unterhalb des BLEU-Wertes der zufälligen Eingabe.

Für die BLEU-Werte zu der Übersetzung ins Deutsche gilt Folgendes. Die Ergebnisse der Experimente nach Perplexität und nach durchschnittlicher Phrasenlänge (beste Hypothese) übertreffen die BLEU-Werte einer zufälligen Eingabe. Hieraus ist ein möglicher Einfluss dieser Merkmale ableitbar. Da sich die Ergebnisse hierbei von den zu *Ziel Französisch* gehörenden unterscheiden, ist es möglich, dass der Einfluss von Merkmalen sprachabhängig ist.

Die Ergebnisse für *Ziel Französisch* lassen sich wie folgend sortieren:
Anzahl an *OOV*-Wörtern, Perplexität, durchschnittliche Phrasenlänge (best), durchschnittliche Phrasenlänge (n-best)
Für *Ziel Deutsch* gilt:
Perplexität, durchschnittliche Phrasenlänge (best), durchschnittliche Phrasenlänge (n-best), Anzahl an *OOV*-Wörtern In beiden Fällen erzielt vor allem das Experiment nach dem Kriterium Perplexität eines der besten Ergebnisse und hat somit wahrscheinlich in beiden Fällen einen Einfluss auf die Einfachheit.

6 Fazit

In dieser Arbeit haben wir verschiedene Kriterien betrachtet, um zu untersuchen, ob sie einen Einfluss auf die Einfachheit von Daten und damit auf die Qualität einer Übersetzung haben. Dabei wurden Perplexität, Anzahl an *OOV*-Wörter und die durchschnittliche Phrasenlänge betrachtet.

Durch die Qualität der genutzten Übersetzungen, ist auch die Qualität der Ergebnisse der Experimente beschränkt. Sowohl durch die Evaluationsmetrik BLEU als auch durch manuelle Bewertungen fiel diese ausreichend und weiterhin verbesserbar aus.

Insgesamt sind bei jedem Experiment Verbesserungen im Vergleich zu dem schlechter bewerteten Text zu erkennen, dabei wurde die Qualität des besser bewerteten Textes allerdings nie erreicht oder sogar überschritten. Zugleich wurde der BLEU-Wert für eine zufällige Eingabe nicht immer erreicht, wodurch nicht erkennbar ist, ob die Verbesserung tatsächlich von dem untersuchten Kriterium verursacht wird.

Insbesondere für das Kriterium Perplexität ist zu vermuten, dass ein niedrigerer Wert die Einfachheit eines Textes erhöht. Da der Einfluss der verschiedenen Eigenschaften in dieser Arbeit vor allem einzeln betrachtet wurde, ist nicht auszuschließen, dass erzielte Ergebnisse auch von anderen Eigenschaften verursacht werden.

Insgesamt erscheint es möglich durch einfachere Eingaben eine bessere Qualität zu erreichen. Auch wenn dies hier vor allem auf den bereits übersetzten Daten untersucht wurde, sollte dies ebenfalls für die Daten vor der Übersetzung möglich sein. Welche Kriterien dabei wie stark ins Gewicht fallen ist nicht klar erkennbar.

7 Ausblick

Betrachtet man die durchgeführten Experimente, so existieren noch einige mögliche Verbesserungsmöglichkeiten. Für die hierbei durchgeführten Experimente wurde als Evaluationsmetrik ausschließlich der BLEU-Score verwendet. Somit sind dieselben Experimente auch mit anderen Evaluationsmetriken durchführbar und auswertbar. Zusätzlich können die bisherigen Experimente durch weitere erweitert werden.

In erster Linie können weitere Eigenschaften untersucht werden, welche einen möglichen Einfluss auf die Einfachheit eines Textes haben, wie bspw. die Satzlänge. Mit der Länge steigt auch die Wahrscheinlichkeit, dass ein Satz eine komplexere Struktur hat, was schwieriger zu übersetzen ist.

Eine weitere Ergänzungsmöglichkeit ist die Kombination mehrerer Kriterien zu betrachten. Wie in dieser Arbeit festgestellt, beeinflussen sich manche gegenseitig. Dabei gilt es zu unterscheiden, ob diese Beeinflussung durch die untersuchten Kriterien oder durch andere Kriterien statt findet.

Dazu passend ist es nötig den Einflussgrad der Kriterien zu untersuchen. Abhängig von diesem kann die Selektion derartig erweitert werden, dass nach mehreren Kriterien auf einmal ausgewählt wird und dabei deren Einflusstärke miteinfließt. So können größere Modelle entwickelt werden, welche später in der Praxis einsetzbar sind.

Des weiteren können alle für diese Arbeit durchgeführten Experimente auch auf der Quellseite ausgeführt und ausgewertet werden. Wir rechnen mit ähnlichen Ergebnissen, da eine einfachere Eingabe zu einer entsprechend einfacheren Übersetzung führen sollte. Dies ist jedoch noch zu überprüfen.

Um aus dieser Arbeit resultierende Ergebnisse praktisch umzusetzen, müssen noch weitere Untersuchungen vorgenommen werden. Es besteht die Aussicht, dass derartige Forschungen zu einer Verbesserung der maschinellen Übersetzung führen.

Literaturverzeichnis

[Barancikova und Tamchyna 2014] BARANCIKOVA, Petra ; TAMCHYNA, Ales: Machine Translation within One Language as a Paraphrasing Technique. In: *CEUR Workshop Proceedings* Bd. 1214, 2014, S. 1–6 9

[Bouillon u. a. 2014] BOUILLON, Pierrette ; GASPAR, Liliana ; GERLACH, Johanna ; PORRO, Victoria ; ROTURIER, Johann: Pre-editing by Forum Users: a Case Study. In: *Workshop (W2) on Controlled Natural Language Simplifying Language Use - 9th International Conference on Language Resources and Evaluation (LREC)*, 2014, S. 3–10 10

[Brown u. a. 1993] BROWN, Peter E. ; PIETRA, Vincent J. D. ; PIETRA, Stephen A. D. ; MERCER, Robert L.: The Mathematics of Statistical Machine Translation: Parameter Estimation, 1993 4

[Chen und Goodman 1998] CHEN, Stanley F. ; GOODMAN, Joshua: An Empirical Study of Smoothing Techniques for Language Modeling, 1998 5

[Denkowski und Lavie 2014] DENKOWSKI, Michael ; LAVIE, Alon: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014 7

[Densmer] DENSMER, Lee: *Machine Translation Pre-Editing To Boost Output Quality*. – URL `http://info.moravia.com/blog/machine-translation-pre-editing-to-boost-output-quality` 7

[Foster u. a. 2006] FOSTER, George ; KUHN, Roland ; JOHNSON, Howard: Phrasetable Smoothing for Statistical Machine Translation. In: *Conference on Empirical Methods in Natural Language Processing*, 2006 13

[Koehn 2009] KOEHN, Philipp: *Statistical Machine Translation*. 2009 3, 4

[Koehn u. a. 2003] KOEHN, Philipp ; OCH, Franz J. ; MARCU, Daniel: Statistical Phrase-Based Translation, 2003 4

[Marton u. a. 2009] MARTON, Yuval ; CALLISON-BURCH, Chris ; RESNIK, Philip: Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In: *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* Bd. 1, 2009, S. 381–390 9

[Mirkin u. a. 2009] MIRKIN ; SHACHAR ; SPECIA, Lucia ; CANCEDDA, Nicola ; DAGAN, Ido ; DYMETMAN, Marc ; SZPEKTOR, Idan: Source-language entailment modeling for translating unknown terms. In: *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* Bd. 2, 2009, S. 791–799 10

[Och 1999] OCH, Franz J.: An efficient method for determining bilingual word classes. In: *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 1999, S. 71–76 13

[Papineni u. a. 2002] PAPINENI, Kishore ; ROUKOS, Salim ; WARD, Todd ; ZHU, Wei-Jing: BLEU: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, S. 311–318 7

[Poulis und Kolovratnik 2012] POULIS, Alexandros ; KOLOVRATNIK, David: To post-edit or not to post-edit? Estimating the benefits of MT post-editing for a European organization, 2012 10

[Schmid und Laws 2008] SCHMID, Helmut ; LAWS, Florian: Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In: *COLING*, 2008 13

[Snover u. a. 2006] SNOVER, Matthew ; DORR, Bonnie ; SCHWARTZ, Richard ; MICCIULLA, Linnea ; MAKHOUL, John: A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of Association for Machine Translation in the Americas*, 2006 7

[Stolcke 2002] STOLCKE, Andreas: *SRILM AN EXTENSIBLE LANGUAGE MODELING TOOLKIT*. 2002 6, 13

[Zhao u. a. 2007] ZHAO, Shiqi ; LIU, Ting ; YUAN, Xincheng ; LI, Sheng ; ZHANG, Yu: Automatic Acquisition of Context-Specific Lexical Paraphrases. In: *Proceeding IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence*, 2007, S. 1789–1794 9