Universität Zürich[UZH]

Institut für Computerlinguistik

Bachelorarbeit an der Philosophischen Fakultät der Universität Zürich

Rule-based Text Simplification for German

Regelbasierte Textvereinfachung für Deutsch

Referent:

Prof. Dr. Martin Volk

Betreuerin:

Sarah Ebling

Verfasserin:

Julia Suter

Matrikelnummer 11-726-262

July 15, 2015

Abstract

Plain Language is a linguistic phenomenon aimed at making texts comprehensible and accessible to everyone, including people with low literacy skills. Plain Language is characterised by reduced lexical and syntactic complexity, explanations of difficult words and a clearly structured layout. During the past few years, Automatic Text Simplification has gained in importance and systems for generating Plain Language have been developed for several languages. However, no such system has been created for the simplification of German texts. Thus, I have developed a rule-based Automatic Text Simplification system that translates standard German to Simple German. Since it is based on the syntactic parsing analysis of the source text, it is focused on syntactic simplification, although it also accesses additional resources and tools to reduce lexical complexity and provide complementary information. For a short example text, the system was able to generate a well-comprehensible translation that is comparable in syntactic complexity to its human reference translation. Further development of the system is required to reduce lexical complexity and improve syntactic simplification.

Zusammenfassung

Leichte Sprache hat zum Ziel, Texte für alle verständlich und zugänglich zu machen, auch für Menschen mit Leseschwierigkeiten. Leichte Sprache weist eine geringe lexikalische und syntaktische Komplexität auf. Schwierige Wörter werden erklärt und der Text wird klar strukturiert dargestellt. In den letzten Jahren hat sich Automatische Textvereinfachung etabliert und für verschiedene Sprachen wurden Systeme zur Automatischen Generierung von Leichter Sprache entwickelt. Es gibt jedoch noch kein System zur Vereinfachung von deutschen Texten. Daher habe ich ein regelbasiertes System zur Automatischen Textvereinfachung entwickelt, das deutsche Alltagssprache in Leichte Sprache übersetzt. Da es auf der syntaktischen Analyse des Ausgangstextes basiert, nimmt es hauptsächlich syntaktische Vereinfachungen vor. Allerdings habe ich weitere Tools und Resourcen eingebaut, die die lexikalische Komplexität verringern und zusätzliche Informationen liefern. Mein System kann für einen kurzen Beispieltext eine gut verständliche Übersetzung erzeugen, die von der syntaktischen Komplexität her mit der Experten-Übersetzung verglichen werden kann. Das System muss noch weiterentwickelt werden, um den Ausgangstext auf lexikalischer Ebene stärker zu vereinfachen und die syntaktische Vereinfachung zu verbessern.

Acknowledgement

I would like to thank all the people who helped and supported me during my Bachelor thesis.
First of all, I would like to thank Sarah Ebling who introduced me to the topic of Plain Language and agreed to supervise my work, even though she is currently working abroad. I am grateful for her swift email responses, her advice and feedback, and her encouraging words. I would also like to thank Martin Volk, who took the time to read the first draft of my work and functions as supervisor of my Bachelor thesis.
Special thanks to Bettina Ledergerber and Patrizia Napoli from Pro Infirmis Zurich who invited me to the Büro für Leichte Sprache for an interesting discussion. They helped me see beyond the linguistic side of Plain Language and introduced me to its social aspects. I would also like to thank France Santi for providing Simple German translations on short notice.
I am grateful for the technical support and patience of Adrian van der Lek and Matthias Fluor. And I would like to thank Jonas Hartmann for encouraging me and proof-reading my thesis, even at late hours.
Last but not least, I would like to thank my mother who took great interest in my thesis and discussed linguistic aspects of Simple German with me over dinner. Thanks to both my mother and brother for supporting and motivating me throughout my studies.

List of Acronyms

ATS	Automatic Text Simplification
CNL	Controlled Natural Language
NLP	Natural Language Processing
PBMT	Phrase-based Machine Translation
POS	Part-of-speech
SMT	Statistical Machine Translation
TS	Text Simplification

Table of Contents

1 Introduction

Information is one of the most important resources of our time. However, this resource is not accessible to everyone. Complex language keeps important information away from people with reading difficulties and excludes them from society. A special kind of language is needed to guarantee access to information for everyone. This is called Plain Language.

Plain Language or Easy-to-read Language is a linguistic phenomenon that aims at making texts understandable for people who have difficulties reading and processing written language. This includes not only cognitively impaired people but also functional illiterates, prelingual deafs, people suffering from dementia or other neurodegenerative diseases, and immigrants. Texts written in complex language form a considerable obstacle and hinder access to important information and knowledge. Accessible information plays an essential role in the inclusion process of people with physical or mental disabilities into society. Therefore, many organisations that support handicapped people encourage the use of Plain Language, provide information about their organisation and everyday life in Plain Language or even offer translation services. The United Nations Convention on the Rights of Persons with Disabilities (*germ.* UN-Behindertenrechtskonvention)[1] states that important information must be accessible to persons with disabilities in a form that enables them to understand their rights, participate in society and make decisions on their own. Plain Language is the key to accessibility in communication and has therefore become a modern research topic in linguistics, cognitive science and social studies.

Plain Language is characterized by low lexical and syntactic complexity. The sentences are short and usually contain only one piece of information at a time. Difficult vocabulary and complex syntactic structures are avoided. The structure of the text is emphasised with visual markers (headlines, indentations) and examples and accompanying pictures explain complicated issues. Most texts written in Plain Language have been translated from an original source written in standard language. Various guidelines suggest rules for generating Plain Language, although these rules are dis-

[1] `http://www.behindertenrechtskonvention.info/`

puted by experts and remain subject of research. Translations into Plain Language are usually performed by a trained translator or person working for an organisation for people with disabilities.

While Automatic Text Simplification has been developed for several languages (e.g. English, Swedish, Portuguese), there exists no equivalent approach for the simplification of German texts. This Bachelor thesis is a first attempt at filling this gap. I aim to develop a rule-based system for Automatic Text Simplification for German that translates a text written in standard German into Simple German, or at least a simplified version of German. My goal is to examine if and how rule-based Text Simplification can produce Plain Language and what rules for lexical and syntactic simplification are most relevant and suitable for the implementation. My Text Simplification approach is built on the basis of syntactic parsing of the source text and rules extracted from the various guidelines of Simple German that simplify the structure of complex sentences. In addition I will include various resources that help simplify the text on the lexical level.

In Chapter 2, I will present the concept of Plain Language in general, discuss the target groups and then focus on Simple German (Leichte Sprache). Chapter 3 demonstrates the guidelines for simplifying German texts. In Chapter 4, I will explain the purpose and challenges of Text Simplification, followed by a brief overview on previous work done on the field. In Chapter 5, I will present my own work on Automatic Text Simplification for German. I will start by introducing the auxiliary tools and then present the simplification rules I selected and briefly explain the implementation of each. I will evaluate my system and discuss the results in Chapter 6, and conclude my work in Chapter 7.

2 Plain Language

2.1 Definition

Plain Language is written communication that is understandable by all, including readers with low literacy skills. It presents information in a well-structured and clearly written way and explains difficult words and concepts using examples and pictures. Plain Language stands out by reduced lexical and syntactic complexity, meaning that only basic vocabulary and simple syntactic structures are allowed. There is preferably only one sentence per line, printed in a large, sans-serif font. Headlines, clear paragraphs and indentations emphasise the structure of the text [Netzwerk Leichte Sprache, 2009], [Maass et al., 2014, 61-74].

Plain Language has gained in importance during the last 50 years, but it is still a very young field of research and the lack of clear definitions can cause confusion when discussing the topic. For example, the terms *Plain Language*, *Easy-to-read Language* and *Simple Language* are often used interchangeably, even though they (can) represent different concepts. *Simple Language* usually refers to language that uses short sentences and simpler grammar and vocabulary than standard language, while retaining the complex information given in the text [Kellermann, 2014]. Simple English Wikipedia[1] is an example for Simple Language: It is written in basic English, a controlled language that only uses the 850 most basic English words [Ogden, 1944] and is aimed at students and English language learners as well as people with reading difficulties. The large number of articles (over 100'000) shows the desire and need for information in Simple Language. Simple Language, however, can still be too difficult to read for people with poor literacy skills. Plain Language shows even further reduced complexity in grammar and vocabulary and explains difficult concepts using additional information and examples. An optically well-structured layout facilitates the reading process. Professor Robert Eagleson, cofounder of the Center for Plain Legal Language at the University of Sydney and author of *Writing in Plain English*, defines Plain English as follows:

[1] `https://simple.wikipedia.org/wiki/Main_Page`

> Plain English is clear, straightforward expression, using only as many words as are necessary. It is language that avoids obscurity, inflated vocabulary and convoluted sentence construction. It is not baby talk, nor is it a simplified version of the English language. Writers of plain English let their audience concentrate on the message instead of being distracted by complicated language. They make sure that their audience understands the message easily [Eagleson, 1990, 4].

Plain Language falls under the category of Controlled Natural Languages (CNL), languages that are restricted in grammar and dictionary in order to reduce complexity and ambiguity. CNL can be divided into two groups: Those aiming at producing texts with enhanced readability for human readers (such as Plain Language), and those focusing on improving performance of Natural Language Processing tasks[2] [Kuhn, 2014].

Although the term *Plain Language* is often used when it comes to simplifying legal, governmental and medical texts, organisations focusing on mentally disabled people seem to prefer the term *Easy-to-read Language*. The Center for Easy-to-Read in Sweden, Inclusion Europe, People First (see Chapter 2.2) all exclusively use the term *Easy-to-read* (in English texts). I could not detect any difference in the definitions of the two terms but consider it important to notice the varying use, depending on the environment.

The German term to describe the concept of Plain Language is *Leichte Sprache*, which translates to *light language* or *easy language*. While *Plain English* seems to be prominent in literature on accessibility in communication, I could not find any mention of the term *Plain German* and will therefore use *Simple German* whenever I refer to *Leichte Sprache* for German, as was done in (Klaper et al. [2013]). However, the use of the term *Simple German* should be discussed further since it can be a source of confusion due to the differing definitions of *Leichte Sprache* and *Einfache Sprache*. Distinguishing between *Leichte Sprache* and *Einfache Sprache* (simple language) presents a similar problem to differentiating Plain and Simple Language. Finding convincing distinction features is very difficult due to lack of research on this topic. We can observe, however, that *Einfache Sprache* is more complex and does not differ much from the original text in structure and content. Also, there are no explicit guidelines as there are for *Leichte Sprache*. *Einfache Sprache* is used when a text should be readable for people with low communication and reading skills, yet must not be altered as much as the principles of *Leichte Sprache* would

[2] *Attempto Controlled English* for example can be automatically translated into first-order logic and serves as a knowledge representation language [Fuchs et al., 2008]

demand [Bock, 2014, 21-22]. Kellermann finds a distinguishing criterion in the target group, saying that *Leichte Sprache* is geared toward people with learning disabilities, while *Einfache Sprache* can be helpful for a larger group of readers, such as people with poor reading and spelling skills, brain lesions, elderly people, deafs, immigrants, language learners and tourists [Kellermann, 2014]. The *Netzwerk für Leichte Sprache* (Network for Simple German) however, considers all these groups potential addressees of *Leichte Sprache* and therefore rejects this definition.

Since an important principle of Plain Language is the inclusion of meaningful examples, I conclude the definition section with Example 2.1, which demonstrates how a text can be simplified using *Leichte Sprache*[3].

(2.1) Original Text:

Die ordentlichen Gerichte unterteilen sich auf Landesebene in Amts-, Land- und Oberlandsgerichte. Sie sind für bürgerlich-rechtliche und strafrechtliche Verfahren sowie die freiwillige Gerichtbarkeit zuständig.

Text in Simple German:

Ordentliche Gerichtsbarkeit
2 Personen streiten sich.
Die Personen gehen zu einem Gericht.
Die Personen gehen zu einem ordentlichen Gericht.
Ordentlich hat hier **nichts** mit sauber zu tun.
Ordentliche Gerichte sind die streitigen Gerichte.
Das Wort streitig kommt von Streit.
Die ordentlichen Gerichte regeln Streite.
Zu den ordentlichen Gerichten gehören zum Beispiel:
- Amts·gerichte
- Und Land·gerichte

2.2 Background

Simplifying text to make it more readable and understandable is not a new idea and has been recognized as a complex task a long time ago. Cicero argues that one should avoid unnecessary words and ornaments to keep the speech as brief, precise and simple as possible, although he admits that achieving this effect is a difficult art (The Orator xxiii, 76-79).

[3]Example derived from [Maass, 2015, 7-8]

> Plainness of style seems easy to imitate at first thought, but when attempted, nothing is more difficult.

Sweden holds a pioneer position in the promotion of Easy-to-Read Language (*swedish* lättläst). The Swedish National Agency for Education released the first book written in Easy-to-read Swedish in 1968. Since 1984, the Center for Easy-to-Read (*swedish* Centrum för lättläst) publishes a weekly newspaper named *8 SIDOR*[4] that consists of 8 pages. The Center also offers translation services and workshops on Easy-to-read Language. Following the example of *Lättläst*, Finland, Norway, Denmark, Belgium, Estonia and the Netherlands established similar newspapers written for people with reduced literacy [Kellermann, 2014].

Since 1980, there have been many organisations and networks that promoted the use of Plain English. However, most of them focus on simplifying legal texts and government documents and do not necessarily specialize in writing for an audience with reading disabilities: The Plain Language Action and Information Network (PLAIN)[5] formed in 1993 is a group of federal employees that support the use of clear communication in government writing in the US. *Clarity*[6] is an example for an international association that promotes plain legal language. The first organisation to develop and promote Easy-to-read English for people with reduced literacy was People First[7]. People First is run by and for people with learning difficulties and provides Easy-to-read services since 1996.

Inclusion Europe[8] did pioneer work in the European Easy-to-read movement. Under the principle of *Information for All*, they developed Easy-to-read rule sets for 16 languages that inspired many other guidelines.

The workshop in Natural Language Processing for Improving Textual Accessibility (NLP4ITA)[9] suggests that the increasing interest in Text Accessibility for people with mental and physical disabilities is due to the growing importance of the web. Information on the web can and should be provided in a way that makes it accessible to everyone. The Web Content Accessibility Guidelines (WCAG)[10] of the Web Accessibility Initiative contain recommendations for making information accessible on the web to people with special needs. Among the most basic guidelines is the

[4] `http://8sidor.se/`

[5] `http://www.plainlanguage.gov/`

[6] `http://www.clarity-international.net/`

[7] `http://peoplefirstltd.com/contact/`

[8] `http://www.inclusion-europe.org`

[9] `http://www.taln.upf.edu/nlp4ita/index.html`

[10] `http://www.w3.org/WAI/intro/wcag`

requirement for alternative texts for pictures, transcripts for podcasts and keyboard-controllable navigation to make web contents accessible to people with sight, hearing or movement disabilities. Readability and understandability of text content constitute another important foundation for web accessibility. Although the term *Plain Language* does not appear in the guidelines, the recommendations to reduce text complexity correspond to the rules for writing Plain Language.

2.3 Target Groups

The target group for Plain Language are people with poor literacy skills. This group, however, is very diverse and every subgroup has its own needs and preferences. Since Plain Language is often initiated by organisations that support people with mental disabilities, the main addressee group seems to be people with learning disabilities[11]. This group is already very heterogeneous and although Plain Language was mainly developed for them, there are many other groups that can benefit from it. A related target group are people with brain lesions and neural diseases such as dementia and aphasia. Dementia patients can use Plain Language especially in early stages of the disease to stay independent and included in society. Another, often neglected group are prelingual deafs, which have lost their sense of hearing before language acquisition. They show very low reading and writing skills because they cannot learn language on a phonological basis like other children. They have difficulties understanding compound sentences and words with morphological information like pronouns, and even adults only possess a passive vocabulary of roughly 2000 words (in German). Their language skills are not sufficient for reading difficult texts written in standard language, but they do not suffer from any mental disabilities and are therefore able to comprehend complex content [Maass, 2015, 14-17].

The largest target group of Plain Language are functional illiterates. They are individuals who have gone through school education and learned to read and write but do not meet a minimum standard of literacy. Their reading abilities are not sufficient to understand written language on textual, sentence or word level [Maass et al., 2014, 57-59]. The number of functional illiterates in modern countries with mandatory schooling in today's information age is surprisingly high. In Germany, there are 7.5 million functional illiterates among the working population [Grotlüschen and Riekmann, 2009]. In Switzerland, 800'000 people (16% of the adult population) have insufficient reading skills to understand a simple text [Bundesamt für Statistik, 2006,

[11] The Network People First states that *people with learning disabilities* is the preferred and self-chosen term for mentally disabled people.

6]. The functional illiteracy rate in the UK is 16% among adults[12]. In Brazil, 7% of the population are illiterates while 21% are literate only on a rudimentary level [INAF, 2009]. Another potential target group are second-language learners. They usually have average or better literacy skills in their native language and can be considered temporary addressees because they might eventually improve their language skills to a level on which they can read standard language texts. Plain Language can support immigrants integrating into society, by both providing information about their new country and helping to master the language [Maass, 2015, 18]. It is still subject of research how exactly and to what degree the mentioned target groups can benefit from Plain Language and how the guidelines should be adapted to match the needs of the diverse groups.

2.4 Simple German

The origin of *Leichte Sprache* in the German speaking area lies in the Easy-to-read movement of People First in the US. The equivalent German organisation *Mensch zuerst* was founded in 2001 and published the first two dictionaries in Simple German. The *Netzwerk Leichte Sprache* was established in 2006 and has taken a leading role in promoting *Leichte Sprache*. The Austrian organisation *atempo* that fights for equality of people developed *capito*, a translation method for Simple German [Bock, 2014, 20-21]. In Switzerland, Simple German is a relatively new topic addressed mainly by aid organisations for people with disabilities, for example Insieme and Pro Infirmis. In January 2015, Pro Infirmis founded the *Büro für Leichte Sprache* (Office for Simple German) that translates texts into Simple German.

Simple German has experienced a boom after the ratification of the Convention on the Rights of Persons with Disabilities by Germany in 2008. The Convention proclaims equal rights for handicapped people and declares accessible information an essential right. Needless to say, the convention is also available in Simple German[13]. In 2002, Germany passed the *Behindertengleichstellungsgesetz* (law of equality for handicapped peopled) that led to the *Barrierefreie-Informatistechnik-Verordnung* called BITV (order for accessible information technology). The revised edition BITV 2.0 of 2011 demands that new web contents published by the government should also be made available in Simple German and German Sign Language, and that older articles should be translated by March 2014. Only a part of the documents were translated by the deadline and the quality of the Simple German texts varies, yet

[12] `http://www.literacytrust.org.uk/adult_literacy/illiterate_adults_in_england`

[13] `http://www.ich-kenne-meine-rechte.de/`

this was an important step to draw attention to the need for accessible information for people with disabilities [Maass, 2015, 20-22]. Switzerland ratified the Convention on the Rights of Persons with Disabilities in April 2014 and the canton St. Gallen was the first Swiss administration office to translate a text into Simple German [Hiller, 2015].

Even though Simple German is usually well-received by people with reduced literacy, it seems to have a negative image in society. The texts are rejected as primitive, childish and over-simplified, which can be a valid criticism for low quality translations but does not apply to all Simple German texts. When Simple German is dismissed as German for disabled people (*Behinderten-Deutsch*)[14] it achieves the contrary of the intended effect and builds a barrier between the people using standard language and those using Simple German, instead of including them into society. It is therefore essential for social inclusion that Simple German finds public acceptance. Simple German is a very new phenomenon and profound research on the concept, its target groups and guidelines is still missing, as well as a clear definition. To distribute the idea of Simple German, increase acceptance and further develop Simple German, systematic linguistic and social research is needed [Bock, 2014, 17,29,33-34].

[14] Article in free newspaper "20 Minuten" from 26.05.2015
`http://www.20min.ch/schweiz/news/story/-Behinderten-Deutsch--sorgt-fuer-rote-Koepfe-16240248`

3 Guidelines

An important foundation for writing Simple German are the various guidelines that present rules for Text Simplification. There are four essential guidelines for Simple German: The first rule set to present information in a readable and understandable way was released by Inclusion Europe in 2009. The extensive brochure contains 44 rules for Text Simplification and inspired the later guidelines for Simple German. It is written in Easy-to-read Language but the text ignores some of the self-set rules, perhaps due to the translation from the original English text. Inclusion Europe believes that the principles of Plain Language are the same for all languages and does not give specific rules for Simple German [Inclusion Europe, 2009]. The *Netzwerk Leichte Sprache* wrote the most popular guidelines, released in 2009 in electronic form and in 2013 as a brochure. The rules are written in Simple German [Netzwerk Leichte Sprache, 2009]. The BITV 2.0 rule set was written in 2011 to help translating governmental web contents into Simple German. This guide was important for the wide distribution of Simple German, although the rules themselves are very unspecific and incomplete. The most recent guidelines were developed by the *Forschungsstelle für Leichte Sprache* (research center for Simple German), founded in 2014 at the University of Hildesheim. The institute conducts research on Simple German and systematically examines the existing guidelines. They have not yet concluded the development of their rule set but already present guidelines based on scientific research with very distinct rules and explanations [Maass, 2015, 26-29]. Since the *Forschungsstelle für Leichte Sprache* provides the most elaborate guidelines with detailed linguistic descriptions for transforming standard German into Simple German, I will build my Text Simplification system upon these rules.

The following rules and examples are derived from Christiane Maass' rule book[1] for Simple German developed at the *Forschungsstelle für Leichte Sprache*. Maass preludes her rule set by pointing out that Simple German forms a bridge to standard language in the sense that some readers (for example functional illiterates) might improve their language skills by reading Simple German and move on to standard

[1] Christiane Maas. *Leichte Sprache - Das Regelwerk.* Lit Verlag Dr. W. Hopf, Berlin, 2015

German texts. Therefore it is essential that Simple German texts do not contain incorrect German, such as fragmented sentences or unnecessarily hyphenated compounds (see Example 3.1). It would be disrespectful and preposterous to let people with reduced literacy read texts written in incorrect German. They want to be taken seriously, which is why texts should never be written in a childish tone and adults have to be addressed with the polite form ("Sie") at all time. Maass concludes that the aim of Simple German is to produce readable and understandable text and that every rule can be dismissed if it conflicts with this goal. The rules are divided into 5 categories according to the level of text they concern: character level, word level, sentence level, textual level and layout.

3.1 Character Level

Special characters are not allowed in Simple German, with the exception of the punctuation marks full stop, question mark, exclamation mark, quotation mark, colon and the newly introduced *Mediopunkt*. Other signs such as the paragraph symbol (§) are forbidden because they are unknown to some readers and can easily be replaced. Note that the comma should not appear in Simple German texts, either; not because it is hard to read but because it implies subordinate clauses and enumerations which are resolved by other rules. The comma becomes dispensable. The *Mediopunkt* (literal translation: middle dot) was introduced by Christiane Maass for compound splitting. German compounds are productive and not multiword units as in English, so they can result in very long words that are hard to segment for inexperienced readers (see Example 3.1 a, b). Other guidelines suggest splitting compounds using hyphens, which helps reading long words but can lead to incorrect spelling or unwanted ambiguity (3.1 c - e). Especially odd (and orthographically incorrect) is compound segmentation of words with a *Fugen-s* (linking element s). Since compound splitting is important for Simple German, yet incorrect spelling should be avoided, the *Mediopunkt* was implemented. It facilitates reading long words and can be combined with real hyphenated compounds (3.1 f - g). It helps learning the correct spelling, disambiguates words (3.1 h) and contributes to the acceptance of Simple German. It has been criticised, however, that the *Mediopunkt* introduces a new sign and spelling rule that is not part of the standard German language.

(3.1) (a) Ost-West-Konflikt
(b) Steuererklärungsfristerstreckungsantragsformular
(c) *Schlag-Anfall

(d) *Rechts-Anwalt
(e) Wasser-Hahn: water-tap or special kind of bird?
(f) Rechts·anwalt
(g) Lotto-Annahme·stelle
(h) Musik·erleben vs. Musiker·leben

Numbers should be written as digits and not words because they improve text comprehension. The word *ein* should only be written with a *1* when it represents a number, not when it takes the role of an indefinite article. Roman numerals must be avoided, large numbers, percentages and year dates should be used sparsely. If possible, year dates should be omitted or paraphrased with expressions such as *x years ago*. In text types that cannot renounce numbers, charts can be used to illustrate important numeric data.

3.2 Word Level

The rules for Simple German on the word level can be readily summarized as: Use easy, short and well-known words. In case a difficult word is needed, it should be explained using simple words (see Example 2.1). Technical terms and foreign words should be avoided, as well as abbreviations. Common acronyms like *CD* or *WC* may be used if their full forms (*compact disc, water closet*) are less known than the acronym. It is not an easy task to determine what vocabulary is basic enough for Simple German, let alone the diverse language abilities of the target groups. Extracting basic vocabulary according to the frequency in a corpus may sound like a reasonable idea but word frequency does not necessarily correlate with understandability [Bock, 2014, 75]. A list of permitted vocabulary for Simple German is not available yet. However, there is an online dictionary named Hurraki[2] that explains difficult words in Simple German. Hurraki counts more than 2200 articles and is constantly growing.

3.3 Sentence Level

Many rules for Text Simplification concern the syntactic structure of a sentence. An important principle is formulating sentences in a way that makes it easy to see *who does what*. Thus, nominalization and passive constructions are forbidden. Especially

[2] `http://hurraki.de/wiki/Hauptseite`

paired with genitive constructions, nominalization can make it hard to extract the content of a sentence (see Example 3.2). Paraphrasing nominalization and passive constructions is challenging because the agent that is needed to formulate an active voice phrase can be missing and has to be reconstructed from context (see Example 3.3). Attributive genitives should also be avoided because the use of the genitive case is increasingly rare and inexperienced readers may not recognize the genitive forms. If possible, the genitive attribute should be transferred into a prepositional phrase using *von* (see Example 3.4).

(3.2) Ein Schwerpunkt der Frauenpolitik als Querschnittsaufgabe liegt im Bereich der Umsetzung der "Gender-Mainstreaming-Strategie der EU".

(3.3) (a) Heute ist die Wahl zum Heim·beirat.
(b) Heute wird der Heim·beirat gewählt.
(c) Heute wählen wir den Heim·beirat. (Simple German)

(3.4) (a) Das Haus des Lehrers.
(b) Des Lehrers Haus.
(c) Das Haus vom Lehrer. (Simple German)

Even though the standard German word order is considered Subject-Verb-Object (SVO), the word order is not fixed and varies depending on other words in the sentence, emphasis and clause type (see Example 3.5). When writing Simple German, the SVO word order should be chosen, unless another word order is more understandable.

(3.5) (a) Er besucht seinen Freund. (SVO)
(b) Morgen besucht er seinen Freund. (VSO)
(c) Seinen Freund besucht er morgen. (OVS)
(d) [Ich weiss,] dass er morgen seinen Freund besucht. (SOV)

Another essential principle of Simple German is that every sentence may only contain one piece of information. This rule is challenging on the semantic and syntactic level because complex information has to be broken down systematically. The principle leads to one of the most demanding rules of Simple German: Coordinate and subordinate clauses are forbidden. Splitting coordinate clauses is not a very difficult task, paraphrasing subordinate clauses into main clauses, however, requires distinct transformation rules. Maass suggests ways for splitting and paraphrasing conditional, causal, modal, temporal, consecutive, concessive, final and relative clauses. They are described in detail in Chapter 5.3.2, where I discuss how I implemented rules for syntactic simplification. Subjunctive mood should be avoided because the forms are rarely used and thus unfamiliar to inexperienced readers. Studies have

shown that the simple past (Präteritum) is less understandable than past perfect (Perfekt) because the often irregular and complex simple past word forms can differ considerably from the original verb and are harder to understand. Past perfect uses the past participle (Partizip II) which can be associated more easily to the lemma of a verb. In sentences in past perfect, information about the past tense, person and number is given in a well-known auxiliary verb instead of a complex verb form (see Example 3.6). Besides, the simple past is relatively rare in spoken language; in Swiss German, the simple past does not exist at all. As a final syntactic rule, negations should be avoided. If needed, it is better to formulate a sentence with *nicht* (not) instead of *kein* (not a) because the *k* of *kein* can easily be overlooked, resulting in *ein* (one), which changes the meaning of the whole sentence. The *nicht* should always be printed in bold (see Example 3.7).

(3.6) (a) Er ass.
(b) Er hat gegessen.

(3.7) (a) Wir haben heute einen Kuchen gebacken.
(b) Wir haben heute keinen Kuchen gebacken.
(c) Wir haben heute **nicht** Kuchen gebacken.

Transparent metaphors like *Leichte Sprache* (which does not literally have low weight) may be used if they can be easily understood. More complex metaphors and idioms should be replaced by literal expressions.

3.4 Textual Level

In Simple German, one should always use the same word for the same issue and refrain from the use of synonyms. Personal pronouns of the 1st and 2nd person (*ich, du, wir, ihr*) may be used, but personal pronouns of the 3rd person (*er, sie, es, ihm, ihnen*) should be replaced by the corresponding noun phrase. Indirect speech has to be changed to direct speech (see Example 3.8). The text may be changed and extended with examples and explanations, which should follow and not precede the explained term. Pictures, charts and graphics should be meaningful and appropriate for the target reader.

(3.8) (a) Peter sagt, er sei krank.
(b) Peter sagt: Ich bin krank.

3.5 Typography and Layout

Simple German is always displayed one sentence per line and if a sentence needs to be split up due to shortage of space it should be segmented at syntactical phrase borders. The text is written in a large sans-serif font type and the structure is emphasised by headlines and indentations.

3.6 Proof Reading

Although not itself a rule but an important principle of Simple German is the proof-reading process. To guarantee high quality, every simplified text needs to be proofread by a member of the target group. The proof-reader points out words and phrases (s)he does not understand, then the text is revised respectively. According to *capito*, a text can only be considered "simple" when the proof-reader is able to read and understand it without help. Usually, more than one proof-reader is consulted to verify the quality of a text. Verified texts written in Simple German are labeled by *capito* with a quality seal named *Leicht Lesen* (easy reading). Inclusion Europe has its own quality seal that is assigned to Easy-to-read texts in all European languages. The seal may only be applied to texts that meet certain conditions and are acknowledged as Easy-to-read by Inclusion Europe.

4 Automatic Text Simplification

Text Simplification (TS) is the process of reducing syntactic and lexical complexity of a text while attempting to preserve its meaning and content information. The aim of TS is to improve both readability and understandability to make a text more comprehensible for the reader, or easier to process by a program. Automatic Text Simplification (ATS) has recently become an established research topic that combines many Natural Language Processing (NLP) tasks [Siddharthan, 2006]. ATS can be performed using (manually crafted) rules for simplification or Statistical Machine Translation (SMT) algorithms.

ATS tackles two generally unrelated fields of research. Firstly, it builds text with enhanced readability that is accessible to a broader audience and contributes to promoting accessible communication (see Chapter 2). Secondly, it breaks down complex sentences which improves the reliability of NLP tasks such as Parsing, Machine Translation, Information Retrieval and Text Summarization [Chandrasekar et al., 1996]. Peng et al. [2012] for example report that syntactic simplification improved their Information Retrieval system to identify sentences about biological events by 20% in recall and 10% in accuracy.

4.1 Challenges

ATS faces a wide array of challenges. First of all, there is not only one solution for simplifying a text. Depending on whether TS is performed as preprocessing for NLP tasks or accessibility aid for reduced literacy readers, the simplification measures can vary. The diverse group of target readers with different abilities poses another problem; what might be an over-simplified text for one reader can be complex for another. The missing definition and poorly researched guidelines for Plain Language present another problem for both manual and automatic TS. Notably, while most publications on ATS list the various target users for simplified texts, hardly any of them use the word *Plain Language* or *Easy-to-read Language*, let alone give a definition of the concept. If ATS is used as an assistive technology, it is essential that

is produces accurate results. While moderate accuracy can be sufficient for other NLP tasks, a low-accuracy ATS system might generate incomprehensible text, which confuses the user more than the original complex text [Shardlow, 2014]. It is also important to notice that rules for Plain Language conflict with stylistic guidelines for scientific and belletristic writing. While word repetition should be avoided and the sentence structure ought to vary in standard language, Plain Language forbids the use of synonyms and demands simple and repetitive grammatical structures. Thus, it is essential for Text Simplification to specify the audience and purpose of the generated text.

Rule-based approaches for TS rely on text analysis, usually in the form of syntactic parsing. Parsing, however, is not considered a solved task in NLP and still produces incorrect analysis for many sentences. Such preprocessing errors present a considerable problem in ATS: Brouwer et al. [2014] described that 89% of all ATS errors in their system for Simplified French are due to preprocessing errors. Approaches that consider ATS a Machine Translation task are confronted with other challenges: SMT is based on parallel or comparable corpora, but they are very rare and small for Standard/Plain Language pairs. The English/Simple English Wikipedia data is so far the only available corpus for Simple English and was first used for ATS by Zhu et al. [2010]. Sentence alignment for monolingual corpora forms another obstacle. There exist many, usually length-based, algorithms for aligning bilingual parallel corpora, however those do not apply to comparable monolingual data. The Plain Language version can be structured completely different and since sentences are much shorter in the simplified version, a standard language sentence may have to be aligned to several Plain Language sentences. Hence, monolingual sentence alignment is usually performed using lexical similarity [Klaper et al., 2013].

A final issue of ATS is the lack of significant evaluation measures for fluency and correctness. The BLEU-score – standard for SMT evaluation – requires at least one human-written reference translation, which is usually missing for rule-based approaches. The BLEU-score has also been criticised as an evaluation metric for TS because there are many different ways to simplify a sentence. Readability metrics based on sentence and word length, sometimes including syntactic and discourse characteristics, are used for evaluation, although it is important to differentiate between readability and understandability. Readability defines how easy a text is to read, based on the complexity of grammar, length of sentences and familiarity of vocabulary. Understandability is the amount of information that can be gained from a text, depending on the reader's knowledge of the topic, its specific vocabulary and the understanding of complex concepts. A high readability text can still be hard to understand if the reader is not familiar with the topic. Plain Language aims at

producing text with both high readability and understandability [Shardlow, 2014]. The best way to evaluate is human judgement. Evaluation done by fluent readers, however, fails to tell us if a text is really comprehensible for low literacy people. Besides, manual evaluation is time-consuming and expensive [Siddharthan, 2014].

4.2 Previous Work

There are rule-based, corpus-based and hybrid approaches for simplifying texts. The rule-based ATS can be divided into three subtasks: Lexical simplification, explanation generation and syntactical simplification. Lexical simplification replaces difficult, unfamiliar words with more common alternatives and reduces lexical density by eliminating synonyms. When using the same words for the same issues, reading takes less cognitive effort. Many systems use word frequency as a measure for difficulty. However, frequent words do not necessarily increase the comprehensibility [Saggion et al., 2011]. Kandula et al. [2010] find user-friendly alternatives to difficult medical terms in the Open Access and Collaborative Consumer Health Vocabulary. PorSimples, one of the most famous ATS projects that aims at simplifying Brazilian Portuguese, has built its own common word dictionary. Difficult words are replaced by synonyms that appear in this dictionary, ranked according to their frequency in a Google search [Aluísio and Gasperin, 2010]. For morphologically rich languages such as German, the synonym needs to be declined to the respective linguistic case and number. Another essential characteristic of Plain Language is the explanation of difficult terms, which can be achieved by including dictionary entries for terms that were identified as difficult and crucial. The FACILITA plug-in of the PorSimples project recognises Named Entities and annotates them with a short explanation derived from Wikipedia, which is shown in a separate box, so not to hinder the reading flow [Watanabe et al., 2010].

Syntactic simplification is the most challenging task of ATS and has been developed inter alia for English, Dutch, Swedish, French and Portuguese. Following a manual for Simple Portuguese, PorSimples developed simplification operations that are applied when any of 22 linguistic phenomena are detected. Appositive, relative, coordinate and subordinate clauses are simplified, as well as passive voice, irregular word order and long adverbial phrases. Since low-literacy readers prefer short texts, yet sentence splitting makes a text longer, Text Summarization is applied at the end [Aluísio and Gasperin, 2010]. Rennes and Jönsson [2015] perform syntactical simplification for Swedish texts using 4 operation types: replacement and deletion of phrases, shifting of the word order and sentence splitting. Using those

operations, they implemented rules for changing passive to active voice, quotation inversion that puts the speaker at the beginning of the sentence (*He said: [quote]* instead of *[quote], he said.*), rearranging the word order and splitting sentences. Siddharthan [2006] suggests a three-level architecture for syntactic simplification: analysis, transformation and regeneration. The analysis process includes resolving 3rd person pronouns, detecting relative clause attachments and clause boundaries, and marking up appositives. During the transformation step, conjoined clauses are split, 3rd person pronouns are replaced by the noun phrase they refer to and relative clauses and appositive phrases are paraphrased. The regeneration module prevents the simplified text from losing cohesion: The newly split sentences are reordered and appropriate cue words that signal the rhetorical relation between sentences are selected. For duplicated noun phrases resulting from the anaphora resolution, a suitable determiner is chosen (definite or indefinite article or demonstrative determiner). Additional adjustments are carried out to create a simplified, cohesive text that contains the same information as before simplification.

Phrase Based Machine Translation (PBMT) approaches for ATS require large corpora, but so far there exist only a few parallel corpora containing Plain Language. English/Simple English Wikipedia is the most prominent one. Zhu et al. [2010] extracted 108'016 sentence pairs; but since the number of Simple English articles is increasing, more recently created corpora would certainly contain more data. The corpus resulting from the PorSimples project is composed of 104 texts from a newspaper plus their simplified versions in two levels, which makes a total of 128'586 words [Caseli et al., 2009]. The Simplext corpus for Spanish contains 200 short news articles [Saggion et al., 2011]. The corpus for German/Simple German by Klaper et al. [2013] consists of about 70'000 tokens in 7000 sentences. Brunato et al. [2015] developed a simplification annotation scheme and tagged an Italian corpus containing original and simplified texts with simplification operations (split, merge, reordering, insert, delete and transformation). They aim to use this annotated corpus as training material for a semi–automatic supervised TS system. Siddharthan [2014] argues that, contrary to rule-based approaches, PBMT systems do not include linguistic knowledge and are therefore not equipped to handle simplifications that require syntactic reordering, morphological changes and insertions. They can only perform lexical substitutions, deletion and simple paraphrasing.

Simplext proposes a hybrid approach for syntactic simplification. A grammar looks for possible target structures that need simplification (such as relative clauses), then the statistical filter classifies them according to whether they should be changed or not. The syntactic simplification itself is based on rules and involves deletion, insertions and copying of syntactic nodes and trees. The classifier prevents the

simplification system from manipulating wrongly detected target structures, such as restrictive relative clauses or complement clauses [Bott et al., 2012].

To the best of my knowledge, no work other than the corpus built by Klaper et al. [2013] has been done on automatically simplifying German texts. Plain Language in general is under-represented in the German speaking area [Matausch and Nietzio, 2012] and systematic research on Simple German has only just begun. The increasing importance of Simple German in society might trigger more work on ATS for German in the future. Another reason for missing ATS systems for German may be the rich morphology of the German language that complicates tasks such as synonym replacement and passive to active voice paraphrasing. When a word is replaced or a sentence rephrased, producing the adequate word forms requires morphological generation, a yet largely unsolved NLP task.

5 Rule-based Text Simplification for German

5.1 Goal

As described in Chapter 4, there have been no attempts of automatically simplifying German texts as of yet. I took a first step towards Automatic Text Simplification for German by developing a rule-based system that reduces the complexity of standard German texts. My goal was not to produce perfect Simple German output but rather a slightly simplified version of the source text. I examined what it takes to simplify German language according to the guidelines described in Chapter 3 and implemented as many rules as possible in my Text Simplification system. My system is based on the output of a dependency parser and therefore focuses on syntactic simplification. However, I also included other resources and tools to reduce the lexical complexity.

5.2 Tools and Resources

5.2.1 ParZu

The basis of my TS system is given by the syntactic parsing output of the source text that is to be simplified. My system accesses the hybrid dependency parser *ParZu* that combines hand-written rules with a probabilistic disambiguation system [Sennrich et al., 2009]. *ParZu* performs sentence segmentation and tokenization, and provides linguistic information for every token in the text: position in the sentence, lemma, part-of-speech, morphological information, grammatical function and the head of the phrase it depends on. This information alone allows various syntactic simplification operations and therefore constitutes the foundation of my system.

5.2.2 Gertwol

I used the morphology tool GERTWOL for the segmentation of compounds [Haapalainen and Majorin, 1995]. GERTWOL returns all possible segmentations for a word and provides morphological analysis (see Example 5.1). The morphological information can be essential for disambiguation. In the sentence *Das kleinste Staubecken ist nur zur Hälfte gefüllt* the word *Staubecken* can only be singular (because of the determiner, adjective and verb), so the option *Staub·ecken* that occurs only in the plural form can be dismissed.

(5.1) Analysis for *Staubecken*

lemma	*morphology*
Stau.becken	S NEUTR SG NOM
Stau.becken	S NEUTR SG AKK
Stau.becken	S NEUTR SG DAT
Stau.becken	S NEUTR PL NOM
Stau.becken	S NEUTR PL AKK
Stau.becken	S NEUTR PL GEN
Stau.becken	S NEUTR PL DAT
Staub.eck e	S FEM PL NOM
Staub.eck e	S FEM PL AKK
Staub.eck e	S FEM PL DAT
Staub.eck e	S FEM PL GEN

5.2.3 Hurraki

Hurraki is an online dictionary in form of a wiki that consists of more than 2200 articles written in Simple German. The articles are structured in three sections: A short definition, a list of synonyms and a more precise description of the term. The definition is written in one or two short sentences and is usually accompanied by a picture. The synonym section (*Gleiche Wörter*) contains not only exact synonyms but also abbreviations and sometimes hypernyms. The length and structure of the descriptions vary. Since every user can compose articles, the quality of the entries is heterogeneous. Article writers are encouraged to follow the guidelines from Inclusion Europe but it is not described if and how the entries are reviewed by an expert. The overall website is designed very user-friendly, even the interface for writing and changing articles. Hurraki is an important project for Simple German and creates a valuable resource for both Plain Language users and translators.

5.2.4 Abbreviation List

Since abbreviations are not allowed in Simple German, a list of common abbreviations and their full form is needed. I found the most extensive list on Wikipedia[1] and derived 405 abbreviations ending with a full stop (for example *zzgl.*) and a list of 278 acronyms ($T\ddot{U}V$) and abbreviations without full stop (*Abo*).

5.2.5 Conjugation

Although many simplification operations can be performed without changing the verb form (for example certain types of sentence splitting), more complex tasks such as passive to active transformation and modification of the tense require adaptation of the verb. I could not find a suitable tool for the generation of verb forms, so I settled for an online resource: The website for verb conjugation[2] created by Andreas Göbel in 2000 gives the full conjugation table for every German verb, in all tenses and modes. It even suggests conjugations for verbs that are unknown to the system or do not exist in German (for example *leichtsprachisieren*), which makes it very robust. If a change in verb form is needed, my system accesses this website and retrieves the correctly conjugated form. So far I have not detected any mistakes in the conjugation tables and will therefore continue to rely on this auxiliary website.

5.2.6 Declension

Not only the conjugation of verbs is needed for Text Simplification but also the declension of nominals. Paraphrasing passive sentences and genitive attributes requires a change in case for nouns, adjectives, determiners and possessive pronouns. Since it worked well for verb forms, I decided to extract declined nominals from an online resource as well. CanooNet[3] is an online dictionary released by Cannoo Engineering in 2000 that contains more than 250'000 manually checked German word entries. It provides not only words forms but also synonyms and hyper- and hyponyms, which could be used for lexical simplification. Although CanooNet also returns conjugation tables for verbs, I continued to extract verb forms from *Verbformen.de* because it specializes on verbs and is more robust.

[1] `https://de.wikipedia.org/wiki/Portal:Abk%C3%BCrzungen/Gebr%C3%A4uchliche_Abk%C3%BCrzungen`

[2] `http://www.verbformen.de/`

[3] `http://www.canoo.net/`

5.3 Text Simplification Rules

My Text Simplification system is based on syntactic parsing output. The parsing results for each sentence are stored individually as my ATS system applies the simplification rules sentence by sentence. There are no rules that go beyond (original) sentence boundaries and the simplification rules can be divided into the same categories as the guidelines for Simple German described in Chapter 3. Figure 1 shows the architecture of my system and the simplification processes at the different levels. On the following pages, I will explain the implemented rules in detail and demonstrate them with example translations[4] by my system.

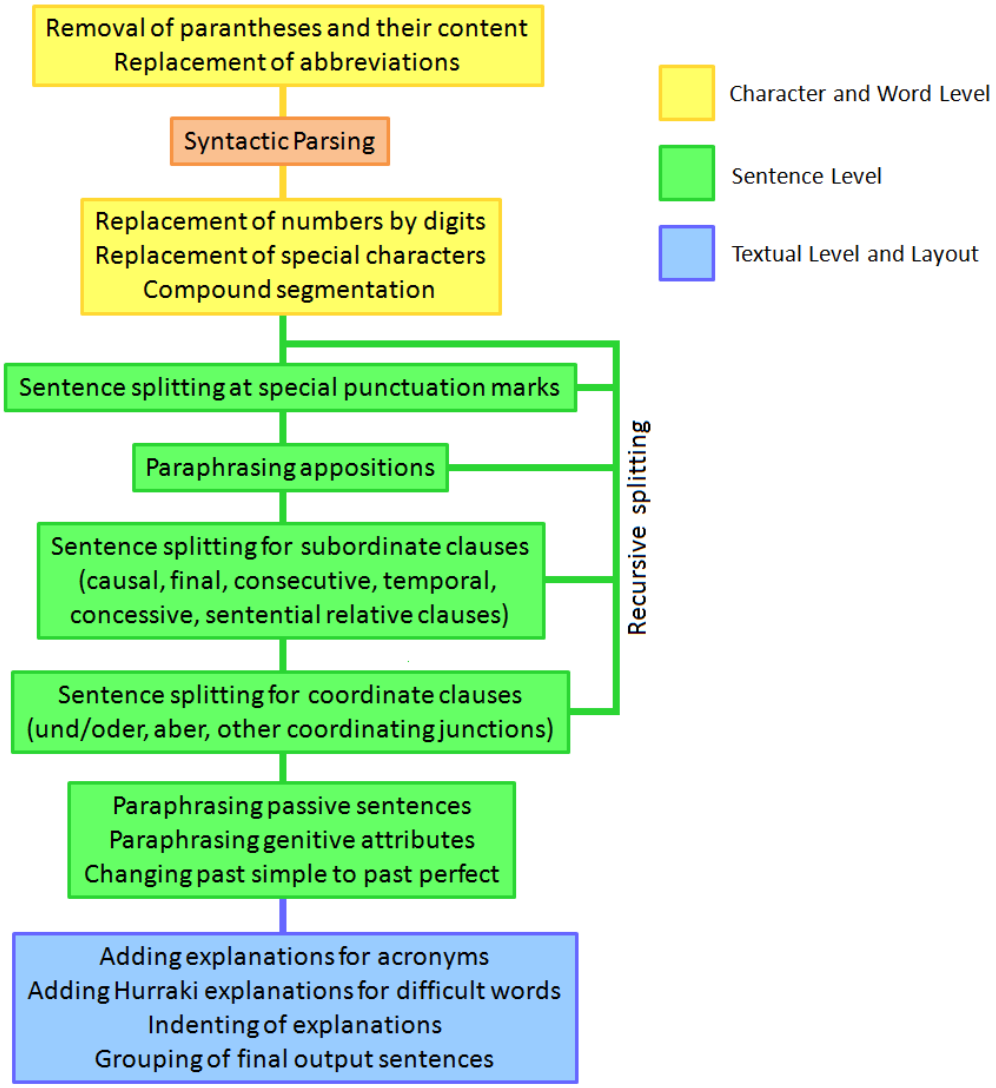

Figure 1: Architecture of rule-based Text Simplification system

[4]O: Original sentence; SG: Simple German translation

5.3.1 Character and Word Level

The first simplification steps are executed before parsing the source text. Since parentheses and their contents increase information density, all parentheses and the tokens in between are removed from the text. This is a radical step but it improves both readability and parser performance. The source text is then tokenized and abbreviations are replaced by their full spelling, using the abbreviation list derived from Wikipedia (see Example 5.2). The replacement of the abbreviations is carried out before the parsing process to prevent wrongly segmented sentences due to full stops in abbreviations. I observed that this preprocessing step improves the performance of the parser. Acronyms are not replaced because their extended spelling could influence the parser performance, interfere with the simplification rules and require case adaption. Besides, full spellings lengthen the sentences and are not necessarily easier to read. So instead, I inserted explanations for acronyms after the first simplified sentence that contains them (see Example 5.15). After preprocessing, *ParZu* performs dependency parsing on the whole source text. Every sentence in the text is searched for cardinal and ordinal numbers written in words, which are then replaced by digits. The part-of-speech tags (POS-tags) help differentiating between *ein* as an indefinite article and as a cardinal number. Special characters such as % and § are also replaced by words using a hand-made dictionary. If a noun (that is not a proper noun) is longer than 5 characters, it is examined as to whether it is a compound, and split accordingly using the *Medio·punkt* (see Example 5.2). If case and number are known, the corresponding match is chosen to resolve ambiguity (see 5.1). Unfortunately, sometimes the wrong segmentation is selected, as for *Töpfer·eibe·trieb* (correct: *Töpferei·betrieb*) or *Hau·sauf·gaben* (correct: *Haus·aufgaben*). Even though the *Medio·punkt* is disputed by experts because it introduces a new symbol to Simple German, I found it a suitable alternative to orthographically incorrect hyphenated spellings. To me, the *Medio·punkt* feels less disturbing in the reading process than hyphens, especially in words that can be segmented into several parts. Test readers from the target group would have to decide which version is easier to read for them but since the *Medio·punkt* is a very new complement to Simple German, there have been no studies yet.

(5.2) O: Prof. Müller kauft sich am Hbf. den siebten Band (den letzten) seiner Lieblingskrimireihe für 8$ 50¢ inkl. MwSt.

SG: Professor Müller kauft sich am Haupt·bahn·hof den 7. Band von seiner Lieblings·krimi·reihe für 8 Dollar 50 Cent inklusive Mehr·wert·steuer.

5.3.2 Sentence Level

On the sentence level, a large number of syntactic simplification rules are executed. These rules split and/or rephrase the sentences and return grammatically correct, complete and independent sentences. Since these resulting sentences might need further simplification, the process of syntactic simplification is structured in a recursive loop. The specific simplification rules are either executed once per loop or triggered by keywords. Various helping functions return information about the target sentence or perform frequent paraphrasing operations. For example, I created functions that return the full noun phrase when given the head (needed for subject extraction and reordering), change the case of a noun phrase (leaving prepositional phrases and genitive attributes untouched) or invert subject and verb to achieve correct word order. Those helping functions facilitate the addition of new rules and prevent the need to solve the same linguistic challenges multiple times.

The syntactic simplification loop starts by looking for semicolons, dashes (not hyphens within words) and colons, and the sentence is split there. The semicolon usually divides independent sentences so splitting is safe. Splitting at dashes can result in a full sentence and an elliptic sentence without subject and/or predicate. However, this does usually not pose a problem. If two dashes are detected – possibly containing an insertion (like here) – the sentence is left alone. If a colon is found, the sentence is split there. If the second part of the sentence contains a predicate (outside relative clauses), it is left unchanged. Otherwise it is likely to be an enumeration and therefore a *nämlich* (particularly) is added (see Example 5.3). A better solution (proposed by Maass) would be to list the items with bullet points. As a next step, appositions are paraphrased. Appositions are detected by a function tag assigned by the parser and are removed from the original sentence. A new sentence is generated in which the noun phrase the apposition refers to forms the subject (X) and the apposition is the predicative noun (Y), yielding an *X is Y* sentence (see Example 5.4). For this, apposition and corresponding noun phrase have to be changed to nominative case.

(5.3) O: Es gibt – gemäss Informatikern – 10 Arten von Menschen: diejenigen, die Binärcode verstehen, und die anderen.

SG: Es gibt – gemäss Informatikern – 10 Arten von Menschen.
Nämlich diejenigen, die Binär·code verstehen, und die anderen.

(5.4) O: Jeden Abend sehe ich Bello, den kleinen Hund meiner Nachbarin, im Garten spielen.

SG: Jeden Abend sehe ich Bello im Garten spielen.
Bello ist der kleine Hund von meiner Nachbarin.

The next simplification functions concern paraphrasing sentences containing subordinate clauses. These rules all have a similar structure: If a subordinating conjunction (such as *weil*, *nachdem*, *obwohl*) is found, it is checked whether it really is a conjunction (and not for example a preposition (*nachdem*) or adverb (*da*)). Then, the sentence is split at the conjunction and both resulting sentences are edited and paraphrased to form independent sentences. Suitable connective words that express the rhetorical relation are added to maintain the original meaning, and the correct word order is restored. The paraphrasing of the sentence sometimes depends on whether the subordinate clause is placed before or after the main sentence. In causal clauses, for example, if the subordinate clause is in first position, the conjunction (e.g. *weil* (because)) is removed and not replaced. Instead, the connective word *deshalb* (therefore) is added to the main clause. The two new sentences retain the information of the original one. If on the other hand, the subordinate clause is in the second position, the conjunction is replaced with the connective word *denn* (because). Maass suggests using *nämlich* (thus/that is to say) instead of *denn*. However, I decided to only use connectives that can be placed at the beginning of the sentence so they can serve as signal words, even if it results in VSO word order. If the sentence starts with *denn*, the reader instantly realizes that an explanation will follow (see Example 5.5). Final clauses are paraphrased using the modal verb *wollen* (want). This might not always be an ideal solution but generally produces understandable output. Final clauses containing *um...zu* constructions pose an additional challenge because the subject is not mentioned in the subordinate clause and has to be retrieved from the main clause (see Example 5.6). Consecutive clauses with *sodass* (so that) are paraphrased using *deshalb* (therefore) (see Example 5.7).

(5.5) O: Weil Anna noch nicht da ist, müssen wir warten.

SG: Anna ist noch nicht da.
Deshalb müssen wir warten.

O: Wir müssen warten, weil Anna noch nicht da ist.

SG: Wir müssen warten.
Denn Anna ist noch nicht da.

(5.6) O: Um sich zu entspannen, nimmt Lisa ein Bad.

SG: Lisa will sich entspannen.
Deshalb nimmt Lisa ein Bad.

(5.7) O: Die Geschäfte laufen schlecht, sodass sie bald schliessen müssen.

SG: Die Geschäfte laufen schlecht.
Deshalb müssen sie bald schliessen.

Splitting temporal clauses while retaining their information content turned out to be especially challenging. I found acceptable connectives to paraphrase temporal clauses containing *nachdem* (after), *bevor* (before), *seit* (since) and *während* (while) (see Example 5.8) but some resulting sentences might need tense adapation because of anteriority or posteriority in the subordinate clause. I also struggled with *solange* (as long as) and decided to simply reorder the sentences and reuse *solange* as connective, which transmits the meaning of the sentence but can result in unnatural language. I could not find a good and general way to paraphrase temporal clauses with the most frequent temporal conjunction *als* (when/as) which shows that sentence splitting is a challenging task, not only for computers but also for humans. Paraphrasing concessive clauses with *obwohl* (although) and similar conjunctions, on the other hand, could simply be arranged by placing the concessive part introduced by the connective *trotzdem* (however) after the main clause (see Example 5.9). Resolving relative clauses is very difficult so I decided only to tackle sentential relative clauses, which do not refer to the preceding noun but the whole sentence or clause. The words referring back are called pronominal adverbs. Such sentences can be split at the pronominal adverb which is then replaced by its cataphoric corresponding adverb. The predicate which is placed at the end of the subordinate clause is pulled forward to generate a correct sentence (see Example 5.10).

(5.8) O: Nachdem er zwei Stunden auf sie gewartet hatte, ging er nach Hause.

SG: Er hatte 2 Stunden auf sie gewartet.
Dann ist er nach Hause gegangen.

O: Seit ich meinen Fernseher weggegeben habe, lese ich viel mehr.

SG: Ich habe meinen Fern·seher weggegeben.
Seitdem lese ich viel mehr.

O: Während die Eltern arbeiten, passen die Grosseltern auf die Kinder auf.

SG: Die Eltern arbeiten.
In dieser Zeit passen die Grosseltern auf die Kinder auf.

O: Solange du einen Lernführerschein hast, darfst du nicht alleine fahren.

SG: Du hast einen Lern·führer·schein.
Solange darfst du nicht alleine fahren.

(5.9) O: Er geht nicht zum Zahnarzt, obwohl er starke Zahnschmerzen hat.

SG: Er hat starke Zahn·schmerzen.
Trotzdem geht er nicht zum Zahn·arzt.

(5.10) O: Lena wurde von Marco, dem Schwarm aller Mädchen, zum Essen eingeladen, worüber sie sich sehr freute.

SG: Marco hat Lena zum Weihnachts·ball eingeladen.
Darüber hat sie sich sehr gefreut.
Marco ist der Schwarm von allen Mädchen.

Splitting coordinate clauses is easier than subordinate clauses, yet not a trivial task. Most sentences containing coordinate conjunctions can simply be split at the conjunction. Sentences with conjunctions such as *und* (and), *oder* (or) and *aber* (but), however, might be elliptic, meaning that the subject is not repeated. In this case, the subject has to be retrieved from the previous clause. Only the necessary parts of the subject noun phrase are extracted, leaving out adjectives, genitive attributes and prepositional phrases to keep the sentences as short as possible (see Example 5.11). I allowed sentences that start with *und* and *oder* to indicate that more information about the previous topic follows, even though some experts argue against the use of such sentences.

(5.11) O: Mein bester Freund an der Universität lernt jeden Tag und nimmt sogar Nachhilfeunterricht, aber seine Noten werden einfach nicht besser.

SG: Mein bester Freund an der Universität lernt jeden Tag.
Und mein Freund nimmt sogar Nach·hilfe·unterricht.
Aber seine Noten werden einfach nicht besser.

Text Simplification is not only about shortening and splitting sentences but also deals with paraphrasing complex syntactical structures. A challenge for untrained readers is the passive voice. If a passive construction is detected by my system, it retrieves the grammatical agent (indicated by a prepositional phrase starting with *von* (by)), the object (the subject of the passive phrase) and the action verb (past participle), and generates an active voice sentence. The action verb is conjugated to the 3rd person, to the number as derived from the agent and to the tense extracted from the auxiliary verb *werden*. Verbs with seperable verb prefixes are not correctly transformed and I did not implement rules to paraphrase sentences with agents of 1st or 2nd person. If the agent was not mentioned, I used the impersonal pronoun *man* (one) as agent (see Example 5.12). Impersonal forms should be avoided in Simple German but I considered passive constructions more challenging than an impersonal

pronoun. Again, test readers from the target group have to be asked which version they prefer. Inserting a meaningful agent would be the best solution but if the agent is not mentioned in the text it can be difficult even for human translators to produce one. To resolve genitive attributes, the whole attribute is transformed to dative and completed with the preposition *von*. If the new noun phrase in dative starts with the article *dem*, the preposition and article are merged to the more naturally sounding *vom* (see Example 5.13). Genitive attribute resolution is executed after passive reconstruction because the resulting *von* prepositional phrases could be misleading when searching for the agent. To remove more genitive forms, one could change the noun phrases after the prepositions *wegen* (because of) and *trotz* (despite) to dative since those prepositions allow both cases.

(5.12) O: Der Geiselnehmer wurde überwältigt und festgenommen.

SG: Man hat den Geisel·nehmer überwältigt.
Und man hat den Geisel·nehmer festgenommen.

O: Die befreiten Geiseln werden von der Polizei befragt und der Täter wird abgeführt.

SG: Die Polizei befragt die befreiten Geiseln.
* Und man führt den Täter.

(5.13) O: Das ist das Zimmer meines kleinen Bruders.

SG: Das ist das Zimmer von meinem kleinen Bruder.

O: Der Dativ ist der Tod des Genitivs.

SG: Der Dativ ist der Tod vom Genitiv.

After splitting the sentences and resolving passive constructions and genitive attributes, the tense of the resulting sentence is determined. If it is past simple, the sentence is changed to past perfect. The past participle of the predicate is retrieved and the predicate is replaced by the corresponding auxiliary *sein* or *haben*. The auxiliary is conjugated to the person and number of the original predicate and the past participle is added to the end of the sentence. Since the sentence is already shortened as much as possible at this time, this method usually works well, with the exception of verbs with separable prefixes (see Example 5.14). Auxiliary verbs *sein* and *haben* or modal verbs such as *wollen* (want) and *können* (can) are not changed to past perfect because they their past simple forms are usually known even to untrained readers. Besides, changing auxiliary and modal verbs to past perfect results in unnatural language.

(5.14) O: Sie kamen am See an und sprangen gleich ins Wasser, weil sie sich abkühlen wollten.

SG: * Sie sind am See an gekommen.
Und sie sind gleich ins Wasser gesprungen.
Denn sie wollten sich abkühlen.

5.3.3 Textual Level and Layout

On the textual level, my simplification rules only include the addition of explanations for difficult words occurring in the text. Difficult words are defined as acronyms and words that are explained in the Hurraki online dictionary[5]. The acronyms are explained as seen in Example 5.15. If a simplified sentence contains a Hurraki word, the Hurraki definition is given (only at the first occurrence). I excluded the synonyms and detailed explanation from the explanation to keep it as short as possible (see Example 5.16). Creating a link to the full Hurraki article would be a nice feature. Since Hurraki does not only explain difficult words and concepts but also contains entries for words such as *man*, *woman* or *car*, I created a list with trivial Hurraki words and did not provide explanations for those. Unnecessary explanations disturb the reading process. To mark automatically added explanations, I indented the explanation paragraphs. This way, the explanation can be skipped easily. When printing the simplified text, all sentences resulting from one original sentence are grouped together in a paragraph to emphasize what information belongs together.

(5.15) O: Die SBB bietet ein vergünstigtes Abo für Studienbeginner an.

SG: Die SBB bietet ein vergünstigtes Abo für Studienbeginner an.

SBB ist die Abkürzung für Schweizerische Bundesbahnen.
Abo ist die Abkürzung für Abonnement.

(5.16) O: Ein Heimittel gegen Aids wird sicher bald entdeckt.

SG: Man entdeckt ein Heil·mittel gegen Aids sicher bald.

Aids ist ein schwieriges Wort.
Hurraki erklärt es so:
Aids ist eine Krankheit.
Diese Krankheit ist ansteckend.
An Aids kann man sterben.

[5]Dictionary word list retrieved on June 24, 2015

6 Evaluation and Discussion

6.1 Evaluation

I evaluated my Automatic Text Simplification system by comparing the output for an example source text to its human translation. I used a short text about the arrival of the Swiss team at the Special Olympics in Korea. The text and its translation to Simple German was provided by France Santi, a trained translator for Simple German. The original text, the human translation and the output of my system can be found in the appendix A. I selected this text because it contains many of the structures that are simplified by my program and is therefore suitable to evaluate my system. It should be pointed out, however, that it was also this text that inspired me to implement rules for paraphrasing appositions and sentential relative clauses. No other adjustments were made to improve the quality of the output text.

The strongest difference between the human translation and machine translation is the lexical complexity. In the human translation, difficult words and expressions such as *Delegation*, *Volunteers* and *unter die Fittiche nehmen* were replaced by easier words. Interestingly, the foreign word *Games* was not changed. My system does not replace difficult words or synonyms but explains some of them: *Botschaft* and *Chef* both triggered Hurraki explanations. The human translation does not provide explanations for those words, probably because they are not difficult or relevant enough. While my system segments the words *Ein·drücke* and *Unter·haltung* with the *Mediopunkt*, they remain unchanged in the human translation. Even the newly introduced and long word *Schweizersportler* is not segmented. My system can correctly segment this word.

In both translations, the sentences were split at the dash sign and two colons. In my system, splitting at the dash results in an elliptic sentence without verb. The apposition was paraphrased very similarly in both translations, although the rest of the sentence was translated differently. The human translation might be more understandable because it links *Timothy Shriver* to the aforementioned *Besucher*. The final clause was resolved in both versions, although in the human translation,

the final clause was removed in favor of an explanation of *Meditation* and the intentional thought behind the final clause is lost. The sentential relative clause is also rephrased by both human and machine translation, using the connective word *das*. I implemented no rule for paraphrasing participle constructions so the third sentence is not split or changed. As a matter of fact, the participle *ermüdet* was tagged as an infinite verb by the parser so a rule paraphrasing participle constructions would not affect this sentence. Since the sentence is not split and *ermüdet* was identified as predicate in present tense, the sentence is not transformed to past perfect later. While my system splits all coordinate conjunctions, there are two unsplit *und* sentences in the human translation. However, they are visually split by a line break.

In both translations, passive constructions and genitive attributes are resolved, although my system naturally returns more literal translations. In the second sentence, the prepositional phrase *von koreanischen Volunteers* was not identified as agent but as attribute to *unsere Gruppe*, due to the results of the dependency parser. This annotation is not necessarily wrong because constructions such as *eine Gruppe von Schülern* are very common and in this sentence, this reading is also possible. Only the context makes it clear that *unsere Gruppe* does not consist of but was taken care of by *koreanische Volunteers*. Apart from that, the passive could be resolved and even the elliptic second part (*und zum Shing Hun Sa Tempel geführt*) was completed and transformed correctly. Note that only *unsere Gruppe* was used as the second subject and not the whole noun phrase including the prepositional phrase. The genitive attributes are also paraphrased in both versions, although in the phrase *Strapazen der langen Reise*, the parser assigns the wrong lemma (*Reis*) to *Reise* which causes an incorrect dative form. In the human translation, the expression *während der Games* contains a genitive that could have been avoided using the accusative case instead. In both translation, past simple forms were changed to past perfect, with the exception of the previously discussed sentence.

The human translator grouped the resulting sentences from one original sentence together, just like my system does. While my system prints every output sentence on a new line, the human translation contains more line breaks. The sentences are split at commas, *und* and at phrase borders so the lines contain fewer words and less information at once. Especially for long sentences, this layout is very helpful to inexperienced readers. Note that the human translation was not proof-read by a member of the target group, probably because of its length and relatively easy content. For better evaluation of my text, proof-reading would be essential. Discussions with Simple German experts and people from the target group are needed to judge the quality of my translation and determine what it takes to transform my output text to Simple German.

Apart from the manual evaluation, I also computed the readability index LIX[1] to compare the complexity of all three texts. LIX is a readability score based on average sentence length and the proportion of long words. Thus, a text with long sentences and many long words is considered difficult to read and is assigned a high LIX value [Smith and Jönsson, 2011]. The original text yields a LIX of 53, which indicates scientific literature or newspaper articles. The LIX of the human translation is 35, labelling the text as easy. The translation generated by my system has a LIX score of 41, which classifies it as a text of average difficulty. Although the scores roughly correlate with my own evaluation, they are insufficient to evaluate the complexity of the evaluation texts. Sentence and word length alone are not sufficient measures to compute readability, especially when segmented words are taken into consideration. Syntactical structures and word density are neglected completely, and the LIX score only evaluates readability, not understandability.

6.2 Discussion

When applied to the evaluation source text, my Automatic Text Simplification system produces a well-readable translation. Especially on the syntactic level, many simplifications are achieved and the output of my system can be compared to the reference translation in syntactic complexity. It has to be noted, however, that the text used for evaluation is relatively simple and was selected because it is suitable to demonstrate the implemented simplification rules. My system still produces incorrect sentences with chaotic word order and wrong word forms for many test sentences; sometimes there is no output at all because the program runs into an error. Gibberish output or runtime errors are caused by many factors. First of all, there are countless linguistic structures that could be simplified and it is impossible to create rules for all of them. Then, the rules I wrote for some of these linguistic phenomena are not robust enough to deal with all varieties of them, let alone the combination of different phenomena. A missing subject or predicate in a main clause for instance, possibly caused by an error in the previous simplification step, has consequences for the rest of the process. Finally, since most rules are based on the parsing output, parser mistakes can be fatal. However, I built my system with an eye on the parser's capabilities and limits and included some rules to verify the correctness of frequently wrongly annotated tags and completed missing annotations by guessing, so parsing errors did not considerably reduce my system's performance.

[1] `http://www.psychometrica.de/lix.html`

There are many structures that are not yet simplified by my system, such as indirect speech, sentences containing *dass* and many other subordinate clauses, or sentences in subjunctive mood. Also, my system does little work on reducing lexical complexity. My goal, however, was not to create an Automatic Simplification System that produces perfect Simple German output. I rather wanted to take a first step in Automatic Text Simplification for German and demonstrate the variety of simplification rules needed to generate Simple German. Producing Simple German requires simplification on five different levels of the text. I implemented various rules for all of these levels and showed that even with a relatively small number of rules a text can be simplified considerably.

When further developing my system, one should put more emphasis on lexical simplification because even a text with short and simple sentences can be hard to read for inexperienced readers if it contains high lexical density and difficult words. Apart from that, syntactic simplification should be extended, refined and implemented more systematically, not exemplary as it was done so far. Some simplification rules may need to be revised to guarantee smooth interaction with other rules. Last but not least, a suitable output format should be chosen (for example XML) so that the text's structure – possibly emphasised by more indentations, line breaks, bold prints and headings – can easily be saved and displayed.

7 Conclusion

In my Bachelor thesis, I have implemented a rule-based system for Automatic Text Simplification that aims at generating Simple German. I have created a variety of simplification rules based on guidelines for Simple German to reduce both lexical and syntactical complexity of a source text. Although my system still produces incorrect output in an number of cases, it was able to generate a simplified version of a short text that is comparable to the human translation, at least on the syntactic level.

I introduced the phenomenon of Plain Language and explained its importance in social inclusion. I gave an overview of Plain Language worldwide and then focused on the movement of *Leichte Sprache* in Germany, Austria and Switzerland. I explained the most important guidelines for Simple German that form the foundation of my system. I discussed the purpose and challenges of Text Simplification in general and gave a brief overview of the previous work. I also presented my system in detail: I introduced the auxiliary tools and resources, explained the simplification rules I implemented and demonstrated them with example translations. In the end, I evaluated my system by comparing its output to an expert translation.

Automatic Text Simplification is a novel topic in Natural Language Processing and this Bachelor thesis is, to my knowledge, the first attempt at simplifying German texts. In this simple proof of concept, I show that even a small number of carefully selected and implemented simplification rules can reduce text complexity and make a text more comprehensible. I hope that this will serve to inspire further work on Automatic Text Simplification for German, especially with the specific aim of generating Simple German. Plain Language is an interdisciplinary topic with many different facets and I hope that interest and research in this fascinating field will increase over the next few years so that more and more information becomes accessible to everyone.

References

S. M. Aluísio and C. Gasperin. Fostering Digital Inclusion and Accessibility: the PorSimples project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics, 2010.

B. M. Bock. Leichte Sprache: Abgrenzung, Beschreibung und Problemstellungen aus Sicht der Linguistik. *Sprache barrierefrei gestalten. Perspektiven aus der Angewandten Linguistik. Berlin: Frank & Timme.(= TransÜD. 69)*, pages 17–52, 2014.

S. Bott, H. Saggion, and D. Figueroa. A Hybrid System for Spanish Text Simplification. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 75–84. Association for Computational Linguistics, 2012.

L. Brouwer, D. Bernhard, A.-L. Ligozat, and T. Francois. Syntactic Sentence Simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 47–56, 2014.

D. Brunato, F. Dell'Orletta, G. Venturi, and S. Montemagni. Design and Annotation of the First Italian Corpus for Text Simplification. In *The 9th Linguistic Annotation Workshop held in conjuncion with NAACL 2015*, page 31, 2015.

Bundesamt für Statistik. Lesen und Rechnen im Alltag. Grundkompetenzen von Erwachsenen in der Schweiz. Nationaler Bericht zu der Erhebung Adult Literacy & Lifeskills Survey. BFS Bundesamt für Statistik, Neuchatel, 2006. URL `http://www.bfs.admin.ch/bfs/portal/de/index/international/22/publ.Document.80535.pdf`. (retrieved June 21, 2015).

H. M. Caseli, T. F. Pereira, L. Specia, T. A. Pardo, C. Gasperin, and S. M.

Aluísio. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *the Proceedings of CICLing*, 2009.

R. Chandrasekar, C. Doran, and B. Srinivas. Motivations and Methods for Text Simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1041–1044. Association for Computational Linguistics, 1996.

R. Eagleson. *Writing in Plain English*. Australian Government Publishing Service, Canberra, 1990.

N. E. Fuchs, K. Kaljurand, and T. Kuhn. Attempto Controlled English for Knowledge Representation. In *Reasoning Web*, pages 104–124. Springer, 2008.

A. Grotlüschen and W. Riekmann. Leo-One Studie zur Grössenordnung des Analphabetismus von Bundesministerium für Bildung und Forschung (BMBF), 2009.

M. Haapalainen and A. Majorin. GERTWOL und Morphologische Disambiguierung für das Deutsche. In *Proc. of the 10th Nordic Conference on Computational Linguistics, Helsinki, Finland*, 1995.

O. Hiller. Leichte Sprache - Die ersten Gehversuche. *Ostschweiz am Sonntag, 25.05.2015*, page 2, 2015.

INAF. INAF Brasil - Indicador de Alfabetismo Funcional, 2009. URL `http://www4.ibope.com.br/ipm/relatorios/relatorio_inaf_2009.pdf`. (retrieved June 21, 2015).

Inclusion Europe. Information für Alle - Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht, 2009. URL `http://www.inclusion-europe.org/images/stories/documents/Project_Pathways1/DE-Information_for_all.pdf`. (retrieved June 21, 2015).

S. Kandula, D. Curtis, and Q. Zeng-Treitler. A Semantic and Syntactic Text Simplification Tool for Health Content. In *AMIA Annual Symposium Proceedings*, volume 2010, page 366. American Medical Informatics Association, 2010.

G. Kellermann. Leichte und Einfache Sprache - Versuch einer Definition. *Aus Politik und Zeitgeschichte. Bundeszentrale für politische Bildung*, pages 19–25, 2014.

D. Klaper, S. Ebling, and M. Volk. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proc. of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, 2013.

T. Kuhn. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1):121–170, 2014.

C. Maass. *Leichte Sprache - Das Regelbuch*. Lit Verlag Dr. W. Hopf Berlin, 2015.

C. Maass, I. Rink, and C. Zehrer. Leichte Sprache in der Sprach-und Übersetzungswissenschaft. *Sprache barrierefrei gestalten. Perspektiven aus der Angewandten Linguistik. Berlin: Frank & Timme.(= TransÜD. 69)*, pages 53–85, 2014.

K. Matausch and A. Nietzio. Easy-to-read and Plain Language: Defining Criteria and Refining Rules, 2012. URL `http://www.w3.org/WAI/RD/2012/easy-to-read/paper11/`. (last accessed June 24, 2015).

Netzwerk Leichte Sprache. Die Regeln für Leichte Sprache, 2009. URL `http://www.leichtesprache.org/images/Regeln_Leichte_Sprache.pdf`. (retrieved June 21, 2015).

C. K. Ogden. *Basic English: A General Introduction with Rules and Grammar*. Number 29. K. Paul, Trench, Trubner, 1944.

Y. Peng, C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker. iSimp: A Sentence Simplification System for Biomedical Text. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

E. Rennes and A. Jönsson. A Tool for Automatic Simplification of Swedish Texts. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, page 317, 2015.

H. Saggion, E. G. Martínez, E. Etayo, A. Anula, and L. Bourg. Text Simplification in Simplext: Making Text more Accessible. *Procesamiento del lenguaje natural*, 47:341–342, 2011.

R. Sennrich, G. Schneider, M. Volk, and M. Warin. A New Hybrid Dependency Parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124, 2009.

M. Shardlow. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*, pages 58–70, 2014.

A. Siddharthan. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109, 2006.

A. Siddharthan. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298, 2014.

C. Smith and A. Jönsson. Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NoDaLiDa-2010), Riga, Latvia*, 2011.

W. M. Watanabe, A. Candido Jr, M. A. Amâncio, M. De Oliveira, T. A. S. Pardo, R. P. Fortes, and S. M. Aluísio. Adapting Web content for low-literacy readers by using lexical elaboration and named entities labeling. *New Review of Hypermedia and Multimedia*, 16(3):303–327, 2010.

Z. Zhu, D. Bernhard, and I. Gurevych. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics, 2010.

A Evaluation Texts

(A.1) **Original Text**

Unsere Delegation ist gut in Korea angekommen - und wurde sogar von Vertretern der Schweizer Botschaft und der Deutschen Schule willkommen geheissen!

Gleich bei der Ankunft wurde unsere Gruppe von koreanischen Volunteers unter die Fittiche genommen und zum Shing Hun Sa Tempel geführt. Von den Strapazen der langen Reise doch einigermassen ermüdet, fanden alle trotz Aufregung und unzähliger Eindrücke sofort einen tiefen Schlaf im Tempel - Frauen und Männer in getrennten Sälen.

Am nächsten Tag wurden sie in verschiedene Meditationstechniken eingeführt, um sich vor den Games noch einmal so richtig zu entspannen. Aber auch die Unterhaltung kam nicht zu kurz: die Delegation hat den Gangnam Style Tanz eingeübt, was bestimmt äusserst amüsant war.

Am Abend kam hoher Besuch in den Tempel: Timothy Shriver, Chef von Special Olympics weltweit, erlebte zusammen mit der Delegation eine Tee-Meditation.

(A.2) **Expert Translation to Simple German**

Schweizersportler... Müde aber glücklich angekommen

Die Schweizersportler sind gut angekommen
in Korea.
Viele Leute haben ihnen Hallo gesagt.
Diese Leute arbeiten
bei der Schweizer Botschaft
und der Deutschen Schule.

Freiwillige Helfer aus Korea
kümmern sich die Schweizersportler.
Die Helfer haben sie
zum Shing Hun Sa Tempel gebracht.
Unsere Sportler wohnen im Tempel
während der Games.

Alle sind müde
wegen der langen Reise.
Es ist alles sehr aufregend und es gibt viele Eindrücke.

Trotzdem sind alle bald eingeschlafen im Tempel.
Frauen und Männer schlafen in verschiedenen Räumen.

Am nächsten Tag haben unsere Sportler gelernt,
wie man eine Meditation macht.
Meditation ist eine Übung
um sich gut zu entspannen.

Es hat aber auch Unterhaltung gegeben:
Die Gruppe hat den Gangnam Tanz-Stil geübt,
das hat sicher viel Spass gemacht!

Am Abend ist ein wichtiger Besucher
in den Tempel gekommen.
Er heisst Timothy Shriver
und er ist der Chef der Special Olympics.
Unsere Gruppe hat zusammen mit Timothy
eine Tee-Meditation erlebt.

(A.3) **Translation generated by my system**

Unsere Delegation ist gut in Korea angekommen.
Und Vertreter von der Schweizer Botschaft und der Deutschen Schule haben unsere Delegation sogar willkommen geheissen.

> Botschaft ist ein schwieriges Wort.
> Hurraki erklärt es so:
> Eine Botschaft ist eine Vertretung eines Staates in einem anderen Staat.
> Die Vertretung der Regierung eines Landes hat ein Haus in einem anderen Land.
> Das heißt Botschaft.
> Im Beispiel, die Botschaft Deutschlands in den USA.
> Deutsche können dort Hilfe in dem fremden Land bekommen

Gleich bei der Ankunft hat man unsere Gruppe von koreanischen Volunteers unter die Fittiche genommen.
Und man hat unsere Gruppe zum Shing Hun Sa Tempel geführt.

Von den Strapazen von der langen Reis doch einigermassen ermüdet fanden alle trotz Aufregung und unzähliger Ein·drücke sofort einen tiefen Schlaf im Tempel.
Frauen und Männer in getrennten Sälen.

Sie wollten sich vor den Games noch einmal so richtig entspannen.
Deshalb hat man sie in verschiedene Meditationstechniken eingeführt.

Aber auch die Unter·haltung ist nicht zu kurz gekommen.
Die Delegation hat den Gangnam Style Tanz eingeübt.
Das war bestimmt äusserst amüsant.

Am Abend ist hoher Besuch in den Tempel gekommen.
Timothy Shriver hat zusammen mit der Delegation eine Tee-Meditation erlebt.
Timothy Shriver ist Chef von Special Olympics weltweit.

> Chef ist ein schwieriges Wort.
> Hurraki erklärt es so:
> Ein Chef ist im Betrieb der Vorgesetzte oder Verantwortliche.

B List of Python Scripts

(B.1) `text_simplification_for_german.py`
Text Simplification program that analyses source text, applies simplification rules and returns simplified text

(B.2) `classes_for_parsing_results.py`
Module that contains classes for easily accessing parsing results on Sentence and Token level

(B.3) `abbreviations.py`
Module that provides full spellings for abbreviations and acronyms

(B.4) `conjugation.py`
Module for conjugation of verbs

(B.5) `declension.py`
Module for declension of nominals

(B.6) `hurraki.py`
Module that provides Hurraki explanations for difficult words

(B.7) `Abbreviations.txt`
List of German abbreviations derived from Wikipedia

(B.8) `Acronyms.txt`
List of German acronyms derived from Wikipedia

(B.9) `hurraki_words.txt`
List of words explained on Hurraki