Erstellung eines parallelen Vereinfachungskorpus für die deutsche Sprache
– Unter Verwendung des HHU Annotationstools TS-anno

Regina Stodden
Heinrich Heine Universität, Düsseldorf
NRW-Forschungskolleg Online Partizipation
regina.stodden@hhu.de

24. Januar 2022
Version 2.0

Inhaltsverzeichnis

1 Die Korpuserstellung im Überblick/Motivation

Diese Instruktionen gelten als Instruktionen und Hilfestellung zum Aufbau eines Vereinfachungskorpus. Der Korpusaufbau beinhaltet die folgenden Schritte:

- Datenauswahl
- Web Scraping
- Alinierung von Satzpaaren
- Bewertung von Satzpaaren
- Annotation der geänderten Wörter während der Vereinfachung und Kategorisierung der Änderung

Im folgenden wird kurz der Gesamtzusammenhang zwischen den einzelnen Schritten aufgearbeitet bevor näher auf die einzelnen Schritte eingegangen wird.

Für den Aufbau eines automatisierten Textvereinfachungssystems werden parallele Daten in vereinfachter und alltäglicher Sprache benötigt. Als Datengrundlage werden in dieser Arbeit inhaltlich vergleichbare Dokumente in zwei verschiedenen Sprachleveln verwendet: in vereinfachter und alltäglicher Sprache. Bei den Dokumenten handelt es sich um Webseitentexte bzw. online zugängliche Texte. Um de Texte weitere verarbeiten zu können, müssen die Webseiten heruntergeladen werden und der darin enthaltene Inhalt von Navigations- und Werbeelement getrennt werden. Nach diesem Prozess des Web Spraping, werden die Texte vorverarbeitet. Zu der Vorverarbeitung zählt die Unterteilung des Textes in Sätze und Token. Die Datenauswahl, das Web Scraping und die Vorarbeitung zählen nicht zu der Aufgabe der Annotation und werden deswegen nicht genauer beschrieben.

Der Prozess der Alinierung und Annotation, sowie die Struktur der Daten wird an einem kleinen Beispiel illustriert. Als Beispieldokumente dienen Dokument A, ein Nachrichtentext zu Thema X in Alltagssprache, und Dokument E, ebenfalls ein Nachrichtentext zu Thema X aber in vereinfachter Sprache. Dokument A und B sind inhaltlich vergleichbar und können deswegen aliniert werden. Die Alinierung kann auf verschiedenen Ebenen stattfinden, wie Dokument-, Paragraph-, Satz-, oder Wortebene. In diesem Korpus werden die Sätze aus inhaltlich vergleichbaren Dokumenten, wie Dokument A und E, aliniert. Das heißt, inhaltlich gleiche oder vergleichbare Sätze werden miteinander verbunden, z.B. Satz A1 aus Dokument A und Satz E1 aus Dokument E. Diese und alle weiteren alinierten Sätze werden als Eingabe für das Textvereinfachungssystem dienen.

Nach der Alinierung wird jedes aliniertes Satzpaar, z.B. das Paar A1E1 aus Satz A1 und Satz E1, hinsichtlich der Satzqualität und Vereinfachungs-

qualität bewertet. Die Bewertung lässt Rückschlüsse darauf zu, ob das bewertete Paar gut für den Korpus geeignet ist und nicht. Paar A1E1 wäre z.B. brauchbar, wenn Satz E1 einfacher und leichter verständlich ist als Satz A1 und E1 trotzdem den Großteil der Bedeutung von Satz A1 beibehalten hat. Paar A1E1 wäre nicht brauchbar, wenn Satz A1 einfacher wäre als Satz E1 oder wenn einer der beiden Sätze schwerwiegende Grammatikalische Fehler aufweisen würde.

Neben der Bewertung wird der Vereinfachungsprozess jedes alinierte Satzpaar in der Annotation beschrieben. Pro Satzpaar wird angegeben, welche Wörter oder Satzteile sich geändert haben und wie diese Veränderung benannt werden kann. Enthält Satz A1 z.B. Konjunktiv und ein langes Kompositum, aber Satz E1 steht im Indikativ und das lange Kompositum wurde aufgelöst, würden eben diese beiden Transformationen benannt werden und die involvierten Wörter markiert werden. Die Transformationen sind dahingehend hilfreich, da sie zur Präzisierung und Verbesserung des Textvereinfachungssystems verwendet werden können. Falls z.B. nur ein paar Sätze vom Vereinfachungssystem unzureichend vereinfacht werden würden, könnte man sich diese Sätze genauer ansehen. Bei der genaueren Analyse könnte dann untersucht werden, ob die unzureichend vereinfachten Sätze z.B. alle vorher im Konjunktiv standen. Daraus würde sich schlussfolgern lassen, dass weitere Regeln oder Feature zur Auflösung des Konjunktiv in dem System benötigt werden würden.

Im Folgenden, wird die Alinierung, die Bewertung und die Transformationsannotation genauer beschrieben. Auf die folgende Beschreibung kann während des gesamten Annotationsprozesses zurückgegriffen werden, um beispielsweise Unklarheiten auszuräumen und eine einheitliche Annotation zu ermöglichen. Zur Unterstützung der Annotation wird das Annotationstool "TS-anno" verwendet. In den folgenden Instruktionen wird die Verwendung des Annotationstool ebenfalls erklärt.

2 Annotationstool

Jedem Nutzenden des Annotationstool können nach Belieben Texte zur Alinierung zugewiesen werden. Die Texte werden zuvor per Korpus sortiert. Jeder Korpus enthält mindestens ein Dokumentenpaar, was wiederum aus einem alltagssprachlichen und einem vereinfachten Dokument besteht. Unter Corpora-Overview wird eine Übersicht über alle verfügbaren Korpora des aktuellen Nutzenden angzeigt. Ein Klick auf einen Korpus leitet zu der Übersicht aller inhaltähnlichen Dokumente bzw. Dokumentpaare des Korpus weiter (Corpus Overview). Mit Klick auf ein Dokumentpaar beginnt entweder die Alinierung oder eine Übersicht über alle alinierten Satzpaare des Dokumtes wird angezeigt (Document Overview).

Auf der linken Seite befindet sich ein Navigationsmenü, das während

aller Schritte des Annotationsprozesses verfügbar ist. Mithilfe des Navigationsmenüs kann jederzeit zu der Document-Overview, Corpus Overview oder zur Corpora Overview gewechselt werden.

Falls zum Zeitpunkt der Nutzung des Navigationsmenüs eine Änderung vorgenommen wurde, so wird diese nicht gespeichert. Eine Speicherung muss immer über den "Save-Button" erfolgen.

Während der Alinierung und der Annotation wird im Hintergrund die Zeit per Aufgabe gestoppt. Dies hilft dabei, um Rückschlüsse ziehen zu können, wie lange bzw. wie schwierig die Bewertung eines Satzpaares oder Alinierung eines Dokumentenpaares war.

Falls Probleme oder Anregungen bei der Nutzung des Annotationstools auftauchen, können diese unter dem Menüpunkt Change Log hinzugefügt werden.

3 Alinierung

Zu Beginn der Annotation steht im Annotationstool mindestens ein Korpus mit mindestens einem Dokumentenpaar AE bestehend aus den inhaltlich vergleichbaren Dokumenten A (Alltagssprache) und E (Einfache Sprache) zur Verfügung. Falls ein Dokumentenpaar noch nicht aliniert wurde, leitet ein Klick auf das Dokument zum Alinierungsinterface weiter. Ansonsten werden alle alinierten Satzpaare mit der Möglicheit zur Bearbeitung der Alinierung, Bewertung der Satzpaare und Annotation der Transformationen angezeigt.

Im Alinierungsinterface werden Dokument A und E parallel nebeneinander angezeigt. Wie in Abbildung zu sehen, wurden beide Dokumente schon in Sätze unterteilt. Im folgenden sollen inhaltsgleiche Sätze miteinander verbunden werden. Dafür muss zunächst die Alinierung freigeschaltet werden. Die Alinierung kann durch einen Klick auf den "Add"- Button gestartet werden.

Es mehrere Möglichkeiten der Alinierung bzw. es können unterschiedlich viele Sätze in die Alinierung einbezogen werden, abhängig davon in oder aus wie vielen Sätzen die Bedeutung übertragen wurde. Die folgenden Optionen sind möglich:

1. 0:1 Alinierung: Einschub bzw. Hinzufügung
2. 1:0 Alinierung: Auslassung bzw. Löschung
3. 1:1 Alinierung
   (a) no-change: Satz A und Satz E sind identisch
   (b) change: Satz A und Satz E unterscheiden sich
4. n:1 Alinierung: Zusammenfassung

5. 1:n Alinierung: Aufteilung

6. n:n Alinierung: Satzübergreifende Verschiebung

Die Optionen müssen bei der Alinierung nicht namentlich erwähnt werden. Die Optionen dienen in der Instruktion lediglich als Anhaltspunkte und zur Illustration verschiedener Alinierungsmöglichkeiten.

3.1 Automatische Alinierung

1:1 no-change Satzpaare werden vom Annotationstool automatisch aliniert. Identische Sätze im alltagssprachlichen und vereinfachten Dokument stehen bei der manuellen Alinierung nicht zur Verfügung, um die Annotation dieser Paare zu beschleunigen und eine versehentliche Alinierung zu vermeiden.

Zudem werden 0:1 und 1:0 Alinierungen ebenfalls vom Annotationstool automatisch vorgenommen. Falls ein Satz aus dem alltagssprachlichen Dokument nicht mit einem anderen aliniert wurde, wird er als Auslassung gehandhabt und mit einer 1:0 Alinierung gespeichert. Ein Satz gilt nur als Auslassung, wenn der Inhalt in keinem der vereinfachten Sätze auftaucht, z.B. wenn der Inhalt nicht relevant genug für eine Vereinfachung erscheint. Ein Satz gilt nicht als entfernt, wenn ein Teil des Inhalt in einem vereinfachten Satz auftaucht. In dem Fall sollte der Satz gemeinsam mit dem Hauptausgangssatz als Satzzusammenfassung (1:n) manuell annotiert werden. Falls hingegen ein Satz aus dem vereinfachten Dokument nicht mit einem anderen aliniert wurde, wird er als Einschub bzw. Hinzufügung und mit einer Alinierung 0:1 gekennzeichnet. Falls ein Teil des Inhaltes in einem oder mehreren Ausgangssätzen enthalten ist, sollte manuell eine Satzzusammenfassung (n:1) annotiert werden. Eine Hinzufügung ist beispielsweise eine Worterklärung oder eine mit Beispielen angereicherte Erläuterung eines Wortes oder Konzeptes.

0:1, 1:0 und no-change Paare werden in die manuelle Bewertung eingeschlossen, um eine Vergleichbarkeit der Grammatikalität und Kontextunabhängigkeit der Ausgangssätze und vereinfachten Sätze mit den anderen Paaren zu ermöglichen. Da allerdings keine Veränderung durch die Vereinfachung bewertet werden kann, wird die Bewertung auf weniger Elemente als die Bewertung der anderen Paare beschränkt. .

Die übrigen 1:1, die n:1, 1:n, und n:n Alinierungen müssen von Hand vorgenommen werden. Dabei ist immer zu berücksichtigen, dass die jeweiligen Seiten des Satzpaares sich möglichst ähnlich sein sollen.

3.2 Manuelle Alinierung

Um ein Satzpaar zu alinieren, müssen die zugehörigen Sätze des alltagsprachlichen Dokumentes und des vereinfachten Dokumentes per Klick auf die jeweiligen Sätze ausgewählt werden. Anschließend die Alinierung mit Klick auf

den Save-Button bestätigt werden. Für eine Alinierung muss mindestens ein Satz des alltagssprachlichen Dokumentes und des vereinfachten Dokumentes ausgewählt werden. Andernfalls wird eine Fehlermeldung angezeigt. Nach erfolgreicher Speicherung wird das Satzpaar unter "List of aligned sentence pairs" angezeigt. Anschließend kann entweder das nächste Satzpaar aliniert werden, das aktuelle Satzpaar bewertet oder das aktuelle Satzpaar mit Transformationen annotiert werden. Falls ein Satzpaar voreilig aliniert wurde und eigentlich noch ein weiterer Satz zu dem satzpaar zugeordnet werden sollte, kann der edit-Button verwendet werden, um die Alinierung anzupassen.

3.2.1 Satzinterne Vereinfachung

Die wahrscheinlich am häufigsten auftretende Alinierungsmöglichkeit ist die 1:1 Alinierung. Sie wird dann annotiert, wenn die Bedeutung eines Ausgangssatzes in genau einen vereinfachten Satz übertragen wird und dabei eine Veränderung bzw. Vereinfachung vorgenommen wird. 1:1 Alinierungen werden annotiert, wenn die Bedeutung eines alltagssprachlichen Satzes in genau einen vereinfachten Satz übertragen wird. Beispiele dafür sind zum Beispiel die Umstellung des alltagssprachlichen Satzes, Weglassung von Worten in der Vereinfachung, Hinzufügung von Worter in der Vereinfachung oder Änderung von Worten in der Vereinfachung (z.B. Zeitform eines Verbs geändert).

Da sowohl n:1, 1:n als auch n:n Alinierungen möglich sind, empfiehlt es sich alle Sätze beider Dokumente schrittweise durch zugehen und schon gespeicherte Alinierungen gegebenfalls anzupassen. Falls die Bedeutung eines alltagssprachlichen Satzes nicht oder nur sehr ungenau in einem der vereinfachten Sätze wiederzufinden ist, wird dieses Satzpaar nicht annotiert. Das gleiche gilt auch andersrum für die vereinfachten Sätze.

3.2.2 Satzübergreifende Vereinfachung

Neben der satzinternen Vereinfachung können auch satzübergreifende Vereinfachungen angewendet worden bzw. zu annotieren sein, z.B. 1:n, n:1, n:n Alinierungen.

1:n Alinierung 1:n Alinierungen bieten sich besonders dann an, wenn Teile eines Satzes auf mehrere aufgeteilt wurden. Dies ist zum Beispiel der Fall, wenn Nebensätze zu eigenständigen Sätzen umgeschrieben werden. Sobald ein Bruchteil des Inhalts des Ursprungssatzes auch in einem vorherigen oder folgenden vereinfachten Satz auftaucht, sollte es mit 1:n annotiert werden. Die Reihenfolge und die Distanz zwischen den vereinfachten Sätzen ist dabei egal. Wenn allerdings ein Wort aus einem alltagssprachlichen Satz in einem vereinfachten Satz aufgegriffen wird, aber nur lediglich weitere Informationen zu dem Wort erfolgen, die nicht in dem alltagssprachlichen Satz enthalten

sind, wird der vereinfachte Satz als 0:1 aliniert und keinem alltagssprachlichen Satz zugeordnet.

n:1 Alinierung n:1 Alinierungen beschreiben die Zusammenfassung von mehreren alltagssprachlichen Sätzen zu einem vereinfachten Satz. Sobald Bruchteile des Inhalts aus einem zweiten alltagssprachlichen Satz in einen vereinfachten Satz integriert wurden, ist es eine n:1 Alinierung und keine 1:1 Alinierung mehr. Dabei werden häufig ein paar Informationen der alltagssprachlichen Sätze weggelassen und nicht mit in den vereinfachten Satz übernommen. Der Inhaltsverlust kann anschließend in der Bewertung angegeben werden. Ein Beispiel für n:1 Alinierungen ist die Zusammenführung zweier Hauptsätze mit Nebensätzen zu einem mit "und" verbundenem vereinfachten Hauptsatz in dem die Nebensätze weggelassen wurden.

n:n Alinierung Falls eine Mischung aus einer Zusammenfassung und Aufteilung mehrerer Sätze in mehrere Sätze vorliegt, werden die Sätze mit einer n:n Alinierung annotiert. Es finden also mehrere satzübergreifende Vereinfachungstransformationen gleichzeitig statt. Zum Beispiel könnte ein Nebensatz von einem ersten komplexen Satz in einen neuen vereinfachten Satz aufgeteilt werden, während gleichzeitig die Hauptinformation eines zweiten komplexen Satzes in den Hauptsatz des ersten komplexen Satz integriert wird. Diese Form der Alinierung kann immer dann vorgenommen werden, wenn keine klare Satzzuordnung getroffen werden kann. Dazu werden im Annotationtstool sowohl mehrere alltagssprachliche als auch mehrere vereinfachte Sätze ausgewählt.

Da die alltagsprachliche Seite eines Satzpaares und die vereinfachte Seite eines Satzpaares mehr als einen Satz enthalten kann, wird im folgenden von alltagssprachlichem Text und vereinfachtem Text geredet, um Uneindeutigkeiten zu vermeiden.

4 Bewertung

Wenn mindestens ein Satzpaar pro Dokument annotiert worden ist, wird in der Dokumentübersicht (Document Overview) die Übersicht aller alinierten Satzpaare dieses Dokumentes angezeigt. Pro Satzpaar kann eine Bewertung sowie die Transformationen der Vereinfachung hinzugefügt werden. Die Bewertung der Satzpaare erfolgt in drei Teilbereichen:

- Bewertung des Ausgangssatzes (Alltagssprache)
- Bewertung des vereinfachten Satzes
- Bewertung des Satzpaares

Für die Bewertung jedes Teilbereiches werden Kriterien aufgestellt. Die Bewertung jedes Kriteriums wird anhand der Zustimmung oder Ablehnung eines Aussagesatzes auf einer Skala mit 5 Werten bestimmt. Die Endpunkte der Bewertungsskala sind "Stimme überhaupt nicht zu." ("strongly disagree") bis "Stimme sehr zu." ("strongly agree"). Das Empfinden, ob ein Text leicht oder schwierig zu verstehen ist, hängt von den Erfahrungen der annotierenden Person, ihrem Wortschatz, ihren Interessen, ihren Sprachkenntnissen und vielem mehr. Daher gibt es keine richtige oder falsche Bewertung. Der zu wählende Wert liegt im Ermessen der annotierenden Person. Die folgenden Beispiele und Hinweise dienen nur zur Orientierung der Bewertung und können sich ebenfalls von Person zu Person unterscheiden.

Die Endpunkte sind auszuwählen, wenn eine klare Entscheidung für oder gegen etwas getroffen werden kann. Die Zwischenpunkte stehen für eine vorhandene, aber nicht deutliche Ablehnung oder Zustimmung. Der mittlere Punkt ist auszuwählen, wenn

1. keine Änderung betreffend des jeweiligen Kriteriums vorliegt oder
2. keine Tendenz vorliegt, also wenn der Aussage weder zugestimmt noch widersprochen werden kann.

Die Interpretation des mittleren Wertes ist abhängig vom jeweiligen Aussagesatzes. Die erste Variante wird im Folgenden als Kriterium mit neutralem Element bezeichnet. Zur Unterstützung der Annotation des neutralen Elementes, erstrecken sich die Skalenpunkte bei Kriterien mit neutralem Element zwischen -2 (negativ), 0 (neutral) und +2 (positiv). Die zweite Variante des mittleren Wertes kann zum Beispiel dann auftreten, wenn Teile des Textes der Aussage widersprechen aber andere die Aussage befürworten. In diesem Fall erstrecken sich die Skalenpunkte von 1 (negativ) bis 5 (positiv).

Besteht der Text der vereinfachten oder der alltagssprachlichen Seite des Satzpaares aus mehr als einem Satz, ist die Bewertung auf den gesamtes Text anzuwenden. Weist einer von zwei Sätzen eines Textes einer Seite beispielsweise Grammatikfehler auf, kann nicht mehr der höchste Wert für Grammatikalität verwendet werden. Überschatten die Fehler zudem die Korrektheit des zweiten Satzes, so ist die Grammatikalität des gesamten Textes mit einem niedrigen Wert zu bewerten. Das gleiche gilt auch für die anderen Kategorien.

Im Annotationstool werden die Bewertungen gruppenweise vorgenommen. Zuerst werden die Bewertungskriterien der alltagssprachlichen Texte angezeigt. Anschließend die der vereinfachten Texte und zum Abschluss die vergleichenden Kriterien zwischen alltagssprachlichen und vereinfachten Texten. Pro Kriterium wird der zu bewertende Aussagesatz angezeigt. Es stehen pro Kriterium 5 Radio-Button zur Verfügung. Bei der Auswahl eines Wertes bzw. Buttons wird der zugehörige Wert angezeigt. Die Anzeige des Wertes unterstützt bei der Unterscheidung zwischen Kriterien mit und ohne neutrales Element.

Eine Übersicht über die Aussagesätze pro Kriterium befindet sich in Tabelle 1. Eine deutsche Übersetzung der Aussagesätze befindet sich im Anhang (s. Tabelle 2).

4.1 Bewertung des Ausgangssatzes und des vereinfachten Satzes

Mit der Bewertung des originalen und des vereinfachten Textes wird angegeben, wie gut verständlich die Texte jeweils sind. Die Unterscheidung ist wichtig, da mit der Bewertung des Originaltextes die Ausgangsqualität des Textes bestimmt werden kann.

Die Bewertung des vereinfachten und des alltagssprachlichen Textes können sich unterscheiden oder auch identisch sein. Es ist sowohl möglich, dass die Werte des vereinfachten Textes sich im Vergleich zu den alltagssprachlichen Textes ins Negative als auch als ins Positive ändern. Beispielsweise kann ein vereinfachter Text mit einem niedrigeren Einfachheitswert und höherem Mehrdeutigkeitswert bewertet werden als der zugehörige alltagssprachliche Text. Damit wäre der vereinfachte Text schwieriger zu verstehen als der als schwieriger angenommene alltagssprachliche Text.

Im Folgenden werden die zu verwendenden Kriterien zur Bewertung des vereinfachten und des alltagssprachlichen Textes erklärt. Eine Übersicht über die zu bewertenden Aussagesätze der Kriterien befindet sich in der oberen Hälfte von Tabelle 1. Bei den Kriterien handelt es sich namentlich um:

1. Grammatikalität,
2. Einfachheit,
3. Kontextunabhängigkeit, und
4. Mehrdeutigkeit.

4.1.1 Grammatikalität

Ein vereinfachter Text sollte möglichst grammatikalisch korrekt sein. Falls ein Text in Alltagssprache schon Grammatikfehler aufweist, könnte es jedoch sein, dass diese in der Vereinfachung weiterhin bestehen geblieben sind. Es ist zu überlegen, ob ein Grammatikfehler ein Fehler des Vereinfachungssystems ist oder eine Folge des fehlerhaften Ausgangssatzes. Um dies abzuschätzen, wird die Grammatikalität des Ausgangstextes und die des vereinfachten Satzes unabhängig voneinander bestimmt. Eine umständliche aber grammatikalisch korrekte Formulierung eines Satzes, wird auch mit einem hohen Wert bewertet.

Die Bewertung soll anhand der folgenden Beispiele erfolgen: Enthält der Text keine Grammatikfehler, kann der Text mit dem höchsten Wert bewertet werden, denn der Aussage der grammatikalischen Korrektheit kann

zugestimmt werden. Enthält der Text zwar ein bis zwei unauffällige Grammatikfehler, kann aber noch sehr flüssig gelesen werden, so kann der Text immer noch mit einem hohen Wert bewertet werden. Der zu wählende Wert liegt im Entscheidungsspielraum der annotierenden Person. Es gilt allerdings zu beachten, dass die Satzstruktur bzw. der Satzbau nicht im Kriterium der Grammatikalität bewertet wird, sondern in dem Kriterium der Syntaktischen Vereinfachung.
Sind einige bis wenige Grammatikfehler in dem Text enthalten, kann ein Zwischenwert gewählt werden. Die Bewertung erfolgt inwiefern der Aussage der grammatikalischen Korrektheit zugestimmt bzw. widersprochen werden kann. Enthält ein Text sehr viele Grammatikfehler, so ist der Text mit dem niedrigsten Wert für Grammatikalität zu bewerten, denn so kann der Aussage der grammatikalischen Korrektheit nicht zugestimmt werden. Das gleiche Prinzip ist auf die weiteren Kriterien übertragbar.

4.1.2 Einfachheit

Falls ein Ausgangssatz schon sehr einfach ist, kann während der Vereinfachung nicht viel verbessert werden. Dementsprechend würde in diesem Fall ein hoher Einfachheitswert nicht auf direkt auf die Güte des Vereinfachungssystem zurückführbar sein. Der Wert würde lediglich angeben, dass das System den Ausgangssatz nicht komplizierter gemacht hat als er zuvor war. Falls ein Ausgangssatz jedoch sehr kompliziert ist und der automatisch generierte Vereinfachung einen hohen Einfachheitswert erhält, spricht dies für die Qualität des Vereinfachungssystems, da eine starke Vereinfachung stattgefunden hat.

Die Bewertung soll wie folgt erfolgen: Falls ein Text mehrmals gelesen werden muss, um ihn zu verstehen, so ist ein niedriger Wert auszuwählen. Falls ein Text beim ersten Lesen klar zu verstehen ist, so ist ein höherer Wert auszuwählen. Der zu wählende Wert liegt im Ermessen der annotierenden Person. Ist der Text allerdings sehr leicht zu verstehen, so ist der höchste Wert für Einfachheit auszuwählen. Zur Bewertung der Einfachheit können viele verschiedene Anhaltspunkte, wie Satzlänge, Satzstruktur, Bekanntheit der Wörter, allgemeine Verständlichkeit und viele mehr zu Rate gezogen werden.

4.1.3 Kontextunabhänigkeit

Desweiteren wird die Kontextunabhängigkeit beider Texte bewertet werden. Da der Fokus in der automatischen Textvereinfachung auf der Vereinfachung von Satzpaaren und nicht auf Textpaaren liegt, spielt der Kontext eines Satzes eine wichtige Rolle. Je weniger Kontext in einem Ausgangstext gegeben ist, desto schwieriger ist es diesen Kontext herzustellen, wenn weitere umliegende Sätze nicht mit angegeben sind. Diese Herausforderung wird durch das

Kriterium der Kontextunabhängikeit mit in die Bewertung eingeschlossen.

Für die Bewertung soll folgender Maßstab verwendet werden: Ist der Text nur verständlich, wenn vorherige und folgende Sätze hinzugezogen werden, so ist der geringste Wert für "Kontextunabhängigkeit" auszuwählen. Falls beispielsweise Pronomen (z.B., er, ihr, dieses) oder deiktische Adverbien (z.B., dort, da, hierhin, unten, gestern, vorher, bald) in dem Text enthalten sind, dessen Bezug innerhalb des Satzes nicht ersichtlich wird, ist ein niedriger Wert auszuwählen. Der zu wählende Wert liegt im Ermessen der annotierenden Person.

4.1.4 Mehrdeutigkeit

Ziel eines vereinfachten Textes ist es möglich eindeutig zu sein und etwaige Mehrdeutigkeiten eines Ausgangstextes aufzulösen. Mithilfe des Kriteriums "Mehrdeutigkeit" fließt dieser Punkt ebenfalls mit in die Bewertung ein.

Hat der Text mehrere Lesarten, so ist er mit dem höchsten Wert für Mehrdeutigkeit zu bewerten. Hat ein Wort des Textes zwar mehrere Lesarten, aber innerhalb des Textes wird dessen Lesart deutlich, so ist ein niedriger Wert auszuwählen. Der auszuwählende Wert liegt im Ermessen der annotierenden Person.

4.2 Bewertung des Satzpaares

Neben der unabhängigen Bewertung des vereinfachten und des alltagssprachlichen Textes des alinierten Satzpaares, kann auch der vereinfachte Text in Abhängigkeit zu dem alltagssprachlichen Text bewertet werden. Die folgenden 5 Kategorien werden bewertet:

1. Sinnerhaltung
2. Informationsgewinn
3. Vereinfachung
4. Syntaktische Vereinfachung
5. Lexikalische Vereinfachung

Entgegen der Einfachheitsbewertung der einzelnen Texte des Paares, wird in dem Kriterium der Vereinfachung die Änderung der Einfachheit durch die Vereinfachung bestimmt. Die Distanz der Bewertung des vereinfachten Textes und der Bewertung des alltagssprachlichen Textes, kann ergänzend als Änderung der Einfachheit berechnet werden. Die Bewertung der Vereinfachung wird nochmal spezifiziert in die Vereinfachung des Satzbaus (Syntaktische Vereinfachung) und der Wortwahl (Lexikalische Vereinfachung). Manche Textvereinfachungssysteme sind entweder auf die lexikalische oder syntaktische Vereinfachung spezialisiert, daher bietet sich die Bewertung der Paare hinsichtlich dieser Bewertungskritierien ebenfalls an.

item	Statement
Grammaticality	The simplified sentence is fluent, there are no grammatical errors.
Grammaticality (original)	The original sentence is fluent, there are no grammatical errors.
Simplicity (simple)	The simplified sentence is easy to understand.
Simplicity (original)	The original sentence is easy to understand.
Coherence (simple)	The simplified sentence is understandable without reading the whole paragraph.
Coherence (original)	The original sentence is understandable without reading the whole paragraph.
Ambiguity (simple)	The simplified sentence is ambiguous. It can be read in different ways.
Ambiguity (original)	The original sentence is ambiguous. It can be read in different ways.
Meaning Preservation	The simplified sentence adequately expresses the meaning of the original sentence, perhaps omitting the least important information.
Information Gain	In the simplified sentence, information is added or get more explicit compared to the original sentence.
Overall Simplicity	The simplified sentence is easier to understand than the original sentence.
Structural Simplicity	The structure of the simplified sentence is easier to understand than the structure of the original sentence.
Lexical Simplicity	The words of the simplified sentence are easier to understand than the words of the original sentence.

Tabelle 1: Englische Bewertungskriterien.

4.2.1 Sinnerhaltung

Nur in Abhängigkeit zum alltagssprachlichen Text kann bestimmt werden, ob sich der Sinn bzw. der Inhalt im vereinfachten Text geändert hat oder nicht. Die Bewertung der Sinnerhaltung wird dadurch erschwert, dass ein Bestandteil der Vereinfachung die Wort- oder Satzteilauslassung sein kann. Falls nur wenige nicht für das Verständnis des Textes relevante Passagen

weggelassen werden, so kann die Sinnerhaltung mit einem hohen Wert bewertet werden. Fehlen jedoch wichtige, unerlässliche Informationen in dem vereinfachten Text, so ist das Paar mit einem niedrigen Sinnerhaltungswert zu bewerten. Welche Informationen als unerlässlich zu betrachten sind, liegt allerdings im Ermessen der annotierenden Person.

4.2.2 Informationsgewinn

Im Bereich der manuellen Textvereinfachung werden besonders oft komplizierte Wörter in einem Einschub oder in einem neuen Satz erklärt. Desweiteren wird der Inhalt, welcher nur zwischen den Zeilen lesbar ist, in der vereinfachten Variante oft expliziter formuliert. Im Kriterium des Informationsgewinn wird daher bewertet, ob neue Informationen zum alltagssprachlichen Text hinzugefügt worden sind, bzw. ob Informationen expliziter formuliert worden sind. Falls eines von beidem der Fall ist, so ist ein hoher Wert auszuwählen. Falls jedoch Informationen weggelassen worden sind oder implizierter formuliert wurden, so ist der Text mit einem niedrigen Wert zu bewerten. Wurden weder Informationen hinzugefügt, noch weggelassen, so ist der mittlere Wert zu wählen, denn bei dem Kriterium des Informationsgewinns handelt es sich um ein Kriterium mit neutralem Element. Ob neue Informationen hinzugefügt wurden oder die vorhandene Information nur umgeschrieben wurde, liegt hier wieder im Ermessen der annotierenden Person.

4.2.3 Vereinfachung

Die Bewertung der Vereinfachung soll nach den gleichen Kriterien wie die Bewertung der Einfachheit erfolgen. Allerdings handelt es sich bei dem Kriterium der Vereinfachung um ein Kriterium mit neutralem Element. Ist der vereinfachte Text deutlich besser zu verstehen als der alltagssprachliche Text, so ist die Vereinfachung mit dem höchsten Wert zu bewerten. Ist der vereinfachte Text gleich kompliziert wie der alltagssprachliche Text, so ist der mittlere Wert auszuwählen. Der niedrigste Wert ist auszuwählen, wenn der vereinfachte Text deutlich schwieriger ist als der alltagssprachliche Text. Die Zwischenwerte sind auszuwählen, wenn der vereinfachte Text etwas schwieriger oder etwas leichter, aber nicht deutlich schwieriger oder leichter ist, als der alltagssprachliche Text. Der zu wählende Wert liegt jedoch im Ermessen der annotierenden Person.

4.2.4 Syntaktische Vereinfachung

Zur Bewertung der syntaktischen Vereinfachung zählen zum Beispiel die Auftrennung von Nebensätzen in neue Hauptsätze, die Umformulierung von Passivsätzen in Aktivsätze, oder die Änderung hinzu der Reihenfolge Subjekt-Verb-Objekt. Falls eine dieser Veränderungen stattgefunden hat und den Satz vereinfacht hat, kann einer der hohen Werte für die Bewertung der

syntaktischen Vereinfachung ausgewählt werden. Wird ein Satz jedoch von Indikativ zu Konjunktiv geändert, ist der Satz vermutlich komplizierter geworden und ein niedriger Wert wäre angemessen. Hat sich die Komplexität des Satzbau oder der Satzstruktur nicht verändert, soll der mittlere Wert gewählt werden. Auch wenn eine Änderung der Satzstruktur stattgefunden hat, aber die syntktische Komplexität sich nicht geändert hat, ist der mittlere Wert zu wählen. Wurden Grammatikfehler behoben, kann der Satzbau und die Satzstruktur des vereinfachten Satzes einfacher sein als die des alltagssprachlichen Satzes, dies ist davon abhängig wie gravierend die Grammatikfehler waren bzw. wie sehr sie den Lesefluss gestört haben. Auch hier liegt der auszuwählende Wert im Ermessen der annotierenden Person. Das Kriterium der syntaktischen Vereinfachung ist, ebenfalls wie das Kriterium der Vereinfachung, ein Kriterium mit neutralem Element.

4.3 Lexikalische Vereinfachung

Die lexikalische Vereinfachung bezieht sich im Kontrast zur syntaktischen Vereinfachung auf die Wortwahl des Satzes und nicht auf die Satzstruktur. Wurden während der Vereinfachung schwierige Worte durch einfachere ersetzt oder wurden schwierige Wörter weggelassen, so ist ein hoher Wert für lexikalische Vereinfachung auszuwählen. Wurden hingegen neue schwierige Wörter hinzugefügt, so ist ein geringerer Wert zu wählen. Falls die Wortwahl insgesamt weder komplizierter noch einfacher geworden ist, so sollte der mittlere Wert ausgewählt werden. Die Entscheidung, ob ein Wort schwierig oder einfach ist, liegt im Ermessen der annotierenden Person. Das Kriterium der lexikalischen Vereinfachung ist, ebenfalls wie das Kriterium der Vereinfachung und der syntaktischen Vereinfachung, ein Kriterium mit neutralem Element.

5 Transformationsannotation

5.1 Allgemein

Bei einem Vergleich bzw. bei der Bewertung der alinierten Satzpaare wird die Veränderung bzw. der Vereinfachungssprozess deutlich. Die einzelnen Schritte der Veränderung werden im Folgenden Transformationen genannt. Die Transformationen helfen dabei zu überprüfen, welche Veränderungen bei einer manuellen Vereinfachung vorgenommen wurden und, zudem, welche Veränderungen bei einer automatischen Vereinfachung möglich wären.

Basierend auf einer Literaturrecherche im Bereich der Textvereinfachung und der Einfachen Sprache, wurden 56 Transformationen als besonders relevant eingestuft. Die Transformationen können zu den folgenden 9 Transformationsklassen zusammengefasst werden:

1. Lexikalische Ersetzung (lexical substitution),

2. Aufteilung (split),
3. Entfernung (deletion),
4. Hinzufügung (insert),
5. Umordnung/Reihenfolgenänderung (reorder),
6. Verbänderungen (verbal changes),
7. Zusammenfügung (merge),
8. Umschreibung/Paraphrasierung (rephrasing, paraphrasing), und
9. keine Veränderung (no operation)

Die Transformationsklassen sind allgemeine Bezeichnungen für lexikalische oder syntaktische Änderungen, die während der Vereinfachung vorgenommen wurden. Die Transformationsklassen können durch 38 Unterklassen spezifiziert werden. Desweiten lassen sich die 9 Transformationsklassen in vier verschiedene Ebenen einteilen. Die Ebenen sind Textebenen, die bei der Vereinfachung betroffen sind. Zu den Kategorien zählen:

1. Wortebene,
2. Satzteilebene,
3. Satzebene, und
4. Absatzebene

Eine Transformationsklasse kann auf mehr als einer Ebene vorkommen, eine Hinzufügung ist beispielsweise auf Wort-, Satzzeil- und Absatzebene möglich.

In Abbildung 1 ist eine Übersicht über alle Transformationsklassen und alle Unterklassen sowie ihre zugehörigen Textebenen. Die 9 Textklassen wurden auf jeder Ebene dargestellt auf der sie auftreten sodass insgesamt 18 Klassen (fett-gedruckt) auf der Abbildung zu sehen sind. Die 38 Unterklassen wurden ihren Oberklassen mit Pfeilen zugeordnet. Die Abbildung enthält zudem auch Informationen über die Relevanz bzw. Nennungshäufigkeit in der Literatur. Je höher der Prioritätswert desto seltener wurde die Transformation genannt. In dem zu erstellendem Korpus werden die Transformationsklassen und ihre Ebene annotiert. Die Unterklassen dienen als Beispiele für die zugehörigen Transformationsklassen. Falls für die Vereinfachung des Satzes mehrere Veränderungen gleichzeitig vorgenommen worden sind, können auch mehrere dieser Optionen gewählt bzw. annotiert werden. Sollte keine der gewählten Optionen zutreffen, kann der Nutzer weitere Optionen hinzufügen.

Neben der Transformation werden auch die dazugeörige Token annotiert. Eenn kein klares Token identifiziert werden kann, wird keins ausgewählt. Es

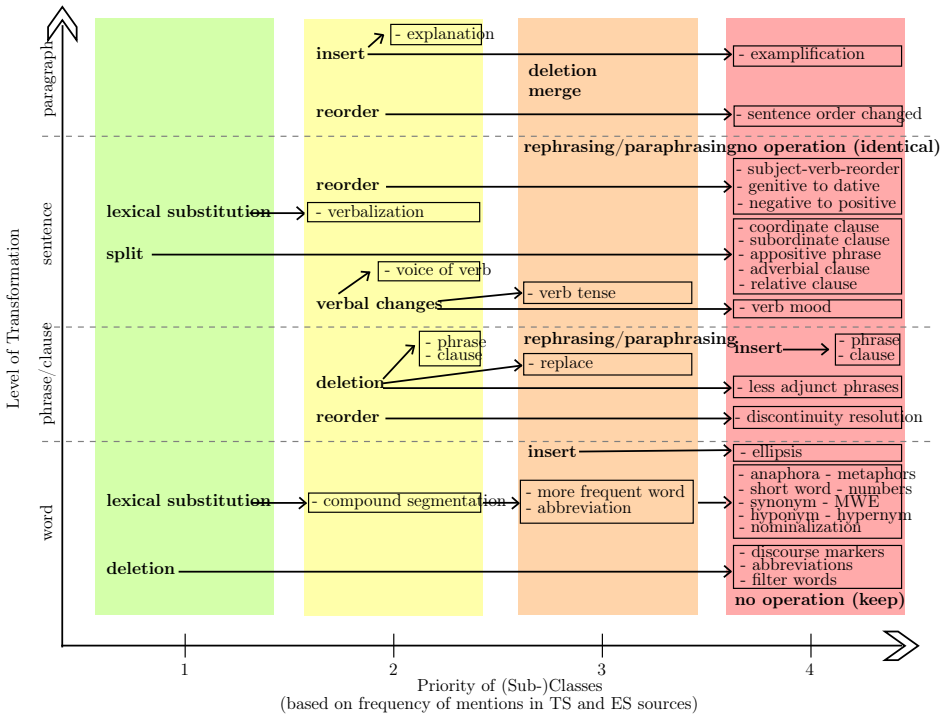

Abbildung 1: classes-subclasses-level-priority

sollen nur offensichtliche Transformationen annotiert werden. Falls der Text vereinfacht wurde aber nicht klar differenziert werden kann, was vereinfacht wurde, muss keine Transformation annotiert werden. Die Annotation der Token erfolgt via Klick auf die zugehörigen Token in dem angezeigten Text. Je nach dem in welchem Text, vereinfacht oder alltagssprachlich, die Veränderung stattgefunden hat, werden die Token markiert. Eine Hinzufügung kann zum Beispiel nur im vereinfachten Satz markiert werden, eine Entfernung hingegen nur im alltagssprachlichen Text. Eine lexikalische Ersetzung oder Veränderung der Verbform wird in beiden Texten annotiert, um die Veränderung nachvollziehbar zu machen auch wenn mehrere Transformationen gleichzeitig stattgefunden haben.

Im Annotationstool werden zunächst die Ebenen angezeigt. Mit Auswahl der Ebene öffnet sich eine Auflistung der Transformationsklassen. Mit Auswahl einer Transformationsklasse öffnet sich erneut eine Auflistung, diesmal eine der einzelnen Transformationen. Die Auflistung enthält zudem ein freies Textfeld über das eine weitere Transformation spezifiziert werden kann. Derzeit öffnen sich die Auflistung nur, wenn der Radiobutton gedrückt wird. Leider öffnet es sich derzeit noch nicht, wenn auf den Namen einer Klasse geklickt wird.
Wurden in einer Vereinfachung zwei Wörter (auf Wortebene) durch leichtere

ersetzt, so sind zwei Transformationen anzulegen, für jede Ersetzung eine Transformation. Handelt es sich um eine Ersetzung einer Mehrworteinheit (z.B. “ins Gras beißen” oder “läuft ... weg”) ist jedoch eine Transformation geeignet, die alle Token der Mehrworteinheit einschließt.

5.2 Transformationen

Die einzelnen Transformation werden im folgenden nach Ebene des Auftretens und Transformationsklasse sortiert erklärt.

5.2.1 Absatzebene

Die Absatzebene enthält Transformationen, die entweder mehr als einen alltagssprachlichen Satz involvieren (n:1 Paar, Umordnung oder Zusammenfügung), keinem alltagssprachlichen Text zugeordnet werden kann (0:1 Paar, Hinzufügung), oder keinem vereinfachten Satz zugeordnet werden kann (1:0 Paar, Entfernung). Alle Transformationen auf Absatzebene werden automatisch annotiert. Eine Zusammenfügung (merge) ist beispielweise daran erkennbar, dass der alltagssprachliche Text aus mindestens zwei Sätzen besteht wohingegen der vereinfachte Texte nur aus einem Satz besteht. Die Hinzufügung (insert) und Entfernung (deletion) werden dann annotiert, wenn sie nach dem Alinierungsprozess keinem Satz im anderen Sprachlevel zugeordnet wurden. Die Umordnung (reorder) wird anhand der Satznummern automatisch annotiert: Wurde beispielsweise ein alltagssprachlicher Satz an Position x (wo $1 \leq x \leq n$ und n = Anzahl aller Sätze des alltagssprachlichen Dokumentes) mit einem vereinfachten Satz an Position x+j (wo $j > i$) annotiert und befinden sich die Sätze eines anderen alinierten Paares an Position x+i (wo $i \geq 1$), so wurde die Reihenfolge der Sätze geändert.

5.2.2 Satzebene

Auf Satzebene befinden sich alle Transformationen, die innerhalb der Grenzen eines Satzes angewendet werden. Die einzige Ausnahme ist die Aufteilung (split). Diese Transformation betrifft einen alltagssprachlichen Satz aber mehrere vereinfachte Sätze. Die Satzebene beinhaltet zudem nur Transformationen, die die gesamte Struktur des Satzes betreffen. Heißt: Änderungen, die nur einem Satzteil oder einem Token zugeordnet werden können, aber nicht den kompletten Satzbau beeinflussen, werden nicht zu dieser Ebene zugeordnet. Die Transformationen dieser Kategorie können größtenteils zur syntaktischen Vereinfachung gezählt werden.

keine Veränderung (no operation) Die speziellste Transformation auf Satzebene ist die, in der keine Veränderung (no operation) stattgefunden hat. Diese Transformation wird automatisch annotiert.

Aufteilung (split) Die wohl am häufigsten auftretende Transformation auf Satzebene ist die Aufteilung (split). Diese Transformation könnte ebenfalls automatisiert annotiert werden, allerdings ist der Grund der Aufteilung von Interesse. Mögliche Gründe sind z.B. die Abtrennung eines Relativsatzes (beginnend mit z.B.: der, die, das, welcher, etc.), eines Adverbialsatzes (beginnend mit z.B.: indem, wenn, worum, obwohl, während), eines untergeordneten Nebensatz (beginnend mit z.B.: weil, dennoch, aber, deshalb, etc.), oder eines nebengeordneten Nebensatz (beginnend mit z.B.: und, oder, sowohl..als...auch, weder..noch, etc.). Bei der Transformation der Aufteilung reicht es aus, die Transformation zu annotieren, eine Markierung der Token ist nicht notwendig.

Umordnung/Reihenfolgenänderung (reorder) Eine Umordnung kann nicht nur auf der Absatzebene sondern auch auf der Satzebene stattfinden. Zu einer Umordnung auf Satzebene zählen zum Beispiel die Verschiebung von einzelnen Satzteilen (general), die Änderung der Reihenfolge von Subjekt, Verb und Objekt (subject verb reorder), die Änderung einer negativen Formulierung hinzu einer positiven (negative to postive), oder die Ersetzung des Genitivs mit dem Dativ (genitive to dative). Die Annotation der zugehörigen Token sollte hier versucht werden durchzuführen. Falls zu viele Token involviert sind oder sie nicht klar identifiziert werden können, kann auf eine Annotation der Token verzichtet werden.

Verbänderungen (verbal changes) Veränderungen des Verbes werden ebenfalls auf Satzebene annotiert, da die Satzstruktur sich beispielsweise durch die Aktivierung eines Satzes stark verändern kann. Neben der Änderung von passiv zu aktiv (voice of verb), befinden sich in dieser Transformationsklasse die Änderung des Modus (mood, Konjunktiv und Indikativ) und die Änderung der Zeitform (verb tense). Es ist davon auszugehen, dass in dieser Transformationsklasse die zugehörigen Token relativ eindeutig sind, sodass zusätzlich mit annotiert werden sollen.

Lexikalische Ersetzung (lexical substitution) Eine weitere Transformationsklasse, die der lexikalischen Ersetzung, enthält Änderungen bezüglich Verben. Häufig werden Sätze im nominal Stil formuliert, wurden diese in der Vereinfachung zu einem verbalen Stil geändert, so kann die Transformation der Verbalisation (verbalization) annotiert werden.

Umschreibung/Paraphrasierung (rephrasing, paraphrasing) Falls keine klare(n) Tranformation(en) der Vereinfachung zu erkennen sind, da beispielsweise zu viele Worte geändert und/oder verschoben wurden, kann die Vereinfachung mit Umschreibung bzw. Paraphrasierung gekennzeichnet

werden. In dieser Transformation wird nicht erwartet, dass die zugehörigen Token annotiert werden.

5.2.3 Satzteil- und Nebensatzeben

Eine Spezifikation der Satzebene ist die Satzteil- und Nebensatzebene. Auf dieser Ebene ist nicht der ganze Satz in die Veränderung involviert, sondern es sind klare, abgegrenzte Wortgruppen identifizierbar. Zu den Wortgruppen zählen vorallem Nebensätze und andere Satzteile.

Entfernung (deletion) Die Satzteile können zum Beispiel komplett weggelassen werden, falls sie nicht relevant genug für den vereinfachten Text sind. Hierbei ist zu unterscheiden, ob der Satzteil komplett weggelassen wurde oder in einem neuen Satz aufgegriffen wird. Im Falle von zweiterem ist die Satzaufteilung die zutreffende Transformation. Bei ersterem jedoch die aktuelle. Beispiele für überflüssige oder wegzulassende Satzteile sind Phrasen, Nebensätze oder andere Einschübe. Die betreffenden Token sollten in dieser Transformation auch annotiert werden.

Hinzufügung (insert) Die gegenteilige Transformation zur Entfernung ist die Hinzufügung. Es können zum Beispiel Phrasen oder Nebensätze hinzugefügt werden. Die betreffenden Token sollten in dieser Transformation auch annotiert werden.

Umordnung/Reihenfolgenänderung (reorder) Ebenso wie auf der Satzebene können TOken innerhalb eines Satzteiles verschoben werden. Dies beinhaltet zum Beispiel auch, wenn verbundene Worte näher aneinander gerückt werden (discontinuity resolution). Ein Beispiel für die Auflösung von Entfernungen ist die Umordnung von Verben und zugehörigen Partikeln. Eine Annotation der betroffenen Token empfiehlt sich auch hier.

Umschreibung/Paraphrasierung (rephrasing, paraphrasing) Eine Umschreibung bzw. Paraphrasierung eines Satzteils ist ebenfalls möglich. Wie auch schon auf Satzebene müssen innerhalb dieser Transformation keine zugehörigen Token annotiert werden.

5.2.4 Wortebene

Die feinste Ebene der Transformationen ist die der Wortebene. In dieser Ebene beziehen sich die Transformationen auf einzelne Worte und Mehrworteinheiten. Die Transformationen dieser Kategorie können größtenteils zur lexikalischen Vereinfachung gezählt werden.

Lexikalische Ersetzung (lexical substitution) Ein weit erforschter Bereich der lexikalischen Vereinfachung ist der der komplexen Worterkennung und -ersetzung. Diese Art der Änderung wird mit der Transformation der lexikalischen Ersetzung gekennzeichnet. Für eine lexikalische Ersetzung können viele Gründe vorliegen. Die folgende List fasst die häufigsten Gründe und häufigsten ausgewählten Alternativen zusammen:

- Kompositazerlegung (compound segementation): Trennung bzw. Unterteilung von Wortteilen innerhalb eines Wortes mit Bindestrichen
- Ersetzung eines seltenen Worted durch ein häufigeres Wort (more frequent word)
- Ersetzung eines komplexten Wortes durch ein einfacheres (simpler word in general)
- Ersetzung einer Abkürzung durch die Langfassung (abbrevation vs. long version)
- Ersetzung von Anaphoren durch die zugehörigen Referenten (anaphora)
- Ersetzung eines langen Wortes durch ein kürzeres Wort (shorter word)
- Ersetzung eines schwierigen Wortes durch ein Synonym (synonym)
- Ersetzung eines schwierigen Wortes durch einen Unterbegriff (hyponym)
- Ersetzung eines schwierigen Wortes durch einen Überbegriff (hypernym)
- Ersetzung eines Begriffs durch seine nominale Version (nominalization)
- Ersetzung von Metaphern und bildlicher Sprache (metaphor)
- Ersetzung von Zahlwörtern durch Ziffen und andere Zahlen betreffende Ersetzungen (number)
- Ersetzung von Mehrworteinheiten durch ein Wort (MWE)

Für jede der genannten Transformationen sollten die zugehörigen Token auch annotiert werden.

Eine lexikalische Ersetzung kann auch als Verbindung der Transformationen der Entfernung eines Wortes und der Hinzufügung eines Wortes gesehen werden. Die Transformation einer Entfernung oder der Hinzufügung wäre allerdings nicht ausreichend, da über die lexikalische Ersetzung ausgedrückt wird, dass ein inhaltsgleicher oder inhaltsähnlicher Austausch vorgenommen wurde.

Entfernung (deletion) Auch auf Wortebene befindet sich die Transformationsklasse der Entfernung. Mögliche Transformationen oder Auslassungen sind z.B.: Abkürzungen (abbreviation removal), Füllwörter (filter words) oder überflüssige komplizierte Wörter (complex words). Hierbei ist zu beachten, dass die Token nicht mehr im vereinfachten Text wieder auftauchen, befindet sie sich nur an einer anderen Stelle, ist die Umordnung-Transformation die richtige. Desweiteren können Diskursmarker, die eine Verbindung zwischen zwei oder mehreren Sätzen herstellen entfernt werden (discourse marker). Das weggelassene Token ist im Annotationstool zu markieren bevor die Bewertung abgeschickt wird.

Hinzufügung (insert) Obwohl kürzere Texte bessre verständlich sind, können in Vereinfachungen auch Wörter hinzugefügt werden. In Ellipsen in alltagssprachlichen Texten könnten zum Beispiel ihre Auslassungen ergänzt werden (filling ellipsis). Zudem ist es möglich, dass ein Satz explizierter formuliert wurde und während dessen neue Wörter hinzugefügt wurden (new words). In allen Transformationsarten wird darum gebeten, die betreffenden Token hinzuzufügen.

keine Veränderung (no operation) Wurde keine Veränderung an einem Wort vorgenommen, so wird dies auch annotiert. Allerdings geschieht dies automatisch. Die Information, dass ein Wort nicht geändert wurde ist dahingehend von Bedeutung, dass das Wort scheinbar verständlich genug ist und nicht ersetzt werden muss.

6 Zusammenfassung

Zusammenfassend wird in diesem Annotationsprojekt ein Korpus erstellt, der zur Erstellung eines Vereinfachungssystem verwendet werden kann. Die Annotation besteht zum einen aus der Alinierung von inhaltsgleichen Textpaaren aus verschiedenen Textleveln. Zum anderen werden die alinierten Paare nach ihrer Qualität beurteilt und der Prozess der Vereinfachung in Form von Transformationsannotationen festgehalten. Zur Unterstützung der Annotation wird ein Annotationstool angeboten. Dieses Dokument gilt als Handreichung zur Bedienung des Annotationstools sowie zur Anleitung der Annotation. Fragen oder Anmerkungen hinsichtliche des Annotationsprozesses oder des Annotationstool können an die Verfassenden dieses Dokument gesendet werden.

Literatur

A Anhang

A.1 Bewertungskriterien

item	Aussage
Grammatikalität	Der vereinfachte Satz klingt flüssig, er enthält keine Grammatikfehler.
Grammatikalität (Ausgangssatz)	Der Ausgangssatz klingt flüssig, er enthält keine Grammatikfehler.
Sinnerhaltung	Der vereinfachte Satz gibt auf angemessene Weise die Bedeutung des Ausgangssatzes wieder. Weniger wichtige Informationen werden dabei möglicherweise weggelassen.
Informations-gewinn	In dem vereinfachten Satz sind neue Informationen enthalten oder deutlicher ausgedrückt. Diese Informationen sind in dem Ausgangssatz nicht ausdrücklich enthalten.
Einfachheit (Allgemein)	Der vereinfachte Satz ist leichter zu verstehen als der Ausgangssatz.
Einfachheit (Satzstruktur)	Der Satzbau und die Satzstruktur im vereinfachten Satz sind leichter zu verstehen als der Satzbau und die Satzstruktur im Ausgangssatz.
Einfachheit (Wortwahl)	Die Wörter im vereinfachten Satz sind leichter zu verstehen als die Wörter im Ausgangssatz.
Einfachheit (Leichter Satz)	Der vereinfachte Satz ist leicht zu verstehen.
Einfachheit (Ausgangssatz)	Der Ausgangssatz ist leicht zu verstehen.
Kontext-unabhängigkeit (Leichter Satz)	Der vereinfachte Satz kann verstanden werden ohne den ganzen Absatz zu lesen.
Kontext-unabhängigkeit (Ausgangssatz)	Der Ausgangssatz kann verstanden werden ohne den ganzen Absatz zu lesen.
Mehrdeutigkeit (Leichter Satz)	Der vereinfachte Satz ist mehrdeutig. Er kann auf verschiedene Weisen verstanden werden.
Mehrdeutigkeit (Ausgangssatz)	Der Ausgangssatz ist mehrdeutig. Er kann auf verschiedene Weisen verstanden werden.

Tabelle 2: Deutsche Bewertungskriterien.