A New Aligned Simple German Corpus

Vanessa Toborek¹, Moritz Busch¹, Malte Boßert¹, Christian Bauckhage¹,², Pascal Welke¹,³
¹University of Bonn, ²Fraunhofer IAIS, ³TU Wien

Email: toborek@cs.uni-bonn.de

Background
- Text in simple language benefits language learners, people with learning difficulties, and children
- Most popular dataset Simple English Wikipedia (Coster and Kauchak 2011)
- In German, there are two forms of simple language: Leichte Sprache (LS) and Einfache Sprache (ES)
- There is no publicly available Simple German resource in any of the above simplifications
- Text simplification requires aligned resources

Our Dataset
- Around 700 Simple German articles from eight different web sources with aligned German text
- Corpus spans different topics from general health information over administrative information to general news
- Seven sources in LS, one extensive source in ES
- Around 250,000 tokens in Simple German
- Article aligned corpus by design, which we extended by sentence alignments

Evaluation of Sentence Alignments
- Two fold evaluation of the combination of eight similarity measures and two matching methods
- (left) Manual creation of gold standard alignments allowed calculation of F1 score
  - sbert MST-LIS F1 score of 0.32
- (right) Manual evaluation of potential matches according to Xu et al. 2015 to calculate the Manual Alignment Classification Accuracy
  - max MST-LS 55.94%
  - We choose max MST-LIS for better precision

Check out
- The biggest corpus in Simple German, completely publicly available
- Over 700 aligned documents from eight different web sources
- Over 10,000 matched sentences with a final F1 score of 0.28 compared to previously 0.085 (Klaper et al. 2013)
- Publicly available gold standard alignments
- Well preserved resources:
  - you can find the code on GitHub
  - the original sites are preserved at web.archive.org

---

*Example sentence pairs aligned between Simple German (LS) and German (AS) and their translations. Examples show successfully, partially, and wrongly aligned sentences.*

<table>
<thead>
<tr>
<th>Source</th>
<th>Simple German</th>
<th>German</th>
<th>a</th>
<th>t</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>apothekel-umschau.de</td>
<td>General health information</td>
<td>168 94 808</td>
<td>166 187 427</td>
<td>ES</td>
<td></td>
</tr>
<tr>
<td>behindenbeauftragte.de</td>
<td>Official office for disabled people</td>
<td>21 5 490</td>
<td>21 8 131</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>branden.de</td>
<td>Various topics</td>
<td>47 9 634</td>
<td>47 9 728</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>lebenshilfe-main-taunus.de</td>
<td>NPO for disabled people</td>
<td>45 6 946</td>
<td>45 9 023</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>mdr.de</td>
<td>State-funded public broadcasting</td>
<td>322 53 277</td>
<td>322 126 191</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>sozialpolitik.com</td>
<td>Explains social policy in Germany</td>
<td>15 5 122</td>
<td>15 11 790</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>stadt-koln.de</td>
<td>Administrative information</td>
<td>82 66 892</td>
<td>82 44 310</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>tu.de</td>
<td>German Newspaper</td>
<td>8 7 924</td>
<td>14 8 171</td>
<td>LS</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>708 250 093</td>
<td>712 404 711</td>
<td></td>
</tr>
</tbody>
</table>

Overview of the websites used for the corpus and some basic statistics. (a) articles and (t) tokens per source. The last column describes the type of Simple German used by the website „Einfache Sprache“ (ES) or „Leichte Sprache“ (LS).

---

*Example sentence pairs aligned between Simple German (LS) and German (AS) and their translations. Examples show successfully, partially, and wrongly aligned sentences.*

---

*Example sentence pairs aligned between Simple German (LS) and German (AS) and their translations. Examples show successfully, partially, and wrongly aligned sentences.*

---

Bibliography: