Creation of a parallel simplification corpus
– Using the annotation tool TS-anno

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

24. Januar 2022
Version 2.0

Inhaltsverzeichnis

1 The corpus creation at a glance/motivation

These instructions are considered as instructions and assistance for building a simplification corpus. The corpus construction includes the following steps:

- data selection
- web scraping
- alignment of sentence pairs
- evaluation of sentence pairs
- annotation of changed words during simplification and categorization of change

In the following, we briefly review the overall context between the steps before going into more detail about each step.

For the construction of an automated text simplification system, parallel data in simplified and everyday language are required. In this work, comparable documents in two different language levels are used as the data basis: in simplified and everyday language. The documents are website texts or texts accessible online. In order to process the texts further, the web pages must be downloaded and the content contained therein must be separated from the navigation and advertising elements. After this process of web scraping, the texts are pre-processed. Pre-processing includes dividing the text into sentences and tokens. Data selection, web scraping and preprocessing are not part of the annotation task and are therefore not described in more detail.

The process of alining and annotation, as well as the structure of the data is illustrated by a small example. As example documents serve document A, a news text to topic X in everyday language, and document E, likewise a news text to topic X however in simplified language. Document A and B are comparable in content and can therefore be aligned. Alining can take place on different levels, such as document, paragraph, sentence, or word level. In this corpus, sentences from documents with comparable content, such as document A and E, are alined. That is, sentences with the same or comparable content are linked together, e.g., sentence A1 from document A and sentence E1 from document E. These and all other alined sentences will serve as input to the text simplification system.

After alining, each alined sentence pair, e.g., the pair A1E1 from sentence A1 and sentence E1, is evaluated in terms of sentence quality and simplification quality. The evaluation allows conclusions to be drawn about whether the evaluated pair is a good fit for the corpus or not. For example, pair A1E1 would be usable if sentence E1 is simpler and easier to understand than sentence A1 and E1 still retained most of the meaning of sentence A1.

Pair A1E1 would not be usable if sentence A1 was simpler than sentence E1 or if either sentence had serious grammatical errors.

In addition to the evaluation, the simplification process of each alined sentence pair is described in the annotation. For each sentence pair, it is indicated which words or parts of sentences have changed and how this change can be named. For example, if sentence A1 contains subjunctive and a long compound, but sentence E1 is in the indicative and the long compound has been resolved, these same two transformations would be named and the words involved would be marked. The transformations are helpful in that they can be used to specify and improve the text simplification system. For example, if only a few sentences were insufficiently simplified by the simplification system, these sentences could be looked at more closely. The closer analysis could then examine whether the insufficiently simplified sentences were, for example, all previously in the subjunctive. From this, it would be concluded that further rules or features would be needed to resolve the subjunctive in the system.

In the following, alinement, evaluation, and transformation annotation are described in more detail. The following description can be referred to throughout the annotation process, for example, to clear up ambiguities and enable consistent annotation. The annotation tool "TS-anno" is used to support the annotation. In the following instructions, the use of the annotation tool is also explained.

2 Annotation Tool

Each user of the annotation tool can be assigned texts for alining at will. The texts are sorted by corpus beforehand. Each corpus contains at least one document pair, which in turn consists of an everyday language document and a simplified document. Under Corpora-Overview an overview of all available corpora of the current user is displayed. Clicking on a corpus redirects to the overview of all content-similar documents or document pairs of the corpus (Corpus Overview). Clicking on a document pair either starts the alining or an overview of all alined sentence pairs of the document is displayed (Document Overview).

On the left side there is a navigation menu that is available during all steps of the annotation process. Using the navigation menu you can switch to the Document Overview, Corpus Overview or Corpora Overview at any time.

If a change has been made at the time of using the Navigation Menu, this change will be not saved. Saving must always be done using the "Save button".

During alining and annotation, time is stopped in the background by task. This helps to draw conclusions how long or how difficult the evaluation

of a sentence pair or alining of a document pair was.

If problems or suggestions arise while using the annotation tool, they can be added under the Change Log menu item.

3 Alignment

At the beginning of the annotation, at least one corpus with at least one document pair AE consisting of the documents A (everyday language) and E (simple language) with comparable content is available in the annotation tool. If a document pair has not yet been alined, a click on the document will redirect to the alining interface. Otherwise, all alined sentence pairs are displayed with the possibility to edit the alignment, evaluate the sentence pairs and annotate the transformations. In the alignment interface, document A and E are displayed side by side in parallel. As shown in Figure , both documents have already been divided into sentences. In the following, sentences with the same content are to be linked. For this, the alignment must be enabled first. The alignment can be activated by clicking on the "Add" button.

There are several options for alining or different numbers of sentences can be included in the alining, depending on in or from how many sentences the meaning was transferred. The following options are possible:

1. 0:1 alignment: insertion or addition
2. 1:0 alignment: omission or deletion
3. 1:1 alignment
   (a) no-change: sentence A and sentence E are identical
   (b) change: sentence A and sentence E are different
4. n:1 alining: summary
5. 1:n alignment: division
6. n:n alignment: shift across sentences

The options do not need to be mentioned by name in the alignment. The options are used in the instruction only as clues and to illustrate different alinement options.

3.1 Automated Alignment

1:1 no-change sentence pairs are automatically alined by the annotation tool. Identical sentences in the everyday language and simplified document are not available for manual alining to speed up the annotation of these pairs and avoid accidental alining.

In addition, 0:1 and 1:0 alignments are also performed automatically by the annotation tool. If a sentence from the everyday language document has not been alined with another, it is handled as an omission and stored with a 1:0 alignment. A sentence is considered an omission only if the content does not appear in any of the simplified sentences, for example, if the content does not seem relevant enough for simplification. A sentence is not considered removed if part of the content appears in a simplified sentence. In that case, the sentence should be manually annotated together with the main source sentence as a sentence summary (1:n). If, on the other hand, a sentence from the simplified document has not been alined with another, it is marked as an insertion or addition and with an aline 0:1. If part of the content is contained in one or more source sentences, a sentence summary (n:1) should be annotated manually. An addition is, for example, a word explanation or an explanation of a word or concept enriched with examples.

0:1, 1:0, and no-change pairs are included in the manual scoring to allow comparability of the grammaticality of the source sentences and simplified sentences with the other pairs. However, since no change due to simplification can be evaluated, the evaluation is limited to fewer items than the evaluation of the other pairs. .

The remaining 1:1, the n:1, 1:n, and n:n alignments must be done by hand. It must always be taken into account that the respective sides of the sentence pair should be as similar as possible.

3.2 Manual Alignment

In order to align a sentence pair, the corresponding sentences of the everyday language document and the simplified document must be selected by clicking on the respective sentences. Then confirm the alignment by clicking on the Save button. For an alignment, at least one sentence of the everyday language document and the simplified document must be selected. Otherwise, an error message will be displayed. After successful saving, the sentence pair is displayed under “List of aligned sentence pairs”. Then either the next sentence pair can be alined, the current sentence pair can be evaluated, or the current sentence pair can be annotated with transformations. If a sentence pair has been aligned prematurely and actually another sentence should be assigned to the sentence pair, the edit button can be used to adjust the alignment.

3.2.1 Within-Sentence Simplification

Probably the most common aliasing possibility is 1:1 aliasing. It is annotated when the meaning of a source sentence is transferred into exactly one simplified sentence and a change or simplification is made. 1:1 alinings are annotated when the meaning of an everyday language sentence is transferred

into exactly one simplified sentence. Examples include rearranging the everyday language sentence, omitting words in the simplification, adding words in the simplification, or changing words in the simplification (e.g., changing the tense of a verb).

Since n:1, 1:n as well as n:n alignments are possible, it is recommended to go through all sentences of both documents step by step and to adjust already stored alignments if necessary. If the meaning of a sentence in everyday language cannot be found in one of the simplified sentences, this sentence pair will not be annotated. The same applies the other way around for the simplified sentences.

3.2.2 Cross-Sentence Simplification

In addition to intra-sentence simplification, cross-sentence simplifications may have been applied or may need to be annotated, e.g., 1:n, n:1, n:n alignments.

1:n alignment 1:n alignments are particularly useful when parts of a sentence have been split among several. This is the case, for example, when subordinate clauses are rewritten as independent clauses. As soon as a fraction of the content of the original sentence also appears in a previous or following simplified sentence, it should be annotated 1:n. The order and distance between the simplified sentences does not matter. However, if a word from an everyday sentence is picked up in a simplified sentence, but only further information about the word is given that is not in the everyday sentence, the simplified sentence is alined as 0:1 and not assigned to any everyday sentence.

n:1 alining n:1 alinings describe the combination of multiple everyday language sentences into one simplified sentence. Once fractions of content from a second everyday language sentence have been integrated into a simplified sentence, it is an n:1 alignment and no longer a 1:1 alignment. In this case, a few pieces of information from the everyday language sentences are often omitted and not included in the simplified sentence. The loss of content can then be indicated in the evaluation. An example of n:1 alignment is the combination of two main clauses with subordinate clauses to a simplified main clause connected with “and” in which the subordinate clauses have been omitted.

n:n alining If there is a mixture of a summary and division of multiple clauses into multiple sentences, the sentences are annotated with an n:n alignment. Thus, several cross-sentence simplification transformations occur simultaneously. For example, a subordinate clause from a first complex sentence might be split into a new simplified sentence, while at the same time

the main information of a second complex sentence is integrated into the main clause of the first complex sentence. This form of aliasing can be done whenever no clear sentence assignment can be made. For this purpose, several everyday language sentences as well as several simplified sentences are selected in the annotation tool.

Since the everyday language side of a sentence pair and the simplified side of a sentence pair can contain more than one sentence, in the following we talk about everyday language text and simplified text to avoid ambiguities.

4 Rating

If at least one sentence pair per document has been annotated, the overview of all alined sentence pairs of this document is displayed in the Document Overview. For each sentence pair, a rating and the transformations of the simplification can be added. The evaluation of the sentence pairs is done in three parts:

- rating of the initial sentence (everyday language).
- evaluation of the simplified sentence
- evaluation of the sentence pair

Criteria are established for the evaluation of each subfield. The rating of each criterion is determined based on the agreement or disagreement of a propositional sentence on a scale of 5 values. The endpoints of the rating scale are "disagree at all." ("strongly disagree") to "strongly agree." ("strongly agree"). The perception of whether a text is easy or difficult to understand depends on the annotator's experience, vocabulary, interests, language skills, and more. Therefore, there is no right or wrong score. The value to be chosen is at the discretion of the annotating person. The following examples and notes are provided only to guide the scoring and may also vary from person to person.

The end points are to be selected when a clear decision can be made for or against something. The intermediate points represent an existing but not clear rejection or agreement. The middle point is to be selected when

1. there is no change concerning the respective criterion or
2. there is no tendency, that is, when the statement can neither be agreed nor disagreed with.

The interpretation of the mean value depends on the particular statement sentence. The first variant will be referred to as the criterion with neutral element in the following. To support the annotation of the neutral element, the scale points for criteria with neutral element range between -2 (negative),

0 (neutral), and +2 (positive). The second variant of the mean value can occur, for example, when parts of the text contradict the statement but others endorse the statement. In this case, the scale points range from 1 (negative) to 5 (positive).

If the text of the simplified or everyday language side of the sentence pair consists of more than one sentence, the rating must be applied to the entire text. For example, if one of two sentences in a text on one side has grammatical errors, the highest grammaticality score can no longer be used. If the errors also overshadow the correctness of the second sentence, the grammaticality of the entire text is to be given a low value. The same applies to the other categories.

In the annotation tool, the ratings are made in groups. First, the evaluation criteria of the everyday language texts are displayed. Then those of the simplified texts and finally the comparative criteria between everyday language and simplified texts. For each criterion, the sentence to be evaluated is displayed. There are 5 radio buttons available per criterion. When a value or button is selected, the corresponding value is displayed. The display of the value supports the distinction between criteria with and without a neutral element.

An overview of the statements per aspect is summarized in Table 1.

4.1 Rating of the Original and the Simplified Sentence

The rating of the original and simplified text indicates how comprehensible the texts are in each case. The distinction is important because the rating of the original text can be used to determine the initial quality of the text.

The scores of the simplified text and the everyday text may differ or may be identical. It is possible that the values of the simplified text change in comparison to the everyday text in the negative as well as in the positive direction. For example, a simplified text may be rated with a lower simplicity value and higher ambiguity value than the corresponding everyday language text. Thus, the simplified text would be more difficult to understand than the everyday language text assumed to be more difficult.

The criteria to be used to evaluate the simplified text and the everyday language text are explained below. An overview of the criteria statements to be rated is provided in the upper half of Table 1. The criteria are namely:

1. grammaticality,
2. simplicity,
3. contextuality, and
4. ambiguity.

4.1.1 Grammaticality

A simplified text should be as grammatically correct as possible. However, if a text in everyday language already has grammatical errors, it could be that these have persisted in the simplification. It is necessary to consider whether a grammatical error is an error of the simplification system or a consequence of the incorrect source sentence. To estimate this, the grammaticality of the source sentence and that of the simplified sentence are determined independently. A cumbersome but grammatically correct formulation of a sentence, is also assigned a high value.

The evaluation should be done on the basis of the following examples: If the text does not contain any grammatical errors, the text can be rated with the highest value, because the statement of grammatical correctness can be agreed with. If the text contains one or two unobtrusive grammatical errors, but can still be read very fluently, the text can still be given a high value. The value to be chosen is up to the annotator to decide. It should be noted, however, that the sentence structure or sentence construction is not evaluated in the criterion of grammaticality, but in the criterion of syntactic simplification.
If there are some to few grammatical errors in the text, an intermediate value can be chosen. The evaluation is done to what extent the statement of grammatical correctness can be agreed or disagreed with. If a text contains very many grammatical errors, the text with the lowest value for grammaticality is to be evaluated, because in this way the statement of grammatical correctness cannot be agreed with. The same principle is transferable to the other criteria.

4.1.2 Simplicity

If an initial set is already very simple, not much can be improved during simplification. Accordingly, in this case, a high simplicity value would not be directly attributable to the quality of the simplification system. The value would only indicate that the system has not made the initial set more complicated than it was before. However, if an output sentence is very complicated and the automatically generated simplification gets a high simplicity value, this speaks for the quality of the simplification system, since a strong simplification has taken place.

The evaluation should be done as follows: If a text has to be read several times to understand it, a low value should be selected. If a text can be clearly understood on the first reading, then a higher value should be selected. The value to be selected is at the discretion of the annotator. However, if the text is very easy to understand, then the highest value for Simplicity should be selected. Many different clues can be consulted to evaluate simplicity, such as sentence length, sentence structure, familiarity of words, general

comprehensibility, and many more.

4.1.3 Context Independence

Furthermore, the context independence of both texts will be evaluated. Since the focus in automatic text simplification is on the simplification of sentence pairs and not on text pairs, the context of a sentence plays an important role. The less context is given in a source text, the more difficult it is to establish this context if other surrounding sentences are not included. This challenge is included in the evaluation by the criterion of context independence.

The following yardstick should be used for scoring: If the text is understandable only if previous and following sentences are included, then the lowest value for "context independence" should be selected. For example, if pronouns (e.g., he, her, this) or deictic adverbs (e.g., there, here, down, yesterday, before, soon) are included in the text whose reference is not apparent within the sentence, a lower value should be selected. The value to be chosen is at the discretion of the annotator.

4.1.4 Ambiguity

The goal of a simplified text is to be as unambiguous as possible and to resolve any ambiguities of a source text. With the help of the criterion "ambiguity" this point also flows into the evaluation.

If the text has multiple readings, it is to be scored with the highest value for ambiguity. If a word in the text has multiple readings, but within the text its reading is clear, a lower value is to be selected. The value to be selected is at the discretion of the annotator.

4.2 Rating of the Alignment Pair

Besides the independent evaluation of the simplified text and the everyday language text of the alined sentence pair, the simplified text can also be evaluated in relation to the everyday language text. The following 5 categories are evaluated:

1. meaning preservation
2. information gain
3. Simplification
4. Syntactic simplification
5. Lexical simplification

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

Tabelle 1: Englische Bewertungskriterien.

Contrary to the simplicity rating of the individual texts of the pair, in the criterion of simplification the change of simplicity is determined by the simplification. The distance of the rating of the simplified text and the rating of the everyday language text, can be computed complementarily as the change in simplicity. The evaluation of simplification is further specified into the simplification of sentence structure (syntactic simplification) and word

choice (lexical simplification). Some text simplification systems specialize in either lexical or syntactic simplification, so the evaluation of pairs with respect to these evaluation criteria is also useful.

4.2.1 Meaning Preservation

Only in dependence on the everyday language text can it be determined whether the meaning or the content in the simplified text has changed or not. The evaluation of the sense is complicated by the fact that a component of the simplification can be the omission of words or parts of sentences. If only a few passages that are not relevant for understanding the text are omitted, the sense retention can be evaluated with a high value. However, if important, indispensable information is missing from the simplified text, the pair should be given a low sense retention score. However, which information is considered indispensable is at the discretion of the annotator.

4.2.2 Information Gain

In the area of manual text simplification, complicated words are often explained in an insertion or in a new sentence. Furthermore, the content, which is only readable between the lines, is often formulated more explicitly in the simplified version. The criterion of information gain therefore evaluates whether new information has been added to the everyday language text, or whether information has been formulated more explicitly. If one of both is the case, a high value is to be selected. If, however, information has been omitted or formulated in a more implicit way, the text is to be given a low value. If neither information has been added nor omitted, the middle value is to be selected, because the criterion of information gain is a criterion with a neutral element. Whether new information was added or the existing information was only rewritten is again at the discretion of the annotator.

4.2.3 Simplicity

The evaluation of simplification is to be carried out according to the same criteria as the evaluation of simplicity. However, the criterion of simplification is a criterion with a neutral element. If the simplified text is much easier to understand than the everyday language text, the simplification is to be rated with the highest value. If the simplified text is as complicated as the everyday text, the middle value is to be selected. The lowest value is to be selected if the simplified text is significantly more difficult than the everyday language text. The intermediate values are to be selected if the simplified text is slightly more difficult or slightly easier, but not significantly more difficult or easier, than the everyday language text. However, the value to be selected is at the discretion of the annotator.

4.2.4 Syntactic Simplicity

The evaluation of syntactic simplification includes, for example, the separation of subordinate clauses into new main clauses, the reformulation of passive clauses into active clauses, or the change to the order subject-verb-object. If any of these changes occurred and simplified the sentence, one of the high values can be selected for the syntactic simplification score. However, if a sentence is changed from indicative to subjunctive, the sentence has probably become more complicated and a low value would be appropriate. If there has been no change in the complexity of sentence construction or structure, the middle value should be selected. Even if there has been a change in sentence structure, but the syntactic complexity has not changed, the middle value should be chosen. If grammatical errors have been corrected, the sentence structure of the simplified sentence may be simpler than that of the everyday language sentence, this depends on how serious the grammatical errors were or how much they disturbed the reading flow. Again, the value to be selected is at the discretion of the annotator. The syntactic simplification criterion, like the simplification criterion, is a criterion with a neutral element.

4.3 Lexical Simplification

In contrast to syntactic simplification, lexical simplification refers to the word choice of the sentence and not to the sentence structure. If difficult words were replaced by simpler ones during simplification, or if difficult words were omitted, a high value for lexical simplification should be selected. If, on the other hand, new difficult words were added, a lower value should be selected. If the overall word choice has become neither more complicated nor simpler, then the middle value should be selected. The decision of whether a word is difficult or simple is at the discretion of the annotator. The lexical simplification criterion, also like the simplification and syntactic simplification criteria, is a neutral element criterion.

5 Annotation of Rewriting Transformations

5.1 General

When comparing or evaluating the alined sentence pairs, the change or the simplification process becomes clear. The individual steps of the change are called transformations in the following. The transformations help to check which changes were made in a manual simplification and, moreover, which changes would be possible in an automatic simplification.

Based on a literature review in the area of text simplification and simple language, 56 transformations were found to be particularly relevant. The transformations can be grouped into the following 9 transformation classes:

1. Lexical substitution (lexical substitution),
2. splitting (split),
3. removal (deletion),
4. addition (insert),
5. reordering (reorder),
6. verbal changes (verbal changes),
7. merge,
8. merge,
9. merge,
10. merge,
11. merge,
12. paraphrasing (rephrasing, paraphrasing), and
13. no change (no operation).

The transformation classes are general names for lexical or syntactic changes made during simplification. The transformation classes can be specified by 38 subclasses. Furthermore, the 9 transformation classes can be divided into four different levels. The levels are text levels that are affected during simplification. The categories include:

1. word level,
2. sentence level,
3. sentence level, and
4. paragraph level

A transformation class can occur at more than one level; for example, an addition is possible at word level, sentence part level, and paragraph level.

In Figure 1 is an overview of all transformation classes and all subclasses as well as their corresponding text levels. The 9 text classes have been represented at each level at which they occur so that a total of 18 classes (in bold) can be seen on the figure. The 38 subclasses have been assigned to their superclasses with arrows. The figure also contains information about the relevance or frequency of mention in the literature. The higher the priority value the less frequently the transformation was mentioned. In the corpus to be created, the transformation classes and their level are annotated. The

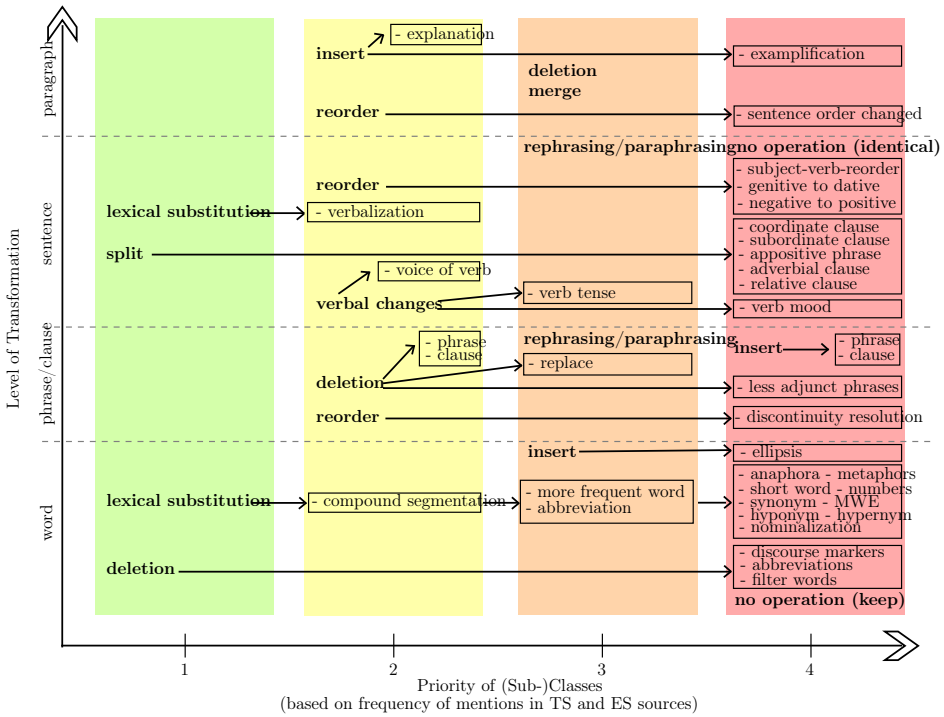

Abbildung 1: classes-subclasses-level-priority

subclasses serve as examples of the associated transformation classes. If several changes have been made at the same time for the simplification of the sentence, several of these options can also be chosen or annotated. If none of the selected options apply, the user can add more options.

In addition to the transformation, the corresponding tokens are annotated. If no clear token can be identified, none is selected. Only obvious transformations should be annotated. If the text was simplified but it cannot be clearly differentiated what was simplified, no transformation must be annotated. The annotation of the tokens is done by clicking on the corresponding tokens in the displayed text. Depending on the text, simplified or everyday language, in which the change has taken place, the tokens are marked. For example, an addition can only be marked in the simplified sentence, a removal only in the everyday text. A lexical substitution or change in verb form is annotated in both texts to make the change traceable even if multiple transformations have occurred simultaneously.

In the annotation tool, the levels are displayed first. Selecting a layer opens a list of transformation classes. With selection of a transformation class again a listing opens, this time one of the individual transformations. The listing also contains a free text field that can be used to specify a further transformation. Currently, the listing only opens when the radio button is

pressed. Unfortunately, it does not currently open when the name of a class is clicked.
If two words (on word level) have been replaced by lighter ones in a simplification, two transformations have to be created, one transformation for each replacement. If it is a replacement of a multi-word unit (e.g. “bite the dust” or “runs ... away”), however, a transformation including all tokens of the multi-word unit is suitable.

5.2 Rewriting Transformations

The individual transformations are explained below sorted by level of occurrence and transformation class.

5.2.1 Paragraph–Level

The paragraph level contains transformations that either involve more than one everyday language sentence (n:1 pair, reordering or merging), cannot be assigned to any everyday language text (0:1 pair, addition), or cannot be assigned to any simplified sentence (1:0 pair, removal). All paragraph-level transformations are automatically annotated. For example, a merge can be recognized by the fact that the everyday text consists of at least two sentences, whereas the simplified text consists of only one sentence. Insertion and deletion are annotated if they have not been assigned to a sentence in the other language level after the alignment process. Rearrangement (reorder) is automatically annotated based on the sentence numbers: For example, if an everyday language sentence was annotated at position x (where $1 \leq x \leq n$ and n = number of all sentences in the everyday language document) with a simplified sentence at position x+j (where $j > i$) and the sentences of another alined pair are at position x+i (where $i \geq 1$), the order of the sentences was changed.

5.2.2 Sentence Level

At the sentence level are all transformations that are applied within the boundaries of a sentence. The only exception is the split. This transformation affects one everyday language sentence but several simplified sentences. The sentence level also only contains transformations that affect the entire structure of the sentence. That is, transformations that can be assigned to only one sentence part or token, but do not affect the complete sentence structure, are not assigned to this level. The transformations of this category can be counted for the most part to syntactic simplification.

No Operation The most special transformation on sentence level is the one in which no change (no operation) has taken place. This transformation is annotated automatically.

Splitting Probably the most common transformation at sentence level is the split. This transformation could also be annotated automatically, but the reason for the split is of interest. Possible reasons are, for example, the separation of a relative clause (starting with e.g.: the, the, that, which, etc.), an adverbial clause (starting with e.g.: by, if, wherefore, although, while), a subordinate subordinate clause (starting with e.g.: because, nevertheless, but, therefore, etc.), or a subordinate subordinate clause (starting with e.g.: and, or, both..and, neither, etc.). For the transformation of the division it is sufficient to annotate the transformation, a marking of the tokens is not necessary.

Reordering Rearrangement can take place not only at the paragraph level but also at the sentence level. Reordering at the sentence level includes, for example, moving individual sentence parts (general), changing the order of subject, verb, and object (subject verb reorder), changing a negative phrase to a positive one (negative to postive), or replacing the genitive with the dative (genitive to dative). Annotation of the associated tokens should be attempted here. If too many tokens are involved or they cannot be clearly identified, annotation of the tokens can be omitted.

Verbal Changes Changes in the verb are also annotated at the sentence level, since the sentence structure can change significantly, for example, when a sentence is activated. Besides the change from passive to active (voice of verb), the change of mode (mood, subjunctive and indicative) and the change of tense (verb tense) are in this transformation class. It can be assumed that in this transformation class the associated tokens are relatively unique, so that in addition should be annotated with.

Lexical Substitution Another transformation class, that of lexical substitution, contains changes concerning verbs. Often sentences are formulated in nominal style, if they were changed in simplification to a verbal style, the transformation of verbalization can be annotated.

Rephrasing & Paraphrasing If no clear transformation(s) of the simplification can be seen, for example because too many words have been changed and/or moved, the simplification can be marked with paraphrase or paraphrasing. In this transformation, the associated tokens are not expected to be annotated.

5.2.3 Clause- and Phrase-level

One specification of the sentence level is the clause and subordinate clause level. On this level, the whole sentence is not involved in the change, but

clear, delimited word groups are identifiable. The word groups are mainly subordinate clauses and other sentence parts.

Deletion For example, the sentence parts can be omitted completely if they are not relevant enough for the simplified text. Here, it is necessary to distinguish whether the sentence part has been omitted completely or is taken up in a new sentence. In the case of the latter, the sentence division is the applicable transformation. In the case of the former, however, the current one. Examples of sentence parts that are superfluous or to be omitted are phrases, subordinate clauses, or other insertions. The tokens in question should also be annotated in this transformation.

Insertion The opposite transformation to removal is addition. For example, phrases or subordinate clauses can be added. The tokens in question should also be annotated in this transformation.

Reordering Just as on the sentence level, TOken can be shifted within a sentence part. This includes, for example, when connected words are moved closer together (discontinuity resolution). An example of distance resolution is the rearrangement of verbs and associated particles. Annotation of the affected tokens is recommended here as well.

Rephrasing & Paraphrasing A paraphrase of a sentence part is also possible. As already on sentence level no associated tokens have to be annotated within this transformation.

5.2.4 Word-Level

The finest level of transformations is that of the word level. In this level, the transformations refer to single words and multi-word units. The transformations of this category can be counted for the most part to the lexical simplification.

Lexical Substitution A widely researched area of lexical simplification is that of complex word recognition and substitution. This type of change is labeled with the transformation of lexical substitution. There can be many reasons for lexical substitution. The following list summarizes the most common reasons and most frequently selected alternatives:

- Compound segementation: separation or subdivision of parts of speech within a word with hyphens.
- Replacement of a rare word by a more frequent word.

- Replacement of a complex word by a simpler one (simpler word in general)
- Replacement of an abbreviation by the long version (abbrevation vs. long version)
- replacement of anaphors by the corresponding referents (anaphora)
- Replacement of a long word by a shorter word (shorter word)
- Replacement of a difficult word by a synonym (synonym)
- Replacement of a difficult word by a subword (hyponym)
- Replacement of a difficult word by a superterm (hypernym)
- Replacement of a term by its nominal version (nominalization)
- Replacement of metaphors and figurative language (metaphor)
- replacement of number words by numerals and other number-related substitutions (number)
- Replacement of multi-word units by one word (MWE)

For each of the above transformations, the associated tokens should also be annotated.

A lexical substitution can also be seen as combining the transformations of removing a word and adding a word. However, the transformation of a removal or the addition would not be sufficient, since via the lexical substitution it is expressed that a content-equal or content-similar exchange was made.

Deletion Also at the word level is the transformation class of distance. Possible transformations or omissions are, for example: Abbreviations (abbreviation removal), filler words (filter words) or superfluous complicated words (complex words). Here it is to be noted that the tokens do not reappear in the simplified text, if it is only in another place, the reordering transformation is the correct one. Furthermore discourse markers, which establish a connection between two or more sentences, can be removed (discourse marker). The omitted token has to be marked in the annotation tool before the evaluation is sent.

Insertion Although shorter texts are more understandable, words can be added in simplifications. For example, in ellipses in everyday language texts, their omissions could be added (filling ellipsis). Moreover, it is possible that a sentence was formulated more explicitly and during that new words were added (new words). In all transformation types it is requested to add the relevant tokens.

No Operation If no change was made to a word, this is also annotated. However, this happens automatically. The information that a word has not been changed is significant in that the word is apparently understandable enough and does not need to be replaced.

6 Summary

In summary, this annotation project creates a corpus that can be used to create a simplification system. On the one hand, the annotation consists of alining pairs of text with the same content from different text levels. On the other hand, the alined pairs are evaluated according to their quality and the process of simplification is recorded in the form of transformation annotations. An annotation tool is provided to support the annotation process. This document is considered as a handout to operate the annotation tool as well as to guide the annotation process. Questions or comments regarding the annotation process or the annotation tool can be sent to the authors of this document[1].

[1]This document was automatically translated with DeepL (https://www.deepl.com/translator).