Easy-to-Read Language Resources and Tools for three European Languages

Margot Madina
margot.madina-gonzalez@h-da.de
Darmstadt University of Applied Sciences (Hochschule Darmstadt)
Darmstadt, Hessen, Germany

Itziar Gonzalez-Dios
itziar.gonzalezd@ehu.eus
HiTZ Zentroa-Ixa, Euskal Herriko Unibertsitatea (UPV/EHU)
Donostia, Gipuzkoa, Spain

Melanie Siegel
melanie.siegel@h-da.de
Darmstadt University of Applied Sciences (Hochschule Darmstadt)
Darmstadt, Hessen, Germany

ABSTRACT

Easy-to-Read (E2R) is a controlled language variant that makes any written text more accessible through the use of clear, direct and simple language. It makes content accessible to anyone with a minimum reading level, but it is mainly aimed at specific user groups. In recent years, important developments in the field of E2R have been carried out in many different languages in addition to English. The main objective of this paper is to present an updated overview of the existing tools and resources for E2R in German, Spanish and Basque. We aim to offer a benchmark for future studies that focus on these languages and their E2R variants.

CCS CONCEPTS

• **Human-centered computing → Accessibility systems and tools**.

KEYWORDS

Easy-to-Read, Leichte Sprache, Lectura Fácil, Irakurketa Erraza, readability

ACM Reference Format:
Margot Madina, Itziar Gonzalez-Dios, and Melanie Siegel. 2023. Easy-to-Read Language Resources and Tools for three European Languages. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3594806.3596530

1 INTRODUCTION

The Convention on the Rights of Persons with Disabilities of the United Nations (CRPD) stated that access to information and communication is a fundamental right of all citizens and that states should facilitate information in accessible ways such as easy to read and understand forms [65]. Easy-to-Read (E2R) is a pivotal part of the inclusion for people with communication disabilities [30]. E2R is a language variant of a standard language. It has a reduced complexity and aims to make content accessible and ensure participation of people with communication impairments by improving the readability and comprehensibility of texts [31, 51]. Here, we will use the term *Easy-to-Read* (E2R) to refer to any of the language E2R variants, as it is the most commonly used term. When talking about a specific language, we will be using the term corresponding to it; *Leichte Sprache* in the case of German, *Lectura Fácil* for Spanish, and *Irakurketa Erraza* for Basque.

E2R is mainly aimed towards people with cognitive or intellectual disabilities; however, other target groups may also benefit from it, such as people with intellectual, cognitive or developmental disabilities, people with auditory disabilities, people with low literacy, migrants or children in need of reading reinforcement [15, 31]. The vocabulary and grammar structures of E2R are limited to the basic vocabulary and grammar of a given natural language [30, 51]. E2R texts are adapted based on a set of rules, and they might also include examples, visual and/or auditory aids that are not present in the source text. When we refer to the E2R adaptation, we are talking about the processes that standard texts undergo in order to be transformed into E2R texts. These processes may include the creation of auditory and/or visual aids and examples, syntactic simplification, lexical simplification or summarization, among others. Automatic Text Simplification (ATS) is an important part of E2R adaptation, as it aims to reduce, at least in part, the efforts required by manual simplification [14].

E2R is not to be confused with what is known as Simple Language or Plain Language (PL), as there are some important differences between them: (1) while PL focuses mainly in the text, E2R covers not only the text, but also the illustrations and layout [25], (2) E2R is usually aimed at people with cognitive or intellectual disabilities (although it may be used by wider audiences), PL might be too challenging for people with cognitive or intellectual disabilities [43], (3) PL was initially focused on improving legal communication, and it is usually employed to communicate information that is usually complex, E2R aims to create a barrier-free communication and therefore be applied to all sorts of texts [7]. In spite of the differences between E2R and PL, this distinction is not yet made in all languages. We can find E2R and PL in English, as well as *Leichte Sprache* and *einfache Sprache* in German. In Spanish, there is *Lectura Fácil* (the equivalent to E2R), *Lenguaje Claro* (the equivalent to PL), *Comunicación Clara* (clear communication) and *Lenguaje Ciudadano* (citizen language). Finally, in Basque, there is only *Irakurketa Erraza.* [1] Depending on the language, the target readers or users of E2R might also vary. For example, it has been shown that *Leichte Sprache* is not suitable to teach learners of German as a second language [1, 34, 37]. However, *Irakurketa Erraza* can be used by learners of Basque as a second language or by people who have difficulties

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07…$15.00
https://doi.org/10.1145/3594806.3596530

[1] This paper will focus on the E2R variant of German, Spanish and Basque; all other variants are beyond the scope of this paper.

understanding Basque[2]. Apart from this, there are some E2R rules that apply to all languages; for example, avoiding special characters such as the parenthesis, or avoiding the use of metaphors. These general rules can be found in the guidelines provided in the Inclusion Europe website[3]. On the other hand, there are some E2R rules that are particular to a language. In German, for example, we see that compound nouns should be split into two or more words by means of bullet points [19, 21, 42].

The primary objective of this paper is to offer an updated overview of the existing tools and resources for E2R in German, Spanish and Basque. This will help set a benchmark for future studies focused on these languages and their E2R variants.[4] We focus on German, Spanish and Basque due to various reasons: (1) these languages have experienced a high contribution rate regarding E2R in recent years, (2) some investigation conducted on these languages has not been published in English, but in the country's language, (3) some of the resources and tools introduced here may be adapted to other languages, (4) most E2R surveys and ATS surveys generally focus on English contributions; however, this paper will explain in depth the German, Spanish and Basque resources, (5) we want to give visibility to other languages besides English.

The paper is structured as follows: the next section will introduce the most important and updated Natural Language Processing (NLP) tools and resources for E2R for German, followed by Spanish and then Basque. Each section might contain different types of tools and resources, as not all languages count with the same amount of E2R aids. The tools and resourced that will be mentioned might include corpora, Readability Assessments (RA), dictionaries, language checkers, and E2R adaptation tools.

2 GERMAN

This section will give a brief overview of the situation of E2R in Germany and will introduce the available resources for *Leichte Sprache*.

In 2002, the establishment of the *Gesetz zur Gleichstellung von Menschen mit Behinderungen* (the German equality law for disabled people) and the *Barrierefreie-Informationstechnik-Verordnung* (the accessible technology enactment) led the German government to provide accessible information to everyone. The *Netzwerk Leichte Sprache* (the plain language association) was founded in 2006, an association sustained by the empowerment movement of people with cognitive impairments and their care service providers [41]. In 2013, the *Netzwerk Leichte Sprache* developed the rules for *Leichte Sprache* [53].[56].

2.1 Corpora

Klaper et al. [38] developed the first parallel, sentence-aligned corpus with German and *Leichte Sprache* texts. The data was crawled from five publicly available webpages, spanning various topics, and is composed of around 70,000 tokens.[7] A multilingual dataset called **CWIG3G2** was created by **Yinam et al.** [69] for Complex Word Identification (CWI) tasks. This dataset contains complex words/phrases for English, German and Spanish, annotated by both native and non-native speakers. It contains a total of 978 sentences from German Wikipedia articles [8]. **Lange** [39] presented the **LeiSa** (Leichte Sprache im Arbeitsleben) collection of texts, which contains 639,826 tokens of *Leichte Sprache*, 779,278 tokens of *einfache Sprache*, and 350,872 tokens of *Leicht Lesen*. They contrasted these text simplification approaches and confirmed that these approaches can be conceptualized as part of a comprehensibility continuum.[9] **Battisti et al.** [8] compiled a corpus from web sources, with the aim to use it in automatic RA and ATS in German. Their corpus consists of both monolingual data (*Leichte Sprache*, consisting of 1,916,045 tokens) and parallel data (German and *Leichte Sprache*, consisting of 347,941 tokens of German and 246,405 tokens of *Leichte Sprache*). It also contains information on text structure, typography and images, which can indicate if a text is simple or complex.[10] **Naderi et al.** [49] offer the **TextComplexityDE** dataset, composed by 1000 sentences, 250 of which have been manually simplified by native speakers. The sentences were taken from 23 Wikipedia articles in 3 different article-genres. It also contains subjective assessment of the simplified sentences (complexity, understandability and lexical difficulty), which was provided by a group of language learners of A and B levels. It is aimed to be used for developing text-complexity predictor models and ATS. [11]. **Säuberli et al.** [56] presented the **APA** (the Austria Presse Agentur) corpus, which is the first parallel corpus for data driven ATS for German. It has a total of 3,616 sentence pairs. The original sentences were manually simplified and aligned into their A2 and B1 equivalents. Spring et al. [59] offer an expanded version of this database, which consists of standard-language news items with their corresponding simplifications between August 2018 and April 2021; this resulted in 2,410 document pairs for B1 and 2,347 for A2.[12] **Jablotschkin and Zinsmeister** [35] developed the **LeiKo** comparable corpus, which contained approximately 50,000 tokens. It is divided into four sub-corpora according to the websites from which they were extracted. It is composed of news texts, systematically compiled and linguistically annotated for linguistic and computational linguistic research.[13]. **Jach** [36] created the **KED** (Korpus Einfaches Deutsch) collection of texts, which contains texts from genres of educational and public discourse in *Leichte Sprache* and *einfache Sprache*. It was scraped from different websites, and containes a total of 3,698,372 words. It

[2] https://lecturafacileuskadi.net/blog/irakurketa-erraza-ezinbesteko/ (last accessed: 2023-02-23)

[3] https://www.inclusion-europe.eu/easy-to-read-standards-guidelines/ (last accessed: 2022-11-23)

[4] Lindholm and Vanhatalo [41] offer an insight to the implementation, principles and practices of E2R in various European countries in 2020 and before, focusing on its background and current situation, terminological definitions, status and stakeholders.

[5] Inclusion Europe developed the first rules for *Leichte Sprache* in 1998. This first set of rules were written in German, however, they were not focused on the German language specifically; they could be applied to any E2R variant of a given standard language, as they were generic E2R rules.

[6] For a more detailed overview of how the legal state of LS has changed over time, refer to Maaß [43]

[7] The corpus can be made available upon request.

[8] The dataset is available under CC-BY (https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/complex-word-identification-dataset.html)

[9] The corpus is not yet accessible through databases, but this is still a future project.

[10] The data is available upon request (https://zenodo.org/record/6576356#.Y39MxS_WdmA)

[11] The dataset is available in GitHub (https://github.com/babaknaderi/TextComplexityDE)

[12] This dataset, aligned with the LHA method, is available online upon request (https://zenodo.org/record/5148163#.Y4co9i8rxqs)

[13] The corpus is available online (https://zenodo.org/record/3923917#.Y3uxdS8rxao)

is divided into different sub-corporas depending on the provider.[14] **Rios et al.** [55] presented the **20m** corpus of 18,305 articles paired with shortened summaries collected from the Swiss news portal *20 Minuten*. The sentences are not aligned, and the corpus does not distinguish different simplification levels, as they do not stick to a simplification standard.[15] The **Geasy** (the German Easy Language) corpus was built by **Hansen-Schirra et al.** [32]. It is a parallel corpus with professional translations from standard German into *Leichte Sprache*. It is aligned at sentence level and currently contains 1,087,643 words of source text and 292,552 words of *Leichte Sprache* translations.[16] **Toborek et al.**[62] created a monolingual corpus consisting of publicly available articles, spanning different topics, of 7 different webpages that publish news articles in German and their corresponding *Leichte Sprache* version. The corpus also has articles from a website in *einfache Sprache* in an aim to achieve a larger vocabulary size. The authors refer to all simplified versions of German as *Simplified German*. It has a total of 250,093 tokens of *Simplified German* and 404,771 tokens of German, contains 708 aligned documents and a total of 5,942 aligned sentences.[17] We can also find a multilingual dataset of raw text from different news providers in different countries; this dataset is named **SNIML** (Simple News in Many Languages) and was crated by **Hauser et al.** [33]. It is a multilingual corpus of news in simplified language, including articles in Finnish, French, Italian, Swedish, English and German, published between 2003 and 2022. The texts have been created according to different simplification guidelines and for different target audiences, as the level of simplification varies depending on the provider. They plan to release a new version of SNIML every month, and their future work may consist of aligning the articles to related articles in standard language. By the time this article is being written, it contains 4,936,181 tokens in total, 123,021 of which are of German.[18] **Aumiller and Gertz** [5] presented the **Klexicon** corpus, which is based on the German children encyclopedia, named *Klexicon*. This corpus contains 2,898 articles from *Klexicon*, with an average of 436.87 tokens each, and 2,898 documents from Wikipedia, with an average of 5,442.83 tokens each. It is aligned on a document level with corresponding articles from Wikipedia; however, it is unlikely that specific sentences are matched.[19]

2.2 Readability assessment

Battisti et al. [9] presented an unsupervised machine learning approach to analyse texts in simplified German, and also exploit structural and typographic characteristics of simplified texts. They showed that there is not just one complexity level in German simplified texts. **Naderi et al.** [50] developed an automated RA estimator based on supervised learning algorithms over German text corpora. They employed the *TextComplexityDE* corpus, and extracted 73 linguistic features and employed feature engineering approaches to select the most informative ones. **Mohtaj et al.** [48] presented a new model for text complexity assessment for German text based on transfer learning. To train the models, they used the *TextComplexityDE* dataset, showing that fine-tuning the BERT model can outperform the other approaches. **Ebling et al.** [23] presented the first sentence-based Neural Machine Translation approach towards automatic simplification of German and the first multi-level simplification approach for German. This paper also offers an overview of four parallel corpora of standard/simplified German, compiled and curated by their group. They report a gold standard of sentence alignments from these four sources. **Weiss et al.** [67] presented a sentence-wise RA model for German L2 readers. They built a machine learning model with linguistic insights and compare its performance based on predictive regression and sentence pair ranking. Their findings show that the model yielded top performances across tasks. **Mohtaj et al.** [47] offer an overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text and the main contributions.[20]

2.3 E2R adaptation aids

We can find two tools devoted to help to adapt standard texts into *Leichte Sprache*. **EasyTalk** [61] is a system for assisted typing in *Leichte Sprache*. It uses a paraphrase generator based on a lexicalized, unification-based Performance Grammar. On the other hand, **SUMM**[21] is the first AI-powered tool that automatically turns any text into E2R. Its founders claim that it can increase adaptation productivity by 85%. It is only available in its Beta version, but it can be tried out by registering on their webpage. Users can adapt any texts and create their own glossary. Apart from the these tools, German also counts with **Hurraki**[22], a dictionary with explanations of German words in *Leichte Sprache*.

On the other hand, it is worth highlightning that the implementation of E2R rules in language checkers could aid in the creation of E2R texts. **Siegel and Lieske** [40, 58] implemented some E2R rules in Acrolinx[23] and LanguageTool [24].

3 SPANISH

We will now introduce *Lectura Fácil* and the available resources for it.

Mayol and Salvador [44] analysed the situation of E2R in Catalonia and Europe back in 1999, and made some recommendations on how to implement it in Catalonia. In 2001, an Easy Language committee was created [41]. The Easy Read Association was legally founded in 2003; since then, it has been focused on the creation of literary works, the promotion of Easy Read Clubs, and providing training to create accessible information. In 2016, the *Red de Lectura Fácil* (Easy Read Network) was created by various regional associations. Some other initiatives have been in place for years [45, 64]; however, they have not been applied effectively [11]. In spite of this, there are some resources available for *Lectura Fácil*.

[14] The corpus can be obtained online (https://daniel-jach.github.io/simple-german/simple-german.html)

[15] It can be obtained through GitHub (https://github.com/ZurichNLP/20Minuten)

[16] An overview of the texts and word counts can be found online (https://seafile.rlp.net/f/a25a64e6dfa54373b5a1/)

[17] The corpus is published under CC BY-SA and the accompanying code under MIT license. The code to build the dataset is available in GitHub (https://github.com/mlai-bonn/Simple-German-Corpus), and the fully prepared dataset is available upon request

[18] All texts in SNIML are shared under an open license that allows for academic research use (https://pub.cl.uzh.ch/projects/sniml/en/read/)

[19] The code and data of this corpus are available in GitHub (https://github.com/dennlinger/klexikon)

[20] Due to space constraints, not all studies that took part in it have been included in the paper at hand.

[21] https://summ-ai.com/en/ (last accesed: 2023-02-28)

[22] https://hurraki.de/wiki/Hauptseite (last accessed: 2023-02.28)

[23] Acrolinx is a software package to support authors of technical documentation

[24] LanguageTool is an open source text checking software developed since 2003

3.1 Corpora

The **Simplext** parallel corpus was created by **Bott et al.** [14] within the *Simplext* project. They compiled a corpus of 200 news texts and created manual simplifications for them by trained experts, based on 28 rules. The texts were automatically aligned on sentence level with a tool they created to this end [13]. This amounted to a total of 1,149 Spanish and 1,808 simplified Spanish sentences; approximately 1,000 aligned sentences. This corpus has "heavy" simplifications, as it has a high number of strong paraphrases, deletions and heavy structural reordering of sentences.[25] The **CWIG3G2** CWI dataset by **Yinam et al.** [69] (introduced in 2.1) contains a total of 1,387 sentences from the Spanish Wikipedia. **Štajner et al.** [60] built new simplification-specific datasets of synonyms and paraphrases using freely available resources.[26] **Mitkov and Štajner [46]** created a corpus consisting of texts from the general literature and news domain, some of them also present in the *Simplext* corpus. Their manual simplifications were obtained using only six main simplification rules (in contrast with the 28 rules used in the *Simplext* project). This resulted in "light" simplifications.[27] The **Newsela** corpus by **Xu et al.** [68] is available both in Spanish and English. It consists of a total of 1,130 news articles (1,301,767 tokens), and 4 simplified versions of each, written by professional editors. Each of the 4 simplified versions corresponds to a different reading level.[28] The **VYTEDU-CW** (Educational Videos and Texts-Complex Word) corpus by **Ortiz-Zambrano et al.** [52] is composed by university educational texts containing difficult words. The corpus is being used to continue researching the ATS of Spanish texts.[29] **Alarcón** [2] developed the **EASIER** dataset, which was annotated by an E2R linguist expert. They annotated 260 documents, from which, 8,100 complex words were gathered. A total of 7,892 synonyms were proposed. The dataset contains about 5,130 instances with at least one proposed substitute per complex word.[30] **ALEXIS**, by **Ferrés and Saggion** [24] is a Spanish Dataset for Lexical Simplification. It contains a total of 381 instances; each instance is composed by a sentence, a target complex word, and 25 candidate substitutes. Their work also describes the evaluation of three kind of approaches to Lexical Simplification: a thesaurus-based approach, a single transformers-based approach, and a combination of transformers.[31] **Sharlow and Alva-Manchego** [57] proposed **Simple TICO-19**, a new language resource containing manual simplifications of the English and Spanish portions of the TICO-19 corpus for Machine Translation of COVID-19 literature [3]. The annotation process consisted on designing an annotation manual and, based on it, four annotators (two native English speakers and two native Spanish speakers) simplified over 6,000 sentences. They also proposed baseline methodologies for automatically generating the simplifications, translations and joint translation and simplifications contained in the dataset.[32] **IrekiaLF_es** is an open-license benchmark for Spanish text simplification developed by **Gonzalez-Dios et al.** [29]. It compiles 288 parallel original and E2R texts; 35 of them were manually aligned at a sentence level, thus creating a test set of 705 sentences. The corpus was built by crawling texts from *Irekia*[33], and they performed a neurolinguistically-based evaluation of the corpus, following the lexicon-unification-Linearity (LeULi) model of neurolinguistic complexity assessment. This evaluation showed that the corpus is suitable for ATS training and evaluation regarding lexical and sentence simplification; however, it may obstruct en users' comprehension in terms of discourse simplification.[34] **Campillos et al.** [16] created **CLARA-MeD**, a comparable corpus composed by 24,298 pairs of professional and simplified texts in the medical domain. They also aligned a subset of 3,800 sentence pairs manually. The data was extracted from CIMA (*Centro de Información de Medicamentos*) a drug-related service and knowledge database.[35]

3.2 Readability assessment

Quispesaravia el at. [54] developed **Coh-Meetrix-Esp**, a tool able to calculate 45 readability indices. They analysed how complexity indices behave in a corpus of 100 texts, divided into "simple" and "complex" categories equally. The "simple" texts were mainly children's fables, and the "complex" ones were stories for adults. **Bengoetxea and Gonzalez-Dios** [10] presented **MultiAzterTest** . This analyzer is multilingual, as it works with English, Spanish and Basque; besides, it can also be adapted to other languages. It analyzes texts on over 125 measures of cohesion, language, and readability, and it can also be used for text analysis, profiling or stylometrics. In **Ventosa Pérez's** bachelor project [66] they implemented a web application for RA, which calculated some indexes such as types of words or punctuation marks, among others. [36] In the context of a hackaton, **De la Rosa et al.** [20] created **BERTIN for RA** [37]. They presented a data-centric technique called "perplexity sampling". Their resulting models classify three levels of texts and have been created by fine-tuning the language model for Spanish. **Torrijos and Oquendo** [63] present **Clara**[38], a tool that allows to check the clarity of administrative documents and service contracts (although it can be used with any type of text). It offers an overall percentage of clarity, a specific percentage of clarity for each of the metrics included, and proposals for improvement.

3.3 E2R adaptation aids

Bott et al. [14] developed **Simplext**, a prototype that performs syntactic simplification with a rule-based approach, also trying to integrate data driven methods whenever it is possible (due to the lack of parallel resources for Spanish). They targeted three groups of problems within the structural simplification: sentence

[25] This dataset is available upon request.

[26] This dataset is available upon request.

[27] This dataset is available upon request.

[28] This dataset is available upon request (https://newsela.com/data/)

[29] The corpus is available upon request.

[30] The dataset is available at GitHub (https://github.com/LURMORENO/EASIER{_}CORPUS), but without explicit license.

[31] The dataset is availble at GitHub https://github.com/LaSTUS-TALN-UPF/ALEXSIS

[32] It can be obtained through GitHub (https://github.com/MMU-TDMLab/SimpleTICO19)

[33] the open-government communication channel of the Basque Government.

[34] The dataset is publicly available (https://github.com/itziargd/IrekiaLF) under CC BY-SA 4.0 license.

[35] This dataset is publicly available (https://github.com/lcampillos/CLARA-MeD). The authors make it clear that the corpus was built for research and educational purposes, and no medical decision should be taken from the data provided.

[36] This paper follows a series of publications by the Universidad Politéctica de Madrid (UPM), related with E2R RA and adaptation aids.

[37] https://huggingface.co/spaces/hackathon-pln-es/readability-assessment-spanish (last accessed: 2023-02-28)

[38] https://comunicacionclara.com (last accessed: 2023-02-27)

splitting, lexical substitution of functional multi-word units and the re-ordering of syntactic units. **LexSIS** by **Bott et al.** [12] was the first approach to lexical simplification in Spanish. It was created based on the analysis of a sample of data from the *Simplext* corpus. It relies on freely available resources, such as dictionaries and the web, as corpus.In order to find a suitable word substitute, *LexSiS* uses three techniques: a word vector model, word frequency, and word length. **Baeza-Yates et al.** [6] presented **CASSA** (Context-Aware Synonym Simplification Algorithm), which generates simpler synonyms of a word. It is a context-aware method for lexical simplification that uses two free language resources (Google Books Ngram Corpus and the Spanish OpenThesaurus) and real web frequencies of the complex word for disambiguation. It does not require alignment of parallel corpora, and it can be extended to other languages. **Cumbicus-Pineda et al.** [17] added syntactic information into the edit-based system *EditNTS* [22]. They extended the system with a graph convolutional network module that mimics the dependency structure of the sentence; this gave the model an explicit representation of syntax. They conducted experiments in Spanish, Italian and English, and confirmed that syntactic information is useful for ATS systems. Finally, there is also a dictionary called **Diccionario Fácil** [39], which aims to define terms and linguistic expressions, as well as proper names or historical events following the guidelines of E2R.[40]

4 BASQUE

In this section, we overview the state of *Irakurketa Erraza* and its available resources.

The Basque Country's E2R Association [41] makes literature, culture and information accessible. They brought the E2R project from Catalonia in 2012. They offer trainings, books and news in E2R. *Irakurketa Erraza* is sometimes also used to teach Basque to those who have a different first language. Most of the existing tools for *Irakurketa Erraza* ATS have been developed by the IXA group[42].

4.1 Corpora

The Corpus of Basque Simplified Texts (CBST) or *Euskarazko Testu Sinplifikatuen Corpusa* (ETSC) in Basque by **Gonzalez-Dios et al.** [28], compiles 227 original sentences and two simplified versions of each sentence. The sentences belong to the topics of social sciences, medicine and technology. The simplified versions of the sentences were created following two different approaches, the structural and the intuitive.

[43]. **The Leveled Basque Science Popularisation Corpus (LBSPC)** by **Gonzalez-Dios et al.** [27] is composed of 400 texts at 2 levels. 200 of them were extracted from *ZerNola* [44], a website to popularise science among children up to 12 years old; this was meant to be the "easy" collection of texts. The other 200 texts were taken from *Elhuyar Aldizkaria* [45], a journal about science and technology in Basque; this was meant to be the "complex" collection of texts.

4.2 Readability assessment

There are only two studies focused on Basque RA. **ErreXail**, created by **Gonzalez-Dios et al.** [27] calculates 94 indices based on global, lexical, morphological, morpho-syntactic, syntactic and pragmatic features. On the other hand, there is also **MultiAzterTest** [10] [46] The corpus employed in both studies was **LBSPC** [27].

4.3 E2R adaptation aids

Aranzabe et al. [4] performed a linguistic study on long sentences taken from two corpora (EPEC and Consumer), and based on it, they proposed a syntactic simplification by using manually written rules applied to a syntax tree. This syntactic simplification approach is based on morphological constituents, which is necessary for high inflection languages like such a Basque. The simplification process consisted of four steps: (1) splitting, (2) reconstruction, (3) reordering, and (4) adequacy and correction. Following the aforementioned work, **Gonzalez-Dios** [26] presented a rule based system for syntactic simplification that simplified texts at different levels.

5 CONCLUSIONS

Taking everything that has been said into account, we could state that E2R and ATS have been a recurrent field of study in these recent years and seem to stay that way. However, there is still much to do, specially regarding the product development of all the tools and resources that are being created. This would help make information accessible to all those people with communication disabilities. This paper shows the existing heterogeneity between different E2R variants. As of the present day, a consensus regarding a unified standard within the community remains elusive. It is imperative to undertake intercultural and interlinguistic endeavors to establish fundamental benchmarks. Subsequently, each language should adhere to its distinct characteristics rather than relying solely on literal translations. There is no consistency within the databases; they might be parallel, comparable, aligned, and can include texts of different complexity levels. This might be due to different reasons: (1) they have been thought to serve for different purposes, or (2) there is not enough data to create the corpora from. Although efforts have been made to develop tools to automate the process of adapting standard texts to E2R, we see that the final texts do not conform to E2R language standards. It would be interesting to work on tools that also take into account the layout of the texts and the automatic creation of images and examples for those terms that are still difficult to understand. On the other hand, summarisation systems are mostly extractive, not abstractive, so that even if the text content is reduced, some sentences and words are still too difficult for readers. Future work could also focus on improving abstractive summarisation systems.

[39] http://diccionariofacil.org

[40] We can also find *arText* [18], the first Spanish-assisted copywriter that helps to write texts in specialised fields and texts in *Lenguaje Claro*. It is aimed to help public administrations to write in a more comprehensible way. It has not been included in the paper because it does not deal specifically with *Lectura Fácil*, and therefore is beyond the scope of this paper (http://sistema-artext.com/lenguaje-claro)

[41] https://lecturafacileuskadi.net/blog/irakurketa-erraza-ezinbesteko/ (last accessed:2023-03-05)

[42] http://www.ixa.eus/ (last accessed: 2023-03-05)

[43] It is available under the CC BY-NC-SA 4.0 license (http://www.ixa.eus/node/13007?language=eu)

[44] http://www.zernola.net/ (last accessed: 2023-02-25)

[45] https://aldizkaria.elhuyar.eus (last accessed: 2023-02-25)

[46] For the generalities of this system, please refer to 3.2.

REFERENCES

[1] Sarah Ahrens. 2022. Easy Language Translations for Second Language Learners–Worthwhile Concept or Didactic Mistake? *Translation, Mediation and Accessibility for Linguistic Minorities* 128 (2022), 175.

[2] Rodrigo Alarcón García. 2022. Lexical Simplification for the Systematic Support of Cognitive Accessibility Guidelines. (2022).

[3] Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online.

[4] Marıa Jesús Aranzabe, Arantza Dıaz De Ilarraza, and Itziar Gonzalez-Dios. 2012. First Approach to Automatic Text Simplification in Basque. In *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*. 1–8.

[5] Dennis Aumiller and Michael Gertz. 2022. Klexikon: A German Dataset for Joint Summarization and Simplification. *arXiv preprint arXiv:2201.07198* (2022).

[6] Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. Cassa: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1380–1385.

[7] Christopher Balmford. 2002. Plain Language: Beyond a 'Movement'. Repositioning Clear Communication in the Minds of Decision-Makers. In *Fourth Biennial Conference of the PLAIN Language Association International, Toronto, Canada*.

[8] Alessia Battisti and Sarah Ebling. 2019. A corpus for Automatic Readability Assessment and Text Simplification of German. *arXiv preprint arXiv:1909.09067* (2019).

[9] Alessia Battisti, Sarah Ebling, and Martin Volk. 2019. An Empirical Analysis of Linguistic, Typographic, and Structural Features in Simplified German Texts.

[10] Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment. *arXiv preprint arXiv:2109.04870* (2021).

[11] María Pilar Castillo Bernal and Marta Estévez Grossi. 2022. *Translation, Mediation and Accessibility for Linguistic Minorities*. Vol. 128. Frank & Timme GmbH.

[12] Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *COLING*.

[13] Stefan Bott and Horacio Saggion. 2011. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. 20–26.

[14] Stefan Bott, Horacio Saggion, and Simon Mille. 2012. Text Simplification Tools for Spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. 1665–1671.

[15] Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen? Orientierung für die Praxis*. Bibliographisches Institut GmbH.

[16] Leonardo Campillos Llanos, Ana R Terroba Reinares, Sofía Zakhir Puig, Ana Valverde, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. (2022).

[17] Oscar M Cumbicus-Pineda, Itziar Gonzalez-Dios, and Aitor Soroa. 2021. A Syntax-Aware Edit-based System for Text Simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. 324–334.

[18] Iria da Cunha. 2022. Un redactor asistido para adaptar textos administrativos a lenguaje claro. *Procesamiento del Lenguaje Natural* 69 (2022), 39–49.

[19] Das Netzwerk Leichte Sprache. 2013. Die Regeln für Leichte Sprache.

[20] Javier de la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo González de Prado Salas, Manu Romero, and Marıa Grandury. 2022. BERTIN: Efficient Pre-Training of a Spanish Language Model using Perplexity Sampling. *Procesamiento del Lenguaje Natural* 68 (2022), 13–23.

[21] Silvana Deilen. 2022. Visual Segmentation of Compounds in Easy Language: Does the Marking of Morpheme Boundaries Reduce Cognitive Processing Costs? *Translation, Mediation and Accessibility for Linguistic Minorities* 128 (2022), 161.

[22] Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: A Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing". In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3393–3402.

[23] Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic Text Simplification for German. *Simple and Simplified Languages* (2022).

[24] Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A Dataset for Lexical Simplification in Spanish. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. 3582–3594.

[25] Óscar García Muñoz. 2013. *Léctura fácil - Métodos de Redacción y Evaluación*. Ministerio de Educación.

[26] Itziar Gonzalez-Dios. 2016. *Euskarazko Egitura Sintaktiko Konplexuen Analisirako eta Testuen Sinplifikazio Automatikorako Proposamena. Readability Assessment and Automatic Text Simplification. The Analysis of Basque Complex Structures*. Ph. D. Dissertation. University of the Basque Country (UPV/EHU).

[27] Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or Complex? Assessing the Readability of Basque Texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*. 334–344.

[28] Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation* 52, 1 (2018), 217–247.

[29] Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar M. Cumbicus-Pineda, and Aitor Soroa. 2022. IrekiaLF_es: a New Open Benchmark and Baseline Systems for Spanish Automatic Text Simplification. In *Proceedings of the TSAR 2022 Workshop*.

[30] Silvia Hansen-Schirra, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, SILVAN DEILEN, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. 2020. Intralingual Translation into Easy Language–or How to Reduce Cognitive Processing Costs. *Easy Language Research: Text and User Perspectives. Berlin: Frank & Timme* (2020), 197–225.

[31] Silvia Hansen-Schirra and Christiane Maaß. 2020. Easy Language, Plain Language, Easy Language Plus: perspectives on comprehensibility and stigmatisation. *Easy Language Research: Text and User Perspectives* 2 (2020), 17.

[32] 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 *New Perspectives on Corpus Translation Studies*. Springer, 281–298.

[33] Renate Hauser, Jannis Vamvas, Sarah Ebling, and Martin Volk. 2022. A Multilingual Simplified Language News Corpus. In *2nd Workshop on Tools and Resources for REAding DIfficulties (READI)*. 25.

[34] Antje Heine. 2017. Deutsch als Fremd-und Zweitsprache–eine besondere Form Leichter Sprache? Überlegungen aus der Perspektive des Faches DaF/DaZ. *Leichte Sprache "im Spiegel theoretischer und angewandter Forschung. Berlin: Frank & Timme* (2017), 401–414.

[35] Sarah Jablotschkin and Heike Zinsmeister. 2020. LeiKo: A Corpus of Easy-to-Read German. (2020).

[36] Daniel Jach. 2020. Korpus Einfaches Deutsch (KED).

[37] Jörg Kilian. 2017. „Leichte Sprache ", Bildungssprache und Wortschatz–Zur sprach-und fachdidaktischen Wertigkeit der Regelkonzepte für „leichte Wörter ". *Eds. Bettina Bock, Ulla Fix, and Daisy Lange."Leichte Sprache" im Spiegel theoretischer und angewandter Forschung* (2017), 189–209.

[38] David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. (2013).

[39] Daisy Lange. 2018. Comparing 'Leichte Sprache','einfache Sprache'and 'Leicht Lesen': A Corpus-Based Descriptive Approach. In *Eds. SUSANNE J. JEKAT, and GARY MASSEY. Barrier-free communication: methods and products: proceedings of the 1st Swiss conference on barrier-free communication. Winterthur: ZHAW Digital Collection*. 75–91.

[40] Christian Lieske and Melanie Siegel. 2014. Verstehen leicht gemacht. *technische kommunikation* 1 (2014), 44–49.

[41] Camilla Lindholm and Ulla Vanhatalo. 2021. *Handbook of Easy Languages in Europe*. Frank & Timme.

[42] Christiane Maaß. 2015. *Leichte Sprache. Das Regelbuch*. Lit-Verlag.

[43] Christiane Maaß. 2020. *Easy Language–Plain Language–Easy Language Plus: Balancing comprehensibility and acceptability*. Frank & Timme.

[44] M. Carme Mayol and Eugènia Salvador. 1999. *Materials de lectura-fàcil: Anàlisi, directrius internacionals i proposta per a elaborar aquests materials a Catalunya*. FUS, Grup de Fundacions.

[45] Ministerio de Justicia. 2011. Informe de la Comisión de Modernización del Lenguaje Jurídico.

[46] Ruslan Mitkov and Sanja Štajner. 2014. The fewer, the better? A Contrastive Study About ways to Simplify. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*. 30–40.

[47] Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*. 1–9.

[48] Salar Mohtaj, Babak Naderi, Sebastian Möller, Faraz Maschhur, Chuyang Wu, and Max Reinhard. 2022. A Transfer Learning Based Model for Text Readability Assessment in German. *arXiv preprint arXiv:2207.06265* (2022).

[49] Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective Assessment of Text Complexity: A Dataset for German Ganguage. *arXiv preprint arXiv:1904.07733* (2019).

[50] Babak Naderi, Salar Mohtaj, Karan Karan, and Sebastian Möller. 2019. Automated Text Readability Assessment for German Language: a Quality of Experience Approach. In *2019 Eleventh International Conference on Quality of Multimedia*

Experience (QoMEX). IEEE, 1–3.

[51] Jean Nitzke, Silvia Hansen-Schirra, Ann-Kathrin Habig, and Silke Gutermuth. 2022. Translating Subtitles into Easy Language: First Considerations and Empirical Investigations. *Translation, Mediation and Accessibility for Linguistic Minorities* 128 (2022), 127.

[52] Jenny Ortiz Zambrano, Arturo MontejoRáez, Katty Nancy Lino Castillo, Otto Rodrigo Gonzalez Mendoza, and Belkis Chiquinquirá Cañizales Perdomo. 2019. VYTEDU-CW: Difficult words as a barrier in the reading comprehension of university students. In *Advances in Emerging Trends and Technologies: Volume 1*. Springer, 167–176.

[53] Daniel M. Pottmann. 2019. „Leichte Sprache and Einfache Sprache"–German Plain Language and teaching DaF German as a Foreign Language. *Studia Linguistica* 38 (2019), 81–94.

[54] Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 4694–4698.

[55] Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A New Dataset and Efficient Baselines for Document-level Text Simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, Online and in Dominican Republic, 152–161.

[56] Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking Data-driven Automatic Text Simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*. European Language Resources Association, Marseille, France, 41–48.

[57] Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple TICO-19: A Dataset for Joint Translation and Simplification of Covid-19 Texts. In *Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France, June. European Language Resources Association.*

[58] Melanie Siegel and Christian Lieske. 2015. Beitrag der Sprachtechnologie zur Barrierefreiheit: Unterstützung für Leichte Sprache. *Zeitschrift für Translationswissenschaft und Fachkommunikation* 8, 1 (2015), 40–78.

[59] Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German Multi-Level Text Simplification. (2021).

[60] Sanja Štajner, Horacio Saggion, and Simone Paolo Ponzetto. 2019. Improving Lexical Coverage of Text Simplification Systems for Spanish. *Expert Systems with Applications* 118 (2019), 80–91.

[61] Ina Steinmetz and Karin Harbusch. 2020. Enabling Fast and Correct Typing in 'Leichte Sprache'(Easy Language). In *Proceedings of The Fourth Widening Natural Language Processing Workshop*. 64–67.

[62] Vanessa Toborek, Moritz Busch, Malte Boßert, Pascal Welke, and Christian Bauckhage. 2022. A New Aligned Simple German Corpus. *arXiv preprint arXiv:2209.01106* (2022).

[63] Carmen Torrijos and Sonia Oquendo. 2021. !Hola! Soy Clara y mido la claridad de tu texto. *Archiletras científica: revista de investigación de lengua y letras* 6 (2021), 119–134.

[64] UNE. 2018. UNE 153101:2018 EX Lectura fácil. Pautas y Recomendaciones para la Elaboración de Documentos. Madrid: Asociación Española de Normalización.

[65] United Nations. 2006. Convention on the Rights of Persons with Disabilities of the United Nations (CRPD).

[66] Antonio Ventosa Pérez. 2022. *Servicio Web de Lectura Fácil: Calculador de Índices de Lecturabilidad*. Bachelor's Thesis. ETSI_Informatica.

[67] Zarah Weiss and Detmar Meurers. 2022. Assessing Sentence Readability for German Language Learners with broad Linguistic Modeling or Readability Formulas: When do Linguistic Insights make a Difference?. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. 141–153.

[68] Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data can Help. *Transactions of the Association for Computational Linguistics* 3 (2015), 283–297.

[69] Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., Varna, Bulgaria, 813–822.