DE-Lite – a New Corpus of Easy German: Compilation, Exploration, Analysis

Sarah Jablotschkin
Universität Hamburg
sarah.jablotschkin@uni-hamburg.de

Elke Teich
Universität des Saarlandes
e.teich@mx.uni-saarland.de

Heike Zinsmeister
Universität Hamburg
heike.zinsmeister@uni-hamburg.de

Abstract

In this paper, we report on a new corpus of simplified German. It is recently requested from public agencies in Germany to provide information in easy language on their outlets (e.g. websites) so as to facilitate participation in society for people with low-literacy levels related to learning difficulties or low language proficiency (e.g. L2 speakers). While various rule sets and guidelines for Easy German (a specific variant of simplified German) have emerged over time, it is unclear (a) to what extent authors and other content creators, including generative AI tools consistently apply them, and (b) how adequate texts in authentic Easy German really are for the intended audiences. As a first step in gaining insights into these issues and to further LT development for simplified German, we compiled DE-Lite, a corpus of easy-to-read texts including Easy German and comparable Standard German texts, by integrating existing collections and gathering new data from the web. We built n-gram models for an Easy German subcorpus of DE-Lite and comparable Standard German texts in order to identify typical features of Easy German. To this end, we use relative entropy (Kullback-Leibler Divergence), a standard technique for evaluating language models, which we apply here for corpus comparison. Our analysis reveals that some rules of Easy German are fairly dominant (e.g. punctuation) and that text genre has a strong effect on the distinctivity of the two language variants.

1 Introduction

The UN Convention on the Rights of Persons with Disabilities (UN-CRPD)[1] states that obstacles to accessibility to "information, communication and other services" should be eliminated by state parties for people with disabilities (article 9). Against this background, many countries have pushed for legislation to reduce the language barrier for people with learning difficulties[2] as one of the core measures in creating equal opportunities. In Germany, different forms of simplified German have emerged including variants of a regulated, "easy" German ('Leichte Sprache') that are intended to make written information accessible for low-literacy readers (Inclusion Europe, n.d.; Netzwerk Leichte Sprache, 2022; Bredel and Maaß, 2016; Bock, 2018; Bundesministerium der Justiz und für Verbraucherschutz, 2017). According to a recent policy of the German Ministry for Work and Social Affairs[3], it is now requested from public institutions to provide information in (regulated) Easy German alongside Standard German. While people with disabilities are the only group whose right to accessible written information is statutory, it is often claimed that non-disabled people such as learners of German or older people, or even all people (Netzwerk Leichte Sprache, 2022), profit from Easy German.

While a long-awaited move in language policy, there are a number of open questions both for the theory and the practice of Easy German. There are several agencies providing guidelines about how to write in Easy German and while there is a fair level of convergence, there is also some conflicting advice. Also, it is unclear whether specific features such as avoiding pronouns or using only simple, paratactic conjunctions (see Section 2.1) are indeed beneficial for comprehension and if so, for which specific target groups. Overall, there is fairly little empirically grounded research about the use of Easy German in particular. This is the motivation of the project we report on in this paper.

[1] https://social.desa.un.org/issues/disability/crpd/

[2] We use this term for people with intellectual and other disabilities because it is considered less stigmatising by self-advocacy groups such as Network People First Germany, see https://www.menschzuerst.de/pages/startseite/wer-sind-wir/verein.php

[3] *Bundesteilhabegesetz und Nationaler Aktionsplan 2.0*: https://www.bmas.de/DE/Leichte-Sprache/leichte-sprache.html

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 106–117
March 21, 2024 ©2024 Association for Computational Linguistics

Our focus is on the exploratory research question: What are the typical features of Easy German in lived practice? This involves empirical studies of authentic productions in Easy German and other variants of simplified German. For this purpose, we have compiled the DE-Lite corpus from preexisting resources of different variants of simplified German and Standard German, and extended it with additional texts from the web.

This paper documents decisions made in the corpus compilation process, including how to address the challenge of duplicate identification. In addition, we present an exploratory, n-gram-based study in which subcorpora of DE-Lite consisting of comparable texts in Easy German and Standard German are compared revealing main characteristics of Easy German. We think that the corpus, its description, and the empirical study are of interest for the development of inclusive language technology, and that insights of the German corpus and its compilation can be transferred to other languages.

Our overarching theoretical approach is rooted in information theory (Shannon, 1948), a mathematical theory of communication, according to which language users modulate the information content of their messages (Crocker et al., 2015), adapting their linguistic encodings to properties of both the channel (e.g. noise) and the recipient (audience design) (see e.g. Vogels et al., 2019; Häuser and Kray, 2021).

The link to Easy German is a natural one: Rules and recommendations for Easy German can be considered intentional measures to reduce the information content (surprisal) of linguistic expressions/units, such as words, sentences or stretches of text. Surprisal being correlated with processing effort, modulation of information content is a measure to adapt to a supposedly lower channel capacity of the target group(s) of Easy German. We thus hypothesise that the information content of linguistic units should be smaller in Easy German compared to standard language, indicated e.g. by a preference for high-frequency words, lower lexical density, lower vocabulary variation and syntactic and cohesive explicitness. To identify the specific properties of Easy German, we compare it with Standard German, employing selected information-theoretic measures, such as relative entropy, a measure widely used in NLP for evaluating language models.

The paper is structured as follows. In related work (Section 2), we sketch the history of Easy German, followed by a brief state-of-the-art on corpus-based work on Easy Language. In Section 3 we introduce the DE-Lite corpus containing texts in simplified variants of German by describing corpus design, the challenge of harmonising existing resources, and the mathematical basis of our language modeling. Section 4 complements the corpus description by presenting an exploratory, comparative analysis of two DE-Lite subcorpora of Easy German and Standard German. We conclude with a summary and discussion (Section 5).

2 Related work

The next section outlines the development of Easy German as a highly restricted variant of German.

2.1 Easy German

Easy German ('Leichte Sprache') only emerged in the late 1990s, while similar concepts have been practised in countries such as Finland, Sweden, and the USA since the 1970s (Netzwerk People First Deutschland e.V.; Tjarks-Sobhani, 2012, 28; Gross, 2015, 81). Today, simplified variants of national languages exist in numerous countries around the globe.[4] The concept originated from the empowerment of people with learning difficulties advocating their right to participation in society. In Germany, they developed relatively rigid rules for creating easily comprehensible text together with their supporters (Inclusion Europe; Netzwerk Leichte Sprache, 2014, 2022). The rule sets also emphasise the importance of letting representatives of the target groups check texts written in Easy German for comprehensibility and partly make this procedure a prerequisite for awarding an official quality seal for Easy German. While Easy German is a concept that has been developed by laypeople and has been in use for a long time, even before it was legally recognised, linguistic research in this area has only increased over the past few years.

Even though there are differing rule sets and guidelines for creating text in Easy German, they overlap with regard to general linguistic principles: All rule sets emphasise the importance of syntactic simplicity, for example by using short sentences, only making one statement per sentence (Inclusion Europe, 16-17; Netzwerk Leichte Sprache, 2022, 30), or using a fixed constituent order with sentence-initial subject (Netzwerk Leichte Sprache,

[4] https://www.easy-plain-accessible.com/home/around-the-world/

Was ist Leichte Sprache?	*What is Easy Language?*
Leichte Sprache ist eine besondere Form der deutschen Sprache.	*Easy language is a special form of the German language.*
Leichte Sprache ist leicht zu lesen und zu verstehen.	*Easy language is easy to read and understand.*
Texte in leichter Sprache haben zum Beispiel:	*Texts in easy language have for example:*
• einfache Wörter	• *simple words*
• kurze Sätze	• *short sentences*
• Bilder	• *pictures*
Deshalb verstehen viele Menschen Texte in leichter Sprache besser.	*That is why many people understand texts in easy language better.*
Dadurch wissen sie mehr.	*So they know more.*
Und sie können mitreden.	*And they can have their say.*
Sie können selbst Entscheidungen treffen.	*They can make decisions for themselves.*

Table 1: Definition of Easy German in Easy German with specific typography (Netzwerk Leichte Sprache, 2021, 209)

2022, 31; Bredel and Maaß, 2016, 419-425) which is not required in Standard German.

On the lexical level, it is commonly recommended to use only frequently used words and avoid technical terms as well as borrowed words (Netzwerk Leichte Sprache, 2022, 13). With regard to morphology, verbs are preferred over nouns, passive voice should be avoided and prepositional paraphrases are considered easier than genitive case (Netzwerk Leichte Sprache, 2022, 16-17). There are also some recommendations on the textual level: Difficult words, if they cannot be avoided, should always be explained (Inclusion Europe, 15), and instead of using pronouns or lexical substitution, the “same words for the same things” (Netzwerk Leichte Sprache, 2022, 14) should be used.

The example shown in Table 1 illustrates some of these characteristics: On a syntactic level, it consists of paratactic structures with the subject or an adverbial connective (*deshalb* ‘therefore’, *dadurch* ‘thereby’) being the first sentence constituent. The coreferring expression *Leichte Sprache* is repeated several times instead of being replaced by a pronoun as would be the coherent way to put it in Standard German. At the same time, the text shows some inconsistencies with respect to the rules mentioned above: The nominal phrase *Menschen* (‘people’) is not repeated, but is referred to anaphorically by the personal pronoun *sie* (‘they’), and the first sentence employs the genitive attribute *der deutschen Sprache* (‘of the German language’) instead of a prepositional paraphrase as is recommended for example by Netzwerk Leichte Sprache (2022).

2.2 Corpus resources and corpus-based studies

Multilingual corpora and corpora that include different intra-lingual variants such as the DE-Lite corpus can be classified according to the relation that texts of the different variants have to each other: In a ‘parallel corpus’ there is a translation relation between individual texts of the different languages or variants (which can be made explicit by aligning on sentence, paragraph, or text level); in a ‘comparable corpus’ texts are sampled for the same genres or text types across variants.[5] If there is neither a translation relation nor a thematic relation, the corpus just contains samples of monolingual subcorpora of different languages or variants.

There are a number of corpora for simplified German which we summarised in Table 3 in Appendix A. While the Geasy corpus (Hansen-Schirra et al., 2021) contains Easy German texts, several other corpora contain different variants of simplified text: LeiKo (Jablotschkin and Zinsmeister, 2023), DEplain (Stodden et al., 2023) and the Simple German Corpus (Toborek et al., 2023) contain Plain German as well as Easy German, APA-RST (Hewett, 2023) is a corpus of Austrian texts that are categorised into different complexity levels (A2, B1 according to Council of Europe, 2001), and both the LeiSa corpus (Lange and Bock, 2016) and WebCorpus (Battisti et al., 2020) sample simplified text without restricting it to a specific simplification

[5] This terminology is broader than the use of *comparable corpus* in translation studies where the term is used for sets of texts originally written in a language L and thematically comparable texts that are translated into L.

method or label. Most of the corpora are (partly) parallel and contain Standard German texts as well. Others are comparable corpora for different variants of simplified German. The corpora also differ with regard to whether they contain sentence alignments and linguistic annotations.

While parallel corpora in the setting of simplified language are especially suited for training automatic simplification algorithms or analysing intralingual translation strategies, comparable corpora allow for the acquisition of larger amounts of data and the detection of linguistic differences between or within language variants, e.g. based on metadata such as text genre or publisher.

Various corpuslinguistic studies investigate specific linguistic characteristics of Easy German, often in order to evaluate the applicability and application of individual rules (e.g. Lange, 2019; Fuchs, 2019). There are also psycholinguistic studies that evaluate characteristics of Easy German with regard to whether they improve text comprehensibility for the recipients (e.g. Lasch, 2017; Bock, 2017a). There are few studies that (like our own) use corpus data to explore characteristics or complexity levels of Easy German inductively (e.g. Bock, 2014). Unlike previous studies, our approach is not restricted to specific linguistic levels such as syntax or morphology. By calculating KLD on every token of the corpus and isolating distinctive types (see Section 4), we take this as a starting point to draw conclusions about the expression of complexity reduction on different linguistic levels such as syntax, morphology or pragmatics of (text genres in) Easy German.

3 Corpus

In order to re-use previously collected data as well as annotations and alignments, we merged parts of different existing corpus resources containing texts in variants of simplified German: DEplain (Stodden et al., 2023), Geasy (Hansen-Schirra et al., 2021), WebCorpus (Battisti et al., 2020) and LeiKo (Jablotschkin and Zinsmeister, 2023). The corpus is still under construction and further existing Easy German corpora will be included, such as APA-RST (Hewett, 2023) and the Simple German Corpus (Toborek et al., 2023). To further expand the corpus, we also collected html text as well as PDFs from additional websites, our main sampling criterion being date of publication: In order to ensure comparability and avoid date of publication as confounding variable, we excluded texts that had been published before 2017. This is motivated by the assumption that Easy German has undergone substantial changes with regard to its linguistic characteristics. One trigger for this has been the publication of linguistically founded rules and recommendations for Easy German texts by *Forschungsstelle Leichte Sprache Hildesheim* (Research Unit Easy German Hildesheim) (Maaß, 2015; Bredel and Maaß, 2016). In addition, there have been research projects that improved the general understanding of what exactly is comprehensible for the target groups of Easy German, such as LeiSa (Bock, 2018).

The collected data comprises different file formats and requires different methods of preprocessing. As for PDFs, we used the Python library `PyMuPDF` to extract text and conducted additional manual cleaning. For webscraping, we used the Python `requests` library, and `BeautifulSoup` in order to parse the downloaded html files. We used the tcf version (Heid et al., 2010) of the WebCorpus data (Battisti et al., 2020) containing primary text as well as annotations and metadata, which we also parsed with `BeautifulSoup`.

3.1 Duplicate identification

An important issue when combining different web-based corpora is near-duplicate cleanup, see Rodier and Carter (2020) for a recent overview. For example, Geasy (Hansen-Schirra et al., 2021), WebCorpus (Battisti et al., 2020) as well as DEplain (Stodden et al., 2023) all made use of the website `einfach-teilhaben.de` by Germany's federal ministry for labour and social affairs (BMAS), which provides official information about topics such as disability, inclusion and social participation. To detect and exclude duplicates, we computed substring edit distances between corpus texts by BatchSED (Adelmann, 2021)[6] following the approach of (Adelmann and Gius, 2020). This approach takes into account the possibility that one text may be fully or partially contained within another text (in our case, for example, due to different web scraping routines). Hence, BatchSED calculates two scores for each pair of texts, by taking text 1 as a substring of text 2 and vice versa. Two texts are considered duplicates if the substring edit distances for both directions, divided by the length of the text to be embedded as substring, is less than

[6] `https://github.com/benadelm/BatchSED`: It calculates word-based distances with insertion costs equal to deletion costs equal to substitution costs equal to one.

Category	Values
Label	Leichte Sprache, Einfache Sprache, children, other
Rule set / agency	Forschungsstelle Leichte Sprache Hildesheim (FLS), capito, Netzwerk Leichte Sprache, Inclusion Europe, other
Complexity level	A1, A2, B1, none
Original corpus	Geasy, WebCorpus, DEplain, LeiKo, DE-Lite
Publisher	[name of publisher], e.g. public broadcasters, governmental institutions, welfare institutions, research institutions, non-profit organisations/NGOs, publishing houses, political parties, private individuals
Verification process	Target group, none
Year of publication	2017 or more recent
Text genre	lexicon, news/newspaper, wiki, blog, election programme, story/novel, technical text, administrative text and others
Origin of text	user-generated, editorial

Table 2: Core metadata of the DE-Lite corpus: Categories and values

15 %. From a pair of texts identified as duplicates, we kept that instance that was aligned to a parallel text in the corpus. If this filter was not applicable, we followed a fixed preference hierarchy, partly motivated by the availability of metadata, to make the provenance of the corpus texts transparent: LeiKo before WebCorpus before DEplain before Geasy before newly crawled material. This method identified about 400 Easy German texts and about 500 Standard German texts as duplicates which we excluded from the merged corpus. The actual number of duplicates was in fact much higher but many instances were filtered manually in advance, during the process of integrating the resources before further processing.

3.2 Metadata annotation

For our corpus, we collect the metadata displayed in Table 2. Our main sampling criterion is year of publication (cf. beginning of Section 3). In addition, we cover a broad range of text genres in order to approximate representativity. Since the underlying rule set or agency might also have an effect on linguistic characteristics, we include texts written according to the non-linguistic rule sets (Inclusion Europe; Netzwerk Leichte Sprache, 2022) as well as texts written according to the rule sets by *Forschungsstelle Leichte Sprache Hildesheim* (Research unit Easy German Hildesheim; FLS) (Bredel and Maaß, 2016). However, for most of the texts it is not clear whether they were written according to a specific rule set.

These data are partly adopted from the existing corpora, which we merged into our corpus. For newly collected texts, we collect the data from the websites or PDFs. For the texts from existing corpus resources, we complete the metadata according to our annotation scheme wherever possible. Since the original websites cannot always be reconstructed, certain metadata cannot be retrieved any more.

As previously mentioned, there are various seals for marking simplified German text. Sometimes, texts labeled as Easy German further contain an indication of their complexity level. This information is contained in the metadata variables *label* and *complexity level*.

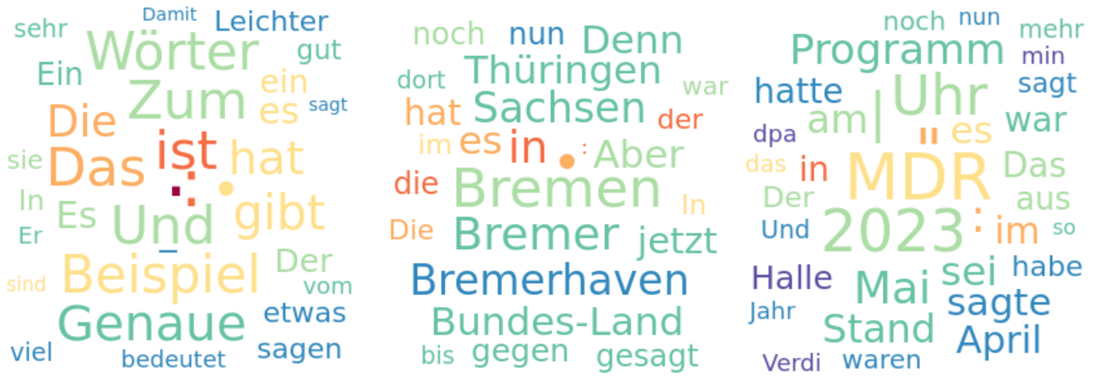
Since some of the rule sets require members of the target groups to verify Easy German texts before they can be labeled as such, *verification process* was also included as metadata variable.

3.3 Language modeling

An effective approach to get a first idea of the differences between language variants is to compute word-based n-gram models (including punctuation) for each variant and compare the models with a divergence measure, such as Jensen-Shannon or Kullback-Leibler Divergence. Here, we use the asymmetric variant, Kullback-Leibler Divergence (KLD). Formally, KLD computes the difference between two probability distributions in terms of the number of additional bits needed to encode a unit x from a distribution A with an optimal encoding for distribution B (see eq. 1). The higher the number of additional bits, the greater the difference.

$$D_{KL}(A\|B) = \sum_{x\in X} A(x) \log\left(\frac{A(x)}{B(x)}\right) \quad (1)$$

While a standard method for evaluating language models, KLD has the advantage of giving us not only an indication of the overall difference between two language variants, but also of the most distinctive linguistic features. The specific features (here: words, punctuation marks) involved in the difference are obtained by ranking the features in terms of pointwise KLD. For inspection we use



Easy vs. Standard

News vs. non-news (Easy)

News vs. non-news (Standard)

Figure 1: Term clouds displaying distinctive terms in the respective subcorpora of DE-Lite v1. Size: Distinctivity by KLD, Colour: Relative frequency

a word cloud visualization (see Figure 1) that encodes the relative *frequency* (colour) and the *distinctivity* (size) of features. For assessing the statistical significance of an observed difference in overall frequencies, a p-value is calculated with an unpaired Welch t-test on the observed probabilities in the individual documents of each corpus. By default, the p-value is set to 0.05 (95 % confidence) (cf. Fankhauser et al., 2014). Note that this method is equivalent to a (relative) frequency-based account combined with a statistical test on a feature distribution but has the advantage that features are not a priori selected but automatically detected and ranked in terms of their contribution to the distinction between language variants.

3.4 DE-Lite v1: Data basis of this study

DE-Lite contains two subcorpora of Easy German texts, a parallel one and a monolingual one. In addition to Easy German texts, the corpus also contains comparable texts in other simplified German variants, such as Plain German and texts addressing children.

For the explorative corpus comparison described in Section 4, we use the subset DE-Lite v1[7] containing 1,195,176 Easy German tokens (from both the parallel and the monolingual subcorpora) and 1,154,226 Standard German tokens. The other variants of simplified German (e.g. texts for children) are not relevant for this study.

4 Corpus comparison: Easy vs. Standard German

For an explorative corpus study, we use DE-Lite v1 (see Section 3.4). We investigate the corpus data with the help of n-gram based KLD computations along two dimensions: Language variant with the two categories Easy and Standard, and text genre with the categories news and non-news. To this end, we compare what specific types contribute significantly to the overall KLD of the respective dimension category. Figure 1 shows a visualisation of the distinctivity (size) and relative frequency (colour) of individual types. In order to illustrate typical uses and functions of the distinctive terms in the respective subcorpora, we additionally draw on concordances and example sentences.[8]

In a first step, we compare the Easy German subcorpus to the Standard German subcorpus without drawing on any additional metadata (see Section 4.1). On the one hand, our data reveal that in Easy German, particular care is given to establishing coherence. On the other hand, we find characteristics that illustrate the ways morphological and syntactic simplicity is ensured in Easy German.

Subsequently, we show that our approach can be used to detect text-genre specific features within Easy and Standard German by comparing Easy German news to Easy German non-news and Standard news to Standard German non-news (see Section 4.2). Our results show that the characteristics that distinguish news from non-news in Easy German only partly overlap with those that distinguish news from non-news in Standard German.

[7] DE-Lite v1 is downloadable from `https://github.com/HeikeZinsmeister/DE-Lite`.

[8] We used the corpus tool AntConc to systematically sift through the contexts of distinctive types (Anthony, 2023).

4.1 Easy vs. Standard

In order to establish local coherence, texts in Easy German typically contain explanations of difficult words and examples to make abstract concepts more concrete. This general observation can be reproduced by our approach: Some of the terms that significantly contribute to the overall KLD between Easy and Standard German data are used for exemplification and explanation or rephrasing: *Zum* ('for'; sentence-initial) and *Beispiel* ('example') are very prominent and typically occur together, as can be shown by a further analysis of concordances (see also examples (1) and (2)). Another very prominent term is sentence-initial *Das* ('that'), which in our data is frequently followed by verbs such as *ist* ('is'), *heißt* ('means') or *bedeutet* ('means'). However, while explanation and rephrasing are important to ensure comprehension, resolving anaphora such as the pronoun *Das* ('that'), which often refers to a preceding clause as its non-nominal antecedent, may also be challenging (Kolhatkar et al., 2018) and therefore should be evaluated with members of the target groups of Easy German. In our Easy German data, *Das* also frequently occurs as a determiner in the phrase *Das Wort* ('the word'). A closer examination of the instances reveals that they all originate from one and the same website, namely Hurraki, a wiki-like site in Easy German. The same is true for phrases like *Gleiche Wörter* ('same words') and *Genaue Erklärung* ('precise explanation'). Entries in Hurraki follow a fixed structure and often contain additional information about the use and meaning of words. While this is another strategy to establish coherence, these specific phrases are not representative of Easy German. Systematically collecting metadata of Easy German corpus texts is thus essential in order to detect biases like this. This observation is relevant because Easy German sites tend to be more structured than standard language sites also by using formulaic sequences (see also the recommendation to use the same words for the same things, Section 2.1).

Not only words, but also punctuation marks significantly contribute to KLD: Colons, full stops and bullet points are distinctive for Easy in comparison to Standard German. The bullet point is frequently preceded by a colon and introduces a list of examples intended to make a concept more graspable (cf. example (2)). As has been shown by Jablotschkin and Zinsmeister (2021), another function of the colon in Easy German is to indicate a syntactic dependency relation between a matrix and a subordinate object clause (cf. example (3)), a function, which is more commonly accomplished by a comma in Standard German. The distinctivity of the full stop in Easy German is not surprising as Easy German uses shorter and therefore more sentences per number of tokens than Standard German (syntactic simplification).

The higher sentence density might also be one of the reasons why some finite verb forms are very prominent in our Easy German data, such as *ist* ('is') and *hat* ('has'). Furthermore, both verbs are not only used as main verbs but also function as auxiliaries in German, so their distinctivity in relation to Standard German also points out the prevalence of periphrastic verb forms in Easy German which are morphologically more simple than alternative synthetic verb forms. In addition, a closer look into concordances shows that *ist* is frequently followed by a nominal phrase with a definite or indefinite article, which illustrates the importance of predications in Easy German, another means to explain words or concepts.

(1) Früher hat sich der Pflege-Dienst um alles gekümmert.

Zum Beispiel hat der Pflege-Dienst die Assistenten ausgesucht und bezahlt. (p_765_easy)

Before, the nursing service took care of everything.

For example, the nursing service chose and paid the assistants.

(2) Ein Behinderten-Verband ist auch ein Sozial-Verband.

Sozial-Verbände vertreten noch mehr Interessen.

Zum Beispiel von:

- Arbeitslosen,
- Rentnern und
- Menschen, die wenig Geld haben. (m_5314_easy)

A disabled people's organisation is a social association, too.

Social associations represent even more interests.

For example of:

- *unemployed persons,*
- *retired persons and*
- *people who have little money.*

(3) Sie denkt:

Viel mehr Menschen sollen das Persönliche Geld benutzen. (p_765_easy)

She thinks:

A lot more people should use the Personal Money.

(4) Und sie hat gesagt:

Ab dem nächsten Schuljahr bekommen die Lehrer mehr Geld. (p_1162_easy)

And she said:

From next school year, the teachers will get more money.

(5) Der Korea-Konflikt geht schon sehr lange.
Er hat im Jahr 1945 angefangen. (m_1193_easy)
The Korean conflict has been lasting for a very long time already.
It started in the year 1945.

(6) Sie arbeiten in Voll-Zeit.
Oder sie arbeiten in Teil-Zeit.
Oder sie machen eine Ausbildung für einen Beruf. (m_3042_easy)
They work full time.
Or they work part time.
Or they train for a profession.

Moreover, personal pronouns such as *Er* ('He'), *sie* ('she'/'they') and *[E|e]s* ('[I|i]t') show a high pointwise KLD value in Easy German. This is a logical consequence of splitting up complex sentences into simple ones, each requiring an individual subject which is often realised by a personal pronoun. In Standard German, parataxis typically contains elliptic structures such as subject ellipsis. In Easy German, instead of dropping the subject, there is a tendency to syndetically or asyndetically conjoin syntactically complete sentences (see (5) and (6)). While examples (5) and (6) illustrate syntactic simplification in Easy German, they may create problems for reference resolution. Firstly, German personal pronouns like *Er* allow reference to animate/human as well as inanimate/non-human referents (such as *Korean conflict*) and secondly, there might be contexts in which there are more than one potential antecedents with the required grammatical features (in this case: singular masculine noun phrases), so personal pronouns bear potential for semantic as well as pragmatic ambiguity. It is still an open question whether avoiding ellipses simplify texts for recipients of Easy German and in what contexts avoiding personal pronouns might be beneficial for comprehension.

4.2 News vs. non-news (Easy vs. Standard)

An open research question up-to-date is how text genres differ within Easy German. Despite the restricted linguistic means of Easy German, it is supposed to achieve various communicative functions. Bock (2017b, 191) emphasises the importance of text adequacy in order to ensure comprehensibility and the ability of the recipient to recognise the communicative function of the text, so different text genres within *Leichte Sprache* should be recognisable based on characteristic linguistic forms. Since we computed KLD not only with regard to language variant (Easy vs. Standard) but also with regard to text genre (news vs. non-news), our approach allows us to identify specific linguistic features that are characteristic for news in Easy German compared to other texts in Easy German (in contrast to news in Standard German compared to non-news in Standard German). Our term clouds show that news in Easy German typically employ a lot of place names (i.e., names of cities) and local as well as temporal adverbials (*dort* ('there'), *[I|i]n* ('[I|i]n'), *jetzt* ('now'), *nun* ('now'), *bis* ('until')) serving as frame-setters. In a corpus study, Fuchs (2017) found out that in short Easy German news texts the text-initial position is frequently used for local frame-setters to establish a "cognitive meeting point". Fuchs (2017, 103) points out that in Easy German, frame-setters are especially important because a Common Ground between author and recipient cannot be presupposed. Apart from frame-setters, in the term clouds for KLD of Easy German news in contrast to Easy German non-news, sentence-initial connectives such as *Denn* ('Because') and *Aber* ('However') stick out. These findings support the findings by Jablotschkin and Zinsmeister (2023), who demonstrate that the sentence-initial position in Easy German news texts is frequently used for discourse connectives and frame-setting adverbials.

When comparing news and non-news in Standard German, similarly to Easy German some linguistic expressions have high distinctivity that potentially serve as frame-setters, such as *nun* ('now'), *im* ('in the'), *in* ('in'), *am* ('at the'/'on the'). However, there are also several finite verb forms that distinguish Standard news from Standard non-news while they are not distinctive of Easy news compared to Easy non-news: *sei* (subjunctive form of 'are'), *habe* (subjunctive form of 'have'), *waren* ('were'), *hatte* ('had'), *sagt* ('says'), *sagte* ('said'). In addition, quotation marks are significantly more frequent in this Standard text genre. These verb forms along with the quotation marks hint at the relevance of (direct and indirect) reported speech in news texts. Reported speech is semantically and pragmatically complex and its use in Easy German is therefore restricted. As a substitute for reported speech marked by subjunctive or quotation marks, Easy German news texts tend to use matrix clauses with a perfect form of the main verb *sagen* ('say') followed by a colon and a subordinate object clause (see example (4)). This observation is supported by the high distinctivity of *gesagt* in our term cloud visualising KLD of Easy German news in contrast to Easy German non-news. Constructions like in example (4) are syntactically and morphologically relatively simple. However, the lack of quotation marks and subjunctive mood in these clauses creates ambiguity and requires the recipient to make additional inferences mainly based on context to determine whether the subordinate clause contains direct or indirect speech.

5 Summary and conclusions

We presented a new corpus documenting the lived practice in simplified German writing. On this basis we built n-gram language models of the strongly regulated variant Easy German and of Standard German. We applied relative entropy to analyse the differences between the Easy German and Standard German models and between text genres within the respective variant. We extracted typical features of Easy German on different linguistic levels and detected text genre differences within Easy and Standard German.

By analysing distinctive types and additionally drawing on sample sentences and concordances, we showed that many of the typical features of Easy German can be traced back to efforts of improving coherence, e.g. by explicitly connecting sentences of a text or explaining difficult words. Some other features of Easy German displayed by our models are a direct consequence of syntactic or morphological simplification. By including metadata into our analysis, we detected overrepresentations of words and phrases in texts by individual publishers that cannot be considered typical features of Easy German. Moreover, we showed that text genre variation is expressed differently in Easy vs. Standard German. Many of these findings are not surprising keeping in mind the rules and recommendations for simplifying text in German. Others, however, such as the distinctivity of potentially ambiguous pronouns in Easy German, are related to simplifications of another aspect, showing that simplifying text with regard to one feature can make it more complex with regard to another. Our approach can thus be used to uncover linguistic features of Easy German that have been overlooked so far.

In a next step, we will use our insights about typical linguistic features of Easy German to design psycholinguistic studies evaluating the comprehensibility of specific linguistic characteristics for people with learning difficulties, one of the main user groups of Easy German. In the future, we will also apply our approach to simplified variants other than Easy German (e.g. Plain German or German texts addressing children) and to further text genres (e.g. lexicons or administrative text). Our findings can be used to classify simplified text found on the web or generated by AI but not carrying any specific label, or to fine-tune simplification algorithms.

6 Acknowledgements

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. We would like to thank Stefan Fischer for his assistance in creating this corpus and computing KLD as well as Nele Benz for her meticulous metadata collection and her help in acquiring and preprocessing text material. Additionally, we would like to thank the reviewers for their insightful comments. Finally, we would like to acknowledge the creators of the other simplified German corpora for making their data available.

References

Benedikt Adelmann. 2021. Batch Substring Edit Distance (Benadelm).

Benedikt Adelmann and Evelyn Gius. 2020. Korpusbereinigung für größere Textmengen. Eine (kurze) Problematisierung und ein Lösungsansatz für Duplikate. In *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, pages 331–334, Paderborn.

Laurence Anthony. 2023. *AntConc (Version 4.2.3) [Computer Software]*. Tokyo, Japan: Waseda University.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Bettina M. Bock. 2014. “Leichte Sprache”: Abgrenzung, Beschreibung und Problemstellungen aus Sicht der Linguistik. In Susanne Jekat, Heike Elisabeth Jüngst, Klaus Schubert, and Claudia Villiger, editors, *Sprache barrierefrei gestalten: Perspektiven aus der Angewandten Linguistik*, 69, pages 17–51. Frank & Timme, Berlin.

Bettina M. Bock. 2017a. Das Passiv- und Negationsverbot "Leichter Sprache" auf dem Prüfstand - empirische Ergebnisse aus Verstehenstest und Korpusuntersuchung. *Sprachreport*, 33(1):20–28.

Bettina M. Bock. 2017b. Texte in "Leichter Sprache"' schreiben. Zwischen Regelerfüllung und Kontext-Angemessenheit. In Dagmar Knorr, Katrin Lehnen, and Kirsten Schindler, editors, *Schreiben im Übergang von Bildungsinstitutionen*, number Band 15 in Textproduktion und Medium, pages 189–213. Peter Lang, Frankfurt am Main [a.o.].

Bettina M. Bock. 2018. *"Leichte Sprache"' - kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt*. Universität Leipzig, Leipzig.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache. Theoretische Grundlagen, Orientierung für die Praxis*. Dudenverlag, Berlin.

Bundesministerium der Justiz und für Verbraucherschutz. 2017. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung - BITV 2.0). ausfertigungsdatum: 12.09.2011. zuletzt geändert: 21.05.2019.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Matthew W. Crocker, Vera Demberg, and Elke Teich. 2015. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz*, 30(1):77–81.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and visualizing variation in language resources. In *Proceedings of the 9th Language Resources and Evaluation Conference*, pages 4125–4128, Reykjavik, Iceland.

Julia Fuchs. 2017. Leichte Sprache und ihr Regelwerk - betrachtet aus der Perspektive der Informationsstruktur. *Sprachwissenschaft*, 42:97–119.

Julia Fuchs. 2019. Leichte Sprache auf dem Prüfstand. Realisierungsvarianten von kausalen Relationen in Leichte-Sprache-Texten. *Sprachwissenschaft*, 44(4):441–480.

Susanne Gross. 2015. Regeln und Standards für leicht verständliche Sprache. Ein Rundblick. In Klaus Candussi and Walburga Fröhlich, editors, *Leicht Lesen. Der Schlüssel zur Welt*, pages 81–105. Böhlau Verlag, Wien [a.o.].

Silvia Hansen-Schirra, Jean Nitzke, and Silke Gutermuth. 2021. An Intralingual Parallel Corpus of Translations into German Easy Language (Geasy Corpus): What Sentence Alignments Can Tell Us About Translation Strategies in Intralingual Translation. In Vincent X. Wang, Lily Lim, and Defeng Li, editors, *New Perspectives on Corpus Translation Studies*, pages 281–298. Springer Singapore, Singapore.

Katja Häuser and Jutta Kray. 2021. Effects of prediction error on episodic memory retrieval: evidence from sentence reading and word recognition. *Language, Cognition and Neuroscience*, pages 1–17.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: The D-SPIN text corpus format and its relationship with ISO standards. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Freya Hewett. 2023. APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 173–179, Toronto, Canada. Association for Computational Linguistics.

Inclusion Europe. *Informationen für alle. Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht*. Brussels.

Inclusion Europe. n.d. *Informationen für alle. Europäische Regeln, wie man Informationen leicht lesbar und leicht verständlich macht*. Brussels. `https://easy-to-read.inclusion-europe.eu/wp-content/uploads/2014/12/DE_Information_for_all.pdf`.

Sarah Jablotschkin and Heike Zinsmeister. 2021. Annotating colon constructions in Easy and Plain German. In *Proceedings of the 3rd Swiss conference on barrier-free communication (BfC 2020)*, pages 125–134, Winterthur (online), June 29–July 4, 2020. Winterthur: ZHAW Zurich University of Applied Sciences.

Sarah Jablotschkin and Heike Zinsmeister. 2023. LeiKo. ein Vergleichskorpus für Leichte und Einfache Sprache. In Marc Kupietz and Thomas Schmidt, editors, *Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022*, pages 71–88. Narr Francke Attempto, Tübingen.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Survey: Anaphora With Non-nominal Antecedents in Computational Linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Daisy Lange. 2019. Der Genitiv in der "Leichten Sprache"' – das Für und Wider aus theoretischer und empirischer Sicht. *Zeitschrift für Angewandte Linguistik*, 70(1):37–72.

Daisy Lange and Bettina M. Bock. 2016. Was heißt "leichte" und "einfache Sprache"'? Empirische Untersuchungen zu Begriffssemantik und tatsächlicher Gebrauchspraxis. In Nathalie Mälzer, editor, *Barrierefreie Kommunikation: Perspektiven aus Theorie und Praxis*, pages 117–134. Frank & Timme, Berlin.

Alexander Lasch. 2017. Zum Verständnis morphosyntaktischer Merkmale in der funktionalen Varietät “Leichte Sprache”. In Bettina M. Bock, Ulla Fix, and Daisy Lange, editors, *“Leichte Sprache”' im Spiegel theoretischer und angewandter Forschung*, pages 275–300. Frank & Timme, Berlin.

Christiane Maaß. 2015. *Leichte Sprache. Das Regelbuch*. Lit, Münster.

Netzwerk Leichte Sprache. 2014. *Leichte Sprache. Ein Ratgeber*. Bundesministerium für Arbeit und Soziales (BMAS), Bonn.

Netzwerk Leichte Sprache, editor. 2021. *Leichte Sprache verstehen: Mit Beispielen aus dem Alltag, Tipps für die Praxis und zahlreichen Texten in Leichter Sprache*. S. Marix Verlag, Wiesbaden.

Netzwerk Leichte Sprache. 2022. *Die Regeln für Leichte Sprache vom Netzwerk Leichte Sprache*.

Netzwerk People First Deutschland e.V. *Wer sind wir? Der Verein*.

Simon Rodier and Dave Carter. 2020. Online near-duplicate detection of news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1242–1249, Marseille, France. European Language Resources Association.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

Marita Tjarks-Sobhani. 2012. Leichte Sprache gegen schwer verständliche Texte. *Technische Kommunikation*, 34(6):23–30.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2023. A new aligned simple German corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11393–11412, Toronto, Canada. Association for Computational Linguistics.

Jorrig Vogels, David M. Howcroft, Elli Tourtouri, and Vera Demberg. 2019. How speakers adapt object descriptions to listeners under load. *Language, Cognition and Neuroscience*, 35(1):78–92.

A Appendix

Due to the formatting, you will find Table 3 on the following page.

	Corpus	Reference	Size	Architecture	Alignments	Annotations
1	APA-RST*	Hewett (2023)	Standard: 9,567 tokens; A2: 1,871 tokens; B1: 2,009 tokens	parallel (Original, A2, B1)	text, sentence	RST
2	DEplain*	Stodden et al. (2023)	1,239 document pairs; 16,562 sentence pairs	parallel (Standard, Plain)	text, sentence	simplification operations; aspects of coherence and simplicity
3	Geasy*	Hansen-Schirra et al. (2021)	Standard: 1,078,643 words; Easy: 292,552 words	parallel (Standard, Easy)	text, sentence	dependencies and tree alignments (in progress)
4	LeiKo*	Jablotschkin/Zinsmeister (2023)	Plain: 16,706 tokens; Easy: 39,653 tokens	comparable	none	lemmas, POS; dependencies; PDTB relations; coreference; text structure; typography; metadata
5	LeiSa corpus	Lange/Bock (2016)	Easy: 1,382,142 tokens; simplified (other): 882,806 tokens	comparable	none	POS; text level: ‘area of communication’
6	Simple German Corpus*	Toborek et al. (2023)	Plain: 94,808 tokens; Easy: 155,285 tokens; Standard: 404,771 tokens	parallel (Standard, simplified)	text, sentence	none
7	WebCorpus*	Battisti et al. (2020)	approx. 6,200 documents; 211,000 sentences	parallel (Standard, Simplified)/comparable	text, sentence	lemmas, morphological units, POS, dependencies, text structure, typography, metadata
8	DE-Lite (under construction)	this text	DE-Lite v1: approx. 8,000 texts/more than 3 million tokens	parallel (Standard, Easy; Standard, Plain; Standard, children)/ monolingual	text	will be published in upcoming versions

Table 3: Related German Corpora (sizes are quoted from the original publications) and our own DE-Lite corpus; *) Corpus texts (partially) included in DE-Lite