Do professionally adapted texts follow existing Easy-to-Understand (E2U) language guidelines? A quantitative analysis of two professionally adapted corpora

Andreea Deleanu and **Constantin Orăsan** and **Shenbin Qian** and
Anastasiia Bezobrazova and **Sabine Braun**
Centre for Translation Studies
University of Surrey, UK
{m.deleanu, c.orasan, s.qian, a.bezobrazova, s.braun}@surrey.ac.uk

Abstract

Easy-to-Understand (E2U) language varieties have been recognised by the UN Convention on the Rights of Persons with Disabilities as a means to prevent communicative exclusion of those facing cognitive barriers and guarantee the fundamental right to Accessible Communication. However, guidance on what it is that makes language 'easier to understand' is still fragmented and vague, leading practitioners to rely on their individual expertise. For this reason, this article presents a quantitative corpus analysis to further understand which features of E2U language can more effectively improve verbal comprehension according to professional practice. This is achieved by analysing two parallel corpora of standard and professionally adapted E2U articles to identify adaptation practices implemented according to, in spite of or in addition to official E2U guidelines analysed by the research team (Deleanu et al., 2024). The results stemming from the corpus analysis, provide insight into the most effective adaptation strategies that can reduce complexity in verbal discourse. This article will present the methods and results of the corpus analysis.

1 Introduction

Accessibility has recently been defined in the European Standard EN 17161 (2019) as the "extent to which products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use". Contexts of use include interaction between people and Accessible Communication, as advocated by the UNCRPD (2006), has therefore called for alternatives to be supplied when users cannot (completely) access information in its original form (Greco, 2016). To date, efforts in Accessible Communication have focused on tackling barriers in written verbal communication (Maaß, 2020; Perego, 2020) and have offered 'Easy-to-Understand language varieties' as a means to overcome verbal communication barriers for a plethora of users (UNCRPD, 2006).

Easy-to-understand (E2U) is an umbrella term that encompasses a wide range of "functional language varieties of different national languages with reduced linguistic complexity, which aim to improve comprehensibility" (Hansen-Schirra and Maaß, 2020). These language varieties thus differ from standard language as they are user-oriented and their main function is to help understand and use information provided (Hansen-Schirra and Maaß, 2020), regardless of individual (dis)abilities or cultural and expert knowledge. This is achieved by adapting content to match users' abilities guarantee its function is fulfilled. E2U varieties enhance written comprehension for a wide range of users, including functional illiterates, vulnerable age groups (Maaß, 2020) and people with diverse cognitive abilities[1]. *Plain Language* and *Easy Language* are the most widely used and known E2U language varieties (Perego, 2020). They deviate from standard language and decrease in complexity, as shown in Figure 1.

Plain Language and *Easy Language* are two distinct language varieties that rely, to different extents, on verbal and non-verbal strategies to make language more accessible and meaning easier to retrieve and perceive (Perego, 2020), thus matching content to end users' abilities. Although there are currently several official guidelines for both

[1]'People with diverse cognitive abilities' and 'cognitively diverse individuals' are used as umbrella terms to identify individuals with temporarily reduced cognitive abilities (due to fatigue, inattention, a learning difficulty, age and/or injury-related cognitive decline) and individuals with permanent impairments. These include, but are not limited to, the conditions identified by the American Psychiatric Association as 'mental disorders' (American Psychiatric Association, 2013)

Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL),
pages 73–93, June 23, 2025

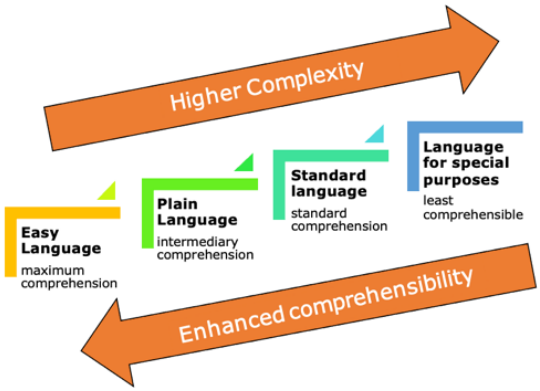

Figure 1: Natural language varieties.

Plain Language and *Easy Language* to be used in context-specific written communication, several issues arise, undermining the success of these two language varieties.

Firstly, the UNCRPD (2006) does not (yet) provide practical guidance on E2U principles to be followed nor specifies which conditions end-users have, leaving signatories to develop guidelines and best practices at company, national[2] or transnational[3] level. Secondly, reception studies with end users in the field of Accessible Communication are scarce and often rely on individual endeavours. This contributes to the absence of an official E2U taxonomy and a growing pool of vague, context-specific or unreliable guidelines created by academia and the public and private sectors. This in turn results in the proliferation of official and non-official guidelines based on intuition or individual preference, leaving professional and amateur content-creators to navigate through a multitude of recommendations, often in contrast with one other, as shown in our guidelines analysis (see Section 2). Thirdly, official guidance regarding the application of *Plain Language* and *Easy Language* principles in spoken interactions, audiovisual and multimodal formats is yet to be established (Maaß and Hernández Garrido, 2020; Maaß, 2020; Perego, 2020) with a few exceptions[4], further excluding people with diverse cognitive abilities from a truly accessible communicative environment.

This research is conducted within the framework of a project in Media Accessibility, with a focus on overcoming cognitive barriers in audiovisual formats. The final goal was to identify best practice and recommendations applicable to audiovisual content, and more specifically, to the adaptation of film narratives for cognitively diverse audiences. This has resulted in the creation of an audiovisual mode called 'Accessible Cues'.

To achieve this, we carried out a review and classified existing official E2U guidelines to identify shared recommendations, discrepancies and grey areas (Deleanu et al., 2024). In this paper, we focus on analysing E2U practice to identify to what extent guidelines are applied in professionally adapted texts. This has been pursued by analysing two professionally adapted parallel standard vs. E2U language corpora, the FIRST corpus (Orasan, Evans and Mitkov, 2018) and the Guardian Weekly corpus (Onestopenglish, 2007).

Our contributions can be summarised as follows:

(1) we conduct a comprehensive quantitative analysis of two professionally adapted English text corpora to identify strategies covered by existing guidelines. The analysis was also conducted to explore how professionals have tackled elements which have been found to be grey areas and discrepant in official E2U guidelines and whether any other strategies not mentioned by the guidelines have been consistently used.

(2) we provide an alternative methodology to analyse standard and adapted corpora, beyond the use of readability indices.

Related work will be reviewed in Section 2, with a focus on readability measures and an overview on the framework used for the guidelines analysis we conducted. This will be followed by Section 3 on the corpus analysis which will focus on presenting the corpora and methodology used. Section 4 will cover the corpus analysis results and discussion. Section 5 will provide conclusions and an overview on future work. Section 6 will conclude with a brief discussion on limitations. Section 7 provides the references while Section 8 provides the links to the resources used for the corpus analysis.

2 Related Work

2.1 Assessing complexity: readability indices

The expected level of difficulty of a text or the appropriate grade level score can be captured by

[2]See UNE 153101:2018 EX, Accessibility Standard on Easy Language (here called easy to read)

[3]See Lindholm & Vanhatalo (2021) for a discussion on the application of E2U language varieties across the EU

[4]See the EU project SELSI (Spoken Easy Language for Social Inclusion) on spoken Easy Language and the EU project EASIT (Easy Access for Social Inclusion Training) on training materials for the adaptation of existing audiovisual access services.

readability[5] indices. Metrics such as Gunning-Fox Index, Flesch-Kincaid Grade level, Flesch Reading Ease scale, Simple Measure of Gobbledygook (SMOG) and Coh-Metrix have been traditionally used to assess the complexity of standard texts and Easy-to-Understand (E2U) texts (Daghio et al., 2006; Pothier et al., 2008; Crossley et al., 2008; Yaneva, 2015; Štajner, 2021; Arfé et al., 2018). In general, readability indices rely on statistical averages and analyse sentence length to determine syntactic complexity, as well as word length, number of syllables, and word frequency to determine semantic difficulty. Their use to assess verbal complexity has, however, often been criticized. For example, the presence of high-frequency words may boost readability but could result in a higher number of polysemic words, while shorter sentences could result in grammatical errors or alteration of meaning, thus increasing complexity (Crossley et al., 2007; Allen, 2009; Fajardo et al., 2014; Saggion, 2018). Moreover, while some official E2U guidelines are in favour of the use of readability indices (Inclusion Europe, 2010), (PLAIN, 2011), others (McGee, 2010) warn against their use, as reading grade levels can differ significantly depending on the formula chosen, proving unreliable.

The corpora investigated in this research have been manually adapted according to professional expertise rather than according to a structural approach based on readability testing and age of acquisition wordlists (Allen, 2009). For this reason, it was deemed more effective to explore a different approach to establish the readability of and identify the strategies adopted in the adapted *FIRST* and *Guardian Weekly* corpora.

2.2 Guidelines Analysis

A set of 10 *Plain Language* and *Easy Language* guidelines have been analysed, classified and compared to identify shared recommendations, discrepancies and grey areas in official E2U guidelines developed for Anglophone countries by organisations such as the *International Federation of Library Associations and Institutions, Inclusion Europe*, the *Plain Language Action and Information Network* and Australian and British disability service providers such as *Scope* and *Mencap*. We have presented a comprehensive analysis of the guidelines in (Deleanu et al., 2024) and have relied on the guidelines classification framework and analysis results to establish the methodology to be used in the corpus analysis for this paper. The categories identified in the guidelines analysis encompassing lexical, syntactic, and adaptation strategies have been used to explore the behaviour of the adapted texts in the *FIRST* corpus (Orasan, Evans and Mitkov, 2018) and the *Guardian Weekly* corpus (Onestopenglish, 2007).

3 Corpus analysis

To gauge the extent to which the above-mentioned guidelines are followed in practice, this research has opted for a corpus analysis to identify expected and unexpected language-dependent phenomena that characterise professionally adapted texts in the Easy-to-Understand (E2U) language varieties.

The FIRST corpus, the code used for the analysis and the corresponding generated data developed as part of this project are available upon request. Please contact the 1st or 2nd author for more information.

3.1 Corpora

Because there are no substantial standard vs. *Plain* or *Easy Language* parallel corpora available – nor audiovisual corpora for that matter – the analysis has focused on data sets that contain a type of adapted language closely related to E2U. The data set includes two plain text corpora, namely the parallel corpus developed for the *A Flexible Interactive Reading Support Tool* (FIRST) project (Orasan, Evans and Mitkov, 2018) and the *Guardian Weekly* (GW) parallel corpus (Onestopenglish, 2007). Neither of the adapted texts in these corpora were *explicitly* created following the official E2U guidelines analysed in previous work (Deleanu et al., 2024), with FIRST and GW content-creators relying on their experience and in-house standards. Nevertheless, the list of adaptation recommendations used in the FIRST project can be found in Table A in Appendix.

The FIRST project[6] addresses the needs of people with Autism Spectrum Disorder, who have been identified as end users of *Easy Language* (IFLA, 2010). The corpus developed in the FIRST project comprises a total of 62 texts, divided into 31 orig-

[5]Readability relates to language-dependent variables that determine text complexity. It represents the degree to which printed information is unambiguous based on the reader's language fluency, the message communicated, and the quantity and the quality of text delivered (Perego, 2020).

[6]See the 2011–2014 EU project FIRST (Flexible Interactive Reading Support Tool)

inal texts and their 31 manually adapted counterparts. The texts were manually adapted by five professionals who work with people with autism. The texts were selected based on the user requirements and include extracts from novels, book and film plot summaries and reviews, scientific articles, news items and leaflets in plain text.

The GW corpus is made up of 300 adapted texts, which are equally divided into three different levels of ascending language proficiency: elementary, intermediate and advanced. These are the adapted versions of 100 original articles from *The Guardian* newspaper, selected and adapted by four experts to provide relevant online material for English learners (Allen, 2009). As the original articles are no longer available and the advanced texts present minor changes compared to their original counterparts (Allen, 2009), the advanced texts have been used as the standard against which to compare the elementary texts. Intermediate texts have not been considered in this analysis in order to mirror the structure of the FIRST corpus, i.e., have only one standard and one adapted version of each article.

3.2 Resources used in the analysis

Five secondary resources related to the creation and evaluation of accessible language were used to support the data analysis. These resources were used to identify any recurring patterns or preferences in the adapted versions at lexical and syntactic level. These resources can be accessed following the URLs provided in section 8.

(1) The *UK Subtlex word frequency database* built on a corpus of words extracted from BBC broadcasts (van Heuven et al., 2014) was used to assign a word frequency score to each type and token in the FIRST and GW subcorpora as an indicator of their difficulty.

(2) *Concreteness ratings* by Brysbaert et al. (2014) were assigned to types and tokens in the subcorpora to understand to what extent abstract words are removed or replaced by experience-based words, as advised by guidelines (Deleanu et al., 2024).

(3) The *English Vocabulary Profile* (EVP) grading database (University of Cambridge et al., 2011) used by Text inspector (Bax, 2012) was used to grade the lexical proficiency of types and tokens in the subcorpora. EVP uses the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) as its reference scale. We assigned a score (1-6) to the proficiency levels (A1 to C2), and 0 to the EVP's *Unlisted* words, to facilitate the analysis.

(4) Over 200 words to be avoided and their preferred *Plain Language* counterparts in domain-specific communication (PLAIN, 2011b) were also checked in the corpora to explore which adaptation strategies have been used for phrases (e.g., *by means of, in accordance with*), phrasal verbs (e.g., *set up, give up*), collocations (e.g., *interpose no objection, pursuant to*), and technical terms (e.g., *notwithstanding, remuneration*).

(5) As far as linking words are concerned, the list provided by PLAIN, (2011a) was used to evaluate the extent to which they are maintained, added or replaced in adapted texts.

3.3 Methodology

The first step was to clean the corpora of special characters, typos, grammatical errors, duplications and encoding problems which would have interfered with our analysis. The two corpora were analysed using corpus linguistic, computational and statistical methods, in line with previous studies (Crossley et al., 2012). A manual analysis was also performed.

Our analysis covers **lexical** and **syntactic** features and **adaptation** strategies (simplification and easification strategies and narrative choices) used by professionals at type and token level. Narrative choices will not be discussed in this paper, as their analysis was conducted in order to identify best practices to inform the creation of 'Accessible Cues' for audiovisual formats. For this reason, they are beyond the scope of this paper. Table B in Appendix provides an overview of the analysed elements per category.

3.3.1 Automatic processing

In order to carry out the analysis of lexical and syntactic features, we used Stanza (Qi et al., 2020) to tokenise, lemmatise and add part-of-speech (POS) information to all texts in the two corpora. We replaced American spelling with British spelling for the comparison with resources in Section 3.2. Sentences were extracted from the processed texts via Stanza. The length of each sentence and the number of sentences in each text were computed thereafter.

Tokenised lemmas were compared with the words in the *UK Subtlex frequency database, the Concreteness* ratings list, the PLAIN lists of content words (2011b) and linking words (2011a) to

count their occurrences for the analysis of lexical frequency and concreteness rating, lists of words to be avoided and linking words respectively. The count of personal pronouns and negations was based on a list of personal pronouns and negative words.

POS labels were used to identify contractions, tenses, passive voice, and clauses. More specifically, we used string matching for contracted formats such as *'s* and checked their POS labels to detect contractions. Labels such as *VBZ* were used to detect tenses and passive voice with the help of auxiliary verbs such as *will*. Words such as *who, when, which* and their corresponding POS labels were used to find types of clauses and count their occurrences in the corpora.

We calculated the Mean (M) and the Standard Deviation (SD) of each text in the corpora for the convenience of comparing standard *vs* adapted versions. Results have been rounded to the first decimal point.

3.3.2 Manual checking

A manual check was conducted when statistical results per article were below or above the average of the subcorpora, and when results were unexpected. We also conducted a manual check to identify and confirm simplification and easification strategies used. This was done by manually consulting each adapted and parallel article and noting the presence of simplification and easification strategies used for each article.

4 Results and Discussion

Although we have explored all phenomena mentioned in Table B in Appendix as part of our project, due to space restrictions and the scope of this paper, the analysis will focus on the following lexical features: lexical frequency and proficiency, concreteness, personal pronouns, tenses and use of passive voice. The following syntactic features will also be presented: sentence counts and clauses. With regard to adaptation strategies, easification and simplification devices will be discussed. Information about data distribution and extensive examples and definitions for each of the analysed lexical, syntactic and easification and simplification features can be found in Tables C, D and E in the Appendix.

4.1 Lexical features

With regard to lexical frequency, the Mean (M) scores of the standard and adapted FIRST subcorpora suggest that the words used belong mainly to the high-frequency range established by van Heuven et al. (2014), with minimal variation between individual texts as shown by the Standard Deviation (SD) in Table 1. The GW subcorpora behave similarly, with a lack of significant difference between the standard and adapted subcorpora.

	FIRST Standard	**FIRST Adapted**	**GW Standard**	**GW Adapted**
M	5.7	5.8	5.8	5.9
SD	1.4	1.4	1.4	1.3

Table 1: Mean and Standard Deviation of lexical frequency of types in the *FIRST* and *GW* corpora

As the results did not provide evidence of a clear division between standard and adapted texts, we have extracted the words that were not present in the *UK Subtlex database* (van Heuven et al., 2014) for each article across all four subcorpora, to clarify whether the lack of difference lays in the database's nature. Surprisingly, we found no notable differences, as words that are *not part* of the *UK Subtlex*, and can therefore be considered too low-frequency, are *still* present in both subcorpora. This suggests that domain-specific and low frequency words can be *kept* in adapted versions as content-creators expect their audiences to cope with both technical and low-frequency terms especially because high-frequency alternatives could prove ambiguous and unsuited, regardless of the "use familiar, high-frequency words" maxim present in all guidelines analysed (Deleanu et al., 2024). While some domain-specific concepts were introduced and terms, foreign words or low-frequency words were explained[7], others were either kept with no further information[8], removed[9], replaced[10] or all of the above within the same text[11], suggesting that word frequency is not a reliable marker for comprehensibility and that multiple strategies can be used simultaneously.

In order to understand whether adapted texts are actually easier to understand based on the CEFR proficiency level (see Section 3.2, EVP) we have analysed the proficiency level of types in all 4 subcorpora using the *English Vocabulary Profile* (EVP)

[7] See Text 1, GW in Table C in Appendix.
[8] See Text 6, FIRST and 37, GW in Table C.
[9] See Text 47, GW in Table C.
[10] See Text 21, GW in Table C.
[11] See Text 32, GW in Table C.

database. The results are shown in Table 2.

	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Unlisted	23.8%	21.2%	37.8%	31.8%
A1	23.7%	25.0%	14.8%	20.5%
A2	19.6%	20.9%	16.4%	21.2%
B1	23.4%	23.7%	25.9%	28.7%
B2	21.1%	20.2%	24.9%	20.7%
C1	7.0%	6.4%	9.6%	5.3%
C2	5.2%	3.9%	8.4%	3.7%

Table 2: Proficiency level of types in the *FIRST* and *GW* corpora according to the EVP

In terms of the distribution of *Listed* types, words tend to belong to B1 and B2 levels for the GW and A1 and B1 for the FIRST in the standard versions. C1 and C2 are also present. On the contrary, a decrease in complexity can be observed in the adapted versions. The percentage of types steadily increases in the *elementary* (A1-A2) and *intermediate* (B1-B2) levels to the detriment of B2, C1 and C2 types for both the GW and the FIRST corpus, (e.g., C2 types *paradoxes* and *albeit* disappear), as advised by Easy-to-Understand (E2U) guidelines (Deleanu et al., 2024). However, *upper intermediate* levels (B2, C1 and C2) do not *completely* disappear in the adapted versions although their numbers do decrease as they are replaced with higher-frequency and therefore lower proficiency level synonyms[12], or removed because they are considered non-relevant information according to content-creators' expertise[13].

In the case of the FIRST corpus, 23.8% of all standard types were unlisted in the EVP database, compared with 21.2% of all adapted types. While the numbers suggest that lexical variety is lower in the adapted version due to the lower number of types and higher number of tokens compared to the standard counterpart[14], the high incidence of *Unlisted* words represents a limitation of the EVP, as differences between the subcorpora could drastically change if a level was allocated to each word. The results are similar in the GW corpus, with 37.8% of types in the standard and 31.8% of types in the adapted subcorpus being *Unlisted*, and thus potentially problematic.

As *Unlisted* words are mainly lexical rather than grammatical in nature (e.g., words such as *nucleotide, Obama, Oscars, Pakistan, plunder* or *punchy* in the FIRST and *fatality, Felix, Lufthansa, Havana, incoming* or *leftist*, in the GW) they can be assumed to belong to *intermediate* and *advanced* levels. These also tend to be removed or explicitated[15], suggesting that when words are perceived as less frequent, and therefore less known, content-creators *have* intervened to contextualise terms, in line with guidelines recommendations.

The extent to which the expertise-based strategies applied to reduce *intermediate, advanced* and *Unlisted* occurrences improve comprehension for end-users, is however not fully confirmed. It can be argued that removal, explanations and replacement depend on content-creators subjective perception of relevance, which can result in bias, information loss, misinterpretation and increased grammatical intricacy and thus text complexity (Halliday, 2008; To, 2017) as lexical units are removed[16]. As a case in point, low-frequency or high-proficiency level words have been kept in many cases[17], suggesting that high-frequency and low-proficiency level words do not necessarily entail more comprehensible output. Often enough higher-frequency and lower-proficiency level words can be polysemic in nature resulting in some texts preferring the use of the specific term to the phrasal verb[18].

In terms of concreteness, there are again no notable differences between the standard and adapted subcorpora. Concreteness ratings (Brysbaert et al., 2014) for both corpora suggest that abstract and concrete words are consistently used across the board and that any topic can undergo adaptation as suggested by 2 out of 10 guidelines analysed. See Table E in Appendix for the distribution.

In order to dispel the vagueness of the guidelines on pronoun use, referencing patterns have been explored for both object and subject personal pronouns, as shown in Table 3. Token occurrences of personal pronouns have been normalized against the total number of tokens per subcorpora. Pronouns can be a hurdle for autistic readers and therefore guidelines provided for the adaptation of

[12]For example, huge (A2) for mammoth (C2) in Text 21, GW and argued (B1) for quarrelled (B2) in Text 5, FIRST. See Table C.

[13]See Text 13, FIRST and 21, GW in Table C.

[14]See Table D in Appendix for type and token distribution.

[15]For example, *destitute children* becomes *poor orphans and street children* in Text 76, GW.

[16]See Text 28, FIRST in Table C for an overview of misinterpretations and mistakes due to adaptation.

[17]See Text 8 FIRST in Table C.

[18]*To take on a case* becomes to *defend a case* in text 28, GW in Table C.

the FIRST corpus suggested their resolution (Jordanova et al., 2014). However, adapted FIRST texts seem to rely more on personal pronouns than their standard counterparts[19], highlighting the inconsistencies between guidelines, expertise-based practice and the needs of end-users (Tavares et al., 2015; Hawthorne and Loveall, 2021). Similar results have been found in the GW adapted subcorpora, suggesting that, contrary to some guidelines, the replacement of pronouns with proper nouns is not consistently carried out as an adaptation technique, with creative alternatives also being preferred[20].

	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Pers. Pron. %	3%	4%	4%	5%

Table 3: Personal pronoun percentage against total tokens in the FIRST and GW corpora

While the number of verbs has increased in the adapted FIRST subcorpus, it has significantly decreased in the adapted GW, as shown in Table 4. Percentages have been obtained by calculating the number of analysed tokens against the total number of tokens identifying verbs for each subcorpora.

	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Total verbs	1410	1605	9822	7981
Simple present	35.9%	41.0%	40.7%	48.2%
Simple Past	28.5%	33.0%	29.2%	33.1%
Simple Future	1.8%	2.1%	2.1%	3.5%
Others	33.8%	22.9%	28.0%	15.3%
Passive	11.5%	10.8%	8.6%	4.0%

Table 4: Distribution of tenses and passives in the *FIRST* and *GW* corpora

The different number of verbs in the adapted subcorpora could be due to different adaptation strategies being used: removal of information, and therefore sentences, in the GW subcorpus as also suggested by sentence counts (see Table 5 in the next Section); and explicitation of nominalised or hidden verbs PLAIN, 2011a and increase in the number of simple sentences in the FIRST subcorpus (see Table 6 in the next Section).

[19] See Text 19, FIRST in Table C.
[20] See text 94, GW in Table C.

Simple tenses are used in abundance in the adapted subcorpora, to the detriment of compound tenses such as auxiliaries, perfects, progressive forms or past participle ('others'), as shown in Table 4. However, 'others' do not disappear, suggesting that *consecutio temporum* is maintained regardless of their numbers being significantly reduced in the adapted versions as advised by guidelines (Deleanu et al., 2024). Also contrary to the guidelines, the simple past is vastly represented in the adapted subcorpora. The same is applicable to the simple future, thus contradicting the ban on future tenses and use of uncertain future[21]. These percentages suggesting that practitioners believe target users to be able to cope with and infer temporal information beyond the simple present, allowing for the production of more natural language in adapted texts[22].

While all guidelines suggest avoiding passives, passive voices have still been kept[23] or introduced[24] in the adapted subcorpora, albeit to a lesser extent (see percentages in Table 4). Passive voices have been significantly reduced in the adapted texts, and especially the GW, with 1 passive out of 2 replaced by an active form[25] or being removed altogether. The presence of passives in the adapted subcorpora could however be justified by a series of reasons, such as the text-type (i.e., articles); the need to improve literacy by gradually introducing passive voices and the underlying pragmatic implications of the original author's intention. Additional reasons are the use of passive to mark order of importance in the sentence and the impossibility of transforming the agent in the performer of the action[26] (Shintani, 1979). These results, once more, highlight how suggestions by official guidelines are ignored in favor of more natural language being produced.

4.2 Syntactic features

In terms of the number of sentences, this increases in the adapted FIRST subcorpus, while it decreases in the adapted GW subcorpus. The figures are presented in Table 5. This could be due to different adaptation strategies being adopted: the GW

[21] Constructed with *might happen* or *should do* ((ILSMH European Association, 1998), PLAIN, 2011a.
[22] See the use of preset and past simple, conditional and present perfect in text 30, FIRST in Table C.
[23] See *has been accused of* in Text 28, GW in Table C.
[24] See Text 5, FIRST in Table C.
[25] See Text 5, FIRST in Table C.
[26] See *to be born* in Text 34, GW in Table C.

content-creators mainly resorted to elimination as preferred E2U strategy while the FIRST project participants have relied on bullet point, extensive text re-organization and explanations to make content more accessible.

	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Sentence count	584	1004	4010	3904

Table 5: Sentence counts in the *FIRST* and *GW* corpora

The total number of verbless clauses, single sentences, coordinate clauses and subordinate clauses has been compared against the total number of clauses in the text. Percentages can be found in Table 6. These numbers were estimated using the part of speech information.

Clause type	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Verbless	3.4%	2.3%	1.5%	1.3%
Simple	27.6%	45.4%	29.3%	31.7%
Coordinate	61.0%	43.9%	69.0%	66.6%
Subordinate	39.7%	27.5%	40.0%	36.3%

Table 6: Distribution of clauses in the *FIRST* and *GW* corpora

Several verbless clauses have been identified in the adapted subcorpora. These are titles, creative devices to maintain engagement[27] or ellipsis of the verb. These elements are surprising, as, intuitively, they could lead to more misunderstandings.

Simple sentences, i.e., independent clauses with one main verb, represent the majority in the adapted versions, in line with guidelines (Deleanu et al., 2024).

Coordinate conjunctions include both syndetic (*or, and, but* and *so*) and asyndetic (commas and semicolon) coordination used in independent clauses. These are largely preferred to subordinate (dependant) clauses, which have been transformed in either coordinates or simple sentences[28] in the adapted texts, as advised by official guidelines.

Interestingly enough, subordinates have not disappeared[29], suggesting that their use is essential for the cohesion and coherence of the overall text as suggested by PLAIN (2011a) and its proposed list of linking words to be used.

[27]For example: *Tense? Angry? Can't get online?* in Text 90, GW and *Rain is our national weather. Snow can cause us problems, yes, and very hot weather, like last summer, causes difficulties, too. But rain?* in Text 99, GW.
[28]See clauses **in bold** in Text 19, FIRST in Table C.
[29]See underlined clauses in Text 19, FIRST in Table C.

4.3 E2U adaptation strategies: easification and simplification

Easification makes texts more accessible by developing in the reader specific learning strategies (Bhatia, 1983). This includes guiding readers, raising awareness of potential ambiguities and difficulties (van den Bos et al., 2007) and restructuring, reorganising or rearranging information in the text at verbal and visual level (Caro, 2020). Simplification is the process of transforming a text into a more understandable equivalent (Saggion et al., 2011) by reducing linguistic complexity (WCAG 2.1, 2019). Easification and simplification strategies are used to different degrees in each adapted subcorpus, with the number of types and tokens and linking words across levels partially indicating whether any elimination, reiteration, exemplification or explanation strategies have been used. Nevertheless, not all strategies have been applied, especially those belonging to easification, as confirmed by our manual checks. Table 7 presents an overview of the E2U adaptation strategies identified in the subcorpora. Ticks indicate strategies that have been used while crosses indicate those that have not. Dashes indicate that the strategy has only been partially applied.

	FIRST Adapted	GW Adapted
Summary	X	X
Introduction	X	✓
Glossary	X	✓
Elimination	✓	✓
Reiteration	–	–
Exemplification	X	✓
Explanation	✓	✓
Context Clue	✓	✓
Definition	X	✓
Paraphrase	✓	✓
Inference	✓	✓

Table 7: Overview of easification and simplification strategies used in the *FIRST* and *GW* corpora

In terms of adaptation strategies, summaries have not been used in the subcorpora, while introductions[30] have been rarely used in the adapted GW. Glossaries are hardly used in the standard texts[31] but existing ones have been, alongside an existing footnote[32] partially adapted and kept at the bottom of the text. No glossaries have been created specifically for adapted versions.

[30]See Text 82, GW in Table C.
[31]See Text 12, GW in Table C.
[32]See Text 85, GW in Table C.

As discussed previously, simplification strategies shared by adapted subcorpora are primarily **elimination**[33], **explanation** (*meaning, meaning that*, definitions, context clues and paraphrase)[34] and **spelling out of implications**[35]. However, practice has not always been consistent between the GW and FIRST adapted subcorpora, with **exemplification**[36] being used in the former rather than the latter. **Reiteration** strategies in the form of repetitions, have not been found in the GW or FIRST corpus. However, reiteration has encompassed a consistent use of lexicon and reiteration of syntactical structures[37].

These results do not mean that all strategies should be simultaneously used in the same text but only when required. Nevertheless, there is a risk of corrupting meaning as personal interpretation can always interfere, as in the following text in Table 8.

In the example in Table 8, 'shells' are a means to *predict* the dissolution of the implant in the original version, rather than a means to *control* it, as suggested in the adapted FIRST subcorpus.

Standard	**Adapted**
Getting the electronics to fade away in a controlled manner relies on two scientific developments – getting the electronics to dissolve at all and **using a shell to control when that happens**.	Electronics melt away in a controlled manner. It relies on two scientific developments. One is to get the electronics to dissolve. **The other is to use a shell to control what happens.**

Table 8: Distortion of meaning in Text 18, FIRST corpus

4.4 Discussion

There are several strategies used by content-creators which have been banned by guidelines. For example, the analysed guidelines have rejected the use of negations, passives and contractions. On the other hand, the analysed adapted subcorpora have instead preserved or added them. This strategy was used by content-creators to maintain or explicitate the meaning intended by the original text and to avoid creating non-grammatical and disconnected sentences, i.e., more complex sentences. However, these opposite strategies provide food for thought. A major problem in Media Accessibility, i.e., the field in which this research was conducted, are time constraints: subtitles for the Deaf and Hard of Hearing and Audio Description are part of the post-production process and therefore depend on the pace and the pauses in the original soundtrack. Resorting to the Saxon Genitive, verb contractions, negations, abbreviations (when previously explicitated, familiar and meaningful in the given context) and pronouns or glosses (when non-ambiguous and reiterated), for instance, could potentially help overcome the media limitation or allow for longer processing time in audiovisual formats, such as films.

Nevertheless, shared patterns have also been identified, such as the elimination of words banned by PLAIN (2011b) or the preference for simple sentences, coordinating conjunctions and elimination and explanation strategies in the adapted subcorpora. Elimination has been the most consistently applied simplification strategy in the analysed adapted subcorpora. The results have shown that higher proficiency terms belonging to *intermediate* and *advanced* levels, alongside *Unlisted* words which can be considered too low-frequency to be graded by the English Vocabulary Profile (EVP) database (Capel, 2010, 2012), have been mostly eliminated during adaptation. It can, however, be argued that adaptation should not be solely guided by a reductive approach, as it is not a matter of subjectively choosing between relevant or irrelevant eliminable information but a matter of identifying which relevant elements can be easily inferred from the available information or the visual aids provided.

As no database of prevalent vocabulary possessed by cognitively diverse individuals has been collated by psycholinguistics (Jordanova et al., 2014), the corpus analysis has relied on the EVP and the *UK Subtlex* (van Heuven et al., 2014) databases to determine which words fall under the category of 'difficult' or 'low-frequency' words to be avoided, as prescribed by the analysed guidelines (Deleanu et al., 2024). In the adapted subcorpora, content-creators have relied on different and often contradictory strategies to address technical or B1 to C2 terms, due to individual expertise-based practice and a lack of a proofreading and validation phase to confirm and unify adaptation strategies within a given text. If both phases had

[33] See Text 33, GW in Table C.
[34] See Text 2, 5, 9 and 28 FIRST in Table C.
[35] See Text 45, GW in Table C.
[36] See Text 3, FIRST in Table C.
[37] See Text 3, FIRST in Table C.

been pursued by content-creators of the FIRST and GW corpora, the E2U strategies used and therefore the results of this analysis might have been different.

5 Conclusions

The effectiveness of Easy-to-Understand (E2U) language varieties is still under-researched, and limitations have been highlighted (Fajardo et al., 2014; Hurtado et al., 2014). Yet, findings from reception studies with individuals with diverse cognitive abilities (Fajardo et al., 2014; Yaneva, 2016; Säuberli et al., 2024) have shown that Easy Language *does* partially address language complexity and thus support comprehension. Language, and especially accessible language, could therefore be instrumental to achieving Accessible Communication for all. However, adapting standard texts into E2U is no easy feat. Often enough, content-creators find themselves juggling different alternatives and having to settle for the one they deem most comprehensible to the majority of their end users. For this reason, it can be argued that adaptation depends on individual instances. This entails that the use of different and often conflicting but valid strategies ought to be acceptable as no universal set of guidelines can be drafted. However, giving content-creators a toolbox of options from which to choose could enable them to adapt standard texts more swiftly and accurately. The existence of a toolbox could also increase awareness and reflexion on what it is about language that makes meaning-making a complex process.

Only a few of the official guidelines that we have analysed (Deleanu et al., 2024)) have stressed the preference for the use of spoken language in adaptations, including to the detriment of natural grammar. Our corpus analysis has highlighted the preference of content-creators for natural language that end-users are familiar with and the often-contradictory presence of elements that guidelines have deemed unapproachable. While improving literacy *does* play an important role in the corpora we have analysed, it could be argued that only end-users can have the final say on *how much* an adapted text has been made accessible. Validation with end-users could help overcome biased interpretations and provide an indication of how much background information is necessary. However, this can prove time consuming and expensive, which is why a toolbox could be the first step towards the production of a higher number of E2U texts. This could include the corpus analysis resources and the guidelines and corpus analysis framework, providing users with an overview of shared *and* contrasting practice.

In the future, we intend to develop a toolbox and make the E2U guidelines recommendations developed in this project available in the future. The toolbox could then be used in the process of training AI models and provide an environment in which standard texts could be efficiently and consistently adapted into E2U. The methodology applied could also be used to assess automatically generated simplified outputs obtained using AI tools, to better understand the 'black box' strategies these tools apply and help detect differences between original and the offered multiple adapted versions. As we conducted this research in the context of a project in Media Accessibility, we also intend to address the gap in Accessible Communication by applying best identified E2U strategies to an audiovisual format.

6 Limitations

We acknowledge that our analysis framework, developed through a qualitative guidelines analysis is, to some extent, subjective and tailored to a project in Media Accessibility. The corpus analysis was conducted using a limited sample of texts (131 standard texts and 131 adapted texts) and only two corpora as our focus was on *professionally* adapted parallel standard vs. E2U texts and few alternatives were available. Although analysed corpora do not specifically fall under the category of either *Plain* or *Easy Language*, the adapted output does represent an E2U language variety meant to reduce verbal complexity for language learners and autistic readers, i.e., primary audiences of *Plain* and *Easy Language* respectively.

References

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of english. *System*, 37:585–599.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edition. American Psychiatric Publishing, Arlington, VA.

Barbara Arfé, Lucia Mason, and Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31:2191–2210.

S Bax. 2012. Text inspector. *Online text analysis tool*. Retrieved from https://languageresearch.cambridge.org/images/pdf/theenglishprofilebooklet.pdf.

Vijay K Bhatia. 1983. Simplification vs. easification — the case of legal texts. *Applied linguistics*, 4(1):42–54.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46:904–911.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1. Retrieved from http://www.englishprofile.org/wordlists.

Annette Capel. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

Rocío Bernabé Caro. 2020. New taxonomy of easy-to-understand access services1. *Monografias de Traduccion e Interpretacion (MonTI)*, 12:345–380.

Council of Europe. 2001. Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Retrieved from https://www.coe.int/en/web/common-european-framework-reference-languages.

Scott A. Crossley, David Allen, and Danielle S. McNamara. 2012. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, 16(1):89–108.

Scott A Crossley, David F Dufty, Philip M McCarthy, and Danielle S McNamara. 2007. Toward a new readability: A mixed model approach. In *Proceedings of the annual meeting of the cognitive science society*, volume 29, pages 197–202.

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.

M. Monica Daghio, Giuseppe Fattori, and Anna V. Ciardullo. 2006. Assessment of readability and learning of easy-to-read educational health materials designed and written with the help of citizens by means of two non-alternative methods. *Advances in Health Sciences Education: Theory and Practice*, pages 123–132.

Maria Andreea Deleanu, Constantin Orăsan, and Sabine Braun. 2024. Accessible Communication: a systematic review and comparative analysis of official English Easy-to-Understand (E2U) language guidelines. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*, pages 70–92. ELRA and ICCL.

European Standard EN 17161. 2019. Design for all. https://universaldesign.ie/about-universal-design/products-and-services/standard-i-s-en-171612019-design-for-all.

Inmaculada Fajardo, Vicenta Ávila, Antonio Ferrer, Gema Tavares, Marcos Gómez, and Ana Hernández. 2014. Easy-to-read texts for students with intellectual disability: linguistic factors affecting comprehension. *Journal of applied research in intellectual disabilities*, 27:212–225.

Gian Maria Greco. 2016. On accessibility as a human right, with an application to media accessibility. *Researching audio description: New approaches*, pages 11–33.

Michael A. K. Halliday. 2008. *Complementarities in language*. Beijing: The Commercial Press.

Silvia Hansen-Schirra and Christiane Maaß. 2020. Easy language, plain language, easy language plus: Perspectives on comprehensibility and stigmatisation. In Christiane Maaß, editor, *Easy Language Research: Text and User Perspectives*, pages 17 – 38. Frank & Timme, Berlin.

Kara Hawthorne and Susan J Loveall. 2021. Interpretation of ambiguous pronouns in adults with intellectual disabilities. *Journal of Intellectual Disability Research*, 65(2):125–132.

Barbara Hurtado, Lara Jones, and Francesca Burniston. 2014. Is easy read information really easier to read? *Journal of Intellectual Disability Research*, 58(9):822–829.

IFLA. 2010. Guidelines for easy-to-read materials. Available at: https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/professional-report/120.pdf.

ILSMH European Association. 1998. Make it simple: European guidelines for the production of easy-to-read information for people with learning disability. Available at: https://docplayer.net/142050357-Ilsmh-europeanassociation-make-it-simple-european-guidelinesfor-the-production-of-easy-to-read-information-for-people-with-learning-disability.html.

Inclusion Europe. 2010. Information for all: European standards for making information easy to read and understand. https://www.inclusion-europe.eu/wp-content/uploads/2017/06/EN_Information_for_all.pdf.

Vesna Jordanova, Richard Evans, and A Cerga Pashoja. 2014. D7.2: Benchmark report (results of piloting task).

Christiane Maaß. 2020. *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability*, volume 3. Frank & Timme, Berlin.

Christiane Maaß and Sergio Hernández Garrido. 2020. Easy and plain language in audiovisual translation. In Christiane Maaß, editor, *Easy Language Research: Text and User Perspectives*, pages 131–161. Frank & Timme, Berlin.

Jeanne McGee. 2010. *Toolkit for making written material clear and effective*. Centers for Medicare and Medicaid Services, Department of Health and Human.

Elisa Perego. 2020. *Accessible Communication: A Cross-country Journey*, volume 4. Frank & Timme, Berlin.

Louise Pothier, Rachael Day, Catherine Harris, and David D Pothier. 2008. Readability statistics of patient information leaflets in a speech and language therapy department. *International journal of language & communication disorders*, 43(6):712–722.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Horacio Saggion. 2018. Text simplification. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, 2 edition. Oxford University Press.

Horacio Saggion, Elena Gómez-Martínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Michiko Shintani. 1979. *The frequency and usage of the English passive*. University of California, Los Angeles.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Gema Tavares, Inmaculada Fajardo, Vicenta Avila, Ladislao Salmerón, and Antonio Ferrer. 2015. Who do you refer to? how young students with mild intellectual disability confront anaphoric ambiguities in texts and sentences. *Research in developmental disabilities*, 38:108–124.

Vinh To. 2017. Grammatical intricacy in efl textbooks. *International Journal of English Language Education*, 5:127–140.

UNCRPD. 2006. United Nations, convention on the rights of persons with disabilities. https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html. Retrieved from the United Nations website. Accessed May 24, 2025.

University of Cambridge, British Council, University of Bedfordshire, and English UK. 2011. English Profile: Introducing the CEFR for English. Retrieved from https://languageresearch.cambridge.org/images/pdf/theenglishprofilebooklet.pdf.

KP van den Bos, H Nakken, PG Nicolay, and EJ Van Houten. 2007. Adults with mild intellectual disabilities: Can their reading comprehension ability be improved? *Journal of Intellectual Disability Research*, 51(11):835–849.

Walter JB van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, 67(6).

WCAG 2.1. 2019. The web content accessibility guidelines. Retrieved from https://www.w3.org/WAI/WCAG21/quickref/.

Victoria Yaneva. 2015. Easy-read documents as a gold standard for evaluation of text simplification output. In *Proceedings of the Student Research Workshop*, pages 30–36.

Victoria Yaneva. 2016. *Assessing text and web accessibility for people with autism spectrum disorder*. Ph.D. thesis, University of Wolverhampton.

7 Language Resource References

7.1 Corpora used in the analysis

Onestopenglish. (2007). *News lessons*. *Macmillan English Campus*. Retrieved from http://www.onestopenglish.com

Orasan, C., Evans, R.J., & Mitkov, R. (2018). Intelligent Text Processing to Help Readers with Autism. *Intelligent Natural Language Processing: Trends and Applications*, Springer.

7.2 Resources used in the analysis

Bax, S. (2012). *Text Inspector.* Online text analysis tool. Retrieved from https://textinspector.com

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meet-ing of the Association for Computational Linguistics:System Demonstrations*, pages 101–108. Association for Computational Linguistics.

PLAIN. 2011a. Federal Plain Language Guidelines. Available at https://www.plainlanguage.gov/guidelines/

PLAIN. 2011b. *Plain Language: Use Simple Words and Phrases.* Available at https://www.plainlanguage.gov/guidelines/words/use-simple-words-phrases/

Walter J. B. van Heuven, Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6).

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46:904–911. https://doi.org/10.3758/s13428-013-0403-5

Table A: FIRST Recommendations	
1.	Detect infrequent words and substitute them with simpler synonyms, definitions or explanations.
2.	Identify figurative expressions such as idioms and metaphors and replace with simpler words or definitions. Provide inferred meaning for non-lexicalized metaphors.
3.	Identify jargon or specialised terms and replace with specific definitions.
4.	Identify phraseological units and polysemic words and replace with specific definitions. Highlight the domain. Avoid using the easier synonym.
5.	Identify and divide long paragraphs.
6.	Detect long sentences and divide them into shorter easier to understand chunks.
7.	Rewrite complicated sentences to make them easier to understand.
8.	Identify and resolve anaphora.
9.	Replace infrequent abbreviations and acronyms with full version.
10.	Use bullet points if necessary to break down the text in easier parts.
11.	Break sentences into segments no longer than 15 words. Prefer simple sentences.
12.	Avoid semicolon, suspension points and special characters.
13.	Substitute infrequent slang with simpler synonym or provide simple definitions to explain slang.
14.	Disambiguate temporal adjectives.

Table B: Analysed elements per each category	
Lexical Features	• Frequency • Concreteness ratings • Proficiency level • Types and tokens distribution • Object & subject personal pronouns • Contractions (Saxon Genitive and Verbs) • Lists of words to be avoided (PLAIN, 2011b) • Compound words • Tenses • Passive voice • Negations
Syntactic Features	• Sentence length • Sentence counts • Clauses counts (verbless clauses with absent or nominalised verbs, single sentences, subordinates, coordinates) • Linking words
Adaptation Strategies	• Elimination • Reiteration • Exemplification • Explanations (context clues, definitions and glossaries) • Summaries • Introductions • Explicitation of inferences

Table C: Examples extracted from the corpora		
	Standard	**Adapted**
TEXT 1, GW (Introducing the topic and explaining low-frequency technical terms)	John Prescott, has been praised by medical experts for his 'brave' admission that he struggled with the **eating disorder bulimia** for **two decades**".	**Anorexia** and **bulimia** are both illnesses where people have problems with **eating problems** known as **eating disorders**. **People who suffer from anorexia** do not eat enough food and they soon become very thin. **People who suffer from bulimia** eat a large amount of food but they usually vomit after eating it. Both these eating disorders can be very dangerous for your health [. . .] John Prescott, has published his autobiography. In the book he says that he suffered from **bulimia** for **twenty years**.
TEXT 6, FIRST (Technical terms kept or adapted)	In **trauma patients** who have undergone heavy **blood loss**, these **molecules** are **in short supply**, and its makers claim MP40X can deliver an **oxygen boost** to organs and tissue in the body reducing the risk of organ failure.	**Patients** who suffered **trauma and lost lots of blood**, have a **shortage** of these **chemicals**. People who made MP40X claim that it can deliver an **oxygen boost** to organs and tissue in the body. In this way the risk of **organ failure** will be reduced.
Text 37, GW (Foreign words and domain-specific terms are *kept* or *wrongly used* to replace other domain-specific ones)	In recent years the US army has been forged into a motivated, effective tool for large-scale military operations overseas. But it has never been suited to combating **insurgency**. **Guerrillas** and **suicide bombers** can impose a deadly corrosion on *conventional forces*.	In recent years the US army has become a very effective army for big military operations in other countries, but it has never been effective against **insurgents**. It is very difficult for *regular armies* to fight against **guerrillas** and **suicide bombers**.
TEXT 47, GW (Removal of foreign words)	Life for the housewife is an endless ***faena***, a round of tasks to ensure the comfort of every (other) member of the family.	Many women, especially older woman, like to serve the rest of the family. They work very hard to make the rest of the family comfortable.
Text 21 GW (Low-frequency words **replaced** by high-frequency yet polysemic ones) (Removal of non-relevant domain-specific information) (Addition of temporal contextualization)	Six years after 9/11, bin Laden is **maddeningly out of reach**. Despite the world's largest manhunt and a $25m **bounty**, he **remains at large**, the Scarlet Pimpernel of jihad. [. . .] The Pakistani army thought it had **cornered** in a village in the lawless North Waziristan tribal agency in 2003. A year later the Spanish newspaper El Mundo claimed to have **located** him inside a Muslim **enclave** of western China. After the **mammoth** earthquake that devastated northern Pakistan, Senator Harry Reid from Nevada **announced** that bin Laden had died under the rubble.	Six years after 9/11, bin Laden is **still free**. The world's largest manhunt and a possible **reward** of $25 million have **not managed to find** him. [. . .] The Pakistani army thought it had **found** him in a village in North Waziristan in 2003. A year later, the Spanish newspaper El Mundo said he **was** in a Muslim **area** of western China. One US senator **said** that bin Laden had died in the **huge** earthquake in Pakistan last year.

Table C: Examples extracted from the corpora (continued)		
	Standard	**Adapted**
Text 32, GW (Simultaneous lexical strategies)	Tourists could also visit some of Britain's **ancient architectural treasures** which, she says, **risk becoming derelict** because of a **lack of funding**. Strawberry Hill, Sir Horace Walpole's **folly in Twickenham**, west London, **which sparked the Gothic revival in the early 19th century**, is **struggling** to raise £8m. One of the oldest **parish churches** in England, St Mary's, in Stow in Lindsey, Lincolnshire, needs £3m for renovations. Another London **landmark**, Battersea power station, becomes more **run-down** every day as government, developers and **local community boards** argue over its future [. . .] West points out that the guidebook's message is not all **gloom**.	Tourists should also visit some of Britain's **ancient architectural treasures** which, she says, are in **danger of falling down** because there is **no money** to save them. Strawberry Hill, Sir Horace Walpole's **building** in west London **needs** £8m. One of the oldest **churches** in England, St Mary's, in Stow in Lindsey, Lincolnshire, needs £3m. Another London **landmark**, Battersea power station, becomes more **run-down** every day as government, property developers and the **local people** argue about its future [. . .] West points out that the guidebook's message is not all **gloom**.
TEXT 13, FIRST (Removal of non-relevant information *including* higher-proficiency words)	To celebrate their first wedding anniversary in April, Jeremy Forrest and his photographer wife Emily spent two **idyllic** weeks in Thailand and Malaysia.	Jeremy Forrest and his wife Emily celebrated their first wedding anniversary on a holiday in Thailand and Malaysia.
Text 28, FIRST (**Bending** of meaning and **mistakes** due to adaptation) (Domain-specific term replaces phrasal verb) (Passive voice kept) (Explanation: paraphrase)	Atticus decides to take on a case involving a black man named Tom Robison who has been accused of raping a very poor white girl [. . .] Because Atticus is defending a black man, **Scout and Jem find themselves whispered at and taunted, and have trouble keeping their tempers**. At a family Christmas gathering, Scout beats up her cloying relative Francis when he accuses Atticus of ruining the family name by being a "nigger-lover". Jem cuts off the tops of an **old neighbour's** flower bushes **after she** derides Atticus, and as punishment, has to read out loud to **her** every day.	Atticus decides to defend a case involving a black man named Tom Robinson. This man has been accused of raping a very poor white girl [. . .] **Scout and Jem gossip about Atticus** defending a black man. **They are very disapproving of him. They become annoyed.** At a family Christmas gathering, Scout beats up her cloying relative Francis. Scout does this because he accuses Atticus of ruining the family name by being sympathetic to black people. Jem cuts off the tops of **her neighbour's** flower bushes. **She** derides Atticus. As punishment, **she** has to read out loud to her every day. **Scout is a female character, Atticus, Jem and Francis are male characters. The neighbour is female.*
Text 8, FIRST (Higher-proficiency words are kept rather than replaced by lower-proficiency ones).	The story starts shortly after the **sudden** death of Barry Fairbrother, a kind-hearted member of the **parish council**.	The story starts shortly after the **sudden** death of Barry Fairbrother, a kind-hearted member of the **parish council**. **Sudden* (EVP: B2) could have been removed or replaced by *fast* (A1) or *quick* (EVP: A2) **Parish* (Unlisted, possible C1/C2) could have been replaced by *community* (EVP: A2) and *council* (EVP: B1) by *group* (EVP: A1)

Table C: Examples extracted from the corpora (continued)		
	Standard	**Adapted**
Text 19, FIRST (Use of anaphoric references) (Dependent and independent clauses **replaced** by main clauses. With the exception of two dependent clauses in the adapted version – a kept infinitive and an added conditional clause).)	Teenager Marty McFly [...] **gets sent back** to 1955 **where** he **runs into** his mother and father and **prevents** their union. **Meeting** up with younger Doc, Marty **hatches** a plan to get his parents together **before** he **gets erased** from existence.	He **sends** teenager Marty McFly back to 1955. He **meets** his mother and father. He **stops** them from getting together. He **meets** Doc as a young man. Marty and Doc **plan** to get his parents together. If they are not together, he **cannot** be born. He would **cease** to exist.
Text 94, GW (Modifying the existing gloss to avoid asides in the adapted version)	In what seems to have been another misjudged remark, **Obama's wife, Michelle**, campaigning for **him** in South Carolina, also brought up race.	On the other side, **Michelle Obama**, campaigning for **her husband** in South Carolina, also mentioned race.
Text 30, FIRST (Kept *consecutio temporum*)	The standardised house design **has led** some to believe that there **was** no hierarchy of rank within the settlement at Skara Brae, and that all villagers **were** equal. Whether or not this **is** true **is debatable**. However, it **is likely** that life here **was probably** quite comfortable for the Neolithic people. The villagers **kept** sheep and cattle and **grew** wheat and barley. They **probably traded** these commodities for pottery. They **would have hunted** red deer and boar for their meat and skins. They **would also have consumed** fish, seal and whale meat, and the eggs of sea birds. The skin and bones of these animals **would have provided** tools such as needles and knives. Flint for cutting tools **would have been traded** or **gathered** from the shore.	The uniform house design **has led** some to believe that there **was** no rank order within the village at Skara Brae. The standardised house design **has led** some to believe that all villagers **were** equal. This **is uncertain**. However, it **is likely** that life here **was probably** quite comfortable for the Neolithic people. The villagers **kept** sheep and cattle. The villagers **grew** wheat and barley. The villagers **probably traded** wheat and barley for pottery. They **would have hunted** red deer and boar for their meat and skins. The villagers **would also have consumed** fish, seal and whale meat, and the eggs of sea birds. The skin and bones of these animals **would have provided** tools such as needles and knives. Flint for cutting tools **would have been traded** or **gathered** from the shore.
Text 5, FIRST (High-proficiency **replaced** by low-proficiency equivalent) (Passive voice added to paraphrase) (Passive voice replaced with active voice)	The children ran wild all over the house; the English governess **quarrelled** with the housekeeper [...] Three days after the **quarrel**, Prince Stepan Arkadyevitch Oblonsky – Stiva, as he was called in the fashionable world – woke up at his usual hour, that is, at eight o'clock in the morning [...]	The children ran around the house and nobody looked after them all day. An English woman who was paid to look after the children **argued** with the woman who looked after the house [...] Three days after the **argument** between the husband and the wife had finished, Prince Stepan Arkadyevitch Oblonsky – Stiva woke up at eight o'clock in the morning. Only people who called him Stepan Arkadyevitch Oblonsky – Stiva were from the upper class [...]
TEXT 34, GW (Passive voice kept)	They know that they are father and daughter, that Ryann **was conceived** thanks to sperm **donated** by Mr Harrison in the 1980s.	They know that Ryann **was born** thanks to sperm **given** by Mr Harrison in the 1980s.

Table C: Examples extracted from the corpora (continued)		
	Standard	**Adapted**
TEXT 82, GW (Introduction added)		Many animals around the world are very rare, gorillas and tigers, for example. There is a danger that in a few years time these animals will disappear from the world forever and we will only see them in photographs. Animals like these are known as endangered species and there are laws which protect endangered species. However, some people are buying and selling rare and endangered animals on the internet.
Text 12, GW (Glossary kept)	Near-Earth objects Comets and asteroids pulled into orbits near the Earth by the gravitational attraction of planets. Most NEOs are made of ice and dust, or are bits of rock from the asteroid belt between Jupiter and Mars. Outside chance Apophis had been tracked since its discovery in June 2004. [. . .] Slight gravitational attraction Everything in the universe that has mass attracts anything else with mass via the force of gravity. If a gravity tractor is placed near an asteroid, the asteroid will move fractionally towards it. [. . .]	Near-Earth objects Comets and asteroids that start to circle very near the Earth. Most NEOs are made of ice and dust, or are bits of rock from the asteroid area between Jupiter and Mars. Outside chance Astronomers discovered Apophis in June 2004. [. . .] Slight gravitational attraction Everything in the universe that has mass attracts anything else with mass because of gravity. If a "gravity tractor" is placed near an asteroid, the asteroid will move very slightly nearer to it. [. . .]
Text 85, GW (Footnote kept)	* Note: In the UK, so-called 'public' schools are not public at all. They are private schools for the children of rich parents.	* Note: In the UK, so-called 'public' schools are not public at all. They are private schools for the children of rich parents.
Text 33, GW (Elimination)	There is a fear that the film will stop people buying all African diamonds, **something both the industry and the campaigners want to avoid**. **"It would be terrible if the film meant that people saw Sierra Leone as a pariah,"** said Sanders. "Quite a few African countries have weak control systems."	Some people are worried that the film will stop people buying all African diamonds. "Quite a few African countries have weak control systems," says Sanders.
Text 2, FIRST (Explanation: definition)	The **twenty-five metre** pool is available for recreational swimming from seven to nine in the morning and twelve thirty to one thirty on weekdays, and ten am to four pm on Saturdays.	**Recreational swimming means not swimming in lanes.** The 25 metre pool is available for **recreational swimming** from 07:00–09:00 and 12:30–13:30 on weekdays. The 25 metre pool is available for recreational swimming from 10:00 – 16:00 on Saturdays.

Table C: Examples extracted from the corpora (continued)		
	Standard	**Adapted**
Text 9, FIRST (Explanation: paraphrase)	About half of this **shrinkage is due to change in distribution and abundance**, the remainder to **changes in physiology**.	Half of the **shrinkage is because of location of the fish and their number**. The other half is because of the **changes in the structure and function of their bodies**.
Text 45, GW (Spelling out implications)	Not since Hamelin has the discovery of a rat provoked so much **alarm**. It was only a single creature, but it had no business being on the island of Santa Fe in the **isolated Galapagos archipelago, where conservationists now strive to keep foreign wildlife at bay as effectively as hundreds of miles of open ocean did for millions of years**.	About 1,000 km west of the coast of Ecuador in the middle of the Pacific Ocean is a group of islands called the Galapagos Islands. Because the Galapagos Islands are so far away from the rest of South America, **the wildlife there is unique and plants and animals found in other parts of the world do not exist on the islands**. There are no rats, for example. But now a rat has been found on the island of Santa Fe and the conservationists who are working to stop foreign wildlife reaching the islands are **very worried**.
Text 3, FIRST (Exemplification)	Students will receive guidance from their **tutors** on how best to conduct research and write it up effectively.	**A tutor is like a teacher who gives private lessons**. Students will be guided by their tutors on how to do research.
Text 3, FIRST (Reiteration of structure)	Our **pre-sessional courses** are ideal for students who have a conditional place at a British university, but who need to achieve a certain level of English in order to be accepted. The **course** aims to provide students with the English language and study skills that they need in order to be successful at university or another academic establishment.	**Pre-sessional courses** are courses to prepare people for university. Our **pre-sessional courses are ideal for students who** have a conditional place at a British university. The **pre-sessional courses are ideal for students who** need to achieve a certain level of English in order to be accepted.

Table D: Type and token distribution

	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Type count	3 089	2 818	10 902	7 026
Token count	10 850	11 171	74 527	61 569

Table E: Concreteness rating distribution

	FIRST Standard	FIRST Adapted	GW Standard	GW Adapted
Mean	3.1	3.1	3.2	3.2
Standard Deviation	1.1	1.1	1.1	1.1