An Intralingual Parallel Corpus of Translations into German Easy Language (Geasy Corpus): What Sentence Alignments Can Tell Us About Translation Strategies in Intralingual Translation

Silvia Hansen-Schirra, Jean Nitzke, and Silke Gutermuth

Abstract Parallel corpora are traditionally interlingual and contain source and target texts in different languages. However, intralingual translations into Easy Language (EL) become more and more common in various countries. First intralingual corpora have been built up and investigated in terms of linguistic and structural features, but a translation-driven corpus linguistic approach is still missing to empirically describe the strategies of Easy Language translation, the characteristics of translated texts as well as to make these parallel corpora usable for professionalising and automatising translation processes. In this paper, we introduce an intralingual parallel corpus of translations into German Easy Language (*Geasy Corpus*). It contains published professional translations from Standard German into German Easy Language, including different text types and various formulation guidelines for German Easy Language. Currently, the corpus contains 1,087,643 words of source text and 292,552 words of Easy Language translations. So far, 93 (of 276) texts have been sentence aligned. We compare descriptive values, investigate the alignments, and describe which translation strategies are revealed to give first empirical evidence on the characteristics of Easy Language translations. Finally, we will discuss the potentials of tree annotations for Easy Language corpora, summarise our findings and give on outlook on future research.

S. Hansen-Schirra · J. Nitzke (✉) · S. Gutermuth
Johannes Gutenberg University, Mainz, Germany
e-mail: jean.nitzke@uia.no

S. Hansen-Schirra
e-mail: hansenss@uni-mainz.de

S. Gutermuth
e-mail: gutermsi@uni-mainz.de

J. Nitzke
University of Adger, Kristiansand, Norway

V. X. Wang et al. (eds.), *New Perspectives on Corpus Translation Studies*,
New Frontiers in Translation Studies, https://doi.org/10.1007/978-981-16-4918-9_11

Keywords Parallel corpora · Intralingual translation · Easy Language · Translation strategies

1 Introduction

Since the implementation of the UN Convention on the Rights of Persons with Disabilities in Germany in 2009, accessible communication, and especially intralingual translations into Easy Language become more and more common. From an applied perspective, Easy Language was developed for people with intellectual disabilities with the aim to reduce linguistic complexity and to enhance comprehensibility. From a scientific-empirical point of view, first parallel corpora have been built and investigated in terms of linguistic and structural features (e.g., Battisti et al. 2019; Klaper et al. 2013). The scientific examination of these corpora, however, still lacks a translation-driven approach to empirically contrast the characteristics of source and target texts as well as to examine the strategies in Easy Language translation. These parallel corpora can help professionalizing the field and automatizing translation processes (as suggested in Hansen-Schirra et al. 2020a). Hence, we want to start exploring the field by investigating sentence alignments and what they can tell us about potential translation strategies.

In this paper, we will first present a brief overview about Easy Language in Germany (Sect. 2). We will then introduce an intralingual parallel corpus of translations into German Easy Language (Geasy Corpus) that we currently build (Sect. 3). The corpus contains published professional translations from Standard German into German Easy Language. The corpus includes different text types and various formulation guidelines for German Easy Language (Bredel and Maaß 2016a, b; Netzwerk Leichte Sprache 2013; Inclusion Europe 2009). Both corpus creation and alignment processes are work in progress. We will compare descriptive values like lexical density, sentences length, type-token ratio, and term density, as well as briefly describe the use of pictures in the corpus to shed light on the differences in information density and structural complexity of German Easy Language. In Sect. 4, we will quantify and investigate the sentence alignment of the texts that have been aligned so far. Section 5 will present how these alignments can be used to potentially reveal translation strategies. Sentence alignment is not very straightforward in Easy Language corpora, as the translation is usually not a direct equivalent of the source text segment. There happen to be considerable changes between source and target texts which are necessary to make the texts readable and comprehensible for the target audience. Therefore, we will also describe the shortcomings of a sentence-based analysis and present the potential benefits of a tree alignment. Finally, we will summarize our findings and present some ideas for future research.

2 A Quick Introduction to German Easy Language

Access to information for people with disabilities has become an important issue in the countries that have ratified the UN Convention on the Rights of People with Disabilities. Easy Language is a linguistic approach to reduce language complexity for the sake of better readability and comprehensibility. It is, therefore, seen as one of the central pillars of communicative inclusion. Accordingly, the legislation concerning Easy Language progressed greatly in recent years. In Germany, the Convention on the Rights of Persons with Disabilities of the United Nations (UN CRPD) has significantly changed the way in which disabilities are addressed in the political and legal discourse. The implementation of the UN CRPD is on the agenda of each German Federal state, each municipality, each regional parliament, and local authority, as well as the Federal Government and its public bodies. Laws and regulations have been passed and implemented at the federal and national level (on the legal situation of Easy Language, see Lang 2019). On this basis, paragraph 11 on "Comprehensibility and Easy Language" [1]was added to the Act on Equal Opportunities of Persons with Disabilities ("Behindertengleichstellungsgesetz") in 2018. For a more detailed historical development and a comparison of different varieties of EL in Europe see Lindholm and Vanhatalo (forthcoming).

As a result, political and public institutions have to face the need to translate existing texts with domain-specific contents into Easy Language. Easy Language ("Leichte Sprache" in German) addresses recipients with cognitive disabilities, prelingual hearing loss, aphasia, dementia type illnesses and Parkinson's disease. They can be regarded as primary target groups, which are legally entitled to Easy Language (roughly 3% of the German population, cf. Maaß et al. forthcoming). In addition, there are other communication impairments which afford Easy Language texts, such as poor reading or language skills (e.g., migrants).

Translating specialized or technical content to the target group means filling the gap between expert knowledge and the knowledge of the recipient. Easy Language is used as the means to create a common ground (Pickering and Garrod 2004). It is a sort of controlled language variety of Standard German, which aims for a better readability and comprehensibility of texts. A controlled language is a subset of a natural language such as German, which is restricted according to certain rules (Lehrndorfer 1996). Controlled languages have traditionally been used for technical documentation to make them more consistent and more comprehensible. Within the context of accessible communication, other text types are relevant for EL translation, too. This especially holds true for administrative and legal texts since they are mentioned in the respective regulations and acts to ensure participation of the target groups (see above). More recently, EL translation has also been considered for multimedia text types (e.g., EL subtitling, cf. Maaß and Hernández Garrido 2020) or for literary texts (cf. Maaß et al. forthcoming). Another area of application is the medical discourse where illnesses and their treatment are explained, or informed

[1]https://www.gesetze-im-internet.de/bgg/__11.html, last accessed 08/05/21

consent documents are translated for the target groups (cf. Maaß 2020). It is hard to quantify the amount of EL texts since they are typically used as printed leaflets and translated by various stakeholders (e.g., translation agencies, empowerment associations, editors in federal ministries or publishing houses, etc.). However, the amount of EL texts is increasing rapidly.

So far, rules and formulation guidelines for Easy Language for German have been based on practical experience (Inclusion Europe 2009; Netzwerk Leichte Sprache 2013) or linguistic theory (Bredel and Maaß 2016a, b). The guidelines and rule sets for Easy Language can vary a lot and can also be rather vague. Further, not all guidelines are freely accessible. In general, they suggest, amongst others:

- limitations in the lexicon: specific terms and lexical variation should be avoided
- reduced complexity on the morphological level: compounds should be avoided, or segmentation aids (hyphen or mediopoint) should be used
- reduced complexity on the phrasal level: complex pre- and post-modifications of phrases (e.g., genitive attributes) should be avoided
- reduced complexity on the syntactic level: subordinate clauses should be avoided and rephrased in main clauses instead
- reduced complexity on the textual level: repetition of nouns should be used instead of pronouns to build cohesive ties; each sentence should include only one proposition
- typographic facilitation: negations should be printed in bold face; each sentence should start on a new line; the integration of pictures should enhance the comprehension of key concepts and terminology.

Maaß (2015) differentiates between general principles of Easy Language, rules regarding the semiotic level, word level, sentences level, text level, rules regarding typography and layout as well as rules regarding translation strategies. Some rules influence the sentence structure and, hence, can be used to interpret the alignment results. One of the general principles for Easy Language texts is an orientation towards activities, which includes amongst others a spread and repetition of information (instead of compensating information and avoiding redundancies, ibid.: 76–81). The information selection for an Easy Language text/translation depends on the topic of the text (ibid.: 129). This implies for Easy Language texts that not all information of the target text necessarily has to be integrated into the target text. Often, information is added that is not explicitly available in the source text to explain or exemplify difficult terminology or phrases. On the other side, information from the source text is reduced to make the amount of text suitable for the target group. A comprehensive comparison of the different rule sets for German can be found in Maaß (2020) and for European EL varieties in Lindholm and Vanhatalo (forthcoming).

The following example presents some of the mentioned characteristics (reduced complexity, added and reduced information, adapted typography—more examples of German Easy Language and its source text can be found in the following sections):

- **SL:** Die UN-Konvention
  Die Unterzeichnung der UN-Behindertenrechtskonvention durch Deutschland während der deutschen EU-Ratspräsidentschaft war ein politisch wichtiges Signal für die anderen Mitgliedstaaten der Europäischen Union.
  (English: The UN Convention
  The signing of the UN Convention on the Rights of Persons with Disabilities by Germany during the German EU Council Presidency was a politically important sign for the remaining member states of the European Union)
  EL: Die UN-Konvention
  Konvention ist ein anderes Wort für **Vertrag**.
  Wir sprechen:
  Kon-wen-zion
  Die **UN-Konvention** ist ein Vertrag zwischen sehr vielen Ländern auf der Welt.
  Die **Vereinten Nationen** haben diesen Vertrag geschlossen.
  Für **Vereinte Nationen** sagen wir auch **UN**.
  Deshalb nennen wir den Vertrag:
  UN-Konvention.
  In dem Vertrag steht:
  Menschen mit und ohne Behinderung haben die gleichen Rechte.
  Menschen mit und ohne Behinderung müssen gleich behandelt werden.
  Die Regierungen müssen alle Hindernisse für Menschen mit Behinderung beseitigen.
  Auch viele Länder in der EU haben die UN-Konvention unterschrieben.
  (English:
  The UN Convention
  Convention is another word for **contract**.
  We pronounce it:
  Con-wen-tion
  The **UN Convention** is a contract between a lot of countries in the world.
  The **United Nations** entered this contract.
  We call the **United Nations** also **UN**.
  Therefore, we call the contract:
  UN Convention
  The contract says:
  People with and without disabilities have the same rights.
  People with and without disabilities have to be treated equally.
  The governments have to eliminate all obstacles for people with disabilities.
  Many countries of the EU have signed the UN conventions, too.

Currently, these rules have been empirically evaluated from a user-based perspective (Bock 2019, Hansen-Schirra et al. 2020b, Deilen 2020, Schiffl 2020, Sommer 2020, Gutermuth 2020) and result in recommendations to optimise or specify the rules.

First corpus studies focus on rather computational-linguistic aspects. Battisti et al. (2019) use their corpus of simplified German texts for a natural language

approach and for unsupervised machine learning to empirically investigate "whether different complexity levels exist in previous German simplification practice in the first place" (ibid.: 3). Klaper et al. (2013) built a parallel corpus with Standard German and Simple German with the aim to have training data for statistical machine translation systems. Developing machine translation systems for intralingual translation is also discussed in Hansen-Schirra et al. (2020a). However, a corpus-based quantification of existing Easy Language usage patterns, translation strategies, and contrastive text characteristics remains a research desideratum, which we try to address in this paper with the help of the Geasy Corpus.

3 The German Easy Language (Geasy) Corpus

In translation studies, we differentiate between two types of corpus designs: the parallel corpus and the monolingually comparable corpus. A parallel corpus is defined as a collection of source language texts and translations of those texts into a target language. In computational linguistics, such corpora are used in bilingual lexicography (Sahlgren and Karlgren 2005) and as training corpora for machine translation systems (Koehn 2005; Artetxe et al. 2017). In translation research, parallel corpora provide information on language-pair specific translation patterns and to investigate translation quality (Zanettin 2000).

Monolingually comparable corpora are collections of translations and original texts in the target language. Comparable corpora "should cover a similar domain, variety of language and time span, and be of comparable length" (Baker 1995: 23); they have

> the potential to reveal most about features specific to translated text, i.e., those features that occur exclusively, or with unusually low or high frequency, in translated text as opposed to other types of text production, and that cannot be traced back to the influence of any one particular source text or language (Kenny 1997).

Comparable corpora are used to test hypotheses on translation-universals or specific features of translations such as explicitation, simplification, normalization/conservatism (Baker 1995, Hansen-Schirra et al. 2013).

The Geasy Corpus is a combination of both corpus types described. It consists of parallel but monolingual subcorpora. The corpus is parallel since it includes source and target texts, which we are currently aligning. However, both of these monolingual subcorpora are in German language. The translation direction is Standard German into Easy Language German, i.e., the Geasy Corpus is a monolingual, parallel corpus containing texts from various genre. The translations were created according to different guidelines because different guidelines are common in Germany as mentioned above. In the following, we will present some basic characteristics of both subcorpora in comparison. The alignment data will be analysed and assessed in Sects. 4 and 5.

Currently, the corpus contains 1,087,643 words of source text and 292,552 words of Easy Language translations. For now, most source and target texts are publicly available on website, usually provided by a public organisation, etc. So far, 93 texts (33.7% of 276 texts in total) from three sources have been sentence aligned. Not all texts in the corpus are suitable for alignment. For example, one source text contains 59,795 words, while the Easy Language translation only consists of 2,728 words. We can assume that the choice was deliberate to reduce the information of the Easy Language text that drastically. However, we decided that this example and other source and target texts were too different for alignment and would not fit our purposes. This also applies, e.g., for leaflets in Easy Language that give additional information to enable the reader to fill out a form.

The source texts of the aligned data contain 33,061 words and 1596 sentences, which results in a medium sentence length of 20.7 words per sentence. The Easy Language translations of the aligned data consist of 41,722 words and 4090 sentences (average sentence length: 10.2 words/sentence). Easy Language translations, therefore, present the contents in more words and more sentences (we will discuss information restructuring and sentences splitting in Sects. 4 and 5). On average, sentences in Easy Language are only half as long as sentences in Standard Language. Both characteristics (more sentences and less words per sentence) point into the direction that the Easy Language texts are less complex than the Standard Language texts, which corresponds to the goals of Easy Language. Certain rules for Easy Language texts encourage sentence splitting, e.g., there should be only one information per sentences, subclauses should be avoided, or information should be spread (Maaß 2015). To further investigate the complexity of Easy Language translations compared to Standard Language texts, we will have a closer look a lexical density and the type-token ratio.

Table 1 shows the distribution of four main lexical word classes (all content words) in the Easy Language and Standard subcorpora. The part-of-speech tagging was carried out with Sketch Engine (Kilgarriff et al. 2014). The general assumption is that the lexical density in Easy Language should be lower compared to Standard Language (Hansen-Schirra and Gutermuth 2018), but this is not the case in our data (t-test: $t(5.82) = 0.41$, $p = 0.7$). In total, the numbers for lexical density are very similar. Interestingly, we find a similar frequency for nouns in the two subcorpora. According to Maaß (2015: 76), however, Easy Language texts should rather be in verbal than in nominal style. One explanation for the high number of nouns might be that the texts often are domain-specific (e.g., explaining legal contents) and, hence, have to use and introduce important concepts and terms. We can observe a

Table 1 Distribution of lexical word classes in the Geasy Corpus

	Easy Language	Standard German
Nouns	13,128 (31.47%)	11,119 (33.63%)
Verbs	7572 (18.15%)	4163 (12.59%)
Adjectives	2403 (5.76%)	3437 (10.40%)
Adverbs	2258 (5.41%)	1122 (3.39%)
TOTAL	25,361 (60.79%)	19,841 (60.01%)

significant shift in the usage of verbs and adjectives ($X^2(1) = 867.34$, $p < 0.0001$). Fewer adjectives and more adverbs are used in Easy Language, which is in line with an increase in verbal style. Further, adjectives are often used in complex nominal phrases, which should be avoided in Easy Language. The frequency of verbs increases in the Easy Language corpus. This corroborates the results of our sentences analysis. As mentioned above, information is split—noun phrases, enumerations, etc. are dissolved—and presented in single sentences as in example 1. In example 1, the people who can get counselling and support are summarised in one sentence in the source text, while they are listed in two separate sentences in the target text. Further, an explanation for chronic diseases is added to the target text.

(1) SL: Die Beratungsstelle bietet Menschen mit körperlichen Behinderungen oder chronischen Erkrankungen und ihre Angehörigen Beratung und Unterstützung […] an.[2] (English: The counselling centre offers counselling and support to people with physical handicaps or chronic diseases and their relatives […].)
EL: Menschen mit Körper-Behinderungen und ihre Familien können Beratung und Hilfe bekommen. Die Beratung ist auch für Menschen mit chronischen Krankheiten. Chronisch heißt in Leichter Sprache: Diese Krankheiten gehen nicht mehr weg. (English: People with physical handicaps and their families can get counselling and help. The counselling is also for people with chronic diseases. Chronic means in Easy Language: The disease is permanent.)

Table 2 presents the type-token ratio of the Easy Language and Standard subcorpora. The assumption concerning this feature is that the easier the text, the lower the type-token ratio becomes, which seems to be true on first sight in the analysed section of the Geasy corpus. The figures suggest that fewer lexical items and terms are introduced in the Easy Language texts, which makes it less lexically diverse and accordingly less complex. This is in line with the findings in Hansen-Schirra and Gutermuth (2018: 16). More sophisticated analyses will shed more light on the complexity of the texts. Some studies already started to explore complexity in EL texts, e.g., Gutermuth (2020) correlated textual complexity with cognitive processing costs while reading EL texts and Battisti et al. (2019) applied unsupervised machine learning techniques to EL texts to cluster and classify complexity levels of EL corpora.

Finally, we address the quantification and characterisation of the use of images and pictures. None of the source texts contained pictures or images. According to Maaß (2015: 86) different semiotic representations help recipients of Easy Language to understand the contents of the texts. Therefore, it is valid to use pictures, figures, or to highlight important terms or phrases in the texts. Only one text of the 93 corpus texts in Easy Language did not contain pictures in the original formatting. In total, 867 pictures were counted in the target texts (mean = 9.42 SD = 8.83). The first two subcorpora used the same set of pictures (*Lebenshilfe*

[2]All examples discussed in the paper are taken from the Geasy Corpus. For further information see https://traco.uni-mainz.de/geasy-korpus/, last accessed 08/05/21.

Table 2 Type-token ratios of the two subcorpora in the Geasy Corpus

	Easy Language	Standard German
Types per 100 tokens	7.35	18.51

Bremen), while the third subcorpus used different pictures. All pictures were designed in a comic style, except for some real-life pictures in the third corpus that referred to real life objects, e.g., the town hall of Hamburg. Interestingly, some pictures recurred in the texts, usually when same or similar concepts are introduced in different texts, probably to increase the memorability of the concepts.

In summary, the analysed subcorpus already shows differences in characteristics between source and target texts. These differences confirm that the Easy Language translations seem to be less complex than the Standard Language source texts. However, more data points might be necessary to get significant result for some aspects.

4 Alignment Characteristics

In this section, we want to analyse the sentence alignment of the subcorpus that was aligned so far. Source and target texts were aligned with the help of the translation memory tool *memsource*. Source and target texts can be automatically prealigned and then be viewed and corrected manually.

The basis for the alignment process was the source text. In total, the subcorpus consists of 1816 alignments. Different alignments can be observed between source and target text. These will be introduced with a brief interpretative approach and quantified in the following:

- 1:1 alignment—one source sentence is aligned with one target sentence
- 1:0 alignment—a source sentence has no equivalent in the target sentence
- n:1 alignment—several sentences in the source texts are aligned with one sentence in the target text
- 0:n alignment—several sentences in the target text have no equivalent in the source text
- 1:n alignment—one source text sentence is represented by several sentences in the target text
- n:m alignment—several sentences are represented by several sentences in the target text
  - n > m (without n:1)—more sentences in the source sentence were aligned with less sentences in the target text
  - n = m (without 1:1)—several sentences in the source text were aligned with an equal number of sentences in the target text
  - n < m (without 0:n and 1:n)—less sentences in the source sentence were aligned with more sentences in the target text

Table 3 Quantification of alignments in absolute numbers and %

	Absolute figures	in %
1:1 alignment	664	36.56
1:0 alignment (including one 2:0 alignment)	17	0.94
n:1 alignment	21	1.16
0:n alignment = > 0:1	146 = > 6	8.04
1:n alignment = > 1:2	868 = > 308	47.80
n:m alignment = > n > m (without n:1) = > n = m = > n < m (without 0:n and 1:n)	100 = > 4 = > 29 = > 67	5.51

The following table shows the absolute and relative figure of our alignments.

Table 3 shows that most alignments are 1:n-alignments, followed by 1:1-alignments. 0:n-alignments and n:m-alignments occurred in the lower medium section, while the least alignments can be counted for n:1 and 1:0 alignments. These findings corroborate the results by Klaper et al. (2013) who report very low scores for automatically aligning an intralingual corpus of Standard and Easy Language. They attribute this low performance to specificities of the language variety, the domain and massive changes from source to target text.

Furthermore, the alignments suggest that, for those texts, Easy Language translations rather add, enrich and restructure information than reduce or delete contents. On the basis of the alignment figures, one could assume that it might be more likely that information is added or enriched than that the information and structure is identical. This seems also to be the case for the n:m-alignments, where n < m-alignments were counted more often than n = m-alignments, and by far as n > m-alignments. To shed more light on this assumption, we will examine the alignments in further detail and interpret them with respect to translation strategies in the next section.

5 Translation Strategies

Let us look at some phenomena in more detail. Different translation strategies can be applied when translating from Standard Language into Easy Language. These strategies might be similar to interlingual translation strategies, might differ, or might deviate to a certain degree. Potentially, different alignments point towards certain translation strategies. In this section, we want to concentrate on translation strategies that focus on the information presentation and structure. We will combine alignment patterns and potential translation strategy and discuss this procedure critically.

Interestingly, the majority of alignments are 1:n-alignment in the subcorpus, which is in line with most of the Easy Language rules presented above. These alignments, further, point to a form of sentence splitting. Sentence splitting is a strategy to restructure and simplify complex sentences in interlingual translation, as well: "[T]ranslating a source sentence by a sequence of independent target sentences aims at reducing informational density of the target text as opposed to the source text by increasing incrementality." (Fabricius-Hansen 1999: 188).

On the other hand, only few instances were found, in which deletion and reduction processes become obvious, although an information selection is supported by the rules to keep the texts short enough for the intended recipients. Another aspect, which can hardly be tested with the sentence alignment is the structure on the macro level. It was observed, however not quantified, that the texts in the analysed subcorpora did not show any evidence concerning a restructuring on a macro-level, meaning that the information of the source text was presented in a similar order in the target text. The reason might be that the source texts we aligned so far were rather short (mean: 355.5 words/text) and accordingly the target texts were still short enough (mean: 448.6 words/text) to be appropriate for the target group.

Further, we want to stress that sentence alignment does not shed light on the information structure within the sentences. In (2), you can see an example in which the source text is reduced. However, the alignment remains a 1:1-alignment.

(2) **ST**: Themen wie Bildung, lebenslanges Lernen und berufliche Anpassung werden in Zeiten einer globalisierten und digitalisierten Arbeitswelt immer wichtiger. (Eng.: Topics like education, life-long learning and career adaptation are becoming more important in times of a globalised and digital work environment.)
EL: Eine gute Schul-Bildung ist immer wichtiger. (English: Good school-education becomes more important.)

"Lebenslanges Lernen", "berufliche Anpassung", and "in Zeiten einer globalisierten und digitalisierten Arbeitswelt" are not represented in the target text sentence illustrating an information reduction strategy on the lexical level (cf. Hansen-Schirra et al. 2020a for a discussion of translation strategies). Further the term "Bildung" was constrained to the education that is imparted in school. These changes are not presented in the sentence alignment, which evokes the impressions that the same information are presented. Hence, it seems plausible to include word or character counts in future research in addition to sentence alignment to represent these changes.

In example (3), the mere alignment figures suggest that information was added since we have a 1:3 alignment. However, when we look closer at the contents, the information was reduced concerning the actors involved, while the information on the "discussion process" itself was exemplified with questions. This exemplification strategy is typical of Easy Language translation (cf. Hansen-Schirra et al. 2020a).

(3) **ST**: Hierzu wurde am 23. Januar 2009 ein Diskussionsprozess mit der Beauftragten für die Belange behinderter Menschen, den Verbänden behinderter Menschen, der Leistungserbringer und den kommunalen Spitzenverbänden angestoßen. (English: To this topic, a discussion process was initiated on the 23rd of January in 2009 including the representative for people with disabilities, associations of people with disabilities, providers, and the community's umbrella organisations.)
EL: Alle haben überlegt: Sind die Vorschläge gut? Was muss anders sein? (English: Everybody reflected on: Are the proposals good? What has to be different?)

Finally, the example in (4) represents another difficulty when it comes to sentence alignment.

(4) ST: Außerdem müssen Prüfungsordnungen die besonderen Belange von Studierenden mit Behinderung zur Wahrung der Chancengleichheit berücksichtigen. (English: Furthermore, exam regulations have to respect the special interests of students with disabilities to keep the opportunities equal.)
EL: Studenten und Studentinnen mit Behinderung müssen bei den Prüfungen die gleichen Chancen haben. Zum Beispiel muss ein blinder Student die Prüfungs-Aufgabe in Blinden-Schrift bekommen. Das muss auch in den Prüfungs-Ordnungen stehen. (English: Male and female students with disabilities have to have the same opportunities in exams. For example, a blind student has to get the exam task in braille. This has to be mentioned in the exam regulations.)

The source sentences do not give any reasons to include the example that is inserted in the target text. The translator probably decided to include the example to make the content more concrete. When analysing the sentence alignment, however, the question comes up whether to describe the alignment as a 1:3-alignment or as a 1:2-alignment plus a 0:1-alignment. In our analysis, we decided that examples that were added and for which there was no indication in the source text were counted as the latter. However, both counting possibilities seem reasonable.

These are only some examples why a mere sentence alignment might not be enough to analyse translation strategies in Easy Language translation. Information reorganisation is not only processed by sentences splitting, additions or deletion of whole sentences, or similar strategies. Often, information is reorganised within the sentences or split sentences are further reduced or enriched. In the next section, we want to discuss another approach.

6 Tree Alignment

As we have discussed in Sect. 4, a sentence-based analysis only gives first impressions of the information restructuring in intralingual translation. According to the rules for German Easy Language (see above), complex sentences with finite and non-finite subordinate clauses have to be split into several main clauses in the EL translation. 1:n or n:m alignments show these translation patterns but they do not reveal where which parts of the clause complex are exactly moved to, whether the information is restructured or preserved in the same order, which subordinate clauses are not translated and which information is added. Hence, we suggest a tree-based alignment in future research, i.e., building up a parallel treebank. A treebank is a corpus which is annotated for syntactic information (i.e., syntactic trees). This means that it includes information on parts-of-speech, morphology, phrase structure, syntactic functions, main and subordinate clauses and their dependency structure. Recently, the need has emerged to build up interlingual parallel treebanks: In computational linguistics, they are employed for multilingual grammar induction, as test suites and gold standards for alignment tools and multilingual taggers and parsers (Volk et al. 2011). Additionally, they are used for the development of corpus-based machine translation systems (cf. Čmejrek et al. 2004). In translation studies, interlingual parallel treebanks are needed as linguistically enriched text basis for empirical research on translations strategies and specific properties of translations (cf. Hansen-Schirra et al. 2012). Following this argumentation, we argue that we need a monolingual parallel treebank for the investigation of translation strategies from Standard into Easy Language. Such a monolingual parallel treebank sheds light on phrase and clause structures, syntactic functions and dependencies, and the alignments thereof. The annotation and alignment of the Geasy Corpus in terms of treebank structures and alignments is still work in progress. Nevertheless, example 5 taken from our corpus illustrates the advantages of such a monolingual parallel treebank for the investigation of n:m alignments.

(5) **ST**: Auf Grundlage der Globalrichtlinie Stadtteilkultur stellt die Behörde den sieben Hamburger Bezirken Fördermittel zur Verfügung. (English: On the basis of the global guidelines for urban district culture, the authority provides funding to the seven Hamburg districts.)
EL: Hamburg ist eine sehr große Stadt. Es gibt 7 große Bezirke in Hamburg. Es gibt besondere Förderung für die Kultur in den Bezirken. (English: Hamburg is a very big city. There are 7 large districts in Hamburg. There is special funding for culture in the districts.)

In Fig. 1, we can see the dependency trees, clause and phrase annotations of the source sentence and the translation, which consists of three sentences. The ParZu parser was used for dependency annotation (cf. Sennrich et al. 2009). The alignment on the sentence level suggested in Sect. 4 results in a 1:3-alignment since one source sentence is translated into three target sentences (s. Table 4). However,

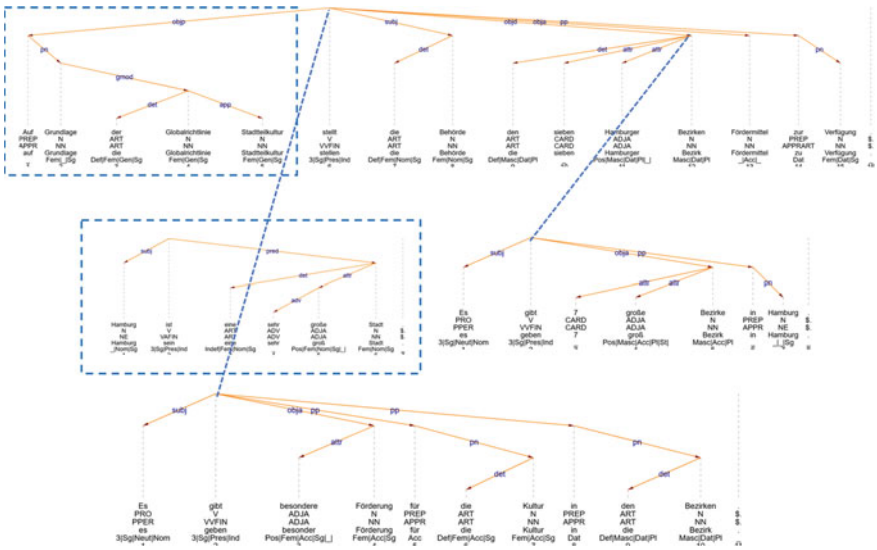

Fig. 1 Parallel treebank of example (5)

Table 4 Linear alignment of clauses and phrases of example (5)

Source text in Standard German		Target text in Easy German
		[A] Hamburg ist eine sehr große Stadt. [clause simplex]
[a] Auf Grundlage der Globalrichtlinie Stadtteilkultur [prepositional phrase]		
[b] stellt die Behörde [predicate, subject] [d] Fördermittel zur Verfügung [direct object]		
	[c] den sieben Hamburger Bezirken [indirect object]	[B] Es gibt 7 große Bezirke in Hamburg. [clause simplex]
		[C] Es gibt besondere Förderung für die Kultur in den Bezirken. [clause simplex]

taking a closer look into this translation reveals that we have several empty links from Standard to Easy Language but also the other way around. Empty links are units in the target text which do not have matches in the source text and vice versa (cf. Hansen-Schirra et al. 2017). They may occur on all language levels. This requires alignments on several levels: sentence level, clause level, phrase level, word level, etc. In the example, the third sentence in the target text can be aligned

with the source sentence (the alignments are indicated by the dotted line in the figure). However, the second sentence in the EL text is equivalent to the indirect object in the source language, which results in a phrase-sentence alignment from source to target text. This sentence further explains the word "Bezirke" (districts) in the aligned source text phrase, but it is also referring to the same word in the third sentence of the EL translation and establishes a co-reference relation between the two EL sentences. On the basis of this annotation and alignment, we will be able to automatically extract phrase-sentence or clause-sentence alignments.

In addition, we can also search for empty alignment links (see the dotted boxes in the figure). The first sentence in the EL translation is an explanation which is added to create further common ground for the target group (cf. Hansen-Schirra et al. 2020b; Pickering and Garrod 2004). It explains that Hamburg is a big city, which is subdivided into districts. This information is not necessary in the source text since it can be taken as general knowledge by the unimpaired reader. This additional explanation results in an empty link when aligning the sentence pairs. Another empty link occurs on phrase level and affects the first prepositional phrase in the source sentence "On the basis of the global guidelines for urban district culture". This prepositional phrase specifies the administrative basis for the funding. This detail is not necessary for the target group of the EL text and therefore left out. Here, information selection is an important translation strategy in order to keep the text short and processable for the EL reader (cf. Hansen-Schirra et al. 2020a). In conclusion, this example shows that we will be able to automatically quantify and extract empty links in the source text resulting from information selection strategies as well as in the target text resulting from explanation or exemplification strategies. An integration of the word alignment into the analyses would shed light on more fine-grained translation strategies. This remains object to future research.

7 Conclusion and Future Research

In this paper, we have shown that the analysis of the sentence alignment of a parallel corpus of intralingual translations gives first indications of intralingual translation strategies. To summarise the results, we found evidences for information splitting and adding of information (mainly explanations), while there were only few evidences of reductions and deletions of information or a restructuring processes on a text level. In future research, these kinds of analyses could benefit from taking word or character counts into consideration to integrate adding, deletion, or restructuring processes within the sentence.

Like all branches of linguistics, also research on Easy Language translation profits from empirical corpus analyses. For a more fine-grained perspective, linguistically enriched corpora are still a desideratum. Especially translation phenomena resulting from translation selection and deselection strategies need—apart from a lexical analysis—a more detailed linguistic enrichment in order to be answered. This linguistic enrichment should cover syntactic functions, phrase

structure and dependency trees but also semantic annotation such as coreference relations (cf. Kunz 2010) or frame semantics (cf. Czulo 2017), which requires of course more comprehensive annotation methods and more sophisticated query facilities.

While state-of-the-art corpus and treebank research in Standard and Easy Language share several problems such as multi-layer annotation and exploitation, they are also different in a number of ways. Monolingual parallel corpora including Easy Language texts raise numerous problems which are specific to this kind of research and which are caused by the specificities of this language variety—especially when it comes to the alignment of source and target texts. This includes the following problems: the lacking comparability of segments, the repetition of concepts and the differences in lexical and idea density. Future research has to address exactly these issues aiming at a comprehensive treebank including Easy Language translations.

Another research gap affects the empirical analysis of text-image integration in Easy Language translations. In our corpus analyses, we have so far not taken into account the role of images, pictures, pictograms, etc. In future research, it would be interesting to investigate in-depth the use of the pictures and images in Easy Language texts. According to Maaß (2015: 143), central concepts of a text can be illustrated by pictures. However, using pictures only to make the texts more appealing should be avoided since there is also the risk of causing distraction and cognitive overload (Bock 2018). Hence, an assessment if the pictures are solely used to make the text more understandable would be desirable. Quantifiable conclusions thereof still remain a research desideratum.

Finally, we will continue to enlarge our corpus and align the data so future analysis can draw on a larger data set. Further, the aligned data, for example, can be used to train machine translation systems for intralingual translation (Hansen-Schirra et al. 2020a).

References

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2017. Unsupervised neural machine translation. arXiv preprint http://arxiv.org/abs/1710.11041.

Battisti, A., S. Ebling, and M. Volk. 2019. An empirical analysis of linguistic, typographic, and structural features in simplified German texts. In *Proceedings of the sixth italian conference on computational linguistics (CLiC-it 2019)*. Bari.

Baker, M. 1995. Corpora in translation studies: An overview and some suggestions for future research. *TargetInternational Journal of Translation Studies* 7 (2): 223–243.

Bock, B.M. 2018. *Leichte Sprache: Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSAProjekt*. Technical report, Universität Leipzig.

Bock, B.M. 2019. *„Leichte Sprache". Kein Regelwerk.* Berlin: Frank & Timme.

Bredel, U., and C. Maaß. 2016a. *Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis*. Berlin: Duden.

Bredel, U., and C. Maaß. 2016b. *Ratgeber Leichte Sprache*. Berlin: Duden.

Čmejrek, M., J. Cuřín, J. Havleka, J. Hajič, V. Kuboň. 2004. Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation. In *4th International conference on language resources and evaluation*. Lisbon, Portugal.

Chesterman, A. 2007. What is a unique item? In *Doubts and directions in translation studies*, ed. Y. Gambier, M. Shlesinger, and R. Stolze, 3–13. Amsterdam: Benjamins.

Czulo, O. 2017. Aspects of a primacy of frame model of translation. In *Empirical modeling of translation and interpreting*, ed. S. Hansen-Schirra, O. Czulo, Oliver and S. Hofmann, 465–490. Berlin: Language Science Press.

Deilen, S. 2020. Visual segmentation of compounds in easy language: Eye movement studies on the effects of visual, morphological and semantic factors on the processing of German Noun-Noun compounds. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 241–256. Berlin: Frank & Timme.

Gutermuth, S. 2020. *Leichte Sprache für alle? Eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*. Berlin: Frank & Timme.

Hansen-Schirra, S., S. Neumann, and E. Steiner. 2012. *Cross-linguistic corpora for the study of translations: Insights from the language pair English-German*. Berlin, New York: de Gruyter.

Hansen-Schirra, S., S. Neumann, and E. Steiner. 2013. *Cross-linguistic corpora for the study of translations*.Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110260328

Hansen-Schirra, S., S. Neumann, O. Čulo, and K. Maksymski. 2017. Empty links and crossing lines: Querying multi-layer annotation and alignment in parallel corpora. In *Annotation, exploitation and evaluation of parallel corpora: TC3 I*, ed. S. Hansen-Schirra, S. Neumann, and O. Čulo, 53–87. Berlin: Language Science Press.

Hansen-Schirra, S., and S. Gutermuth. 2018. Modellierung und Messung Einfacher und Leichter Sprache. In *Barrieren abbauen, Sprache gestalten*, ed. S. Jekat, M. Kappus and K. Schubert, 7–23. Winterthur: ZHAW.

Hansen-Schirra, S., J. Nitzke, S. Gutermuth, C. Maaß, and I. Rink. 2020a. Technologies for the translation of specialised texts into easy language. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 99–127. Berlin: Frank & Timme.

Hansen-Schirra, S., W. Bisang, A. Nagels, S. Gutermuth, J. Fuchs, L. Borghardt, S. Deilen, A.K. Gros, L. Schiffl, and J. Sommer. 2020b. Intralingual translation into Easy Language—or How to reduce cognitive processing costs. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 197–225. Berlin: Frank & Timme.

Inclusion Europe. 2009. *Informationen für alle. Europäische Regeln, wie man Informationen leicht lesbar und verständlich macht.* http://easy-to-read.eu/wp-content/uploads/2014/-12/DE_Information_for_all.pdf

Kenny, D. 1997. (Ab)normal Translations: a German-English Parallel Corpus for Investigating Normalization in Translation. In *Practical applications in language corpora. PALC '97 proceedings*, ed B. Lewandowska-Tomaszczyk, and P.J. Melia, 387–392.

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel. 2014. The sketch engine: Ten years on. *Lexicography* 1: 7–36.

Klaper, D., S. Ebling, and M. Volk. 2013. Building a German/Simple German parallel corpus for automatic text simplification. In *Proceedings of the ACL workshop on predicting and improving text readability for target reader populations*, 11–19. Madison/USA: Omnipress.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. *MT Summit* 5: 79–86.

Kunz, K. 2010. *English and German nominal co-reference: A study of political essays*. Frankfurt/Main: Peter Lang.

Lehrndorfer, A. 1996. *Kontrolliertes Deutsch: Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der technischen Dokumentation.* Tübingen: Narr.

Lindholm, C., and U. Vanhatalo (eds.) Forthcoming. *Easy language in Europe*. Berlin: Frank & Timme.

Maaß, C. 2015. *Leichte Sprache. Das Regelbuch*. Münster: Lit.

Maaß, C. 2020. *Easy language—plain language–easy language plus*. Berlin: Frank & Timme.

Maaß, C., and S. Hernández Garrido. 2020. Easy and plain language in audiovisual translation. In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra, and C. Maaß, 131–161. Berlin: Frank & Timme.

Maaß, C., I. Rink, and S. Hansen-Schirra. Forthcoming. Easy language in Germany. In *Easy language in Europe*, ed. C. Lindholm and U. Vanhatalo. Berlin: Frank & Timme.

Netzwerk Leichte Sprache. 2013. *Leichte Sprache. Ein Ratgeber, Bundesministerium für Arbeit und Soziales.* Paderborn: Bonifatius GmbH.

Pickering, M.J., and S.C. Garrod. 2004. Toward a mechanist psychology of dialogue. *Behavioral and Brain Sciences* 27: 169–226.

Sahlgren, M., and J. Karlgren. 2005. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering* 11 (3): 327–341.

Schiffl, L. 2020. Hierarchies in lexical complexity: Do effects of word frequency, word length and repetition exist for the visual word processing of people with cognitive impairments? In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 227–239. Berlin: Frank & Timme.

Sennrich, R., G. Schneider, M. Volk, and M. Warin. 2009. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology* 115: 124.

Sommer, J. 2020. A study of negation in German easy language—does typographic marking of negation words cause differences in processing negation? In *Easy language research: Text and user perspectives*, ed. S. Hansen-Schirra and C. Maaß, 257–272. Berlin: Frank & Timme.

Volk, M., T. Marek, and Y. Samuelsson. 2011. Building and querying parallel treebanks. *Translation: Computation, Corpora, Cognition (Special Issue on Parallel Corpora: Annotation, Exploitation and Evaluation)* 1 (1): 7–28.

Zanettin, F. 2000. Parallel corpora in translation studies: Issues in corpus design and analysis. *Intercultural Faultlines*, 105–118.

Silvia Hansen-Schirra Professor for English Linguistics and Translation Studies, Johannes Gutenberg University Mainz, Faculty of Translation Studies in Germersheim, Germany. Director of the Translation & Cognition (TRA&CO) Center, Head of the Research Group "Simply complex —Easy Language".

Jean Nitzke Associate Professor for Translation with a focus on Translation Technology at the University of Agder, Norway. Former Lecturer and Post-Doc for English Linguistics and Translation Studies, Johannes Gutenberg University Mainz, Faculty of Translation Studies in Germersheim, Germany.

Silke Gutermuth Lecturer for English Linguistics and Translation Studies, Johannes Gutenberg University Mainz, Faculty of Translation Studies in Germersheim, Germany. Research Assistant at the Translation & Cognition (TRA&CO) Center, Manager of the on-site eyetracking lab.