Universität Zürich[UZH]

Bachelorarbeit

zur Erlangung des akademischen Grades

Bachelor of Arts

der Philosophischen Fakultät der Universität Zürich

Erstellung und Alignierung eines parallelen Korpus auf der Basis von “The Times in Plain English”

Verfasserin: Anja Ryser

Matrikel-Nr: 17-704-461

Betreuer: Prof. Dr. Martin Volk

Co-Betreuer: Mathias Müller

Institut für Computerlinguistik

Abgabedatum: 15.06.2020

Abstract

At the moment there are few sources of parallel texts for the training of automatic text simplification. The alignment is mostly done by hand and is very time consuming. In the last years there was some effort to automatically create alignments for simplified texts. Alignment tools like Cats, Transformer and LHA were created, but only a few like Cats are specialised for the use in text simplification. In 2019, Facebook published a new alignment tool called Laser, which achieved very promising results in aligning sentences from different languages. Machine translation systems which were trained with the alignments of Laser beat other 'state-of-the-art' systems. The question comes up, whether this alignment tool can also be used for aligning simplified texts. To answer this question a corpus was created with articles from 'The Times in Plain English' and their original articles. The data was cleaned and alignments were generated with all four alignment tools. These alignments were then evaluated and compared with each other.

Zusammenfassung

In der automatischen Textvereinfachung gibt es noch sehr wenige parallele Daten, auf denen automatische Textvereinfachungssysteme trainiert werden können. Die Alignierung von Sätzen erfolgt meist von Hand und ist sehr zeitaufwändig. In den letzten Jahren gab es verschiedene Versuche, die Alignierung von vereinfachten Texten zu automatisieren. So sind Systeme wie Cats, Transformers und LHA entstanden. Nur sehr wenige, darunter Cats sind darauf spezialisiert, Alignierungen für vereinfachte Texte zu erstellen. 2019 brachte Facebook Laser heraus, ein neues System, das Alignierungen von verschiedenen Sprachen generiert. Mit ihren Daten konnten die Autoren Übersetzungssysteme verbessern und überholten Systeme auf dem neusten Stand der Technik. Es stellt sich die Frage, ob Laser auch für die Alignierung von vereinfachten Sätzen eingesetzt werden könnte. Um diese Frage zu beantworten wurde ein Korpus aus Artikeln von 'The Times in Plain English' und deren Originalen erstellt. Die Daten wurden bereinigt und mit allen vier Systemen aligniert. Diese Alignierungen wurden dann evaluiert und miteinander verglichen.

Danksagung

Ich danke Martin Volk, meinem Betreuer, für die Unterstützung, die hilfreichen Vorschläge und das Interesse für diese Arbeit. Trotz den speziellen Bedingungen hatte er immer ein offenes Ohr für Probleme, die auftauchten. So konnten die Probleme besprochen und gelöst werden und es fand sich immer ein Weg, wie es weitergehen konnte.

Ich danke Mathias Müller, Stefan Vrankovich und Marek Kostrzewa für das zur Verfügung stellen ihrer eigenen Arbeiten und das schnelle Beantworten von Fragen, die dazu auftauchten. Sie lieferten wichtige Beiträge für diese Arbeit.

Ich danke Valentin Huber, Hanna Gfeller und meinen Eltern dafür, dass sie sich durch alle Schreibfehler und den Formulierungs-Dschungel geschlagen haben und mir halfen, dass diese Arbeit verständlicher wird (Das hoffe ich zumindest).

Und ich danke all denen, die in irgendeiner Form mitgeholfen haben, dass diese Arbeit entstanden ist, die ich hier vergessen habe zu erwähnen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

In Schaffhausen lasen die Schüler und Schülerinnen einer fünften Klasse zwei Wochen lang jeden Tag die Zeitung. Nach den zwei Wochen schrieben die Schüler einen Leserbrief an die Zeitung und präsentierten ihre Vorschläge, was die Schaffhauser Nachrichten in Zukunft besser machen könnte. Ein Schüler schrieb:

'Liebe Zeitungsmacher/innen, manchmal finde ich die Zeitung langweilig, aber manchmal auch spannend. Man könnte spannende Ereignisse einfacher erklären.' (Ruh [2020])

Wir alle mussten schon einmal einen Text zwei, drei oder viermal lesen, weil er so kompliziert geschrieben ist, dass man ihn nicht versteht. Im besten Fall kostet das ein paar Sekunden Zeit, bis wir zum Beispiel einen Zeitungsartikel verstehen. Im schlimmsten Fall verstehen wir einen Text gar nicht, der wichtige Anweisungen enthält und verpassen darum zum Beispiel einen Abgabetermin für ein wichtiges Formular. In unserer unpersönlichen, digitalisierten Welt ist es immens wichtig geworden, schriftliche Texte verstehen zu können. Wurden vor ein paar Jahren viele Aufgaben des täglichen Lebens noch von Mensch zu Mensch ausgeführt, erledigen wir diese heute digital von unseren Endnutzergeräten aus. Dafür erhalten wir schriftliche Anweisungen und müssen die Aufgaben ohne die Hilfe von Fachpersonen, die uns dabei aktiv unterstützen, selbst erledigen. Trotz der Wichtigkeit, dass Texte verstanden werden, sind viele Texte so geschrieben, dass ein durchschnittlicher Leser sie nur schwer versteht. Ein Lösungsansatz für dieses Problem ist die vereinfachte Sprache. Es werden Richtlinien erarbeitet, wie Texte geschrieben werden können, dass sie verstanden werden. Das Umschreiben schon bestehender Texte ist ein zusätzliches Problem. Durch die schiere Menge an Texten ist der zeitliche und finanzielle Aufwand diese manuell umzuschreiben, schlichtweg zu hoch. Eine Lösung dafür ist die Textvereinfachung zu automatisieren.

1.1 Motivation

Die Vorlesungen bei Sarah Ebling haben mein Interesse an computerlinguistischen Anwendungen in der Barrierefreiheit geweckt. Dieses Interesse verstärkte sich in einem Seminar zum Thema 'automatische Textvereinfachung'. In der automatischen Textvereinfachung vereinen sich Probleme der automatischen Übersetzung mit ganz eigenen Problemen wie zum Beispiel: Wie können Inhalte durch einen Text einfach vermittelt werden? Was sind die wichtigen Informationen in einem Text? Wie können diese extrahiert werden? Ich entschied mich, dieses Thema in meiner Bachelor-Arbeit weiter zu vertiefen. Um Systeme für die automatische Textvereinfachung zu trainieren, braucht es viele Daten, also Original-Texte und deren Vereinfachungen. Diese Daten werden meist von Hand vereinfacht und aligniert. Dieses Verfahren ist sehr aufwändig. Um mehr Daten zu generieren und so auch die automatische Textvereinfachung zu verbessern, werden neue Ansätze gesucht, wie die zusammengehörigen Satzpaare automatisch aligniert werden können. 2018 veröffentlichten Schwenk et al. ein neues Alignierungssystem 'Laser' (Language Agnostic Sentence Representation). (Schwenk [2018]). Es erbrachte vielversprechende Ergebnisse bei der Alignierung von Übersetzungen. Dabei erstellte es Satzembeddings, welche die Bedeutung eines Satzes in einer Matrix repräsentieren. Es stellt sich die Frage, ob Laser auch für die Alignierung von vereinfachten Texten geeignet sein könnte. Ich entschied mich, dieser Frage in meiner Arbeit nachzugehen. Ich erstellte ein Korpus und versuchte dieses mit Laser zu alignieren.

1.2 Fragestellung

Ich baue anhand von einer Zeitung ('The Times in Plain English'), welche Artikel von verschiedenen, grossen Zeitungen inhaltlich und formell vereinfacht zur Verfügung stellt, ein Korpus auf. Dazu gehören die entsprechende Vorverarbeitung und einige Analysen der Daten.

Die Fragen, die ich in dieser Arbeit beantworten möchte, sind: Kann Laser benutzt werden, um Texte in einfachem und normalem Englisch zu alignieren? Erreicht Laser in diesem Gebiet, wie in der automatischen Übersetzung, bessere Resultate als bisherige Systeme?

1. Was mache ich?
   Ich erstelle ein Korpus von 'The Times in Plain English'-Artikeln und ihren

Originalen. Diese Sätze aligniere ich mit Laser und vergleiche die Ergebnisse mit Ergebnissen von Cats, Transformer und LHA.

2. Wie mache ich das?
   Ich lade die Artikel von der Webseite der Zeitung herunter und bereite sie auf, indem ich übriggebliebene HTML-Elemente lösche. Mithilfe von Skripts von Mathias Müller aligniere ich die Sätze mit Laser. Zusätzlich aligniere ich die Daten mit Cats. Marek Kostrzewa alignierte die Daten mit Transformer und LHA. Danach vergleiche ich alle Resultate. Ich versuche herauszufinden, welches System für die Alignierung am besten geeignet ist.

3. Warum mache ich das?
   Kann Laser für die Alignierung von vereinfachten Texten eingesetzt werden, können parallele Daten in Zukunft schneller aligniert und sogar neue Satzpaare aus nicht-parallelen Daten generiert werden.

1.3 Struktur der Arbeit

In Kapitel 1 beschreibe ich meine Motivation und das gewählte Vorgehen.
In Kapitel 2 stelle ich Plain-Englisch als eine Vereinfachung von Englisch vor und führe die Lesbarkeitsformel für den 'Flesch-Kincaid-Grade-Level' ein. Dazu folgen einige Gedanken zur Lesbarkeit der in dieser Arbeit verarbeiteten Daten.
Kapitel 3 behandelt die Daten und das daraus entstandene Korpus. Ich erläutere, wie die Daten gewonnen und vorverarbeitet wurden. Dazu stelle ich das Korpus in einigen Analysen des Korpus-Profils dar.
In Kapitel 4 beschreibe ich das Vorgehen bei der Alignierung mit den verschiedenen verwendeten Systemen. Der Fokus liegt dabei wegen der gewählten Fragestellung auf Laser.
In Kapitel 5 präsentiere ich die erzeugten Resultate, erkläre und diskutiere sie.
In Kapitel 6 fasse ich zusammen, was ich gemacht habe und mache mir einige Gedanken zu noch offene Fragen.

Im Anhang finden sich ausführliche Tabellen und Auszüge aus einem Artikel im Korpus und dessen Alignierungen.

Mein Code, den ich im Rahmen dieser Arbeit geschrieben habe, wurde mit der Arbeit abgegeben.

Das gesamte Korpus mit den entstandenen Alignierungen von Cats, Transformers und LHA wurde separat abgegeben.

2 Leichte Sprache im Englischen

Schon vor einiger Zeit fiel verschiedenen Institutionen auf, dass das Nicht-Verstehen von Texten und schriftlichen Anweisungen fatale Folgen haben kann. So stürzte 1989 ein chinesisches Flugzeug ab. Auf der Aufnahme der Blackbox konnte man eine Konversation aus dem Cockpit hören: 'Was heisst 'pull up?' Das Flugzeug hatte an Höhe verloren und im Cockpit blinkte die Warnmeldung 'Pull up' auf. Die chinesische Besatzung verstand diese nicht. Im internationalen Funkverkehr wird der Begriff 'climb' verwendet, muss ein Flugzeug an Höhe gewinnen. Der Begriff 'Pull up' wird nur im amerikanischen Kontext verwendet und die chinesische Besatzung konnte die Bedeutung des zusammengesetzten Verbes nicht erschliessen. (Thrush [2001]). Doch auch bei Personen, die Englisch als Muttersprache sprechen, geschehen kostspielige Fehler. So beschreibt Kincaid solche Fehler im US-amerikanischen Militär: *The effects of faulty communication are well known and disastrous. One recent Air Force study (...) has traced many costly errors to the reading difficulty level of the instructions in manuals to be followed. The more difficult the material, the more mistakes were made.* (Kincaid et al. [1975], Seite 1. Die erwähnte Studie ist nicht publiziert.) Es liegt im Interesse verschiedenster offizieller Stellen, wichtige Informationen einem durchschnittlichen Bürger verständlich mitzuteilen.

Englisch war eine der ersten Sprachen, in der eine systematische Vereinfachung angestrebt wurde. Schon 1943 entwickelte Flesch eine erste Lesbarkeitsformel, um darauf aufbauend Texte vereinfachen zu können. Flesch [1948]

Da meine Arbeit auf Artikeln, Lesbarkeitsniveaus und Richtlinien für Plain-Englisch aus den USA basiert, behandle ich im weiteren Plain-Englisch aus einer US- amerikanischen Perspektive. Von all den verschiedenen Lesbarkeitsmassen wird nur der Flesch-Kincaid-Grade-Level besprochen, da dieser verwendet wird, um die Lesbarkeit von 'The Times in Plain English'-Texten zu messen.

2.1 Plain-Englisch

Als Plain-English wird eine Form des Englisch bezeichnet, die so klar, präzise und übersichtlich geschrieben ist, dass ein Leser einen Text beim ersten Lesen versteht. Ziele des Plain-Englisch sind, dass ein Leser eine gesuchte Information schnell findet und so versteht, dass er sie optimal umsetzen kann. Momentan wird Plain-Englisch in den USA vor allem als Standard für Regierungsinformationen benutzt (Plain Language Action and Information Network [c]). Plain-Englisch findet aber auch Verwendung in der Medizin, in der Krisenkommunikation und im Militär (Johns and Wheat [1978]) und an vielen anderen Orten, wo es wichtig ist, dass Informationen und Anweisungen schnell verstanden und umgesetzt werden. Obwohl Bemühungen in diese Richtung zielen, gibt es momentan keine allgemein gültige Definition, was Plain-English ist. (Schriver et al. [2010])

Erst in den 1970er Jahren begann die amerikanische Regierung, ihre Texte verständlicher zu verfassen; so verfügte Präsident Nixon 1978, dass das 'Federal Register' (Amtsblatt, indem die Regierungen Vorschriften und Bekanntmachungen veröffentlicht), in 'Layman's Terms' geschrieben wird. Der Begriff 'Layman's Terms' war der Vorgänger des Begriffes Plain-Englisch und soll bedeuten, dass ein 'Layman' (ein Laie), den Text versteht. Das Erziehungsdepartement initiierte darauf Forschungen dazu und erste Texte wurden vereinfacht. In der folgenden Zeit hielten sich erste Departemente an die Regelungen, andere nicht, und lange wurde kein Fortschritt gemacht. (Plain Language Action and Information Network [b])

In den 1990er Jahren gründeten Angestellte der Regierung das Netzwerk PLAIN (The Plain Language Action and Information Network). PLAIN verfolgt das Ziel, sämtliche Kommunikation der amerikanischen Behörden und Regierung in Plain-Englisch zu verfassen(Plain Language Action and Information Network [a]). 1998 veröffentlichte PLAIN die ersten Richtlinien für Plain-Englisch. Es wurden mehrere Verordnungen verabschiedet, die den Gebrauch von Plain-Englisch in der Regierung forderten. Plain Language Action and Information Network [b]

2010 unterschrieb Präsident Obama den 'Plain Writing Act of 2010' und machte so die Vorgaben darin zum Gesetz. (Plain Language Action and Information Network [b])

2010 wird Plain-Englisch im 'Plain Writing Act of 2010' offiziell definiert als: *'writing that is clear, concise, well-organized, and follows other best practices appropriate to the subject or field and intended audience.'*, (Section 3,3). Das Gesetz forderte von allen Behörden, bis Mitte 2011 Zuständige innerhalb der Behörden zu bestimmen, die die Einhaltung des Gesetzes durchsetzen, die Angestellten informieren und aus-

bilden. Die Webseiten sollten einen Plain-Englisch Abschnitt beinhalten. Ein Jahr nach Verabschiedung des Gesetzes sollten alle neuen Dokumente in Plain-Englisch geschrieben werden. (Printing Office U.S.-Government [2010])

PLAIN erstellte die Richtlinien, die die Behörden benutzen sollen und bietet halbtägige Schulungen für Beamte der Regierung an. Die Richtlinien sind denen im 'Regelbuch Leichte Sprache' von Christiane Maass (Maass [2015]) sehr ähnlich. Unter anderem sollen in Plain-Englisch-Texten:

- der Text an die angesprochene Gruppe und deren Vorwissen und Interessen angepasst werden
- der Lesende direkt angesprochen werden und indirekte Anreden vermieden werden (z.B. 'Wenn sie sich bewerben wollen sollten sie...' statt 'Der Bewerbende sollte...')
- klare, kurze Titel gesetzt werden
- Abschnitte kurz und prägnant ein einzelnes Thema beschreiben
- klare, prägnante Worte, die der Leser kennt, benutzt und Fachjargon vermieden werden
- schwierige Formen wie passiv oder nominalisierte Verben vermieden werden; wann immer möglich sollte die einfachste Form eines Wortes benutzt werden
- Abkürzungen vermieden werden
- für ein wiederkehrendes Element derselbe Begriff verwendet werden
- Sätze so kurz und so einfach wie möglich gehalten werden
- doppelte Negationen und Ausnahmen von Ausnahmen vermieden werden
- doppeldeutige Sätze und Wörter vermieden werden
- Beispiele, Listen, Tabellen und Bilder benutzt werden, um das Gelesene zu vertiefen

(The Plain Language Action and Information Network [2011])

Auch von Online-Zeitschriften wie 'The Times in Plain English' wird Plain-Englisch eingesetzt, um einfache Texte zu erstellen. Diese Texte sollen Menschen, die Englisch nicht als Muttersprache sprechen, helfen, an wichtige Informationen zu kommen. Auch in Schulen kann Plain-Englisch verwendet werden, um Kinder angemessen an Nachrichten und Politik heranzuführen.

Plain-Englisch wird auch kritisiert. So sind die uneinheitlichen Begrifflichkeiten immer wieder Ursprung von Problemen und Uneinigkeiten bei Forschenden. Möchte man seine Webseite nach dem Standard von Plain-Englisch gestalten, findet man viele Richtlinien, die sich unterscheiden und damit für Verwirrung sorgen. Damit sich Plain-Englisch weiter verbreiten kann, müssen die Definitionen und Richtlinien standardisiert werden.
Ein weiteres Problem ist, dass die meisten Forschungen zum Thema von Personen gemacht werden, deren Muttersprache Englisch ist. Diese Ergebnisse werden dann an Personen getestet, deren Muttersprache ebenfalls Englisch ist. Da Plain-Englisch auch immer mehr zum Mittel der Kommunikation werden soll für Menschen, deren Muttersprache nicht Englisch ist, repräsentieren diese Studien nicht die Realität. Um also die Einfachheit von Texten zu verifizieren müssen Erfahrungsstudien mit Personen durchgeführt werden, deren Erstsprache nicht Englisch ist. (Schriver et al. [2010])

In den letzten Jahren scheint das Interesse an Studien zu Plain-Englisch im generellen und zur Lesbarkeit von Zeitungen zurückgegangen zu sein. So findet man sehr viele ältere, aber kaum neuere Studien zu diesen Themen.

2.1.1 Vereinfachung in 'The Times in Plain Englisch'

Die Autoren von 'The Times in Plain English' haben keine Richtlinien veröffentlicht, an die sie sich halten. Es ist nur bekannt, dass sie ihre Texte mit dem Flesch-Kincaid-Grade-Level bewerten. (Schiff [2015])

Während der Arbeit mit den Daten konnte ich einige Beobachtungen machen. Die Artikel in Plain-Englisch beruhen auf Original-Artikeln, werden aber neu geschrieben. (Dies wird in Kapitel 3 detaillierter besprochen.)

Die folgenden Beispiele stammen aus den Artikeln, die aus dem erstellten Korpus stammen. Es wurden keine zusätzlichen Anführungszeichen gesetzt. Inkonsistenzen wurden direkt aus den Daten übernommen.

Die Vereinfachung passiert zu grossen Teilen durch Weglassen von Informationen. So sind die Original-Artikel meist Erfahrungsberichte, die anhand eines Schicksals ein Thema behandeln. Dadurch finden sich viele Informationen in den Artikeln, die nicht direkt zum Thema gehören. Diese werden in den Artikeln in Plain-English meist weggelassen.
So wird in Artikel 1191, in dem es um Hitzewallungen von Frauen geht, der Artikel mit einem Beispiel einer Betroffenen eingeleitet:

Original: *"It's miserable, I'll tell you what," said Sharon Brown, 57, of Winston-Salem, N.C., who has endured hot flashes for six years. At her job at a tax and accounting office, she has had to stop wearing silk. "I keep one of the little fans with me at all times — one in my purse, a couple in my desk, some in just random places in the office," she said. "I'll be so glad when they stop — if they ever stop." Over all, black and Hispanic women experienced hot flashes for significantly longer periods than white or Asian women.*
Es werden der Name, das Alter, der Job und die Symptome dieser Frau beschrieben und wie diese sie im Alltag einschränken.
Im Artikel in Plain-English wird ihr Bild verwendet und ihre Geschichte mit der Bildunterschrift *Sharon Brown uses fans in her job.* erzählt; mehr erfährt der Leser nicht.

Dafür werden Umstände, die in den Original-Artikeln als Vorwissen vorausgesetzt werden, erklärt. So wird in Artikel 0122 das Problem der Verfolgung der Muslimischen Minderheit in Assam eingeleitet mit:
Plain Englisch: *Most of India is Hindu. There is a long and ancient history of struggle between the two sects. India has 1.3 billion people. The backgrounds of its many peoples are diverse. India is famous for the way its many different people live in a democracy. That is changing.*
Diese explizite Einleitung und Erklärung findet man im Original-Artikel so nicht.

Lange, komplexe Sätze werden teils nur in zwei Sätze umformuliert; der Wortlaut bleibt so weit wie möglich erhalten.
Artikel 0122 Original: *'Many of the people whose citizenship is now being questioned were born in India and have enjoyed all the rights of citizens, such as voting in elections.*

Plain-Englisch: *Many of the people whose citizenship is under review were born in India. They have enjoyed all the rights of citizens, such as voting in elections.'*

Direkte Reden werden eins zu eins übernommen, meistens aber gekürzt oder als indirekte Rede wiedergegeben.
Artikel 0050 Original: *While the roundup "gave the American people their jobs back," said Cortez McClinton, 38, a former construction worker who was hired at the plant hours after the raids, "how they handle the immigration part is that they're still separating kids from their families." (...)*

Plain-Englisch: *How they handle the immigration part is that they are still separating kids from their families ." Another worker denounced the federal agents.*

Während die originalen Artikel gerne mit impliziten Informationen und Emotionen

sowie mit Ironie ihre Leser an ein Thema heranführen, verzichten die Artikel in Plain-Englisch meist darauf und bestehen grösstenteils aus trockenen Fakten. So findet man in Artikel 1997 den Satz:
Original: *'WASHINGTON — The government has some thoughts on how to make the federally financed school lunch program more nutritious: A quarter-cup of tomato paste on pizza will no longer be considered a vegetable.'*
Auf eine solche Bemerkung wird im Artikel in Plain-Englisch verzichtet.

Findet man in den Original-Artikeln meist verschiedene Meinungsbilder, die alle Ansichten auf ein Thema beleuchten, fehlt das bei den Artikeln in Plain-Englisch. Dadurch ist die Gefahr gross, dass Artikel in Plain-Englisch ein Thema einseitig behandeln, da nur eine Meinung dazu wiedergegeben wird.

In wie weit ein Artikel vereinfacht wird und wie nahe er am Thema des Original-Artikels liegt, ist von Artikel zu Artikel unterschiedlich.

2.2 Flesch-Kincaid-Grade-Level

Rudolf Flesch entwickelte bereits 1943 eine Formel, um die Lesbarkeit eines Textes objektiv zu messen. 1948 verbesserte er diese Formel. Er teilte die Lesbarkeit in zwei Kategorien auf; Formel A, die rein die Lesbarkeit eines Textes misst anhand der gewichteten durchschnittlichen Satzlänge und Anzahl der Silben pro 100 Wörter. Formel B versucht, das Interesse eines Menschen an einem Text zu messen. Dazu wurde gezählt, wie viele persönliche Wörter (wie Namen oder Orte) ein Text beinhaltet und wie viele Sätze, die den Leser direkt ansprechen oder die direkte Rede beinhalten. Je persönlicher ein Text ist und je mehr Menschen darin vorkommen, desto interessanter wird er wahrgenommen. Diese Formel liefert einen Wert auf einer Skala von 0 bis 100. Ein Wert von 100 bedeutet bei beiden Formeln, dass ein zehnjähriges Kind, diesen Text gut versteht. Ein Wert von 0 bedeutet, dass ein Akademiker den Text versteht. Die Lesbarkeit war dabei umgekehrt proportional zur Klassenstufe, in der ein Kind den Text gut versteht und Fragen dazu beantworten kann. Die Klassenstufe kann anhand einer Tabelle abgelesen werden. (Flesch [1948]) Die Flesch-Formel wurde bald zum Standard für die Berechnung der Lesbarkeit. Sie wurde zuerst für Englisch entwickelt, wird aber mittlerweile in vielen Sprachen verwendet.

1975 verbesserten Kincaid et al. [1975] verschiedene Formeln für den Gebrauch bei der US-Marine. Die neuen Formeln sollten eingesetzt werden, um Anleitungen und Schulungsmaterial der Marine dem Leseniveau der Soldaten anzupassen. Im Vor-

feld sind immer wieder kostspielige Fehler passiert, bei denen Anweisungen nicht richtig verstanden wurden. Da viele andere Formeln den Grade-Level, ab dem ein Schüler einen Text gut versteht und rezitieren kann, direkt messen, passte Kincaid die Formel A von Flesch an; der Wert entsprach nun ebenfalls dem Grade-Level. Die Flesch-Formeln sind schwierig und zeitintensiv in der Anwendung, da die Gewichtung der zwei Faktoren auf drei Kommastellen genau sein musste. Es zeigte sich, dass deswegen bei der Anwendung der Flesch-Formel sehr häufig Fehler passierten und die Ergebnisse deswegen nicht verlässlich waren. So wurden die Gewichtungen von Kincaid nur auf zwei Dezimalstellen genau angegeben, eine weitere Stelle trug nicht zur Verbesserung der Resultate bei. In den Texten der Marine kommen viele Fachwörter vor, was den Grade-Level eines Textes erhöht. Dies ist allerdings für die Angehörigen der Marine kein Problem, da sie diese Wörter kennen und auch Wiederholungen kein Problem darstellen. Aus diesem Grund wurden die Formel und ihre Gewichtungen spezifisch auf solche Texte ausgelegt; der berechnete Level ist ca. 1.5 Grade-Level tiefer als bei der ursprünglichen Flesch-Formel. (Kincaid et al. [1975])

$$Grade\ Level = 0.39 * \frac{Total\ Wörter}{Total\ Sätze} + 11.8 * \frac{Total\ Silben}{Total\ Wörter} - 15.59$$

Abbildung 1: Flesch-Kincaid-Grade-Level

Ein durchschnittlicher Soldat der US-Marine versteht Texte bis Grade-Level 9 gut. Danach sinkt die Verständlichkeit von Texten. Auch heute gilt noch: ein Text, der verstanden werden muss sollte Grade-Level 9, also das Niveau eines 14-Jährigen, nicht übersteigen. (Kincaid et al. [1975])

Das Grade-Level System der USA geht von 0 bis 12, wobei Grade-Level 0 Vorschule und Grade-Level 12 Abschluss an einer High School bedeutet. Die Grade-Level und das entsprechende Alter werden in Tabelle 1 dargestellt. Wikipedia [2020]

Die Werte der Flesch-Kincaid-Grade-Level-Formel können diese Skala aber auch überschreiten, es gibt kein oberes Limit. Allerdings werden mit natürlichem Text sehr selten sehr hohe Werte erreicht. Worte können theoretisch eine beliebig grosse Zahl an Silben haben und Sätze können eine beliebig grosse Zahl an Wörter haben. Beides kommt allerdings sehr selten vor.

So hat der vorhergehende Satz ('Die Werte...') einen Flesch-Kincaid-Grade-Level von 13, ist also eindeutig zu schwer für einen durchschnittlichen Text.

Der Satz *'This is a text'* erreicht einen Wert von -2.2, (Grade-Level 0) kann also auch von einem 5-jährigen Kind verstanden werden.

Grade-Level	Schule	Alter in Jahren
0	Preschool	bis 5
1	Primary School	6
2	Primary School	7
3	Primary School	8
4	Primary School	9
5	Primary School	10
6	Middle School	11
7	Middle School	12
8	Middle School	13
9	High School	14
10	High School	15
11	High School	16
12	High School (Abschlussjahr)	17

Tabelle 1: Das Grade-Level System der amerikanischen Schulen nach Alter und Klassenstufe

Der Satz von Bryan Garner: *'If at any time the Federal Energy Regulatory Commission should disallow the inclusion in its jurisdictional cost of gas, cost of service, or rate base at any portion of the cost incurred because of this gas purchase or the full amount of an costs incurred by Buyer for any field services or facilities with respect to any well subject hereto, whether arising from any term or provision in this Agreement or otherwise, including but not limited to price and price adjustments, the prices provided for herein, then Seller agrees that the price will be reduced to the maximum price for gas hereunder which the Federal Energy Regulatory Commission will allow Buyer to include in its jurisdictional cost of gas, cost of service, or rate base and Seller shall promptly refund with interest all prior payments for gas purchased hereunder which exceed the amount Buyer is permitted to include in said cost of gas cost of service, or rate base.'*, wird gerne als schlechtes Beispiel genutzt. Der Satz beinhaltet 159 Wörter. (Schriver et al. [2010]) Er erreicht einen Flesch-Kincaid-Grade-Level von 65.2, was schlichtweg zu kompliziert ist.

Die Flesch-Kincaid-Grade-Level-Formel wird heute vielerorts benutzt. Neuere Versionen von Microsoft Word können den Flesch-Kincaid-Grade-Level automatisch berechnen - dies ist auch der Grund, warum 'The Times in Plain English' diese als Richtlinie benützt. Die Funktion ist für das Englische vorhanden, für Deutsch beispielsweise nicht.

Zusätzlich gibt es Online die Möglichkeit, den Grade-Level zu berechnen. Alle Grade-Level für englische Texte wurden mithilfe von Readability Checker von WebFX berechnet. Deutsche Beispiele berechnete ich von Hand, da diese Webseite nur englische Texte berechnen kann.

2.3 Lesbarkeit von 'The New York Times' und 'The Times in Plain English'

Um einen Überblick über die Lesbarkeit der Artikel, die in dieser Arbeit benutzt werden, zu erlangen, testete ich fünf zufällig ausgewählte Artikel in Plain-Englisch und deren Original-Artikel. Die Berechnungen, die ich in Tabelle 2 darstelle, erfolgten mit dem oben genannten Online-Werkzeug.

Artikel-ID	PLAIN-ENGLISCH	ORIGINAL
0076	4.8	9.7
0275	5.6	10.4
0634	4.8	9.8
1960	10.4	12.9
2085	8.2	10.7
Durchschnitt	6.76	10.7

Tabelle 2: Lesbarkeit Artikel in Flesch-Kincaid-Grade-Level

Die Artikel in Plain-Englisch sind, wie erwartet, alle einfacher als ihr Original. Je schwieriger jedoch das Original war, desto schwieriger ist auch die Vereinfachung. Generell sind die Original-Artikel über dem empfohlenen Wert für die Allgemeine Bevölkerung. Die Artikel in Plain-Englisch entsprechen dieser Empfehlung, allerdings variiert die Schwierigkeit hier erheblich. Die Plain-Englisch Version von Artikel 1960 befindet sich im Wertebereich der anderen originalen Artikel. Der Original-Artikel hat sogar einen noch höheren Wert. Ich überprüfte den Artikel manuell und stellte fest, dass in diesem Artikel viele längere Fachbegriffe benutzt werden, obwohl der Artikel sehr kurz ist. Die Sätze sind eher lang, da auch Aufzählungen vorkommen.

Beispielsatz aus Artikel 1960 in Plain-Englisch: *'The Obama administration is trying to improve the system by improving conditions in centers located in the south and in areas with large populations of undocumented immigrants such as Chicago'*

3 Das Korpus

Für diese Arbeit erstellte ich ein Korpus aus Zeitungsartikeln in einfachem Englisch (Plain-English) und normalem Englisch. Diese Daten wurden von den jeweiligen Webseiten heruntergeladen, vorverarbeitet und als XML-Dateien abgespeichert. Dazu erstellte ich ein Korpusprofil und verglich die beiden Sprachen miteinander.

3.1 'The Times in Plain English'

Die Artikel in Plain-Englisch stammen von der Webseite 'The Times in Plain Englisch'. Der Editor ist Arthur Schiff.

Arthur Schiff hat in den 1990er Jahre das Magazin 'City Family' gegründet. Dies ist eine Zeitschrift für Immigranten, die frisch in die USA gekommen sind. Die Idee war, dass diese Menschen andere Themen interessieren als die Bürger mit hohen Einkommen. Schiffs Idee war, dass Immigranten ein Magazin zur Verfügung haben, in dem sie neben den klassischen Themen von Zeitschriften wie Artikel über Prominente und Produktvorstellung, die ihren Bedürfnissen und Einkommen angepasst werden, auch praktische Tipps fürs Leben finden. Die Zeitschrift wurde in einem sehr einfachen Englisch geschrieben, das auch gerade erst in Amerika angekommene Immigranten verstehen sollten. Das Magazin wurde auch in Schulen gerne zur Einführung in Medien genutzt. Nebst der englischen gab es auch eine spanische Version. Auch diese wurde gerne in Schulen als Unterrichtsmaterial genutzt. Das Magazin wurde kostenlos an Orten aufgelegt, wo die Zielgruppe sich aufhielt. (Pogrebin [1996])
Heute scheint das Magazin nicht mehr zu existieren.

'The Times in Plain Englisch' veröffentlichte ihre ersten Artikel im Jahr 2011. Auch hier ist das Ziel, Nachrichten in einem einfachen Englisch zur Verfügung zu stellen. Die Artikel sind kostenlos und nur online erhältlich. Die Artikel in Plain-Englisch beruhen auf Original-Artikeln aus verschiedenen US-amerikanischen Zeitschriften wie zum Beispiel der 'New York Times', der 'Washington Post' und Blogs. Diese gelten als inhaltliche Vorlage, die Artikel in einfachem Englisch werden aber neu geschrieben. Schiff und seine Editoren halten sich dabei an den Flesch-Kincaid-

Grade-Level für einfaches Englisch, da dieser standardmässig in Microsoft Word verfügbar ist. (Schiff [2015])

Auch das Layout wird übersichtlicher gestaltet und es wird viel mit Bildern gearbeitet. Die Artikel stellen vor allem Fakten dar, auf einzelne Geschichten und Erlebnisberichte, wie sie bei der New York Times üblich sind, wird meist ganz verzichtet. Des Weiteren fällt auf, dass die Artikel meist sachlicher gehalten werden als ihre originalen Gegenstücke. Da sehr wenige Fakten in wenigen Worten wiedergegeben werden, werden meist nicht alle Ansichten erläutert, die in den Original-Artikeln besprochen werden. Das kann zu Einseitigkeit in der Berichterstattung führen. Bei neueren Artikeln in einfachem Englisch wird der originale Artikel als Quelle verlinkt. Die Artikel erscheinen zu aktuellen Themen einzeln, meist etwa zwei Mal pro Woche. Die Artikel können auf der Webseite einfach als PDF ausgedruckt werden. Dort stehen Übersetzungen in verschiedenen Sprachen zur Verfügung, die mithilfe von 'gtranslate' von Google erstellt werden.

Nebst den Zeitungsartikeln erscheinen auch regelmässig 'Tipps for Teachers'. Dies sind Fragen zum aktuellen Geschehen inklusive Antworten. Diese können von Lehrern benutzt werden, um ihre Klassen auf geeignete Weise an Medien heranzuführen.

Ein Auszug aus einem Artikel in Plain-Englisch ist im Anhang (B) zu finden.

3.2 Original-Artikel

Die originalen Artikel, auf denen die Artikel in einfachem Englisch beruhen, stammen aus den verschiedensten Quellen. Die meisten Artikel stammen von 'The New York Times'. Es werden aber auch Artikel von anderen amerikanischen Zeitungen verwendet, unter anderem der 'Washington Post', des 'Wallstreet Journals' und der 'Los Angeles Times'. Daneben werden auch einzelne Artikel von Blogs und kleineren Zeitschriften verwendet. Dementsprechend haben die Original-Artikel kein einheitliches Layout, verschiedene Autoren und verschiedene Themen. Die 'The New York Times' ist nicht frei verfügbar und der Zugriff auf Artikel ist für Nicht-Abonnenten pro Tag und Monat beschränkt.

Die Artikel sind meist Erlebnisberichte von Betroffenen und versuchen, verschiedene Sichtweisen aufzuzeigen. Das führt dazu, dass die Artikel sehr lang sind. Daneben wird vorausgesetzt, dass der Leser schon etwas über das Thema weiss und viele Informationen werden nur implizit mitgeteilt. Einige Artikel spielen bei gewissen Themen mit Emotionen, was sehr subtil in der Sprache eingebaut ist. Neben dem

eigentlichen Thema sind meist noch Hintergründe beschrieben, die zur Erklärung der Lage dienen.

Mit einem VPN-Zugang der Universität Zürich kann über das interne Netz auf das Archiv von 'The New York Times' zugegriffen werden. Dort lassen sich pro Tag 200 Artikel herunterladen. Dies muss allerdings von Hand gemacht werden: Der Titel wird gesucht und der Text von Hand kopiert. Auf 'Factiva' kann ebenfalls nach dem Titel gesucht werden. Hier können Artikel im PDF- oder RTF-Format gespeichert werden. Es gibt nach Auskunft der Zentralbibliothek keine Möglichkeit, mehrere Artikel gleichzeitig nach Titel zu suchen und herunterzuladen.

Ein Auszug aus einem Artikel ist im Anhang (B) zu finden.

3.3 Datengewinnung

Die Artikel in Plain-Englisch habe ich mithilfe eines Crawlers aus der Python-library Scrapy heruntergeladen. Dafür erstellte ich eine Liste aller zu einem Zeitpunkt verfügbaren Artikel mithilfe des Archivs der 'The Times in Plain English' und benutzte diese als Ausgangspunkt. Da alle Artikel eine einheitliche HTML-Struktur hatten, konnte ich die Artikel mit wenig Aufwand herunterladen. Nebst dem eigentlichen Artikel speicherte ich zu jedem Artikel den Titel, den Link des Artikels, das Veröffentlichungsdatum, die weiterführenden Links im Text und an welchem Datum ich den Artikel heruntergeladen habe. Zusätzlich wies ich jedem Artikel eine vierstellige ID zu, um den einfachen Text und den Original-Text später einander zuordnen zu können. Layout-Informationen, die nicht aus Zeilenumbrüchen bestehen, und Bilder habe ich nicht gespeichert. Sämtliche Artikel werden von Scrapy automatisch in eine Json-Datei geschrieben. Zum Zeitpunkt des letzten Scrapings kamen so 2103 Artikel zusammen.

Aus den mitgespeicherten HTML-Dateien der Artikel wurde der letzte vorkommende Link des Artikels als Quelle des Artikels gespeichert, wenn es keine Email-Adresse war. Einige frühere Artikel und die 'Tipps for Teachers' hatten keine Quelle. So kamen 1488 Links zustande. Die Darstellung der Quellen in den Artikeln hat sich über die Jahre einige Male verändert. Deshalb führt meine Methode der Linkextrahierungen zu einigen Links, die fälschlicherweise als Quelle erkannt werden. Dieses Problem konnte ich umgehen, da bei der Weiterverarbeitung diese Links herausgefiltert werden und ich nur ausgewählte Zeitschriften als Quellen für die Original-Artikel verwende.

Mit Hilfe dieser Liste versuchte ich, die originalen Artikel herunterzuladen. Da

sich die HTML-Struktur je nach Zeitschrift und Artikel unterschied und eine Beschränkung des Zugriffes bestand, gelang mir das nicht. Mit Hilfe von Stefan Vrankovic gelang es, einige Artikel von den drei häufigsten Quellen in meiner Liste herunterzuladen: 'The New York Times', 'The Washington Post' und 'Wall Street Journal'. Viele Original-Artikel waren nicht mehr zugänglich; der Link war ungültig. Die verschiedenen Webseiten der Zeitungen sind in ihren HTML-Strukturen verschieden. Zusätzlich hat 'The New York Times' auch keine konsistente HTML-Struktur für ihre Artikel, jede ist etwas anders. Dem Crawler muss für jede Webseite mitgegeben werden, wo und welche Teile in der HTML-Struktur gespeichert werden sollen. Deshalb war es uns nicht möglich, die Original-Artikel in einer ähnlich guten Qualität wie die Artikel in Plain-Englisch herunterzuladen. Es fehlen teilweise Wörter oder auch ganze Sätze. Ausserdem konnte ich Layout-Informationen nicht wie beim einfachen Englisch speichern. Ich versuchte, dieselben Metadaten zu erheben. Nicht bei allen Artikeln sind die Metadaten komplett und die weiterführenden Links fehlen ganz; grösstenteils sind die Metadaten aber vorhanden. Insgesamt kamen 761 Artikel zusammen.

Der benutzte Code, erstellt von Stefan Vrankovich, ist zu finden unter: Scrape URLs. Dem Programm wird eine Text-Datei mit den herunterzuladenden Links mitgegeben. Die Ausgabe wird mit den Metadaten in einer Json-Datei gespeichert.

3.4 Vorverarbeitung

Die Daten extrahierte ich aus den Json-Dateien. bereinigte sie und speicherte sie einzeln in XML-Dateien ab. Die Bereinigung bestand darin, dass übriggebliebene Reste der HTML-Struktur gelöscht wurden. Dazu wurden ungültige Unicode-Zeichen, doppelte Leerschläge und Zeilenumbrüche entfernt. Die Sätze wurden mit dem PunktSentenceTokenizer von nltk voneinander getrennt. Durch diesen Schritt wurden auch die Antworten von den Fragen in den 'Tipps for Teachers'-Artikeln getrennt. Sätze mit direkter Rede sollten so richtig tokenisiert werden. Das funktionierte aber nicht in allen Fällen. Der Tokenisierer kennt nicht alle Abkürzungen, die vorkommen, was eine weitere Fehlerquelle darstellt. Danach wurden die Artikel einzeln in je eine XML-Datei geschrieben. Dabei wurde das Layout des einfachen Englisch beibehalten und in einem eigenen XML-Tag dargestellt. Die Sätze wurden einzeln in einen XML-Tag gespeichert. Da die Original-Artikel keine Absätze enthielten, wurden durch die Implementation die Artikel als Ganzes in einem Abschnitt gespeichert. Die Metadaten wurden ebenfalls in diesen Dateien gespeichert. Die zusammengehörenden Artikel können über die ID im Dateinamen ermittelt werden.

3.5 Schwierigkeiten

Bei der Gewinnung und Vorverarbeitung der Daten sind einige Schwierigkeiten aufgetaucht. Die Original-Artikel waren nicht frei zugänglich und konnten nur mit Aufwand heruntergeladen werden. Die gespeicherten Original-Artikel sind verrauscht, das heisst es fehlen Teile des Textes. Dies führt zu Fehlern beim Tokenisieren. Dazu macht der Tokenisierer selbst Fehler. So addierten sich im Verlauf der Verarbeitung die Fehler und verstärkten sich.

3.6 Korpusprofil der gesamten Daten

Nach der Vorverarbeitung besteht das Korpus aus 761 Artikeln in Original-Englisch und 2103 Artikeln in einfachem Englisch. Das Korpus besteht aus fast 90'000 Sätzen, wobei etwa zwei Drittel in einfachem Englisch vorliegen. Im Korpus wurden einerseits Wörter gezählt, das heisst Texteinheiten getrennt durch Leerschläge. Andererseits wurden die Tokens gezählt, die mithilfe des Wort-Tokenisierers 'word_tokenize' von nltk erzeugt wurden. Es wurden die Anzahl Wörter und Tokens, deren Typen (alle voneinander unterscheidbare Formen, die im Text vorkommen), die durchschnittliche Satzlänge und die durchschnittliche Dokumentenlänge berechnet. Kommazahlen wurden auf die nächste ganze Zahl gerundet (Siehe Tabelle 3).

Was	Original-Englisch	Plain-Englisch
Anz. Dokumente	761	2103
Anz. Sätze	36652	65078
Durchschn. Sätze/Artikel	48	31
Anz. Wörter	879126	680007
Anz. Wörter Typen	70112	46618
Durchschn. Wörter/Satz	24	10
Durchschn. Wörter/Artikel	1155	323
Anz. Token	1019519	782900
Anz. Token Typen	42499	26999
Durchschn. Token/Satz	28	12
Durchschn. Token/Artikel	1340	372

Tabelle 3: Korpusprofil Gesamt

Es gibt etwa halb so viele Artikel in Original-Englisch als in Plain-Englisch. Die Sätze in Original-Englisch sind etwa doppelt so lang wie die in Plain-Englisch, es sind aber etwa halb so viele vorhanden. Im Durchschnitt enthält ein Original-Artikel

mehr Sätze als ein Artikel in Plain-Englisch. Dadurch enthalten beide Teile schlussendlich etwa gleich viele Wörter.
Es gibt insgesamt mehr Tokens als Wörter, da Satzzeichen häufig vorkommen und beim Tokenisieren abgetrennt und einzeln gezählt werden. Allerdings gibt es mehr Typen von Wörtern als von Tokens, da die Einheiten, welche ein Satzzeichen enthalten, als eigener Typ gezählt werden. Das ist bei den Token nicht der Fall. Trotz der Verrauschung in den originalen Artikeln zeigt sich wie erwartet, dass Artikel in Plain-Englisch weniger Sätze enthalten und diese kürzer sind. Wider Erwarten stellte die Verrauschung beim Erstellen des Profils kein Problem dar.

3.6.1 POS-Tagging

Gemäss Theorie in Kapitel 2 erwarte ich, dass die Artikel in Plain-Englisch gewisse Wortarten weniger enthalten. Sie sollten weniger Adjektive enthalten, da diese den Satz länger machen. Ich erwarte weniger Wörter, die einen Nebensatz einleiten und auch weniger Pronomen, da die Personen nicht referenziert werden sondern neu genannt werden sollten. Ausserdem erwarte ich in Plain-Englisch mehr Nomen. Genitive und komplexe Verbformen sollten selten vorkommen. Im Englischen ist es jedoch nicht einfach, nur anhand der Verbform zu erkennen, ob ein Verb einfach oder komplex ist, da sie in vielen Fällen die gleichen Formen annehmen.
Des Weiteren erwarte ich in den Artikeln in Plain-Englisch mehr Eigennamen. Diese sollte in den Artikeln in Plain-Englisch mehrfach wiederholt werden, während sie in den Original-Artikeln eher durch Personalpronomen referenziert werden. Die vollständige Zählung aller Wortarten ist im Anhang zu finden (Tabelle 8). Hier bespreche ich lediglich ausgewählte Wortarten, die für die Vereinfachung eines Textes relevant sind.

Um diese Theorien zu überprüfen bestimmte ich die Wortarten mit der pos_ tag-Funktion von nltk. Die benutzten POS-Tags stammen aus dem Penn Treebank Tag-Set. Da dies allerdings 36 Wortarten zur Verfügung hat, und sich so die Prozentzahl der meisten Wortarten unter einem Prozent befindet, entschied ich mich, diese in Kategorien zusammenzufassen. Dazu verwendete ich die gängige Aufteilung in neun Wortarten.

Dabei enthalten:

- Satzzeichen die Tags #, $, ’, (,), ,, ., :, LS und SYM
- Nomen die Tags NN, NNP, NNPS und NNS
- Adjektive die Tags JJ, JJR und JJS

- Verben die Tags MD, VB, VBD, VBG, VBN, VBP und VBZ
- Adverbien die Tags RB, RBS, RBR und WRB
- Pronomen die Tags CD, EX, PRP, PRP$, WP und WP$
- Partikel die Tags CC, IN, RP und TO
- Interjektionen den Tag UH
- Artikel, Determinative die Tags DT, PDT und WDT
- so gehören die Tags FW und POS zu keiner Gruppe.

Das Tag FW (Fremdwörter) kann keiner der oben genannten Kategorien zugeteilt werden, da Fremdwörter aller Wortarten unter diesem Tag zusammengefasst werden. Der Tag POS (Genitiv-Endung) konnte ebenfalls keiner Kategorie zugeordnet werden. Die Anzahl POS-Tags einzelner Kategorien sind in Tabelle 4 zu finden. Die prozentualen Anteile sind zusätzlich in Abbildung 2 und Abbildung 3 dargestellt.

Wortart	Plain-Englisch	In Prozent	Original	In Prozent
Satzzeichen	85139	10.9	98266	9.6
Nomen	232480	29.7	304514	29.9
Adjektive	55806	7.1	78118	7.7
Verben	140169	17.9	167780	16.5
Adverb	28282	3.6	40705	4.0
Pronomen	53903	6.9	63653	6.2
Partikel	110740	14.2	171906	16.9
Interjektion	68	0.009	83	0.008
Artikel, Determinative	76137	9.7	93748	9.2

Tabelle 4: Wortarten

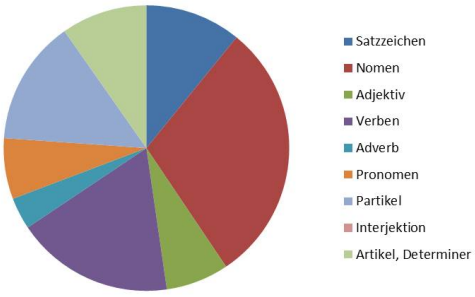

Abbildung 2: Plain-Englisch Wortarten

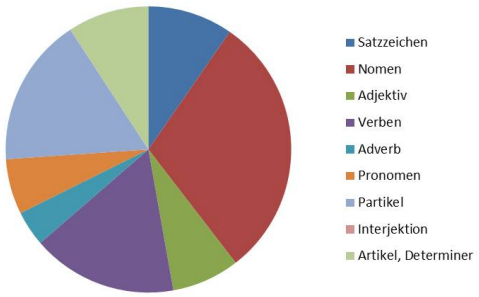

Abbildung 3: Original-Englisch Wortarten

Die Resultate in Abbildung 2 und Abbildung 3 zeigen nicht die erwarteten Resultate. Die beiden Teile des Korpus haben fast die genau gleiche Zusammensetzung aus Wortarten. Der einzig erkennbare Unterschied ist, dass die Artikel in Original-Englisch weniger Adverbien enthalten.

Um die anderen Erwartungen zu überprüfen befinden sich in Tabelle 5 die einzelnen Zahlen einiger ausgewählter Wortarten.

POS-Tag	Plain-Englisch	In Prozent	Original-Englisch	In Prozent
FW	144	0.02	438	0.04
POS	23	0.003	193	0.02
CC	17180	2.2	29439	2.9
IN	71811	9.2	114346	11.2
PRP	25974	3.3	29209	2.9
PRP$	7281	0.9	11978	1.2

Tabelle 5: Auswahl POS-Tags

Wie erwartet kommen in den einfachen Texten weniger Fremdwörter (FW) und Genitiv-Endungen (POS) vor. Bei den koordinierenden Konjunktionen (CC) und den subordinierenden Konjunktionen (IN) ist der Unterschied allerdings weniger gross als angenommen. Eigennamen (PRP und PRP$) kommen in den Originalartikeln sehr viel mehr vor. Das widerspricht den Erwartungen.

Generell ist der Unterschied in den Wortarten nicht so gross wie erwartet, was mich erstaunt hat. Es stellt sich die Frage, ob die Zählung der Wortarten der richtige Weg ist, um die Komplexität eines Tests zu zeigen. Hier sind weitere Untersuchungen nötig. Zusätzlich könnten weitere POS-Tagger getestet werden, um herauszufinden, welcher für diese Daten am besten geeignet ist.

3.7 Korpusprofil der parallelen Dateien

Da die Datenmenge der beiden Teile des Korpus sehr unterschiedlich sind und so ein direkter Vergleich der Daten schwierig ist, habe ich zusätzlich zum Korpusprofil der gesamten Datenmenge ein Korpusprofil erstellt, das nur die zusammengehörenden, parallelen Daten enthält (Tabelle 6).

Was	Original-Englisch	Plain-Englisch
Anz. Dokumente	761	761
Anz. Sätze	36652	26456
Durchschn. Sätze/Artikel	48	35
Anz. Wörter	879126	273885
Anz. Wörter Typen	70112	27146
Durchschn. Wörter/Satz	24	10
Durchschn. Wörter/Artikel	1155	360
Anz. Token	1019519	313667
Anz. Token Typen	42499	17808
Durchschn. Token/Satz	28	12
Durchschn. Token/Artikel	1340	412

Tabelle 6: Korpusprofil Parallel

Die Werte der Original-Artikel sind dieselben wie in Tabelle 3. Jeder Artikel in Original-Englisch hat eine Entsprechung in Plain-Englisch, das ist der Art der Datengewinnung geschuldet. Von den Artikeln in Plain-Englisch wird nur noch etwa die Hälfte gezählt. Dementsprechend sind die Zahlen kleiner.

Die Artikel in Plain-Englisch beinhalten im Gesamten wie auch in den Durchschnitten pro Artikel weniger Sätze, die jeweils weniger Wörter enthalten. Das war schon in Tabelle 3 ersichtlich. Die Durchschnittswerte sind sehr ähnlich wie in Tabelle 3. Die durchschnittlichen Sätze pro Artikel und Worte pro Artikel sind etwas höher im direkten Vergleich. Dennoch lässt sich sagen, dass die Werte über das Gesamtkorpus die Werte der parallelen Artikel gut wiederspiegeln.

3.7.1 POS-Tagging

In den vorhergehenden Abbildungen 2 und 3 des gesamten Korpus konnte bei den Wortarten kein Unterschied zwischen Original-Artikeln und Artikeln in Plain-Englisch festgestellt werden. Im Folgenden stelle ich mir die Frage, ob das anders ist, wenn nur die parallelen Artikel gezählt werden.

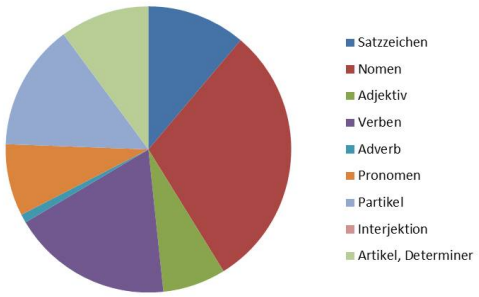

Abbildung 4: Plain-Englisch Wortarten der parallelen Artikel

Abbildung 4 ist sehr ähnlich zu der Darstellung der gesamten Daten in Plain-Englisch. Die parallelen Artikel beinhalten prozentual weniger Adverbien als die Gesamtheit der Artikel. So haben die originalen Artikel erkennbar mehr Adverbien als die Artikel in Plain-Englisch.

3.8 Erstellung des Goldstandards

Um später die Alignierungen evaluieren zu können, erstellte ich einen Goldstandard. Schwenk [2018] und Artetxe and Schwenk [2019] trainierten mit den alignierten Daten ein System und verglichen dieses mit einem, das auf anderen Daten trainiert wurde. Ein solches Training hätte den zeitlichen Rahmen dieser Arbeit überschritten. Später stellte sich heraus, dass Laser zu wenige Daten aligniert hat, um ein System trainieren zu können. Deshalb evaluierte ich die Ergebnisse anhand eines Goldstandards.

Für den Goldstandard wählte ich zufällig 7 Artikel aus, bei denen beide Versionen, also im Original und in Plain-Englisch, vorhanden sind. Das ergab ca. 200 Sätze.

Von Hand sichtete ich die beiden Artikel und suchte Paare, die zusammenpassten. Dafür ging ich vom Original-Artikel aus und suchte Entsprechungen im Artikel in Plain-Englisch. Dabei liess ich auf beiden Seiten Sätze aus, die keine Entsprechungen hatten. So kamen schlussendlich etwa hundert alignierte Paare zusammen.

Während der Evaluation fielen mir verschiedene Dinge an den Daten auf:

- In den Original-Daten fehlen mitten in den Sätzen Wörter; das ist der Datengewinnung geschuldet. Dies waren meist Eigennamen oder Nomen. So fehlt ein sehr wichtiger Teil des Satzes, was das Alignieren schwierig macht.

Beispiel Artikel 0442 Original: *Mr. Trump, who has previously teased his (...), entered the fray as he was scheduled to sign a less ambitious proposal, one that would establish a framework for directing commercial traffic in space and monitoring debris.* Das fehlende Wort, bzw. die fehlende Phrase ist *desire to create a space force.* Dieser ist ein inhaltlich wichtiger Teil des Satzes.

- Trotz Anpassung des Satz-Tokenisierers wurden Sätze falsch tokenisiert. Dies betrifft vor allem Sätze, die mit einer direkten Rede, also einem Anführungszeichen endeten. Der Tokenisierer müsste diese Strukturen allerdings richtig verarbeiten können.
  Beispiel Artikel 0050 Original: *"It's like I stole it," he said, "and I really don't like what I stole." The story of poultry work tracks closely with the 20th-century story of race relations in Mississippi.* Dieser Abschnitt wird als ein Satz erkannt und nicht richtig segmentiert.

- Die beiden Artikel behandelten die gleichen Themen nicht in der gleichen Reihenfolge. Das erschwert eine lokale Alignierung.

- Viele Sätze waren auch von Hand schwierig zu alignieren. Da gab es Sätze, die verschiedenste Themen auf einmal ansprechen. Korrekterweise müssten sie mit 4 Sätzen in Plain-Englisch aligniert werden. Andere Paare haben grundlegend die gleiche Bedeutung, sind aber semantisch so unterschiedlich, dass ich das Satzpaar nicht alignieren kann. Der folgende Beispielssatz ist sehr lang. Zusätzlich wurde der folgende Satz nicht abgetrennt, was den Satz, der aligniert wird zusätzlich verlängert. So wird eine korrekte Alignierung unmöglich: Artikel 0050 Original: *In Canton, African-Americans called for a boycott of the local chicken plant over its refusal to hire black workers, according to Angela Stuesse, an associate professor of anthropology at the University of North Carolina and author of the 2016 book "Scratching out a Living: Latinos, Race, and Work in the Deep South." By the end of the 1960s, black workers predominated on the lines.*

- Gesamthaft gesehen können Sätze in Plain-Englisch häufiger aligniert werden als Original-Sätze. So wurden in den meisten Artikeln in Plain-Englisch fast alle Sätze aligniert, in Original-Artikeln nicht. Trotzdem konnten nur etwa die Hälfte aller Sätze aligniert werden.

4 Alignierung der Sätze

Nachdem ich die Daten heruntergeladen und vorverarbeitet habe, versuchte ich die Daten zu alignieren. Ich wollte herausfinden, ob Laser (Language Agnostic Sentence Representation) auch für die Alignierung von vereinfachten Daten benützt werden kann. Die Laser-Alignierungen verglich ich dann mit Alignierungen von Cats und den Resultaten von Marek Kostrzewa von Transformer und LHA.

4.1 Laser

Für das Training von automatischen Übersetzungssystemen ist die Qualität und Quantität der Daten, auf denen trainiert wird, entscheidend. So gibt es mittlerweile grosse Korpora für bestimmte Sprachpaare, die eine gute Qualität haben. Andere Korpora wurden automatisch erstellt und enthalten eine grosse Menge an Satzpaaren, diese können aber verrauscht sein. Es wurde gezeigt, dass neuronale Übersetzungssysteme mit verrauschten Daten (die Schreibfehler und nicht korrekte Paare beinhalten), sehr schlecht umgehen können (Belinkov and Bisk [2018]). Aus diesem Grund werden Methoden gesucht, die verrauschte Daten aus bestehenden Korpora herausfiltern können und selbst korrekte Daten erstellen können.

Um das zu erreichen verfolgten verschiedene Forschungsteams unterschiedliche Ansätze. So versuchten z.B. Axelrod et al. [2011] aus einem Datensatz, der Daten mit verschiedenen Domänen enthält, einzelne herauszufiltern. So werden Satzpaare, die aus unterschiedlichen Domänen kommen, und somit wahrscheinlich falsch aligniert sind, herausgefiltert. Allerdings werden falsche Satzpaare innerhalb der Domänen, die zwar ähnlich, aber keine korrekte Übersetzungen voneinander sind, nicht herausgefiltert.

Es gab auch verschiedene Ansätze, um Satzpaare mit Hilfe von Satz-Embeddings zu finden. So haben zum Beispiel Grégoire and Langlais [2017] für zwei Sprachen je ein Satz-Embedding-Encoder trainiert. Ein Klassifizierer entschied dann, ob zwei Repräsentationen aus den verschiedenen Embedding-Räumen übereinstimmen. Dieses System konnte jedoch nur für ein Sprachpaar benutzt werden.

Laser beruht auf der Arbeit von Schwenk and Douze [2017]. Das Ziel dieser Arbeit war, Sätze aus sechs Sprachen des 'UN-Dev Korpus' so in einem multidimensionalen Raum abzubilden, dass Sätze, die, unabhängig von der Sprache, ähnliche semantische Inhalte hatten, am gleichen Ort abgebildet werden. So wären Sätze, die sich semantisch ähnlich sind, näher beieinander als Sätze, die keine inhaltlichen Ähnlichkeiten haben. Um dies zu erreichen, trainierten Schwenk et al. ein Übersetzungssystem mit einem Enkodierer und einem Dekodierer für jede Sprache. Das System kann mit verschiedenen Kombinationen von Quell- und Zielsprachen trainiert werden. So kann das System mit Satzpaaren in einer Ziel- und einer Quellsprache trainiert werden. Jeder Durchgang des Trainings passiert mit verschiedenen Sprach-Kombinationen. Hat man parallele Daten für mehr als zwei Sprachen, kann das System auch mit mehreren Quell-Sprachen, die in eine Zielsprache übersetzt werden, trainiert werden. Der Nachteil hierbei ist, dass die Zielsprache nicht als Quellsprache verwendet werden kann. Umgekehrt kann auch ein Training stattfinden, bei dem ein Quellsatz in alle möglichen Zielsprachen übersetzt wird. Diese Strategie erbrachte die besten Ergebnisse. Zusätzlich können bei dieser Methode für jede Sprache Satz-Embeddings generiert werden. Die Embeddings hatten eine fixe Grösse, das kann für lange Sätze (mehr als 50 Tokens) ein Problem werden. Sätze dieser Länge kommen aber eher selten vor. Nach dem Training werden die Dekodierer verworfen und nur die Enkodierer für jede Sprache behalten. Das System sollte nun imstande sein, die semantische Bedeutung von Sätzen sprachunabhängig zu repräsentieren.

Schwenk [2018] benutzte die oben beschriebene Methode, um verrauschte Daten aus einem parallelen Korpus zu filtern und um neue Satzpaare zu alignieren. Laser wurde mit Daten aus dem 'Europarl'-Korpus in neun Sprachen trainiert. Anders als in Schwenk and Douze [2017] wurde für alle Sprachen derselbe Enkodierer trainiert. Dieser bekommt keine Information über die Sprache des Quell-Satzes. Dazu wurden ein 20'000 Wörter grosses BPE-Vokabular (Byte-Pair-Encoding) über alle Sprachen benutzt. Die Sätze wurden durch 1024-dimensionale Vektoren repräsentiert. Die Architektur des Enkodierers ist dieselbe wie in Schwenk and Douze [2017] beschrieben; sie erzielte die besten Resultate. Anders als in Schwenk and Douze [2017] konnten bei Laser bessere Resultate mit einem tieferen neuronalen Netzwerk erzielt werden. Auch hier wurden nach dem Training die Dekodierer verworfen. Laser sucht die nächsten Nachbarn eines Satzes mithilfe der Kosinus-Distanz. Nur Paare, die eine Ähnlichkeit über dem Schwellenwert habe, werden behalten. Dies kann verwendet werden, um bestehende Alignierungen zu Filtern. Sollen neue Alignierungen generiert werden, wählt Laser den Satz aus, der dem Quellsatz am nächsten ist. Die detaillierte Beschreibung der Vorgehensweise findet sich in Schwenk and Douze [2017].

Die Autoren testeten Laser. Dazu wurden die Alignierungen des 'Common Crawl-Korpus' mit Laser gefiltert und zwei Übersetzungssysteme mit den Daten trainiert: Eines mit den gesamten, ungefilterten Common Crawl Daten, eines mit den von Laser gefilterten. Die gefilterten Daten beinhalteten nur noch etwa die Hälfte der Paare, trotzdem erzielte das mit diesen Daten trainierte Übersetzungssystem bessere Resultate. Das System, das nur auf von Laser generierten Daten trainiert wurde, erreichte schlechte Resultate. Allerdings können diese Daten in Kombination mit vorhandenen Daten Resultate verbessern.
Laser annotiert Sätze, die Aufzählungen von Eigennamen haben, mit anderen Aufzählungen von Eigennamen. Allerdings stimmen die Eigennamen nicht überein. So entschieden die Autoren, nebst Sätzen, die länger als 50 Tokens sind, auch Sätze zu ignorieren, die mehr als drei Kommas beinhalten. Aus einem zufällig gemischten 'Europarl-Korpus' konnten 95% der ursprünglichen Paare korrekt wiederhergestellt werden. (Schwenk [2018])

Artetxe and Schwenk [2019] stellen fest, dass der fixe Schwellenwert zu Problemen führte. Da die Kosinus-Ähnlichkeit nicht global konsistent ist, können Paare mit keiner korrekten Übersetzung einen höheren Wert erreichen als Paare, die korrekte Übersetzungen beinhalten. Die Autoren schlagen vor, statt dem fixen Schwellenwert ein Schwellenwert zu nehmen, der relativ zu den nächsten Nachbarn eines Satzes ist. Die Methode mit den besten Resultaten ist dabei die, die das Verhältnis zwischen dem Kandidaten und der durchschnittlichen Kosinus-Distanz der nächsten Nachbarn verwendet. Dieser Ansatz verbesserte die Resultate von Schwenk [2018]. Der Code kann unter Laser-contra gefunden werden.

4.1.1 Benutzte Skripts

Für die Alignierung der Daten benutzte ich Skripts von Mathias Müller (Repository momentan noch privat: Laser). Diese Skripts sind für Texte in fünf Sprachen geschrieben. Ich nahm leichte Anpassungen vor, behielt aber die empfohlenen Parameter von Laser bei.

Die Skripte von Mathias Müller richten Laser ein und stellen sicher, dass die richtigen Versionen von Werkzeugen benutzt werden.

Die Daten werden gesamthaft in einer Text-Datei, ein Satz pro Linie, gespeichert und an Laser weitergegeben. Diese Dateien werden vorgefiltert: Die Sprach-Tags werden überprüft und Sätze unter einer Minimallänge (Standard: 3) und über einer Maximallänge (Standard: 100) werden herausgefiltert.

Danach werden von den vorgefilterten Daten Satz-Embeddings erstellt. Aus den Embeddings werden dann mit Hilfe des Schwellenwertes (Standard: 1.1) die Satzpaare aligniert.

Die Alignierung führte ich auf dem s3it-sciencecluster-server der Universität Zürich durch. Dazu wurde eine NVIDIA Tesla V100-GPU mit 48GB RAM verwendet. Die Alignierung dauerte nur ein paar Sekunden.

4.2 Cats

Um die Resultate vergleichen zu können alignierte ich meine Daten zusätzlich mit Cats (Customised Alignment of Text Simplification).

Schon vor Cats gab es verschiedene Systeme, die vereinfachte Texte alignieren konnten. Diese Systeme waren aber nicht robust. Sie waren entweder darauf angewiesen, dass die parallelen Sätze in beiden Dokumenten in der gleichen Reihenfolge vorkamen und/oder konnten einem Satz nur einen einzigen anderen zuordnen. Das ist ein Problem bei der Textvereinfachung, wo viele komplexe Sätze in mehrere einfache umgeschrieben werden.

Das von Stajner et al. [2017] entwickelte Cats ist spezifisch für die Alignierung von vereinfachten und originalen Sätzen implementiert. Es kann Dokumente alignieren, deren Reihenfolge nicht gleich ist und kann einem Satz mehrere Entsprechungen zuordnen. Cats ist sprachunabhängig, im Moment gibt es Implementationen für Englisch und Spanisch. Cats beinhaltet verschiedene Alignierungs-Strategien und Ähnlichkeitsmethoden. Ich beschreibe hier nur diese, die für die Alignierung meiner Daten benutzt wurden. Die benutzten Parameter entsprechen alle den von den Autoren empfohlenen.

Die verwendete Alignierungs-Methode ist 'Most Similar Text (MST)', da in den Zeitungsartikeln die Reihenfolge der Sätze nicht parallel ist.

Als Ähnlichkeitsmethode benützte ich C3G. Diese Methode benutzt das Charakter-N-gram Ähnlichkeits-Modell von Mcnamee and Mayfield [2004]. Dieses wurde mit TF-IDF (Term Frequency and Inverse Document Frequency) gewichtet und die Vektoren der potenziellen Paare mit Kosinus-Distanz verglichen. Diese Methode ist die einzige, die lexikalisch und nicht semantisch vorgeht und scheint über verschiedene Grade der Vereinfachung am robustesten zu sein.

Mit den alignierten Daten wurde ein automatisches Textvereinfachungs-System trainiert und auf dem 'Newsela-Korpus' getestet. Diese Resultate habe ich manuell eva-

luiert. Das System erzielte in der Grammatikalität und in der Erhaltung von Informationen bessere Leistungen als andere Systeme, die zu diesem Zeitpunkt erhältlich waren.

Der Code ist frei zugänglich unter Cats.

4.2.1 Benutzte Skripts

Ich benutzte oben genanntes Repository, das von den Autoren zur Verfügung gestellt wird mit den von ihnen empfohlenen Parametern. Die Artikel musste ich für Cats umbenennen; im Dateinamen findet Cats die Information, welcher Artikel mit welchem aligniert wird. So konnte ich die Alignierung nur dokumentweise machen, im Gegensatz zu Laser, wo sie artikelübergreifend ist. Ich hatte einige technische Schwierigkeiten, bis ich Cats anwenden konnte. Nachdem diese überwunden waren, erwies sich Cats als sehr benutzerfreundlich.

4.3 Transformers und LHA

Zusätzlich bot sich die Gelegenheit, meine Daten von Marek Kostrzewa mit zwei weiteren Systemen zu alignieren. Zum einen mit Transformer, das mit dem SBERT-Modell Satz-Embeddings generiert, die semantische Informationen enthalten (Reimers and Gurevych [2019]). Zum anderen mit LHA (Large-scale Hierarchical Alignment), das aufgrund verschiedener Embeddings-Varianten zuerst parallele Dokumente und dann pseudo-parallele Sätze aus verschiedenen Schreibstilen aligniert (Nikolov and Hahnloser [2019]). Das gibt weitere Einsichten in mögliche Fehlerquellen in meiner Alignierung mit Laser.
Für Transformer verwendete Marek Kostrzewa einen fixen Schwellenwert; nur Sätze, die eine Ähnlichkeit von über 0.8 haben wurden abgebildet.

4.4 Alignierung mit Laser

4.4.1 Erster Durchgang

Im ersten Durchgang alignierte ich nur die Artikel, die in Plain-Englisch und in Original-Englisch vorhanden waren. Die Alignierung erfolgte zwischen 35663 Originalsätzen und 23156 Sätzen in Plain-Englisch. Eine manuelle Sichtung zeigte, dass

die Resultate sehr schlecht waren. Die Satzpaare, die aligniert wurden, gehörten nicht zusammen.

Nachdem ich ausgeschlossen hatte, dass das schlechte Resultat an einem Anwenderfehler oder an einem technischen Problem lag, versuchte ich mit einigen Anpassungen an meinen Daten bessere Resultate zu erzielen.

4.4.2 Zweiter Durchgang

In einem zweiten Durchgang, bei dem ich schon etwas vertrauter mit Laser war, bestand der Input aus allen Daten, die mir zur Verfügung standen. Die Daten beinhalteten nun also auch die Artikel in Plain-Englisch, die keine originale Entsprechung hatten. Die Hoffnung war, dass durch die grössere Datenmenge auch mehr entsprechende Satzpaare gefunden werden können. Es wurden 36652 Originalsätze mit 65078 Sätzen in Plain-Englisch aligniert. Es gibt einen Unterschied zwischen der Anzahl originaler Sätze im ersten und im zweiten Durchgang. Dieser Unterschied entstand, da zwischen den beiden Durchgängen die Vorverarbeitung verbessert wurde. Auch hier zeigte eine manuelle Sichtung, dass die Resultate sich nicht verbessert hatten. Der Verdacht kam auf, dass der geringere Informationsgehalt in den Artikeln in Plain-Englisch das Problem sein könnte.

4.4.3 Dritter Durchgang

Um dem entgegenzuwirken nahm ich in einem dritten Durchgang zusätzlich zu den einzelnen Sätzen noch Satz-Bi- und Trigramme der Artikel in Plain-Englisch hinzu. Dazu bildete ich innerhalb eines Paragraphen alle möglichen Satzpaare. Die Informationen aus einem Satz in Original-Englisch sind nicht immer in zwei aufeinanderfolgenden Sätzen in Plain-Englisch enthalten. So alignierte ich 146608 Original-Sätze mit 292212 Sätzen in Plain-Englisch. Auch hier kommt die unterschiedliche Anzahl der Original-Sätze von weiteren Verbesserungen an der Vorverarbeitung.

4.4.4 Weitere Versuche

Um weitere Einblicke in die Vorgehensweise von Laser zu erhalten machte ich einige Tests mit Laser. Die Erkenntnisse daraus verwendete ich, um in Kapitel 5 die Ergebnisse von Laser besser einordnen zu können.

Ich alignierte die Datei mit den originalen Artikeln mit sich selber. Zuerst in der gleichen Reihenfolge, dann mit der Original-Datei, in der die Sätze zufällig gemischt wurden. Ziel davon war es zu sehen, ob Laser Sätzen, die identisch sind, dasselbe Embedding zuordnet und diese alignieren kann.

Ich alignierte einen einzelnen Artikel mit Laser. Denselben Artikel alignierte ich mit einem Schwellenwert von 0.5 statt 1.1. Die Hoffnung war, dass Laser bei dieser beschränkten Anzahl von Sätzen, die klar einem anderen zugewiesen werden können, besser abschneidet. Zusätzlich konnte ich so testen, ob die Resultate verbessert werden könnten, wenn der Schwellenwert tiefer ist.

5 Resultate

Die detaillierten Ergebnisse, die in diesem Kapitel besprochen werden, befinden sich im Anhang in den Tabellen 9 bis 12. Auszüge aus den verschiedenen Alignierungen finden sich im Anhang (B)

5.1 Vorgehen

Anfangs plante ich, die Resultate alle mithilfe des Goldstandards zu evaluieren. Allerdings alignierte Laser sehr wenige Satzpaare. Dadurch, dass von den insgesamt 438820 Sätzen nur 101 Satzpaare im Goldstandard sind und Laser nur 1537 Satzpaare aligniert hat, ist die Wahrscheinlichkeit, dass ein Satzpaar in beiden vorkommt, sehr gering. Eine Überprüfung bestätigte diesen Verdacht. Aus diesem Grund entschied ich mich, Laser anders zu bewerten. Ich suchte mit Pythons Random-Modul 100 Sätze zufällig aus und bewertete diese mit 'Ja' und 'Nein' für korrekte, bzw. inkorrekte Paare. Dafür wurde die Alignierung von Original-Englisch zu Plain-Englisch verwendet, da diese mit der Alignierung von Plain-Englisch zu Original-Englisch übereinstimmt. Die anderen Systeme bewertete ich mithilfe des Goldstandards. Deshalb ist der Vergleich von Laser und den anderen Systeme etwas erschwert.

Beim Vergleich mit dem Goldstandard zählte ich zuerst die Satzpaare, die in beiden Dateien vorkamen. Dabei zeigten sich zwei Probleme:

- Im Goldstandard fehlen einige korrekte Paare
- Im Goldstandard habe ich einen Satz in Original-Englisch mit mehreren Sätzen in Plain-Englisch aligniert und als ein Paar gezählt. Der Gedanke dahinter war, dass Laser durch die Bi- und Trigramme auch mehrere Sätze alignieren kann. Diese würden dann als ein Paar gezählt werden. Die anderen Systeme haben daraus aber zwei Paare gemacht.

Um die im Goldstandard fehlenden, korrekten Paare zählen zu können, zählte ich neben den Übereinstimmungen mit dem Goldstandard auch die korrekten Satzpaare, die nicht im Goldstandard vorhanden sind.

Ich entschied mich, für die Übereinstimmung die Paare im Goldstandard zu zählen. Pro Artikel werden dadurch zwei bis drei Paare, die korrekt sind, als eines gezählt.

Satzpaare zählte ich als richtig, wenn der Satz in Plain-Englisch eine Vereinfachung des Satzes in Original-Englisch war. Das war auch dann der Fall, wenn eine Information fehlte. War der Original-Satz allerdings sehr lang (beinhaltete mehr als drei Phrasen), und der Satz in Plain-Englisch kurz (nur eine Phrase), und wenn aus diesem langen Satz nur ein Paar gebildet wurde, zählte ich das Paar als nicht korrekt.

So alignierte Transformer in Artikel 0122 folgendes Satzpaar (Paar 5):
Original: *Less than two weeks ago, Mr. Modi unilaterally wiped out the statehood of India's only Muslim-majority state, Jammu and Kashmir, removing its special autonomy and turning it into a federal territory without any consultation with local leaders — many of whom have since been arrested.*
Plain-Englisch: *He put the people of Kashmir into total lockdown*
Die Vereinfachung ist inhaltlich richtig, allerdings fehlt sehr viel Information. Dieser Satz zählte ich als nicht korrekt.

Beim Vergleichen der Sätze fiel mir auf, dass derselbe Satz in verschiedenen Dokumenten nicht genau gleich gespeichert wurde. Ich nehme an, dass die Systeme verschiedene Vorverarbeitungen unternehmen. Aus diesem Grund liess sich die Evaluation nur von Auge, unterstützt von der Suchfunktion von Notepad++, machen. Ein System mit allen 8 Artikeln zu evaluieren dauerte auf diese Art ca. eineinhalb Stunden.

Transformer und LHA alignieren von Original-Englisch zu Plain-Englisch anders als von Plain-Englisch zu Original-Englisch. Da die Evaluation zeitaufwändig ist, entschied ich mich, nur die Alignierungen von Original-Englisch nach Plain-Englisch zu bewerten, da dies auch die Richtung ist, die in der automatischen Textvereinfachung benutzt wird. Ausserdem ist der Goldstandard auch so aufgebaut, dass zuerst der Originalsatz und danach der Plain-Englisch Satz folgt, was die Evaluation etwas erleichtert.
Ich nehme an, dass die Alignierung von Sätzen in Original-Englisch mit Sätzen in Plain-Englisch bessere Ergebnisse erzielt als umgekehrt, da Informationen weggelassen werden. In die andere Richtung muss von wenig Information auf viel Information geschlossen werden. Um diese These zu überprüfen müssten aber noch weitere Untersuchungen gemacht werden.

5.2 Laser auf Credit-Suisse-Daten

Laser erreichte mit denselben Parametern auf dem 'Credit Suisse Bankenmagazin'-Korpus, das in fünf Sprachen erhältlich ist, sehr gute Resultate. (Alignierungen von Mathias Müller, noch nicht publiziert) Ich sichtete manuell 100 zufällige Sätze aus der Alignierung zwischen der deutschen und englischen Ausgabe des CS-Bulletins. Von diesen 100 Sätzen stufte ich 100 als korrekt aligniert ein. Die Übersetzungen waren Wort für Wort gleich. Nur bei zwei Sätzen fehlten gewisse, sehr kleine Teile; Laser alignierte sie trotzdem richtig.

Sätze, denen Teile fehlten:
De: In dieser Funktion suchte ich nach einer einfachen Methode, die es ermöglichte, 100 verschiedene südamerikanische Holzarten mit einer Lupe voneinander zu unterscheiden.
En: In this position, I had to look for a simple method to differentiate between 100 different types of trees in South America.
In der englischen Übersetzung fehlt der Teil mit der Lupe.

De: Ein Teil der Diskrepanz lässt sich durch die höheren Einkommen und Lebenskosten erklären, aber als Hauptfaktoren gelten die allgemein höheren Kosten für Gesundheitsleistungen wie rezeptpflichtige Medikamente, Spitalaufenthalte und Arztbesuche.
En: They attribute part of the discrepancy to higher incomes and cost of living in the US, but the primary factors are the overall higher prices for health services, such as prescription drugs, hospital stays and doctor visits.
Im deutschen Satz wird die USA nicht erwähnt.

5.3 Laser

Laser ordnet die Alignierungen nach berechneter Ähnlichkeit; das Paar mit dem höchsten Wert steht zuoberst.

Im ersten Durchgang alignierte Laser 1609 Paare. Die Ähnlichkeit reicht von 1.1 (Schwellenwert) bis 1.85. Das Satzpaar mit der höchsten Ähnlichkeit ist:
Original: *It fell apart early in the George W. Bush administration.*
Plain-Englisch: *And as long as the U.S. thought China would keep North Korea in check, the U.S. would work with China.*
Von den hundert getesteten Satzpaaren ist keines korrekt.
Etwa ein Drittel der Satzpaare handeln von einem gemeinsamen Thema.

Zum Beispiel Original: *Some psychiatrists have said it is irresponsible to throw around medical terms without an examination.*
Plain-Englisch: *Aides to Lyndon B. Johnson were so troubled that they sought out three psychiatrists.* Ähnlichkeit: 1.345

Das Satzpaar, das am ehesten als korrekt betrachtet werden könnte, ist folgendes.
Original: *, Daegu Elementary, in Ms. Hwang's district, has.*
Plain-Englisch: *How about enrolling older villagers who wanted to learn to read and write? Ms. Hwang started attending classes last month.*
Die Ähnlichkeit ist bei 1.1. Es wird in beiden Sätzen von der gleichen Person gesprochen. Allerdings fehlt im ersten Satz die Mehrheit der Wörter. Dadurch kann es kaum als korrekt bewertet werden.

Der zweite Durchgang zeigte ähnliche Resultate und keinerlei Verbesserung. Eine genauere Bewertung konnte ich nicht machen. Die Alignierung ging durch ein technisches Problem verloren.

Im dritten Durchgang alignierte Laser 1537 Satzpaare. Die Ähnlichkeit reichte von 1.45 bis 1.1. Das Satzpaar mit der höchsten Ähnlichkeit ist:
Original: *"When money flows into an index or index-related E.T.F., the manager generally buys into the securities in an index in proportion to their current market capitalization (often to the capitalization of only their public float, which interestingly adds a layer of distortion, disfavouring companies with large insider, strategic, or state ownership),"he wrote*
Plain-Englisch: *His hope was that the problem will go away sooner or later. He wanted millions to get legal status.*
Von den 100 Paaren waren 0 korrekt.

Zwei Paare zeigten semantische Ähnlichkeiten, waren aber keine Entsprechungen. So wurde eine Frage im Original, *'To wash or not to wash?'* mit einem Satz in Plain-Englisch aligniert, in dem es um eine Frage geht: *The question of Trump's involvement in these cases is not yet known.*
Nur etwa 10 der 100 Artikel handelten von einem gemeinsamen Thema.

Die meisten Satzpaare scheinen aber ziemlich zufällig zu sein. So wurde zum Beispiel ein Satz über den Sekretär der amerikanischen Homeland Security, mit einem Satz aligniert, wie man Pilze zubereitet:
Orginal: *By that afternoon the first few approvals were issued, Mr. Mayorkas said, with several dozen more on Tuesday.*
Plain-Englisch: *Using a damp paper towel, wipe the mushrooms clean. Slice the remainingg mushroom caps and set aside.*

Bei beiden Durchgängen fällt auf, dass die Ähnlichkeit bei falsch alignierten Paaren sehr hoch ist. Es könnte sein, dass die Satzpaare, die ähnlicher sind, einen geringfügig tieferen Wert als die haben, die sich überhaupt nicht ähneln. Aber auch bei den tieferen Ähnlichkeiten sind viele Paare falsch aligniert.

5.4 Weitere Versuche mit Laser

Die Alignierung der Original-Daten mit sich selber ergab 35'557 Satzpaare (24.2% aller Sätze). Eine manuelle Sichtung zeigte, dass die Alignierung mehrheitlich korrekt erfolgte; nur sehr wenige Sätze wurden nicht mit sich selber aligniert. Die höchste Ähnlichkeit zwischen zwei Sätzen war 1.976.

Die Alignierung der Original-Daten mit den Original-Daten, die zufällig gemischt wurden, war die einzige Alignierung mit Laser, für die die beiden Richtungen der Alignierung unterschiedlich viele Satzpaare ergaben. So resultierte eine Richtung in 36202 Paaren (24.7%), die andere in 37500 Paaren (25.6%). Bei beiden fand ich bei einer manuellen Sichtung kein Paar, bei dem die Sätze unterschiedlich waren. Die höchste Ähnlichkeit betrug bei beiden Richtungen 1.976, der höchst gewertete Satz war bei beiden derselbe.

Die Alignierung des Artikels 1142 mit einem Schwellenwert von 1.1 ergab 3 Alignierungen. Keine dieser drei war korrekt.
Die Alignierung mit einem Schwellenwert von 0.5 ergab 25 Satzpaare, davon stimmte eines mit dem Goldstandard überein und 4 andere könnten als potenzielle Paare gesehen werden, sind aber nicht hundertprozentig korrekt.
So zum Beispiel im Original: *The authors found that voters in congressional districts hardest hit by Chinese imports tended to choose more ideologically extreme lawmakers.*
Plain-Englisch: *These same voters are supporters of Donald Trump.*
Der Satz in Plain-Englisch ist eine Vereinfachung, allerdings nehme ich an, dass es ein zufälliger Treffer war.

Die detaillierte Evaluation ist im Anhang in Tabelle 12 zu finden.

5.5 Cats

Für die Evaluation von Cats verglich ich die acht automatisch alignierten Artikel mit ihrem manuell alignierten Gegenstück aus dem Goldstandard.

Insgesamt alignierte Cats 245 Paare, im Goldstandard sind 101 Paare aligniert. Von diesen 101 wurden 64 korrekt aligniert, von den 245 alignierten Satzpaaren sind 112 korrekt.

Cats fand die meisten Paare, die im Goldstandard fehlten.

Bei den Alignierungen mit Laser war die falsche Satz-Tokenisierung ein Problem. Cats trennte Sätze voneinander und alignierte sie korrekt. Auch diese fehlten im Goldstandard.
Cats hat einige einzelne Worte aligniert. Diese wurden mit einem anderen einzelnen Wort aligniert, allerdings konnte keine semantische Verwandtschaft festgestellt werden. Die Alignierung funktioniert nicht auf Wortebene.

Lange Sätze aligniert Cats selten. Wenn aber alle Teile des Satzes in Original-Englisch auch in Plain-Englisch vorhanden sind, aligniert Cats diese korrekt. So wird in Artikel 2084 der originale Satz: *'But with food prices rising sharply in recent months, many experts are calling on countries to scale back their headlong rush into green fuel development, arguing that the combination of ambitious biofuel targets and mediocre harvests of some crucial crops is contributing to high prices, hunger and political instability.'*
Dreimal mit Plain-Englisch aligniert: *'Many experts are urging countries to reduce their green fuel efforts.', 'They say that the mediocre harvests of some crops contribute to high prices, hunger and political instability.', 'Food prices have been rising in recent months.'*

Viele Fehler, die Cats macht, sind nachvollziehbar. Das Thema ist dasselbe, es kommen dieselben Worte vor, nur ist der Inhalt des Satzes unterschiedlich oder es fehlt zu viel Information. So sind viele Fehler von Cats im Gegensatz zu anderen Systemen nicht völlig falsch.
Zum Beispiel in Artikel 2084 Original: *Each year, an ever larger portion of the world's crops — cassava and corn, sugar and palm oil — is being diverted for biofuels as developed countries pass laws mandating greater use of nonfossil fuels and as emerging powerhouses like China seek new sources of energy to keep their cars and industries running*
Plain-Englisch: *The crops are cassava and corn, sugar and palm oil.*

Die detaillierten Ergebnisse finden sich im Anhang in Tabelle 9.

5.6 Transformer

Transformer alignierte insgesamt 176 Paare. Davon waren 62 im Goldstandard zu finden. 28 Paare wurden korrekt aligniert, waren aber nicht im Goldstandard zu finden.

Die falsch alignierten Paare von Transformer ähnelten den Paaren von Laser: Sie hatten meist kein gemeinsames Thema und ich konnte nicht nachvollziehen, warum diese Sätze aligniert wurden. Zusätzlich neigt Transformer dazu, einen Satz mit vielen anderen zu alignieren, wobei meist alle Alignierungen falsch sind. Das führt zu vielen falschen Paaren. Generell werden längere Sätze eher falsch annotiert.

Bei einigen falschen Sätzen stimmt die Struktur der alignierten Sätze überein. Allerdings sind sie semantisch unterschiedlich. Es scheint, dass Transformer Eigennamen und Zahlen nicht voneinander unterscheiden kann. Das kann, werden die Daten für das Training eines Vereinfachungssystem benutzt, zu Problemen führen, da das System lernen könnte, dass Zahlen und Eigennamen, in diesem Fall Länder, austauschbar sind. Das würde zu falschen Vereinfachungen führen, in denen der Inhalt nicht mehr stimmt.
Beispiel aus Artikel 2084 Original: *The price of corn in Rwanda rose 19 percent last year.*
Plain Englisch: *In the United States, the price of corn rose 73 percent during the second half of 2010.*

Die detaillierten Ergebnisse finden sich im Anhang in Tabelle 10.

5.7 LHA

LHA alignierte insgesamt 73 Sätze, 45 davon sind im Goldstandard zu finden. 12 korrekte Paare, die im Goldstandard fehlen, sind vorhanden.

Generell scheint LHA sehr restriktiv zu alignieren; die meisten Alignierungen sind Sätze, die Wort für Wort übereinstimmen. So kommt es zu weniger Fehlern, aber auch generell zu weniger Alignierungen.

Auch LHA hat verschiedene Sätze falsch aligniert, die sich in der Struktur ähneln, aber andere Eigennamen und Zahlen beinhalten. Das Beispiel zu Transformer findet sich auch in den Alignierungen von LHA.
Original: *The price of corn in Rwanda rose 19 percent last year.*

Plain-Englisch: *In the United States, the price of corn rose 73 percent during the second half of 2010.*

Die detaillierte Evaluation findet sich im Anhang in Tabelle 11.

5.8 Vergleich der Systeme

Um die Systeme miteinander vergleichen zu können berechnete ich für alle die Präzision und den Recall über alle Artikel. Die Resultate sind in Tabelle 7 zu finden. Die Präzision bezeichnet hier, wie viele der automatisch alignierten Paare richtig sind (in Tabellen 9 bis 11 als 'Übereinstimmung autom./Goldstandard' und 'Sonst korrekt' bezeichnet). Der Recall zeigt, wie viele Paare des Goldstandards gefunden wurden.

Durch das Zählen von mehreren korrekten Paaren als eines ist die Präzision bei allen Systemen etwas zu tief.

Vollständigkeitshalber habe ich Laser in die Tabelle 7 eingefügt, obwohl ich keine konkreten Zahlen wie bei den anderen Systemen habe. Da aber alle Resultate von Laser 0 sind, ist der Vergleich trotzdem möglich.

System	Präzision	Recall	F1 score
Laser	0	0	–
Cats	0.44	0.63	0.518
Transformer	0.51	0.61	0.555
LHA	0.78	0.44	0.563

Tabelle 7: Präzision und Recall

Bei allen Systemen sind die Werte tiefer als in den Studien, die mit ihnen veröffentlicht wurden. Das könnte dafür sprechen, dass die Daten nicht ideal für die Alignierung sind. Den Zahlen nach erbringt LHA die beste Leistung. Cats schliesst am schlechtesten ab, was mich erstaunt hat, hatte ich doch bei der manuellen Evaluation das Gefühl, dass Cats die meisten richtigen Paare aligniert hat. Im Grossen und Ganzen unterscheiden sich die Resultate nur geringfügig. Laser schneidet sehr schlecht ab. Laser ist in der Alignierung von mehrsprachigen Korpora sehr vielversprechend. Für diese Anwendung ist Laser nicht geeignet.

5.9 Diskussion

Laser liefert auf meinen Daten Resultate mit einem hohen Sicherheitswert, die Alignierungen sind aber konsistent schlecht. (Siehe Kapitel 5.3) Dadurch können die Alignierungen nicht verbessert werden, indem man den Schwellenwert erhöht. Das kann daran liegen, dass die Ähnlichkeit relativ zu allen nächsten Nachbarn ist (Margin-Based Ansatz von Artetxe and Schwenk [2019]). Hat ein Satz also nur schlechte Nachbarn, wird der Wert des nächsten Nachbarn sehr gross. Da die Satzpaare, die die höchsten Ähnlichkeiten haben, alle falsch sind und die besseren Sätze eine tiefere Ähnlichkeit aufweisen, könnte man versuchen, zusätzlich zum unteren einen oberen Schwellenwert hinzuzufügen.

Generell fällt auf, dass Laser sehr wenige Satzpaare aligniert. So wurden nur etwa 25% aller Sätze aligniert. Das könnte aufzeigen, dass in den Daten sehr wenige geeignete Paare vorhanden sind.
Allerdings findet Laser auch in der Alignierung einer Datei mit sich selber nur etwa 25% der Paare. 100% sind hier durchaus erreichbar. Auch Sätze, die in beiden Korpusteilen vorhanden sind, aligniert Laser nicht. Bei den anderen Systemen werden diese sehr häufig richtig aligniert. Weitere Untersuchungen könnten zeigen, warum das nicht der Fall ist. In Schwenk [2018] gelang es den Autoren, aus dem 'Newsela-Korpus' 95% aller Paare wiederherzustellen. Das spricht dafür, dass die Konfiguration von Laser, die ich in dieser Arbeit verwendet habe, noch verbessert werden müsste. Generell stellt sich die Frage, ob Laser für diese Aufgabe geeignet ist.

Dieselben Skripts, die ich für diese Arbeit benutzt habe, wurden von Mathias Müller benutzt, um das 'Credit-Suisse Bankenmagazin'-Korpus zu alignieren. Auf diesen Daten erzeugte Laser sehr gute Resultate (Siehe Kapitel 5.2. Diese Daten entsprechen der Aufgabe, für die Laser eigentlich entwickelt wurde. Da Laser hier so gute Resultate erreicht, kann ich ausschliessen, dass an den Skripten oder an Laser selbst ein grundsätzliches Problem besteht. Diese Resultate verhärten den Verdacht, dass Laser für die Alignierung von vereinfachten Sätzen weniger gut bis gar nicht geeignet ist.

Die Alignierung von einzelnen Artikeln mit einem tieferen Schwellenwert erreichte bessere Ergebnisse als die mit dem Standard-Schwellenwert von 1.1 (Siehe Kapitel 5.4). Dies könnte ein Indiz dafür sein, dass die Resultate mit einem tieferen Schwellenwert verbessert werden könnten. Allerdings generiert Laser so auch mehr falsche Paare. Ob ein tieferer Schwellenwert wirklich bessere Resultate liefert, müsste in weiteren Untersuchungen ermittelt werden.

Da auch die anderen Systeme weniger gut abgeschnitten haben als erwartet (Siehe

Kapitel 5.5 bis 5.7), sind wahrscheinlich auch die Daten und deren Vorverarbeitung nicht ideal. So fehlen in den Original-Artikeln vielerorts Wörter oder sogar Phrasen. Auch lassen sich viele Fehler der Satzsegmentierung in den Original-Daten feststellen. Durch ein vollständigeres Herunterladen der Daten und eine verbesserte Vorverarbeitung könnten auch die Resultate verbessert werden. Es stellt sich ausserdem die Frage, ob genug Daten vorhanden sind, um gute Ergebnisse zu erzielen. Das in dieser Arbeit erstellte Korpus ist dafür scheinbar zu klein (Mehr dazu in Kapitel 3).

Bei genauerem Betrachten der Tabellen 9 bis 11, fällt auf, dass Artikel 0122 in allen Systemen überdurchschnittlich gut aligniert wurde. Andere Artikel werden in allen Systemen eher schlecht aligniert. Das zeigt, dass die Leistung der Systeme vom Artikel abhängig ist. Die Artikel wurden von Hand von verschiedenen Leuten editiert und sind unterschiedlich nahe am Original-Artikel (Siehe Kapitel 3.1). Das führt zu grossen inhaltlichen und formellen Unterschieden zwischen den einzelnen Artikeln.

Insgesamt führte ich 7 Versuche zu Alignierungen mit Laser durch. Davon fanden 3 mit den gesamten Daten statt. 2 waren Versuche, wie Laser mit identischen Sätzen umgeht und 2 waren Alignierungen über einen einzelnen Artikel, um zu sehen, ob Laser besser abschneidet, wenn die Sätze, die aligniert werden können, sehr beschränkt sind.

6 Fazit

In dieser Arbeit habe ich zuerst Daten von 'The Times in Plain English' und ihre originalen Artikel heruntergeladen. Die Artikel wurden vorverarbeitet und einzeln gespeichert. Danach habe ich die Daten mit Laser aligniert. Ich habe gesehen, dass die Alignierungen nicht korrekt waren. Deswegen versuchte ich, die Vorverarbeitung zu verbessern und verwendete einen anderen Satz-Tokenisierer, um die Sätze besser voneinander trennen zu können. Die verbesserten Daten alignierte ich wiederum mit Laser. Auch diese Alignierungen waren nicht korrekt. Mit einigen Versuchen versuchte ich festzustellen, wie diese schlechten Resultate entstanden sein könnten. Um Laser mit anderen Systemen vergleichen zu können, alignierte ich dieselben Daten mit Cats. Zusätzlich alignierte Marek Kostrzewa die Daten mit Tranformer und LHA.

Es gelang mir im Rahmen dieser Arbeit ein kleines Korpus aus Artikeln in Plain-Englisch und deren originalen Gegenstücken aufzubauen. Ich alignierte die Sätze und stellte dabei fest, dass Laser für diese Aufgabe nicht geeignet ist.

Für die schlechten Ergebnisse von Laser gibt es verschiedene Erklärungsmöglichkeiten:

- Die Daten sind ungeeignet für die Alignierung. Die Art der Vereinfachung von Hand ergibt Sätze, die nicht parallel zu den Sätzen in Original-Englisch sind. Viele Informationen werden in den Artikeln in Plain-Englisch weggelassen. so verändert sich die Semantik der Texte.
- Im Korpus sind zu wenige Daten vorhanden, als dass Laser gute Übereinstimmungen finden kann.
- Die Vorverarbeitung ist zu fehlerhaft, um gute Ergebnisse erzielen zu können.
- Die Konfiguration von Laser ist für diese Daten nicht ideal.
- Laser ist von seiner Arbeitsweise für diese Aufgabe nicht geeignet.

Leider sind die generierten Alignierungen von Laser so nicht zu gebrauchen. Im

Resultat erscheinen zu wenige Satzpaare, die zusätzlich meist keine Entsprechungen sind.

Die anderen Systeme (Cats, Transformer und LHA) haben bessere Ergebnisse erzielt. Mit diesen und zusätzlichen Daten könnte ein System trainiert werden.

6.1 Weiteres Vorgehen

In Zukunft kann das Korpus mit neu erschienenen Plain-Artikeln und deren Originale erweitert werden. Die Original-Artikel müssen so heruntergeladen werden, dass die Artikel vollständig sind, eventuell kann der Zugriff auf 'The New York Times Archive' genutzt werden. Das Herunterladen könnte automatisch stattfinden; Artikel, die zu sehr verrauscht sind könnten automatisch herausgefiltert werden, dies könnte mithilfe eines Sprachmodells geschehen, das Texte, die nicht natürlich sind, herausfiltert. Die Alignierung könnte davon profitieren, wenn Artikel von 'The New York Times' dem Korpus hinzugefügt werden, auch wenn sie keine Entsprechung in Plain-Englisch haben.

Die Vorverarbeitung muss verbessert werden, vor allem die Segmentierung der Sätze. Hier könnten andere Satz-Tokenisierer eingesetzt werden.

Es müssen weitere Untersuchungen mit Laser gemacht werden. Ein Ansatz wäre, die Parameter von Laser zu verändern und so bessere Resultate zu erzielen. Ich konnte die schlechten Resultate nicht abschliessend erklären, versuchte aber, die Probleme mit verschiedenen Ansätzen zu erklären. Man müsste Laser auch auf anderen vereinfachten, schon bestehenden Korpus anwenden. So könnte herausgefunden werden, ob das Alignieren von einfachen Sätzen an sich das Problem ist oder doch die Daten, die in dieser Arbeit verwendet wurden.

6.2 Persönliches Fazit

Die Resultate liegen unter der Erwartung, was eine gewisse Unzufriedenheit zurücklässt. Trotzdem konnte ich viel über die Methodik lernen, wie ein solches Projekt angegangen werden kann. Zusätzlich lernte ich, wie Alignierungen funktionieren und was mögliche Probleme dabei sind.

Literatur

M. Artetxe and H. Schwenk. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1309. URL `https://www.aclweb.org/anthology/P19-1309`.

A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the conference on empirical methods in natural language processing*, pages 355–362. Association for Computational Linguistics, 2011.

Y. Belinkov and Y. Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018.

R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948. Publisher: American Psychological Association.

F. Grégoire and P. Langlais. Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 46–50, 2017.

J. L. Johns and T. E. Wheat. Newspaper readability. *Literacy Research and Instruction*, 18(2):141–147, 1978.

J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel. 1975. URL `https://stars.library.ucf.edu/istlibrary/56`. Publisher: Institute for Simulation and Training, University of Central Florida.

C. Maass. *Leichte Sprache: das Regelbuch*. Number 1 in Barrierefreie Kommunikation. Lit, Münster, 2015. ISBN 978-3-643-12907-9.

P. Mcnamee and J. Mayfield. Character n-gram tokenization for european language text retrieval. *Information retrieval*, 7(1-2):73–97, 2004. Publisher: Springer.

N. Nikolov and R. Hahnloser. Large-scale hierarchical alignment for data-driven text rewriting. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*, 2019.

Plain Language Action and Information Network. plainlanguage.gov: About, a. URL `https://plainlanguage.gov/about/`.

Plain Language Action and Information Network. plainlanguage.gov: History and timeline, b. URL `https://plainlanguage.gov/about/history/`.

Plain Language Action and Information Network. plainlanguage.gov: What is plain language?, c. URL `https://plainlanguage.gov/about/definitions/`.

R. Pogrebin. Slick magazine offers immigrant readers practical advice. *The New York Times*, Nov. 1996. ISSN 0362-4331. URL `https://www.nytimes.com/1996/11/06/nyregion/slick-magazine-offers-immigrant-readers-practical-advice.html`.

Printing Office U.S.-Government. Plain writing act of 2010, 2010. URL `https://www.govinfo.gov/app/details/PLAW-111publ274`.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019.

L. Ruh. Wenn 5.-Klässler zu Zeitungsexperten der «Schaffhauser Nachrichten» werden. *Schaffhauser Nachrichten*, June 2020. URL `https://www.shn.ch/region/stadt/2020-06-10/wenn-5-klaessler-zu-zeitungsexperten-der-schaffhauser-nachrichten-werden`.

A. Schiff. About us. *The Times in Plain English*, May 2015. URL `https://www.thetimesinplainenglish.com/about-us/`.

K. Schriver, A. L. Cheek, and M. Mercer. The research basis of plain language techniques: Implications for establishing standards. *Clarity*, 63:26–32, 2010.

H. Schwenk. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia,

July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2037. URL https://www.aclweb.org/anthology/P18-2037.

H. Schwenk and M. Douze. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2619. URL https://www.aclweb.org/anthology/W17-2619.

S. Stajner, M. Franco-Salvador, S. P. Ponzetto, P. Rosso, and H. Stuckenschmidt. Sentence alignment methods for improving text simplification systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 97–102, 2017. doi: 10.18653/v1/P17-2016. URL https://doi.org/10.18653/v1/P17-2016.

The Plain Language Action and Information Network. *Federal Plain Language Guidelines*. 2011. URL https://plainlanguage.gov/media/FederalPLGuidelines.pdf.

E. Thrush. Plain English? a study of plain English vocabulary and international audiences. *Technical Communication*, 48(3):289–296, 2001. ISSN 0049-3155. URL https://www.jstor.org/stable/43090432. Publisher: Society for Technical Communication.

Wikipedia. Bildungssystem in den Vereinigten Staaten, 2020. URL https://de.wikipedia.org/w/index.php?title=Bildungssystem_in_den_Vereinigten_Staaten&oldid=199583906.

Lebenslauf

Persönliche Angaben

Anja Ryser
05.06.1997
Kirchstrasse 17
8214 Gächlingen
anja.ryser@uzh.ch

Schulbildung

2004-2010	Primarschule in Gächlingen
2010-2012	Sekundarschule in Neunkirch
2012-2016	Kantonsschule in Schaffhausen
seit 2017	Bachelor-Studium Computerlinguistik und Sprachtechnologie an der Universität Zürich

Berufliche und nebenberufliche Tätigkeiten

2016-2017	Praktikum in der Kindertagesstätte Spatzenescht in Hallau
seit 2013	Leiterin in der Ameisli Gä-Lö-Oha
seit 2018	Mitglied im Samariterverein Chläggi Samariter

A Tabellen

Part of speech	Plain-Englisch	In Prozent	Original	In Prozent
#	9	0.00114957	25	0.002452353
$	612	0.0781709	1083	0.106235942
’	12	0.00153276	169	0.016577908
(	1103	0.14088645	620	0.06081836
)	1123	0.14344105	626	0.061406925
,	19168	2.44833312	56632	5.555266723
.	60560	7.73534296	37010	3.63046372
:	2550	0.3257121	2017	0.197855859
CC	17180	2.019440542	29439	2.887793068
CD	12983	1.658322162	15278	1.498682105
DT	73743	9.419121063	87965	8.628850072
EX	2222	0.283816579	1451	0.14233458
FW	144	0.018393154	438	0.042965229
IN	71811	9.172435816	114346	11.21667129
JJ	49168	6.280240133	71706	7.033937626
JJR	4370	0.558181121	4200	0.411995342
JJS	2268	0.28969217	2212	0.216984214
LS	11	0.001405033	1	0.0000980941
MD	11417	1.458296079	9156	0.898149847
NN	105423	13.46570443	141487	13.87904405
NNP	62563	7.991186614	91456	8.971296677
NNPS	2686	0.343083408	3272	0.320963991
NNS	61808	7.894750287	68299	6.699730928
PDT	413	0.052752587	521	0.051107041
POS	23	0.002937795	193	0.018932167
PRP	25974	3.317665091	29209	2.865231419
PRP$	7281	0.930003832	11978	1.174971479
RB	22755	2.906501469	34433	3.37767515
RBR	1408	0.179844169	1904	0.186771222
RBS	461	0.058883638	619	0.060720266

RP	2321	0.296461873	3126	0.306642248
SYM	0	0	2	0.000196188
TO	19428	2.481542981	24995	2.451862758
UH	68	0.008685656	83	0.008141813
VB	31928	4.078170903	31262	3.066618666
VBD	20969	2.678375271	41694	4.089936621
VBG	13785	1.760761272	20387	1.999845011
VBN	10971	1.401328394	22596	2.216534943
VBP	25156	3.21318176	21515	2.110495189
VBZ	25943	3.313705454	21170	2.076652714
WDT	1981	0.253033593	5262	0.516171308
WP	5405	0.690381913	5469	0.536476792
WP$	38	0.004853749	268	0.026289227
WRB	3658	0.467237195	3749	0.36775489
“	0	0	106	0.010397978
Total	782900	100	1019429	100

Tabelle 8: Alle POS-Tags

Artikel ID	Sätze Plain	Sätze Original	Satzpaare Goldstandard	Automatisch Aligniert	Übereinstimmung autom./Goldst.	Sonst Korrekt	Total Korrekt
0050	43	95	15	53	7	10	17
0122	46	98	29	49	17	8	25
0442	30	24	9	34	8	7	15
1142	33	52	6	37	4	2	6
1191	25	46	14	29	8	9	17
1997	14	42	10	15	7	4	11
2048	6	6	5	5	3	1	4
2084	21	51	13	23	10	7	17
Total	218	414	101	245	64	48	112

Tabelle 9: Evaluation Cats

Artikel ID	Sätze Plain	Sätze Original	Satzpaare Goldstandard	Automatisch Aligniert	Übereinstimmung autom./Goldst.	Sonst Korrekt	Total Korrekt
0050	43	95	15	30	8	4	12
0122	46	98	29	34	17	7	24
0442	30	24	9	16	7	2	9
1142	33	52	6	15	2	2	4
1191	25	46	14	22	6	6	12
1997	14	42	10	26	8	2	10
2048	6	6	5	3	3	0	3
2084	21	51	13	30	11	5	16
Total	218	414	101	176	62	28	90

Tabelle 10: Evaluation Transformer

Artikel ID	Sätze Plain	Sätze Original	Satzpaare Goldstandard	Automatisch Aligniert	Übereinstimmung autom./Goldst.	Sonst Korrekt	Total Korrekt
0050	43	95	15	4	4	0	4
0122	46	98	29	17	12	2	14
0442	30	24	9	3	2	1	3
1142	33	52	6	2	2	0	2
1191	25	46	14	20	4	7	11
1997	14	42	10	11	8	1	9
2048	6	6	5	2	2	0	2
2084	21	51	13	14	11	1	12
Total	218	414	101	73	45	12	57

Tabelle 11: Evaluation LHA

Art der Alignierung	Sätze Plain	Sätze Original	Satzpaare Goldstandard	Automatisch Aligniert	Übereinstimmung autom./Goldst.	Sonst Korrekt	Total Korrekt
laser 1.0	33	52	6	3	0	0	0
laser 0.5	33	52	6	25	1	4	5
Cats	33	52	6	37	4	2	6
Transformer	33	52	6	15	2	2	4
LHA	33	52	6	2	2	0	2

Tabelle 12: Evaluation Artikel 1142

B Beispiele

Hier finden sich Auszüge aus den Daten, die während dieser Arbeit entstanden sind. Es werden immer die ersten Sätze/Satzpaare gezeigt.

B.1 Auszug aus Artikel 1142, Original

Hier ein Auszug aus dem Original Artikel ” (url:), extrahiert aus der XML-Datei.

COURTLAND, Ala. — In this forlorn Southern town whose once-humming factories were battered in recent years by a flood of Asian imports, Rhonda Hughes, 43, is a fervent supporter of Donald Trump. Her 72-year old mother is equally passionate about Senator Bernie Sanders. Disenchantment with the political mainstream is no surprise. But research to be unveiled this week by four leading academic economists suggests that the damage to manufacturing jobs from a sharp acceleration in globalization since the turn of the century has contributed heavily to the nation's bitter political divide. Ms. Hughes avoids discussing the election with her mother, but their neighbor Benjamin Green, 83, knows just what Washington needs. “It'll take a junkyard dog to straighten this country out,” he said. Cross-referencing congressional voting records and district-by-district patterns of job losses and other economic trends between 2002 and 2010, that areas hardest hit by trade shocks were much more likely to move to the far right or the far left politically. “It's not about incumbents changing their positions,” said David Autor, an influential scholar of labor economics and trade at the Massachusetts Institute of Technology and one of the paper's authors. “It's about the replacement of moderates with more ideological successors.” Mr. Autor added: “In retrospect, whether it's Trump or Sanders, we should have seen in it coming. The China shock isn't the sole factor, but it is something of a missing link.” In addition to Mr. Autor, the research was conducted by David Dorn of the University of Zurich; Gordon Hanson, a professor at the University of California, San Diego; and Kaveh Majlesi of Lund University in Sweden.

B.2 Auszug aus Artikel 1142, Plain-Englisch

Ein Auszug aus dem Artikel ” (url:), extrahiert aus der XML-Datei

MIRA OBERMAN/AFP/GettyImages You may live in a town where the factory has closed. It may have gone out of business. There is a good chance it moved its production to Asia. For years, trade pacts were with advanced countries such as Japan. They did not result in the importing of lower-cost goods. It changed with China entering the World Trade Organization. Jobs began to go to China. Lower-priced products began to flow into this country. The same was true with Mexico after the Nafta agreement.

B.3 Auszug aus dem Goldstandard von Artikel 1142

(Nacheinander auf je einer Zeile: Originalsatz, Satz in Plain-Englisch, ein Satzpaar pro Abschnitt)

But the collapse of the apparel industry here in the first decade of the 21st century, following China’s entry into the World Trade Organization in 2001, reversed that process.
It changed with China entering the World Trade Organization. Jobs began to go to China.

“But the nature of globalization changed after the end of the Cold War and it took a while for academics to catch up.” Until the Nafta agreement with Canada and Mexico in 1994, and especially the entry of China into the W.T.O., trade deals were mostly multilateral and the rise in manufacturing imports to the United States came primarily from other advanced industrial nations like Germany and Japan.
For years, trade pacts were with advanced countries such as Japan. They did not result in the importing of lower-cost goods.

The authors found that voters in congressional districts hardest hit by Chinese imports tended to choose more ideologically extreme lawmakers.
The people who lost their jobs voted for these right wing candidates.

Those hard-hit districts became, on average, far more conservative: the ideological equivalent of moving from Marco Rubio to Ted Cruz.
The people who lost their jobs voted for these right wing candidates.

But the benefit of free trade is “10 times the size of the losses,” he said.
An economist said, “The benefit of free trade is ten times the size of the losses.

B.4 Auszug aus der Alignierung von Laser

Hier ein Auszug aus der Alignierung von Laser aus dem dritten Durchgang. Zu sehen sind die 10 Alignierungen mit den höchsten Ähnlichkeiten: (Nacheinander auf je einer Zeile: Ähnlichkeit, Originalsatz, Satz in Plain-Englisch, ein Satzpaar pro Abschnitt)

1.4484383037526178
“When money flows into an index fund or index-related E.T.F., the manager generally buys into the securities in an index in proportion to their current market capitalization (often to the capitalization of only their public float, which interestingly adds a layer of distortion, disfavoring companies with large insider, strategic, or state ownership),” he wrote.
His hope was that the problem would go away sooner or later. He wanted millions to get legal status.

1.436887695985032
Mr. Murstein attended business school and started his career at Bear Stearns and Salomon Brothers, the investment banks.
Around 65 degrees helps sleep. A cool body core and warm skin are best for sleep.

1.4321725516709178
Now other Turkish businesses are feeling the downturn.
The government accused her of insulting the Prophet Muhammad.

1.4045318155843194
Mr. Díaz-Canel’s slow and steady climb up the ranks of the bureaucracy has come through unflagging loyalty to the socialist cause — he “is not an upstart nor improvised,” Mr. Castro has said — but he largely stayed behind the scenes until recent years.
Jordan has already taken in more than half a million refugees from Syria.

1.3749503672800418
But no one knows the mechanisms that Mitnick, Trump’s accountant, used.
Children reported that they liked a Bluetooth model. It is best to use the BT2200 without a cord.

1.3734270201788878
“It doesn’t hurt to try,” she said. Syrian refugees are the biggest group.
It does not look like this will happen soon.

B.5 Auszug aus der Alignierung von Laser, Artikel 1142

Artikel 1142 wurde als einzelner Artikel in Laser aligniert. Zudem wurden verschiedene Schwellenwerte ausprobiert. (Nacheinander auf je einer Zeile: Ähnlichkeit, Originalsatz, Satz in Plain-Englisch, ein Satzpaar pro Abschnitt)

B.5.1 Alignierung Schwellenwert 1.1 (Standard)

1.3634051551790474
“Free trade really helps working-class people in terms of lower prices for products.
Source: The New York Times April 25, 2016

1.2718211755894353
But the benefit of free trade is “10 times the size of the losses,” he said.
When facing the voters, few politicians will make that their campaign slogan.

1.1720415169876401
“But those days are never coming back.”
There is another side.

B.5.2 Auszug Alignierung Schwellenwert 0.5

1.3634051551790474
“Free trade really helps working-class people in terms of lower prices for products.
Source: The New York Times April 25, 2016

1.2718211755894353
But the benefit of free trade is “10 times the size of the losses,” he said.
When facing the voters, few politicians will make that their campaign slogan.

1.1720415169876401
“But those days are never coming back.”
There is another side.

1.087395623921767
But the collapse of the apparel industry here in the first decade of the 21st century, following China’s entry into the World Trade Organization in 2001, reversed that

process.
Jobs began to go to China.

B.6 Auszug aus der Alignierung von Cats, Artikel 1142

(Nacheinander auf je einer Zeile: Satz in Plain-Englisch, Originalsatz, ein Satzpaar pro Abschnitt)

MIRA OBERMAN/AFP/GettyImages But research to be unveiled this week by four leading academic economists suggests that the damage to manufacturing jobs from a sharp acceleration in globalization since the turn of the century has contributed heavily to the nation's bitter political divide.

You may live in a town where the factory has closed. Hanson emphasize that trade is only one factor among many that have contributed to a polarizing Congress (income inequality is another, as are attitudes toward immigrants).

It may have gone out of business. Brooks Jr., who has emerged as one of the most right-wing members of Congress.

There is a good chance it moved its production to Asia. "Free trade really helps working-class people in terms of lower prices for products.

For years, trade pacts were with advanced countries such as Japan. Until the Nafta agreement with Canada and Mexico in 1994, and especially the entry of China into the W.T.O., trade deals were mostly multilateral and the rise in manufacturing imports to the United States came primarily from other advanced industrial nations like Germany and Japan.

They did not result in the importing of lower-cost goods. But it has been an important one, particularly over the last decade, when Chinese imports ramped up.

It changed with China entering the World Trade Organization. But the collapse of the apparel industry here in the first decade of the 21st century, following China's entry into the World Trade Organization in 2001, reversed that process.

Jobs began to go to China. "I was in China two weeks ago and they are going to clean our plow if we don't act."

Lower-priced products began to flow into this country. "Free trade really helps working-class people in terms of lower prices for products.

The same was true with Mexico after the Nafta agreement. Until the Nafta agree-

ment with Canada and Mexico in 1994, and especially the entry of China into the W.T.O., trade deals were mostly multilateral and the rise in manufacturing imports to the United States came primarily from other advanced industrial nations like Germany and Japan.

B.7 Auszug aus der Alignierung von Transformer, Artikel 1142

(Nacheinander auf je einer Zeile: Original-Satz, Satz in Plain-Englisch, ein Satzpaar pro Abschnitt

“Exposure to import competition is bad for centrists,” Mr. Hanson said.
The people hurt in this economy are against trade and immigration.

We’ve now found a mechanism for how economic changes create further political divisions.” Parker Griffith experienced the move away from the political middle firsthand.
Could the government have done more to soften the blow where jobs were lost?

But that wasn’t enough to save his seat.
Observers say it did not do much.

“If you’re under economic stress and you can’t provide for your family, the easiest answer is to find someone to blame,” said Dr. Griffith.
Could the government have done more to soften the blow where jobs were lost?

While whites hit hard by trade tend to move right, nonwhite voters move left, eroding support for moderates in both parties, the study concluded.
Older white workers began voting for candidates who blamed Washington for the trade pacts.

“China and the W.T.O.
Jobs began to go to China.

“There are these concentrated pockets of hurt,” Mr. Autor said, “and we’re seeing the political consequences of that.” Mr. Autor and Mr. Hanson emphasize that trade is only one factor among many that have contributed to a polarizing Congress (income inequality is another, as are attitudes toward immigrants).
Older white workers began voting for candidates who blamed Washington for the trade pacts.

This trade-induced polarization has had a significant effect on the overall ideological makeup of Congress.
It changed the face of the Congress.

Some very conservative members of Congress have been sympathetic to free trade arguments in the past, but Representative Brooks, who has welcomed support from the Tea Party, doesn't mince words about where he stands.
When facing the voters, few politicians will make that their campaign slogan.

B.8 Alignierung von LHA, Artikel 1142

(Nacheinander auf je einer Zeile: Original-Satz, Satz in Plain-Englisch, ein Satzpaar pro Abschnitt) But the benefit of free trade is "10 times the size of the losses," he said.
An economist said, "The benefit of free trade is ten times the size of the losses.

"Free trade really helps working-class people in terms of lower prices for products.
Free trade really helps working-class people in terms of lower prices for products."