Building a German/Simple German Parallel Corpus for Automatic Text Simplification

David Klaper **Sarah Ebling** **Martin Volk**
Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zurich, Switzerland
david.klaper@uzh.ch, {ebling|volk}@cl.uzh.ch

Abstract

In this paper we report our experiments in creating a parallel corpus using German/Simple German documents from the web. We require parallel data to build a statistical machine translation (SMT) system that translates from German into Simple German. Parallel data for SMT systems needs to be aligned at the sentence level. We applied an existing monolingual sentence alignment algorithm. We show the limits of the algorithm with respect to the language and domain of our data and suggest ways of circumventing them.

1 Introduction

Simple language (or, "plain language", "easy-to-read language") is language with low lexical and syntactic complexity. It provides access to information to people with cognitive disabilities (e.g., aphasia, dyslexia), foreign language learners, Deaf people,[1] and children. Text in simple language is obtained through *simplification*. Simplification is a text-to-text generation task involving multiple operations, such as deletion, rephrasing, reordering, sentence splitting, and even insertion (Coster and Kauchak, 2011a). By contrast, *paraphrasing* and *compression*, two other text-to-text generation tasks, involve merely rephrasing and reordering (paraphrasing) and deletion (compression). Text simplification also shares common ground with grammar and style checking as well as with controlled natural language generation.

Text simplification approaches exist for various languages, including English, French, Spanish, and Swedish. As Matausch and Nietzio (2012) write, "plain language is still underrepresented in the German speaking area and needs further development". Our goal is to build a statistical machine translation (SMT) system that translates from German into Simple German.

SMT systems require two corpora aligned at the sentence level as their training, development, and test data. The two corpora together can form a *bilingual* or a *monolingual* corpus. A bilingual corpus involves two different languages, while a monolingual corpus consists of data in a single language. Since text simplification is a text-to-text generation task operating within the same language, it produces monolingual corpora.

Monolingual corpora, like bilingual corpora, can be either *parallel* or *comparable*. A parallel corpus is a set of two corpora in which "a noticeable number of sentences can be recognized as mutual translations" (Tomás et al., 2008). Parallel corpora are often compiled from the publications of multinational institutions, such as the UN or the EU, or of governments of multilingual countries, such as Canada (Koehn, 2005). In contrast, a comparable corpus consists of two corpora created independently of each other from distinct sources. Examples of comparable documents are news articles written on the same topic by different news agencies.

In this paper we report our experiments in creating a monolingual parallel corpus using German/Simple German documents from the web. We require parallel data to build an SMT system that translates from German into Simple German. Parallel data for SMT systems needs to be aligned at the sentence level. We applied an existing monolingual sentence alignment algorithm. We show the limits of the algorithm with respect to the language and domain of our data and suggest ways of circumventing them.

The remainder of this paper is organized as follows: In Section 2 we discuss the methodologies pursued and the data used in previous work deal-

[1] It is an often neglected fact that Deaf people tend to exhibit low literacy skills (Gutjahr, 2006).

Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 11–19,
Sofia, Bulgaria, August 4-9 2013.

ing with automatic text simplification. In Section 3 we describe our own approach to building a German/Simple German parallel corpus. In particular, we introduce the data obtained from the web (Section 3.1), describe the sentence alignment algorithm we used (Section 3.2), present the results of the sentence alignment task (Section 3.3), and discuss them (Section 3.4). In Section 4 we give an overview of the issues we tackled and offer an outlook on future work.

2 Approaches to Text Simplification

The task of simplifying text automatically can be performed by means of rule-based, corpus-based, or hybrid approaches. In a rule-based approach, the operations carried out typically include replacing words by simpler synonyms or rephrasing relative clauses, embedded sentences, passive constructions, etc. Moreover, definitions of difficult terms or concepts are often added, e.g., the term *web crawler* is defined as "a computer program that searches the Web automatically". Gasperin et al. (2010) pursued a rule-based approach to text simplification for Brazilian Portuguese within the *PorSimples* project,[2] as did Brouwers et al. (2012) for French.

As part of the corpus-based approach, machine translation (MT) has been employed. Yatskar et al. (2010) pointed out that simplification is "a form of MT in which the two 'languages' in question are highly related".

As far as we can see, Zhu et al. (2010) were the first to use English/Simple English Wikipedia data for automatic simplification via machine translation.[3] They assembled a monolingual comparable corpus[4] of 108,016 sentence pairs based on the interlanguage links in Wikipedia and the sentence alignment algorithm of Nelken and Shieber (2006) (cf. Section 3.2). Their system applies a "tree-based simplification model" including machine translation techniques. The system learns probabilities for simplification operations (substitution, reordering, splitting, deletion) offline from the comparable Wikipedia data. At runtime, an input sentence is parsed and zero or more simplification operations are carried out based on the model probabilities.

Specia (2010) used the SMT system *Moses* (Koehn et al., 2007) to translate from Brazilian Portuguese into a simpler version of this language. Her work is part of the *PorSimples* project mentioned above. As training data she used 4483 sentences extracted from news texts that had been manually translated into Simple Brazilian Portuguese.[5] The results, evaluated automatically with BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) as well as manually, show that the system performed lexical simplification and sentence splitting well, while it exhibited problems in reordering phrases and producing subject–verb–object (SVO) order. To further improve her system Specia (2010) suggested including syntactic information through hierarchical SMT (Chiang, 2005) and part-of-speech tags through factored SMT (Hoang, 2007).

Coster and Kauchak (2011a; 2011b) translated from English into Simple English using English/Simple English Wikipedia data. Like Specia (2010), they applied *Moses* as their MT system but in addition to the default configuration allowed for phrases to be empty. This was motivated by their observation that 47% of all Simple English Wikipedia sentences were missing at least one phrase compared to their English Wikipedia counterparts. Coster and Kauchak (2011a; 2011b) used four baselines to evaluate their system: input=output,[6] two text compression systems, and vanilla *Moses*. Their system, *Moses-Del*, achieved higher automatic MT evaluation scores (BLEU) than all of the baselines. In particular, it outperformed vanilla *Moses* (lacking the phrase deletion option).

Wubben et al. (2012) also worked with English/Simple English Wikipedia data and *Moses*. They added a post-hoc reranking step: Following their conviction that the output of a simplification system has to be a modified version of the input,[7] they rearranged the 10-best sentences output by *Moses* such that those differing from the

[2] `http://www2.nilc.icmc.usp.br/wiki/index.php/English`

[3] English Wikipedia: `http://en.wikipedia.org/`; Simple English Wikipedia: `http://simple.wikipedia.org/`.

[4] We consider this corpus to be comparable rather than parallel because not every Simple English Wikipedia article is necessarily a translation of an English Wikipedia article. Rather, Simple English articles can be added independently of any English counterpart.

[5] Hence, the corpus as a whole is a monolingual parallel corpus.

[6] The underlying assumption here was that not every sentence needs simplification.

[7] Note that this runs contrary to the assumption Coster and Kauchak (2011a; 2011b) made.

input sentences were given preference over those that were identical. Difference was calculated on the basis of the Levenshtein score (edit distance). Wubben et al. (2012) found their system to work better than that of Zhu et al. (2010) when evaluated with BLEU, but not when evaluated with the Flesch-Kincaid grade level, a common readability metric.

Bott and Saggion (2011) presented a monolingual sentence alignment algorithm, which uses a Hidden Markov Model for alignment. In contrast to other monolingual alignment algorithms, Bott and Saggion (2011) introduced a monotonicity restriction, i.e., they assumed the order of sentences to be the same for the original and simplified texts.

Apart from purely rule-based and purely corpus-based approaches to text simplification, hybrid approaches exist. For example, Bott et al. (2012) in their *Simplext* project for Spanish[8] let a statistical classifier decide for each sentence of a text whether it should be simplified (corpus-based approach). The actual simplification was then performed by means of a rule-based approach.

As has been shown, many MT approaches to text simplification have used English/Simple English Wikipedia as their data. The only exception we know of is Specia (2010), who together with her colleagues in the *PorSimples* project built her own parallel corpus. This is presumably because there exists no Simple Brazilian Portuguese Wikipedia. The same is true for German: To date, no Simple German Wikipedia has been created. Therefore, we looked for data available elsewhere for our machine translation system designated to translate from German to Simple German. We discovered that German/Simple German parallel data is slowly becoming available on the web. In what follows, we describe the data we harvested and report our experience in creating a monolingual parallel corpus from this data.

3 Building a German/Simple German Parallel Corpus from the Web

3.1 Data

As mentioned in Section 1, statistical machine translation (SMT) systems require parallel data. A common approach to obtain such material is to look for it on the web.[9] The use of already available data offers cost and time advantages. Many websites, including that of the German government,[10] contain documents in Simple German. However, these documents are often not linked to a single corresponding German document; instead, they are high-level summaries of multiple German documents.

A handful of websites exist that offer articles in two versions: a German version, often called *Alltagssprache* (AS, "everyday language"), and a Simple German version, referred to as *Leichte Sprache* (LS, "simple language"). Table 1 lists the websites we used to compile our corpus. The numbers indicate how many parallel articles were extracted. The websites are mainly of organizations that support people with disabilities. We crawled the articles with customized Python scripts that located AS articles and followed the links to their LS correspondents. A sample sentence pair from our data is shown in Example 1.

(1) **German:**
Wir freuen uns über Ihr Interesse an unserer Arbeit mit und für Menschen mit Behinderung.
("We appreciate your interest in our work with and for people with disabilities.")

Simple German:
Schön, dass Sie sich für unsere Arbeit interessieren.
Wir arbeiten mit und für Menschen mit Behinderung.
("Great that you are interested in our work. We work with and for people with disabilities.")

The extracted data needed to be cleaned from HTML tags. For our purpose, we considered text and paragraph structure markers as important information; therefore, we retained them. We subsequently tokenized the articles. The resulting corpus consisted of 7755 sentences, which amounted to 82,842 tokens. However, caution is advised when looking at these numbers: Firstly, the tokenization module overgenerated tokens. Secondly, some of the LS articles were identical, either because they summarized multiple AS articles or because they were generic placeholders. Hence, the

[8] http://www.simplext.es/

[9] Resnik (1999) was the first to discuss the possibility of collecting parallel corpora from the web.

[10] http://www.bundesregierung.de/Webs/Breg/DE/LeichteSprache/leichteSprache_node.html (last accessed 15th April 2013)

Short name	URL	No. of parallel art.
ET	www.einfach-teilhaben.de	51
GWW	www.gww-netz.de	65
HHO	www.os-hho.de	34
LMT	www.lebenshilfe-main-taunus.de	47
OWB	www.owb.de	59

Table 1: Websites and number of articles extracted

actual numbers were closer to 7000 sentences and 70,000 tokens.

SMT systems usually require large amount of training data. Therefore, this small experimental corpus is certainly not suitable for large-scale SMT experiments. However, it can serve as proof of concept for German sentence simplification. Over time more resources will become available.

SMT systems rely on data aligned at the sentence level. Since the data we extracted from the web was aligned at the article level only, we had to perform sentence alignment. For this we split our corpus into a training set (70% of the texts), development set (10%), and test set (20%). We manually annotated sentence alignments for all of the data. Example 2 shows an aligned AS/LS sentence pair.

(2) **German:**
In den Osnabrücker Werkstätten (OW) und OSNA-Techniken sind rund 2.000 Menschen mit einer Behinderung beschäftigt.
("In the Osnabrück factories and OSNA-Techniken, about 2.000 people with disability are employed.")

Simple German:
In den Osnabrücker Werkstätten und den Osna-Techniken arbeiten zweitausend Menschen mit Behinderung.
("Two thousand people with disability work in the Osnabrück factories and Osna-Techniken.")

To measure the amount of parallel sentences in our data, we calculated the *alignment diversity measure* (ADM) of Nelken and Shieber (2006). ADM measures how many sentences are aligned. It is calculated as $\frac{2*matches(T1,T2)}{|T1|+|T2|}$, where $matches$ is the number of alignments between the two texts $T1$ and $T2$. ADM is 1.0 in a perfectly parallel corpus, where every sentence from one text is aligned to exactly one sentence in another text.

ADM for our corpus was 0.786, which means that approximately 78% of the sentences were aligned. This is a rather high number compared to the values reported by Nelken and Shieber (2006): Their texts (consisting of encyclopedia articles and gospels) resulted in an ADM of around 0.3. A possible explanation for the large difference in ADM is the fact that most simplified texts in our corpus are solely based on the original texts, whereas the simple versions of the encyclopedia articles might have been created by drawing on external information in addition.

3.2 Sentence Alignment Algorithm

Sentence alignment algorithms differ according to whether they have been developed for bilingual or monolingual corpora. For bilingual parallel corpora many—typically length-based—algorithms exist. However, our data was monolingual. While the length of a regular/simple language sentence pair might be different, an overlap in vocabulary can be expected. Hence, monolingual sentence alignment algorithms typically exploit lexical similarity.

We applied the monolingual sentence alignment algorithm of Barzilay and Elhadad (2003). The algorithm has two main features: Firstly, it uses a hierarchical approach by assigning paragraphs to clusters and learning mapping rules. Secondly, it aligns sentences despite low lexical similarity if the context suggests an alignment. This is achieved through local sequence alignment, a dynamic programming algorithm.

The overall algorithm has two phases, a training and a testing phase. The training phase in turn consists of two steps: Firstly, all paragraphs of the texts of one side of the parallel corpus (henceforth referred to as "AS texts") are clustered independently of all paragraphs of the texts of the other

side of the parallel corpus (henceforth termed "LS texts"), and vice versa. Secondly, mappings between the two sets of clusters are calculated, given the reference alignments.

As a preprocessing step to the clustering process, we removed stopwords, lowercased all words, and replaced dates, numbers, and names by generic tags. Barzilay and Elhadad (2003) additionally considered every word starting with a capital letter inside a sentence to be a proper name. In German, all nouns (i.e., regular nouns as well as proper names) are capitalized; thus, this approach does not work. We used a list of 61,228 first names to remove at least part of the proper names.

We performed clustering with scipy (Jones et al., 2001). We adapted the hierarchical complete-link clustering method of Barzilay and Elhadad (2003): While the authors claimed to have set a specific number of clusters, we believe this is not generally possible in hierarchical agglomerative clustering. Therefore, we used the largest number of clusters in which all paragraph pairs had a cosine similarity strictly greater than zero.

Following the formation of the clusters, lexical similarity between all paragraphs of corresponding AS and LS texts was computed to establish probable mappings between the two sets of clusters. Barzilay and Elhadad (2003) used the boosting tool Boostexter (Schapire and Singer, 2000). All possible cross-combinations of paragraphs from the parallel training data served as training instances. An instance consisted of the cosine similarity of the two paragraphs and a string combining the two cluster IDs. The classification result was extracted from the manual alignments. In order for an AS and an LS paragraph to be aligned, at least one sentence from the LS paragraph had to be aligned to one sentence in the AS paragraph. Like Barzilay and Elhadad (2003), we performed 200 iterations in Boostexter. After learning the mapping rules, the training phase was complete.

The testing phase consisted of two additional steps. Firstly, each paragraph of each text in the test set was assigned to the cluster it was closest to. This was done by calculating the cosine similarity of the word frequencies in the clusters. Then, every AS paragraph was combined with all LS paragraphs of the parallel text, and Boostexter was used in classification mode to predict whether the two paragraphs were to be mapped.

Secondly, within each pair of paragraphs mapped by Boostexter, sentences with very high lexical similarity were aligned. In our case, the threshold for an alignment was a similarity of 0.5. For the remaining sentences, proximity to other aligned or similar sentences was used as an indicator. This was implemented by local sequence alignment. We set the mismatch penalty to 0.02, as a higher mismatch penalty would have reduced recall. We set the skip penalty to 0.001 conforming to the value of Barzilay and Elhadad (2003). The resulting alignments were written to files. Example 3 shows a successful sentence alignment.

(3) **German:**
Die GWW ist in den Landkreisen Böblingen und Calw aktiv und bietet an den folgenden Standorten Wohnmöglichkeiten für Menschen mit Behinderung an – ganz in Ihrer Nähe!
("The GWW is active in the counties of Böblingen and Calw and offers housing options for people with disabilities at the following locations – very close to you!")

Simple German:
Die GWW gibt es in den Landkreisen Calw und Böblingen.
Wir haben an den folgenden Orten Wohn-Möglichkeiten für Sie.
("The GWW exists in the counties of Calw and Böblingen. We have housing options for you in the following locations.")

The algorithm described has been modified in various ways. Nelken and Shieber (2006) used TF/IDF instead of raw term frequency, logistic regression on the cosine similarity instead of clustering, and an extended version of the local alignment recurrence. Both Nelken and Shieber (2006) and Quirk et al. (2004) found that the first sentence of each document is likely to be aligned. We observed the same for our corpus. Therefore, in our algorithm we adopted the strategy of unconditionally aligning the first sentence of each document.

3.3 Results

Table 2 shows the results of evaluating the algorithm described in the previous section with respect to precision, recall, and F1 measure. We introduced two baselines:

Method	Precision	Recall	F1
Adapted algorithm of Barzilay and Elhadad (2003)	27.7%	5.0%	8.5%
Baseline I: First sentence	88.1%	4.8%	9.3%
Baseline II: Word in common	2.2%	8.2%	3.5%

Table 2: Alignment results on test set

1. Aligning only the first sentence of each text ("First sentence")

2. Aligning every sentence with a cosine similarity greater than zero ("Word in common")

As can be seen from Table 2, by applying the sentence alignment algorithm of Barzilay and Elhadad (2003) we were able to extract only 5% of all reference alignments, while precision was below 30%. The rule of aligning the first sentences performed well with a precision of 88%. Aligning all sentences with a word in common clearly showed the worst performance; this is because many sentences have a word in common. Nonetheless, recall was only slightly higher than with the other methods.

In conclusion, none of the three approaches (adapted algorithm of Barzilay and Elhadad (2003), two baselines "First sentence" and "Word in common") performed well on our test set. We analyzed the characteristics of our data that hampered high-quality automatic alignment.

3.4 Discussion

Compared with the results of Barzilay and Elhadad (2003), who achieved 77% precision at 55.8% recall for their data, our alignment scores were considerably lower (27.7% precision, 5% recall). We found two reasons for this: language challenges and domain challenges. In what follows, we discuss each reason in more detail.

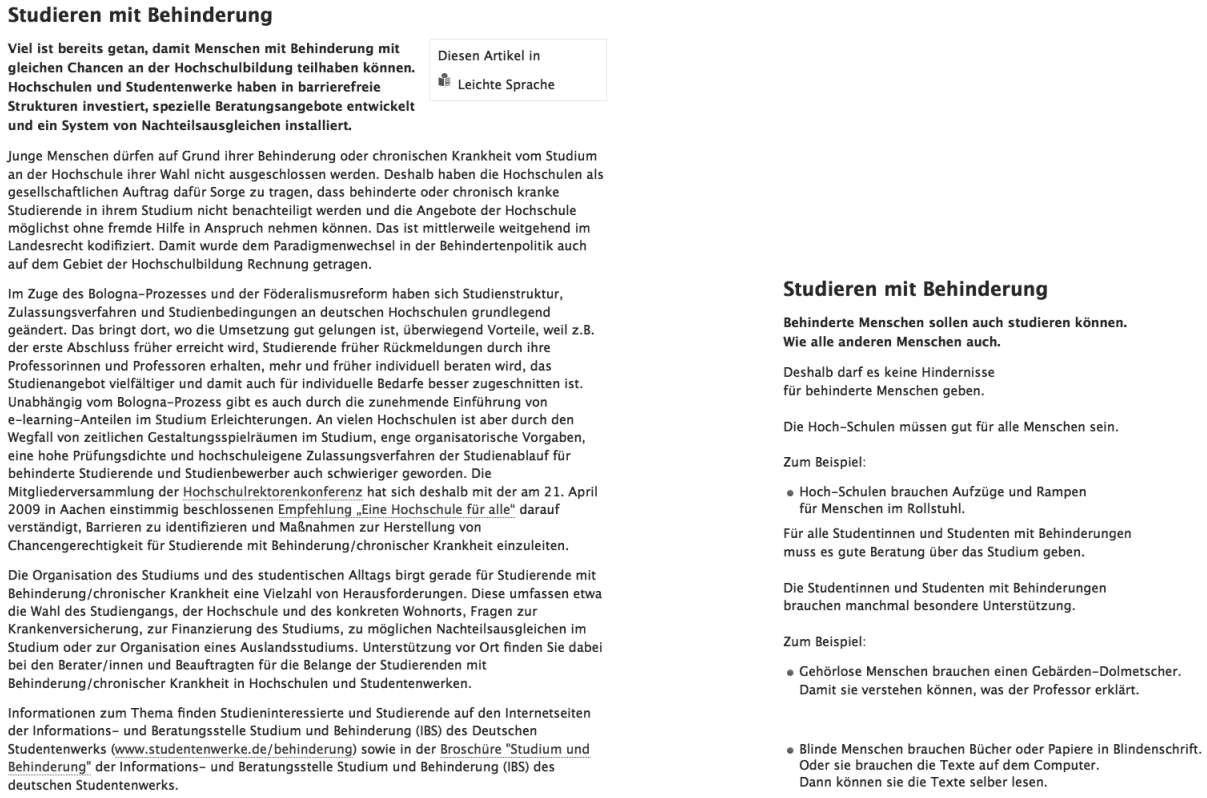
While Barzilay and Elhadad (2003) aligned English/Simple English texts, we dealt with German/Simple German data. As mentioned in Section 3.2, in German nouns (regular nouns as well as proper names) are capitalized. This makes named entity recognition, a preprocessing step to clustering, more difficult. Moreover, German is an example of a morphologically rich language: Its noun phrases are marked with case, leading to different inflectional forms for articles, pronouns, adjectives, and nouns. English morphology is poorer; hence, there is a greater likelihood of lexical overlap. Similarly, compounds are productive in German; an example from our corpus is *Seniorenwohnanlagen* ("housing complexes for the elderly"). In contrast, English compounds are multiword units, where each word can be accessed separately by a clustering algorithm. Therefore, cosine similarity is more effective for English than it is for German. One way to alleviate this problem would be to use extensive morphological decomposition and lemmatization.

In terms of domain, Barzilay and Elhadad (2003) used city descriptions from an encyclopedia for their experiments. For these descriptions clustering worked well because all articles had the same structure (paragraphs about culture, sports, etc.). The domain of our corpus was broader: It included information about housing, work, and events for people with disabilities as well as information about the organizations behind the respective websites.

Apart from language and domain challenges we observed heavy transformations from AS to LS in our data (Figure 1 shows a sample article in AS and LS). As a result, LS paragraphs were typically very short and the clustering process returned many singleton clusters. Example 4 shows an AS/LS sentence pair that could not be aligned because of this.

(4) **German:**
Der Beauftragte informiert über die Gesetzeslage, regt Rechtsänderungen an, gibt Praxistipps und zeigt Möglichkeiten der Eingliederung behinderter Menschen in Gesellschaft und Beruf auf.
("The delegate informs about the legal situation, encourages revisions of laws, gives practical advice and points out possibilities of including people with disabilities in society and at work.")

Simple German:
Er gibt ihnen Tipps und Infos.



Studieren mit Behinderung

Viel ist bereits getan, damit Menschen mit Behinderung mit gleichen Chancen an der Hochschulbildung teilhaben können. Hochschulen und Studentenwerke haben in barrierefreie Strukturen investiert, spezielle Beratungsangebote entwickelt und ein System von Nachteilsausgleichen installiert.

Diesen Artikel in Leichte Sprache

Junge Menschen dürfen auf Grund ihrer Behinderung oder chronischen Krankheit vom Studium an der Hochschule ihrer Wahl nicht ausgeschlossen werden. Deshalb haben die Hochschulen als gesellschaftlichen Auftrag dafür Sorge zu tragen, dass behinderte oder chronisch kranke Studierende in ihrem Studium nicht benachteiligt werden und die Angebote der Hochschule möglichst ohne fremde Hilfe in Anspruch nehmen können. Das ist mittlerweile weitgehend im Landesrecht kodifiziert. Damit wurde dem Paradigmenwechsel in der Behindertenpolitik auch auf dem Gebiet der Hochschulbildung Rechnung getragen.

Im Zuge des Bologna-Prozesses und der Föderalismusreform haben sich Studienstruktur, Zulassungsverfahren und Studienbedingungen an deutschen Hochschulen grundlegend geändert. Das bringt dort, wo die Umsetzung gut gelungen ist, überwiegend Vorteile, weil z.B. der erste Abschluss früher erreicht wird, Studierende früher Rückmeldungen durch ihre Professorinnen und Professoren erhalten, mehr und früher individuell beraten wird, das Studienangebot vielfältiger und damit auch für individuelle Bedarfe besser zugeschnitten ist. Unabhängig vom Bologna-Prozess gibt es auch durch die zunehmende Einführung von e-learning-Anteilen im Studium Erleichterungen. An vielen Hochschulen ist aber durch den Wegfall von zeitlichen Gestaltungsspielräumen im Studium, enge organisatorische Vorgaben, eine hohe Prüfungsdichte und hochschuleigene Zulassungsverfahren der Studienablauf für behinderte Studierende und Studienbewerber auch schwieriger geworden. Die Mitgliederversammlung der Hochschulrektorenkonferenz hat sich deshalb mit der am 21. April 2009 in Aachen einstimmig beschlossenen Empfehlung „Eine Hochschule für alle" darauf verständigt, Barrieren zu identifizieren und Maßnahmen zur Herstellung von Chancengerechtigkeit für Studierende mit Behinderung/chronischer Krankheit einzuleiten.

Die Organisation des Studiums und des studentischen Alltags birgt gerade für Studierende mit Behinderung/chronischer Krankheit eine Vielzahl von Herausforderungen. Diese umfassen etwa die Wahl des Studiengangs, der Hochschule und des konkreten Wohnorts, Fragen zur Krankenversicherung, zur Finanzierung des Studiums, zu möglichen Nachteilsausgleichen im Studium oder zur Organisation eines Auslandsstudiums. Unterstützung vor Ort finden Sie dabei bei den Berater/innen und Beauftragten für die Belange der Studierenden mit Behinderung/chronischer Krankheit in Hochschulen und Studentenwerken.

Informationen zum Thema finden Studieninteressierte und Studierende auf den Internetseiten der Informations- und Beratungsstelle Studium und Behinderung (IBS) des Deutschen Studentenwerks (www.studentenwerke.de/behinderung) sowie in der Broschüre "Studium und Behinderung" der Informations- und Beratungsstelle Studium und Behinderung (IBS) des deutschen Studentenwerks.

Studieren mit Behinderung

Behinderte Menschen sollen auch studieren können. Wie alle anderen Menschen auch.

Deshalb darf es keine Hindernisse für behinderte Menschen geben.

Die Hoch-Schulen müssen gut für alle Menschen sein.

Zum Beispiel:

- Hoch-Schulen brauchen Aufzüge und Rampen für Menschen im Rollstuhl.

Für alle Studentinnen und Studenten mit Behinderungen muss es gute Beratung über das Studium geben.

Die Studentinnen und Studenten mit Behinderungen brauchen manchmal besondere Unterstützung.

Zum Beispiel:

- Gehörlose Menschen brauchen einen Gebärden-Dolmetscher. Damit sie verstehen können, was der Professor erklärt.

- Blinde Menschen brauchen Bücher oder Papiere in Blindenschrift. Oder sie brauchen die Texte auf dem Computer. Dann können sie die Texte selber lesen.

Figure 1: Comparison of AS and LS article from `http://www.einfach-teilhaben.de`

("He provides them with advice and information.")

Figure 2 shows the dendrogram of the clustering of the AS texts. A dendrogram shows the results of a hierarchical agglomerative clustering. At the bottom of the dendrogram every paragraph is marked by an individual line. At the points where two vertical paths join, the corresponding clusters are merged to a new larger cluster. The Y-axis is the dissimilarity value of the two clusters. In our experiment the resulting clusters are the clusters at dissimilarity $1 - 1^{-10}$. Geometrically this is a horizontal cut just below dissimilarity 1.0. As can be seen from Figure 2, many of the paragraphs in the left half of the picture are never merged to a slightly larger cluster but are directly connected to the universal cluster that merges everything. This is because they contain only stopwords or only words that do not appear in all paragraphs of another cluster. Such an unbalanced clustering, where many paragraphs are clustered to one cluster and many other paragraphs remain singleton clusters, reduces the precision of the hierarchical approach.

4 Conclusion and Outlook

In this paper we have reported our experiments in creating a monolingual parallel corpus using German/Simple German documents from the web. We have shown that little work has been done on automatic simplification of German so far. We have described our plan to build a statistical machine translation (SMT) system that translates form German into Simple German. SMT systems require parallel corpora. The process of creating a parallel corpus for use in machine translation involves sentence alignment. Sentence alignment algorithms for bilingual corpora differ from those for monolingual corpora. Since all of our data was from the same language, we applied the monolingual sentence alignment approach of Barzilay and Elhadad (2003). We have shown the limits of the algorithm with respect to the language and domain of our data. For example, named entity recognition, a preprocessing step to clustering, is harder for German than for English, the language Barzilay and Elhadad (2003) worked with. Moreover, German features richer morphology than English, which leads to less lexical overlap when working on the word form level.

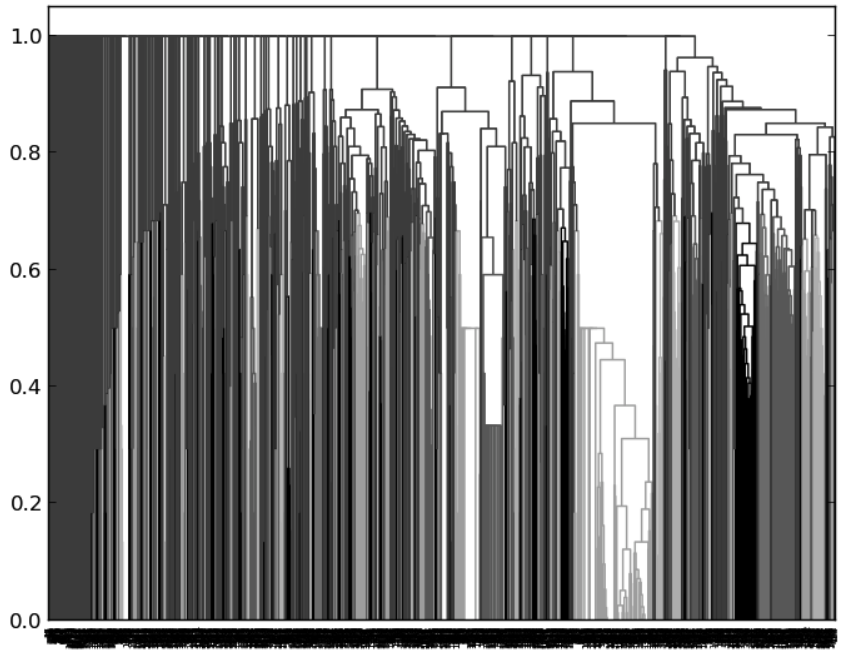


Figure 2: Dendrogram of AS clusters

The domain of our corpus was also broader than that of Barzilay and Elhadad (2003), who used city descriptions from an encyclopedia for their experiments. This made it harder to identify common article structures that could be exploited in clustering.

As a next step, we will experiment with other monolingual sentence alignment algorithms. In addition, we will build a second parallel corpus for German/Simple German: A person familiar with the task of text simplification will produce simple versions of German texts. We will use the resulting parallel corpus as data for our experiments in automatically translating from German to Simple German. The parallel corpus we compiled as part of the work described in this paper can be made available to interested parties upon request.

References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of EMNLP*.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA.

Stefan Bott, Horacio Saggion, and David Figueroa. 2012. A Hybrid System for Spanish Text Simplification. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 75–84, Montréal, Canada, June.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2012. Simplification syntaxique de phrases pour le français. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 211–224.

David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *ACL-05: 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, University of Michigan, Ann Arbor, Michigan, USA.

William Coster and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 1–9, Stroudsburg, PA, USA.

William Coster and David Kauchak. 2011b. Simple English Wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 665–669, Stroudsburg, PA, USA.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *HLT 2002: Human Language Technology Conference, Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, California.

Caroline Gasperin, Erick Maziero, and Sandra M. Aluisio. 2010. Challenging choices for text simplification. In *Computational Processing of the Portuguese Language. Proceedings of the 9th International Conference, PROPOR 2010*, volume 6001 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 40–50, Porto Alegre, RS, Brazil. Springer.

A. Gutjahr. 2006. *Lesekompetenz Gehörloser: Ein Forschungsüberblick*. Universität Hamburg.

Hieu Hoang. 2007. Factored Translation Models. In *EMNLP-CoNLL 2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. 2001. SciPy: Open Source Scientific Tools for Python.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Kerstin Matausch and Annika Nietzio. 2012. Easy-to-read and plain language: Defining criteria and refining rules. `http://www.w3.org/WAI/RD/2012/easy-to-read/paper11/`.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 161–168.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 311–318, Philadelphia, PA, USA.

Chris Quirk, Chris Brocket, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings Empirical Methods in Natural Language Processing*.

Philip Resnik. 1999. Mining the Web for Bilingual Text. In *37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 527–534, University of Maryland, College Park, Maryland, USA.

Robert E. Schapire and Yoram Singer. 2000. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2–3):135–168.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language. Proceedings of the 9th International Conference, PROPOR 2010*, volume 6001 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 30–39, Porto Alegre, RS, Brazil. Springer.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 1015–1024, Jeju Island, Korea.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 365–368.

Z. Zhu, D. Bernhard, and I. Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the International Conference on Computational Linguistics*, pages 1353–1361.