Reading Between the Lines: A dataset and a study on why some texts are tougher than others

Nouran Khallaf, Carlo Eugeni, Serge Sharoff

University of Leeds, UK

N.Khallaf, C.Eugeni, S.Sharoff @leeds.ac.uk

Abstract

Our research aims at better understanding what makes a text difficult to read for specific audiences with intellectual disabilities, more specifically, people who have limitations in cognitive functioning, such as reading and understanding skills, an IQ below 70, and challenges in conceptual domains. We introduce a scheme for the annotation of difficulties which is based on empirical research in psychology as well as on research in translation studies. The paper describes the annotated dataset, primarily derived from the parallel texts (standard English and Easy to Read English translations) made available online. we fine-tuned four different pre-trained transformer models to perform the task of multiclass classification to predict the strategies required for simplification. We also investigate the possibility to interpret the decisions of this language model when it is aimed at predicting the difficulty of sentences. The resources are available from https://github.com/Nouran-Khallaf/why-tough

1 Introduction

The Universal Declaration of Human Rights, in its Article 19, affirms everyone's right to seek and receive information. Similarly, Article 21 of the UN Convention on the Rights of Persons with Disabilities underscores the need for accessible formats, ensuring that individuals with disabilities can access public information without additional cost. For people with intellectual disabilities—those with limitations in cognitive functioning, including difficulties in reading and understanding, an IQ below 70, and challenges in conceptual domains (American Association on Intellectual and Developmental Disabilities (AAIDD), n.d.)—language simplification is crucial for ensuring accessibility and equality, making it essential for them to fully enjoy their human rights.

Text Simplification (TS) research aims to make text easier to read while preserving its meaning and key information (Saggion, 2017). Earlier studies involved lexical, syntactic and semantic modifications, while modern research benefits from the use of Large Language Models (LLMs), with still unclear cost-to-performance benefits, as they do not outperform smaller Pre-trained Language Models (PLMs), such as BERT, on text classification tasks (Edwards and Camacho-Collados, 2024).

Computational studies often overlook insights from translation studies, particularly the various strategies proposed (Vinay and Darbelnet, 1971; Newmark, 1988; Chesterman, 1997; Zabalbeascoa, 2000; Molina and Hurtado Albir, 2002; Gambier, 2006), focusing on the systematic processes involved in translating a source text into a target text across languages. Translation studies provide a complementary lens by examining strategies used in intralingual translation, where a source text is converted into a target text in the same language. Eugeni and Gambier (2023, 82) argue that such shifts often achieve full correspondence between source and target texts. Of particular relevance are two types of intralingual translation. Diamesic Translation involves shifting communication modes (e.g., spoken to written) while retaining the same language (Eugeni, 2020).

Diastratic Translation, on the other hand, involves register shifts within the same language, such as from Standard English (SE) to Easy to Read (E2R) English, i.e. the variation of language that is easy to read and understand for people with reading difficulties, including people with intellectual disabilities, people with little command of the language, people with poor literacy and so forth (Inclusion Europe, 2009; Bernabé Caro, 2017). Compared to standard language E2R language is a simplified version for the sake of readability for specific audiences (Bernabé Caro, 2017). As a result, it forms the foundation of diverse and adaptable translation strategies designed to make information accessible to people with intellectual disabilities.

Previous studies in text simplification have primarily focused on lexical simplification, where individual words or phrases are simplified without considering the broader sentence structure or context. For instance, Saggion and Specia (2015) developed datasets and tools specifically tailored for lexical simplification tasks, emphasising word-level transformations. While this approach has proven effective for specific applications, it often overlooks the interplay between lexical and syntactic features within a sentence.

Other notable resources, such as the ASSET corpus (Alva-Manchego et al., 2020), have focused on sentence simplification but rely on predefined, fine-grained operations at the word or phrase level. Similarly, corpora like WikiLarge (Zhang and Lapata, 2017) offer paired datasets for simplification but lack explicit annotations for the strategies applied during simplification. These resources are invaluable for training machine learning models but are limited in their ability to capture a comprehensive view of the simplification process.

In contrast to the resources mentioned above, our dataset adopts a holistic approach to sentence simplification, focusing on sentence-level transformations that encompass lexical, syntactic, and semantic changes, while focusing on the reason to make these changes. Unlike lexical simplification datasets, which isolate individual words or phrases, our dataset explicitly annotates entire sentences with six predefined categories representing diverse simplification strategies. This allows for better understanding of the simplification process, capturing how different strategies interact within a sentence to enhance its readability and accessibility.

Furthermore, by annotating SE and E2R sentence pairs, our dataset provides a unique resource for exploring context-sensitive simplification strategies. This makes it particularly valuable for tasks that require an integrated understanding of sentence-level transformations.

This study explores strategies to make information more accessible through text simplification. Our contributions concern: (1) the development of an extended taxonomy of translation strategies that integrates insights from Text Simplification research, (2) the annotation of a parallel corpus of complex and simplified texts sourced from diverse public services in Scotland (see Section 2), (3) the investigation of setting to train transformer-based models to predict the application of specific simplification strategies, and (4) an investigation into interpretability of their predictions using Explainable AI (XAI) techniques to explain the model's decision-making process. While Large Language Models (LLMs) demonstrate impressive performance, their *"black-box"* nature often makes it challenging to understand their predictions. To address this, we employ Integrated Gradients (Sundararajan et al., 2017), an XAI method grounded in axiomatic attribution principles. IG identifies the most influential words in the input by analysing gradient variation. By aligning these attributions with human judgments, we enhance the interpretability of the model and build trust in its application.

2 Dataset

The original corpus consists of over 76 parallel texts, primarily sourced from the Scottish care service, political manifestos for the 2024 UK general election, and newsletters from the national charity Disability Equality Scotland. These texts span a diverse range of topics, including health care services, environmental policies, the legal system, waste management, disability advocacy, and linguistic accessibility.

Table 1 compares information about the original documents ("complex") with their simplified versions in terms of the number of words and sentences in each corpus part as well as the Inter-Quartile Range of the sentence lengths measured in words. The overall word count and average sentence length have significantly decreased for the simplified version compared to the complex texts, in spite of some of the strategies aimed at explanation and sentence splitting. This increase in the number of sentences, coupled with the reduction in word count, reflects a structural adjustment typical of simplification strategies, which often involves breaking down longer sentences into shorter, more accessible ones to enhance readability.

Table 2 lists the general strategies for simplification, while Table 3 lists the fine-grained annotation categories used for annotation. A detailed breakdown of macro typology frequencies within their corresponding main strategies showcases the distribution of techniques and methods employed to simplify texts. The prominence of semantic and explanation categories reflects a strong emphasis on clarity and enhancing reader accessibility.

In the field of Translation Studies, many taxonomies have been developed to identify the strategies professional translators apply when producing

Table 1: Snapshot of Scottish Government Dataset Statistics

Source	#Texts	Complex			Simple		
		#Words	#Sentences	IQR	#Words	#Sentences	IQR
Health	21	183677	7258	(15.0-31.0)	30253	1519	(10.0-21.0)
Public info	4	12217	527	(12.0-30.5)	3378	217	(9.0-18.0)
Politics	9	113412	4824	(15.0-29.0)	12474	832	(9.0-17.0)
Data selection	–	4166	155	(12-27)	3259	161	(9-20)

Table 2: Macro-Strategies and Corresponding Strategies for Simplification

Macro-Strategy	Strategies
Transcription	No simplification needed.
Synonymy	**Pragmatic:** Acronyms spelled out; Proper names to common names; Contextual synonyms made explicit. **Semantic:** Hyperyms; Hyponyms; Stereotypes. **Grammatical:** Negative to positive sentences; Passive to active sentences; Pronouns to referents; Tenses simplified.
Explanation	Words given for known; Expressions given for known; Tropes explained; Schemes explained; Deixis clarified; Hidden grammar made explicit; Hidden concepts made explicit.
Syntactic Changes	Word → Group; Word → Clause; Word → Sentence; Group → Word; Group → Clause; Group → Sentence; Clause → Word; Clause → Group; Clause → Sentence; Sentence → Word; Sentence → Group; Sentence → Clause.
Transposition	Nouns for things, animals, or people; Verbs for actions; Adjectives for nouns; Adverbs for verbs.
Modulation	Text-level linearity; Sentence-level linearity: Chronological order of clauses; Logical order of complements.
Anaphora	Repetition replaces synonyms.
Omission	Useless elements: Nouns; Verbs; Complements; Sentences. Rhetorical constructs; Diamesic elements.
Illocutionary Change	Implicit meaning made explicit.
Compression	Grammatical constructs simplified; Rhetorical constructs simplified.

a target text. Most of these strategies have been developed in the field of interlingual translation, first from a written text into another written text (Nida, 1964; Vinay and Darbelnet, 1971; Chesterman, 1997; Molina and Albir, 2002), and then from a spoken text into a written text (Gottlieb, 1992; Lambert and Delabastita, 1996; Ivarsson and Carroll, 1998; Lomheim, 1995; Kovačič, 2000). The study of intralingual translation strategies is relatively more recent and mainly focuses on **Diamesic Translation** (Neves, 2005; Eugeni, 2007; Brumme, 2008; Gambier and Lautenbacher, 2010; Eugeni and Gambier, 2023). Rarer is the number of authors who have tried to define strategies for the translation of written texts within the same language (Korning Zethsen, 2009; Ersland, 2014). To our knowledge, only Hansen-Schirra et al. (2020) and Maaß and Rink (2020) have addressed intralingual translation practices into E2R.

However, none of these taxonomies completely satisfy our need to account for all the simplification strategies we identified in our corpus, as too little detail was provided. The opposite happens in the completely different field of Automatic Text Simplification (ATS), where details are, instead, provided. Here, the focus of typologies is on linguistic descriptions and string edits. A significant contribution in ATS has been provided by Cardon et al. (2022), whose typology essentially focuses on operations that mainly deal with adding, deleting, replacing, and moving words. However, texts translated in E2R language clearly show that professionals in the field apply many more operations that pertain to the field of pragmatics and semiotics, focused on how concepts are distributed and or explained to help the user understand them. It is in this context that this section will try to illustrate the annotation framework that we have developed and used in this study. Because the form of translation we are focussing on in this paper is diastratic (from SE to E2R), we used Inclusion Europe's pioneering guidelines Inclusion Europe (2009) as a basis for our annotation framework, which was then used to identify the strategies used in our corpus.

The principle of Inclusion Europe's guidelines is language simplification, further subdivided into three levels: lexical, syntactical, and semantic. The lexical level mainly focuses on the use of nouns, verb tenses, adjectives, and adverbs. In particular, the guidelines require to only use basic vocabulary

Table 3: A subset of strategies in dataset annotations and their annotation labels

Macro-Strategy	Strategies
Omission	OmiSen, OmiWor, OmiClau, OmiRhet (on the level of sentences, words, clauses or rhetorical structures)
Compression	SinGram, SimGram, SinSem, SinPrag
Explanation	ExplWor, ExplCont, ExplExpr, HidCont, HidGram, WordExpl
Syntactic Changes	SynChange, Clause2Word, WordsOrder, GroupOrder, LinearOrderSen, LinearOrderCla
Substitution	Anaph, SynSem, SemStereo
Transposition	TranspNoun
Modulation	ModInfo

words. For the English language, the Basic Vocabulary (Ogden, 1932) – that has evolved into projects like Voice of America's Word Book of around 1500 words – contains 850 commonly used word roots, like thing, do, good, or very. The syntactical level mainly focuses on the use of the order of words and clauses in a sentence, and that of sentences in the text. In particular, the guidelines require to only use a (chrono-)logically linear word, clause, and sentence order. The semantic level mainly focuses on the distribution of concepts in the text. In particular, the guidelines require one concept per sentence. *Information for all* also add other pieces of information, like the use of pictograms to reinforce the information provided in the text. However, these will not be considered in the present study.

Based on these principles, and a qualitative analysis of the illustrated corpus, we came up with the following nine macro-strategies, that easily adapt to our heterogenous corpus. Macro-strategies are further subdivided into strategies and micro-strategies. The macro-strategies have been thought as points in a continuum between two poles: those resulting in most addition of text (explanation) to those resulting in the most deduction of text (omission), the middle being constituted by transcription, with no addition or deduction of text (Figure 1). Examples are taken from our corpus.

1. *Explanation*, which includes the explicitation of hidden grammar or content (e.g. "wherever they live" → "wherever they live in Scotland"), or the explanation of a word or expression that is given for known (e.g. "**co-design** services with people with experience of accessing and delivering them" → "**co-design** services with people who use or work in them and their carers. **Co-design means** you can share your ideas and experiences with us.").

2. *Modulation* is the distribution of information in a linear order in the text and in a sentence, according to the principle that one sentence should contain one piece of information only. This means that one sentence is turned into more sentences (e.g.

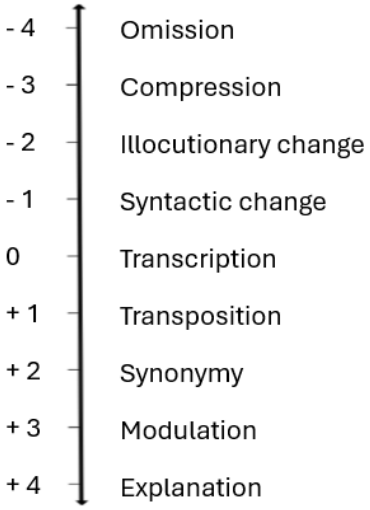

Figure 1: Diastratic Translation Strategies distributed along a continuum, from most deduction of text (-4) to most addition of text (+4)

"He joins in community activities as much as possible, supported by his assistants and his family." → "He likes to take part in activities where he can meet people. He gets support from his assistants and his family.") or words are redistributed within the sentence (e.g. "The NCS will make collaboration and **information sharing** between these services easier" → "The NCS will make working together and **sharing information** easier for services.").

3. *Synonymy*, whereby a complex, technical, or abstract word is replaced by a more common and concrete one. Synonymy includes pragmatic synonyms that depend on the context (e.g. "sir Keir Starmer" → "the new Prime Minister"), as well as semantic synonyms (e.g. "conversation" → "talk"), and grammatical synonyms (e.g. "The money does not have to be paid back" → "You do not have to pay the money back") that depend on grammar.

4. *Transposition*, or word class change, whereby the class of a word is changed depending on the principle that nouns should ideally stand for things, animals, or people, and verbs stand for actions (e.g.

"our aim is" → the Scottish Government wants").

5. *Transcript*, by which the words of the source text are left unchanged because no simplification is needed (e.g. "I love music").

6. *Syntactic change*, whereby a word, group, clause, or sentence is turned into one of the other three syntactic levels (e.g. citizens → people living in Scotland).

7. *Illocutionary change*, by which what is implied is said (e.g. "I like to say that we, the dancers, must gather information about our body's library → "The dancers must know their own body.").

8. *Compression* of grammatical or semantic constructs (e.g. "The moderator asks questions and shows slides, pictures or videos **to guide the group**" → "The moderator asks questions and shows slides, pictures, or videos **to the group**").

9. *Omission* of rhetorical or diamesic constructs (e.g. "I was nervous, **of course**, but it was interesting and fun!" → "I was worried, but it was interesting and fun!"), or of what is considered useless for understanding an idea at the noun, verb, complement or sentence level (e.g. "**Sir Keir Rodney Starmer KCB KC** is a British politician" → **Starmer** is a British politician").

3 Classification Model: Multiclass Text Classification with Transformers

This experiment investigates the application of pretrained transformer-based models for multiclass text classification, focusing on the prediction of simplification strategies need to simplify the respective SE sentences.

For this experiment, seven categories were manually annotated for a selection of 155 complex sentences and their 161 corresponding simplified sentences, randomly selected from various texts see Table 1. The seven categories—*Explanation*, *Grammatical Adjustments*, *Modulation*, *Omission*, *Substitution*, *Transposition*, and *Syntactic Changes*—were applied to ensure coverage of multiple topics and simplification strategies. This selection was designed to create a balanced dataset that represents diverse contexts and simplification strategies. These labels are not hierarchical but independent categories reflecting distinct simplification strategies.

The annotation process consisted of a first analysis of the parallel texts, and a review of the existing typologies used to illustrate translation operations, both in the field of computational linguistics and translation studies. Thanks to these contributions, we came to the definition of the typology provided in Table 1.

The training dataset consists of Standard English sentences paired with their simplified counterparts. Each simplified counterpart was designed to include precisely one simplification strategy, where a single complexity was restored to its original form. This design ensures that the relationship between a sentence and its simplified version highlights specific simplification strategies, allowing the model to associate each sentence with different parts of the complexity being resolved. To streamline classification, these fine-grained simplification strategies were mapped to broader macro-categories based on a predefined hierarchical structure, simplifying the labels while preserving their semantic distinctions.

3.1 Model and Training Procedure

We fine-tuned four different pre-trained transformer models to perform the task of multiclass classification, predicting the most likely simplification typology for each Standard English sentence.

Cross-Validation and Early Stopping We employed *Stratified 5-Fold Cross-Validation* to ensure robust evaluation and generalizability. The dataset was split into four folds, maintaining the proportional distribution of typologies across training and validation sets. For each fold, the model was trained on four folds and validated on the remaining fold, and this process was repeated for all five folds. The validation results were averaged across all folds to compute the final scores.

We used early stopping, where training was terminated if the validation loss did not improve for the patience period. This ensured efficient use of resources while retaining the best model.

Class Imbalance and Weighted Loss Function Class imbalance in the dataset, where certain typologies were underrepresented, posed a challenge during training. To address this, we utilised a *weighted cross-entropy loss function*. Class weights were calculated based on the inverse frequency of each category:

$$w_c = \frac{1}{\text{freq}_c} \cdot \frac{N}{2}, \tag{1}$$

where w_c is the weight assigned to class c, freq_c is the frequency of class c, and N is the total number of samples. This approach ensured that underrepresented classes contributed more significantly

to the overall loss, improving the model's ability to predict these minority classes.

Gradient Clipping Additionally, gradient clipping was applied during training to stabilise the optimisation process. Gradient clipping limits the maximum value of gradients during backpropagation, preventing excessively large updates to model parameters that could destabilise training or lead to divergence. Following best practices in training transformer-based models (Devlin et al., 2019), we used a clipping threshold of 1.0. This ensures that gradients exceeding the threshold are scaled proportionally while gradients below the threshold remain unchanged. Mathematically, gradient clipping can be expressed as:

$$g_{\text{clipped}} = \min\left(g, \frac{g_{\text{threshold}}}{\|g\|}\right), \tag{2}$$

where g represents the original gradient vector, $g_{\text{threshold}}$ is the clipping threshold (in this case, 1.0), and $\|g\|$ is the norm of the gradient vector. Gradient clipping ensures consistent updates to model parameters, improving training stability.

Transformer Models and Training Configuration Each of the four transformer models was fine-tuned for the task, using the same training configuration. The hyperparameters and training configuration are summarised in Table 4.

Table 4: Hyperparameters and Training Configuration

Parameter	Value
Pre-trained Models	`bert-large-cased`, `bert-base-multilingual cased`, `roberta-base`, `roberta-large`
Max_Sequence_Length	512 tokens
Tokenisation	Pre-trained tokenizer
Loss Function	Weighted Cross-Entropy Loss
Class Weights	Inverse frequency of categories
Gradient_Clipping Threshold	1.0
Learning Rate	5×10^{-6}
Batch Size	8
Weight Decay	0.01
Number of Epochs	Up to 20 (early stopping)
Cross-Validation	Stratified 5-Fold
Early Stopping Patience	3 epochs
GPU	NVIDIA Tesla T4 ((15 GB memory)), & Occasionally P100/V100

3.2 Evaluation Metrics and Results

To evaluate the performance of our models, we first established a baseline using a majority-class prediction approach. This naive model assigns the most frequent class, "Explanation," to all samples. The baseline achieved an accuracy of 24.5% and a weighted F1-score of 9.6%. Its macro F1-score, reflecting performance across all classes equally, was only 5.6%, highlighting its inability to handle class imbalance effectively. These results demonstrate the need for a robust machine learning model to capture the nuances of the dataset.

In contrast, our fine-tuned model (mBERT) significantly outperformed the baseline. It achieved an accuracy of 70% and a weighted F1-score of 72%. The macro F1-score of the multilingual model reached 65%, reflecting its ability to generalise across minority classes.

In contrast, the other models demonstrated varying degrees of performance. While `roberta-base` and `roberta-large` produced reasonable results for specific classes, their overall weighted F1-scores lagged behind at 0.52 and 0.50, respectively. Similarly, `bert-large-cased` delivered moderate results with a weighted F1-score of 0.50 and accuracy of 0.53. The instability observed in the training of `roberta-base` and `roberta-large`, as evident from Figure 2, likely contributed to their lower overall scores.

The `mBERT` model excelled in identifying simplification strategies for the *Explanation* (F1-score: 0.93), *Substitution* (F1-score: 0.67), and *Syntactic Changes* (F1-score: 0.80) categories. These results highlight its ability to capture the relationships inherent in these categories. However, underrepresented classes like *Grammatical Adjustments* and *Transposition* remained challenging for all models, with low F1-scores across the board. This indicates the need for a more balanced dataset.

Figure 2 illustrates the evaluation loss progression during training, where the `mBERT` model exhibited a smooth and consistent reduction in loss, indicating stable convergence. In contrast, `roberta-base` and `roberta-large` displayed oscillatory behavior, suggesting instability in their training dynamics.

The progression of the F1-score, as shown in Figure 3, further reinforces these observations. The `mBERT` model achieved the highest F1-scores early in training and maintained steady improvement, outperforming its competitors consis-

tently. Interestingly, increasing model size (e.g., `bert-large-cased` and `roberta-large`) did not consistently improve F1 performance, as both larger models underperformed compared to the smaller `mBERT` model. This finding suggests that model architecture and multilingual capabilities may have a more significant impact on F1 performance than size alone, underscoring the need to tailor models to the specific requirements of multilingual simplification tasks.

The `mBERT` model's performance aligns seamlessly with the project's primary aim of fostering multilingual accessibility, underscoring the critical importance of leveraging multilingual models to address diverse linguistic contexts and ensure inclusivity in simplification strategies.

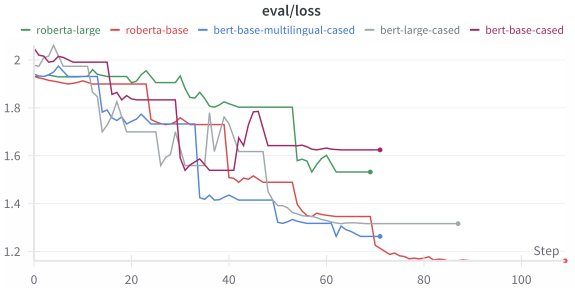

Figure 2: Evaluation Loss Progression During Training

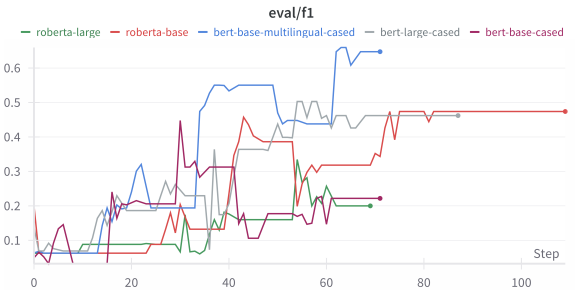

Figure 3: F1-Score Progression During Training

4 Interpretability of predictions

We have trained a classifier for predicting the difficulty of sentences by means of collecting simple and difficult sentences from Wikipedia and fine-tuning mBERT (Devlin et al., 2019).

By means of the implementation of the Integrated Gradients in the Captum library (Miglani et al., 2023), we can:

1. detect which words or syntactic constructions commonly affect readability, as well as
2. which of them align with human annotation.

We utilised the Integrated Gradients (IG) method to identify the tokens in a sentence that contributed most significantly to the model's predictions. IG achieves this by calculating the gradients of the model's output with respect to its input, thereby highlighting the importance of individual features. **For Example:** Consider the following sentence from our dataset:

> *"Provide financially sustainable care, giving security and stability to people and their carers."*

The Integrated Gradients approach offered actionable insights by attributing importance scores to specific words, revealing their influence on the model's predictions. For this sentence, the prediction probabilities are: **Simple:** 0.02, and **Complex:** 0.98.

- **High-impact words:** The IG method highlighted domain-specific and content-heavy words such as *"sustainable," "security," and "stability"*, which were crucial for determining that the sentence was *"Complex."*
- **Stopwords:** Words with minimal semantic content (e.g., *"and," "to," "their"*) were assigned near-zero attribution scores, as expected.
- **Prediction Analysis:** Based on the probabilities, the sentence was classified as *Complex* with a high confidence of 98%.

By applying the IG method, we identified a total of 1303 complex words from the original sentences. These words were then compared against their corresponding simplified, E2R versions to determine which complex words were removed during simplification. This comparison yielded 877 removed words, representing 67.31% of the total complex words identified. The removed words are indicative of tokens that were deemed complex by both the model and human editors, as their removal from the E2R versions suggests that they were perceived as difficult or unnecessary for simplified comprehension. This alignment between the model-predicted complex words and those removed in human-curated simplifications demonstrates the model's effectiveness in predicting words that are likely to be complex and corroborates the utility of

Table 5: Classification Report for Typology Prediction

Class	**bert-large-cased**			**bert-base-multilingual-cased**			**Support**
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
Explanation	0.67	0.50	0.57	1.00	0.88	0.93	8
Grammatical Adjustments	0.00	0.00	0.00	0.00	0.00	0.00	4
Modulation	1.00	0.33	0.50	0.00	0.00	0.00	3
Omission	0.50	0.50	0.50	0.80	1.00	0.89	4
Substitution	0.46	1.00	0.63	0.50	1.00	0.67	6
Syntactic Changes	0.50	1.00	0.67	1.00	0.67	0.80	3
Transposition	0.00	0.00	0.00	1.00	1.00	1.00	2
Avg (Macro)	0.45	0.48	0.47	0.62	0.70	0.65	
Avg (Weighted)	0.48	0.53	0.50	0.68	0.75	**0.72**	
Accuracy	*0.53*			**0.70**			34
Training Time (s)	395.22			*300.55*			
Class	**roberta-base**			**roberta-large**			**Support**
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
Explanation	1.00	0.50	0.67	0.00	0.00	0.00	8
Grammatical Adjustments	0.00	0.00	0.00	0.00	0.00	0.00	4
Modulation	1.00	0.33	0.50	1.00	0.67	0.80	3
Omission	0.75	0.75	0.75	1.00	0.25	0.40	4
Substitution	0.43	1.00	0.60	0.25	0.40	0.31	6
Syntactic Changes	0.60	1.00	0.75	0.67	0.67	0.67	3
Transposition	0.25	0.50	0.33	0.00	0.00	0.00	2
Avg (Macro)	0.47	0.51	0.48	0.28	0.28	0.27	
Avg (Weighted)	0.50	0.53	*0.52*	0.30	0.35	0.32	
Accuracy	*0.53*			0.30			34
Training Time (s)	**219.30**			587.21			

Table 6: Word-level Attributions for the Example Sentence

Word	**Attribution**	**Contribution**
Provide	0.18	Moderately Complex
financially	-0.10	Slightly Easy
sustainable	0.30	Highly Complex
care	0.15	Slightly Complex
giving	0.10	Slightly Complex
security	0.25	Highly Complex
and	-0.02	Neutral
stability	0.28	Highly Complex
to	-0.03	Neutral
people	0.12	Slightly Complex
and	-0.04	Neutral
their	0.05	Neutral
carers	-0.08	Neutral

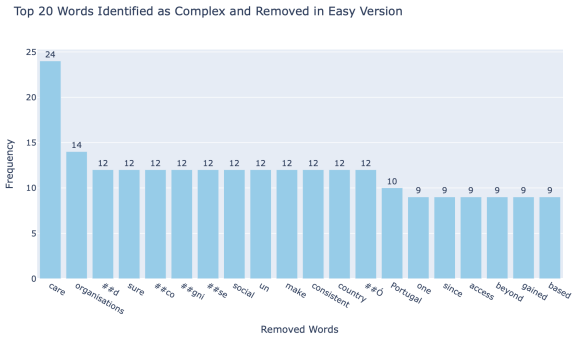

Figure 4: Top 20 Words Identified as Complex and Removed in Easy Version

the IG method for interpretability in text simplification tasks. As shown in **Figure 4**, the most frequently removed complex words included meaningful content terms such as *"care," "organisations,"* and *"consistent."*

5 Findings and Contributions

The findings demonstrate that transformer-based models are capable of handling the complexities of typology classification, especially when supported by preprocessing techniques and loss weighting strategies. The model exhibits moderate success in identifying phenomena that require simplification. However, it encounters notable challenges with underrepresented classes and specific simplification strategies, such as *"grammatical adjustments"* and *"omission."*

In summary, while transformer-based models hold considerable potential for simplifying texts to improve accessibility, addressing class imbalance through the use of comprehensive, balanced datasets is crucial. Leveraging the complete dataset further enhances the model's reliability and enables it to generalise effectively across all simplification categories.

One of the critical findings of this study is the utility of the IG framework for interpretability. IG provides insights that align closely with human annotations regarding complexity. For example, IG effectively identifies tokens contributing to diffi-

culty, such as *"sustainable"* or *"stability"*, while assigning minimal importance to semantically neutral words like *"and"* or *"to."* This alignment bridges the gap between machine predictions and human reasoning, enabling iterative improvements in model development.

The alignment of the model's predictions with the removal of complex words by human editors demonstrates its capability to predict readability effectively. In particular, 67.31% of the complex words identified by IG were removed in the human-simplified versions, highlighting the model's predictive accuracy in real-world applications.

Moreover, the study shows the close connection between linguistic complexity and simplification practices. Frequent removal of meaningful content words, such as *"care," "organisations,"* and *"consistent,"* highlights the importance of meaning and context in making texts easier to understand for different audiences.

6 Conclusions

Building on the annotation framework, several key insights emerge regarding the challenges and strategies involved in translating texts into E2R English. First, intralingual translation facilitates a more straightforward comparison between source and target texts due to the inherent isomorphism between the source and target languages. Second, the choice of translation strategies must be tailored to the specific type of intralingual translation, ensuring that the target text aligns with its intended function. For example, in diastratic translation—specifically the transformation of standard English into E2R English—the focus lies on simplifying vocabulary, syntax, and semantic structures while maintaining fidelity to the source text and accessibility for the target audience.

Moreover, the proposed taxonomy, encompassing 9 macro-strategies, 33 strategies, and 15 micro-strategies, illustrates the cognitive complexity of intralingual translation. These challenges underscore the limitations of current automation tools, as computational analyses reveal the nuanced skills required for transcription and modification strategies. Even in the era of generative artificial intelligence, text simplification remains a non-trivial task due to its intricate linguistic demands.

The novelty of our approach lies not only in the dataset itself but also in the methodology, which bridges translation studies and text simplification by categorizing transformations into well-defined categories. This integration offers new insights into the strategies employed in simplification and provides a robust framework for developing models that can generalise across multiple types of linguistic transformations.

The results highlight the significant progress achieved with our approach, as the fine-tuned `mBERT` model outperformed the baseline majority-class strategy, which achieved an accuracy of 24.5% and a weighted F1-score of 9.6%. In contrast, `mBERT` achieved 70% accuracy, a weighted F1-score of 72%, and a macro F1-score of 65%, demonstrating its ability to generalise across majority and minority classes.

Employing Integrated Gradients (IG) enhances the interpretability of model predictions, ensuring closer alignment with human annotations. IG offers a clearer understanding of the input data elements the model prioritises, thereby elucidating its decision-making processes. Our primary results align with the identification of complex words that were either modified or removed in the simplified versions. In particular, 67.31% of the complex words identified by IG were removed in the human-simplified versions, highlighting the model's accuracy in applications. This transparency is critical for identifying strengths and weaknesses, guiding iterative improvements, and fostering trust in machine-generated outputs. Additionally, IG serves as a tool to validate the predictions of the LLM model against expert judgments, ensuring reliability and consistency in its reasoning, and ensuring that it makes the right predictions for the right reasons (Schramowski et al., 2020).

Future research should prioritise addressing class imbalance through advanced techniques such as hierarchical annotations, domain-specific embeddings, or data augmentation. Incorporating multiple annotators would also enable the calculation of agreement metrics, improving the evaluation of annotation reliability. Expanding the interpretability framework to cross-linguistic simplifications presents another promising avenue. Leveraging the full Scottish Government dataset and employing advanced machine learning techniques could further enhance performance across all linguistic categories. This work ultimately contributes to the broader goal of creating accessible, inclusive texts while promoting trust and transparency in AI-driven systems.

Acknowledgments

This document is part of a project that has received funding from the European Union's Horizon Europe research and innovation program under the Grant Agreement No. 101132431 (iDEM Project). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

The University of Leeds (UOL) was funded by **UK Research and Innovation (UKRI)** under the UK government's Horizon Europe funding guarantee (Grant Agreement No. 10103529).

References

Fernando Alva-Manchego, Joachim Bingel, Gustavo H. Paetzold, and Horacio Saggion. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679.

American Association on Intellectual and Developmental Disabilities (AAIDD). n.d. FAQs on Intellectual Disability. https://www.aaidd.org/intellectual-disability/faqs-on-intellectual-disability. Accessed: 2024-12-09.

Rocío Bernabé Caro. 2017. Propuesta metodológica para el desarrollo de la lectura fácil según el diseño centrado en el usuario. *Revista Española de Discapacidad*, 5(2):19–51. Discusses simplification strategies for specific audiences.

Jenny Brumme. 2008. *Intralingual Translation: Concepts and Applications*. University of Granada Press.

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter, Magali Norré, Adeline Müller, Patrick Watrin, and Thomas François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew Chesterman. 1997. *Memes of Translation: The Spread of Ideas in Translation Theory*. John Benjamins Publishing Company, Amsterdam.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Anlaug Ersland. 2014. Is change necessary? a study of norms and translation universals in intralingual translation. Master's thesis, University of Bergen.

Carlo Eugeni. 2007. Respeaking the news for the deaf: for a real special needs-oriented subtitling. *Studies in English Language and Literature*, 21.

Carlo Eugeni. 2020. Human-computer interaction in diamesic translation. multilingual live subtitling. In Carlo Eugeni Daniel Dejica and Anca Dejica-Cartis, editors, *Translation Studies and Information Technology - New Pathways for Researchers, Teachers and Professionals*, Translation Studies Series, pages 19–31. Editura Politehnica, Timișoara.

Carlo Eugeni and Yves Gambier. 2023. *La traduction intralinguistique – Les défis de la diamésie*. Editura Politehnica, Timişoara.

Yves Gambier. 2006. La traduction audiovisuelle : une traduction sélective. In Jorma Tommola and Yves Gambier, editors, *Translation and Interpreting – Training and Research*, pages 21–37. University of Turku, Department of English Translation Studies, Turku.

Yves Gambier and Brigitte Lautenbacher. 2010. Intralingual translation: Expanding the field. *Translation Studies*, 3(2):175–187.

Henrik Gottlieb. 1992. Subtitling: A new university discipline. *Cinemas*, pages 161–170.

Silvia Hansen-Schirra, Walter Bisang, Arne Nagels, Silke Gutermuth, Julia Fuchs, Liv Borghardt, Silvana Deilen, Anne-Kathrin Gros, Laura Schiffl, and Johanna Sommer. 2020. Intralingual translation into easy language – or how to reduce cognitive processing costs. In Silvia Hansen-Schirra and Christiane Maaß, editors, *Easy Language Research: Text and User Perspectives, Easy – Plain – Accessible*, volume Volume 2, pages 197–226. Frank & Timme, Berlin.

Inclusion Europe. 2009. *Information for All: European Guidelines for the Production of Easy-to-Read Information*. Inclusion Europe, Brussels, Belgium. Available online at http://www.easy-to-read.eu.

Jan Ivarsson and Mary Carroll. 1998. *Subtitling*. TransEdit.

Karen Korning Zethsen. 2009. Intralingual translation: An attempt at description. *Meta*, 54(4):795–812.

Irena Kovačič. 2000. Quality assessment of subtitles. *Translation Journal*, 4(3).

José Lambert and Dirk Delabastita. 1996. Film and translation. *Meta: Translators' Journal*, 41(1):85–98.

Sylfest Lomheim. 1995. L'écriture sur l'écran: Stratégies de sous-titrage à nrk. In Yves Gambier, editor, *Communication audiovisuelle et transferts linguistiques*, volume 14, pages 288–293. Translatio, FIT Newsletter/Nouvelles de la FIT.

Christiane Maaß and Isabel Rink. 2020. Scenarios for easy language translation: How to produce accessible content for users with diverse needs. In Silvia Hansen-Schirra and Christiane Maaß, editors, *Easy Language Research: Text and User Perspectives*, pages 41–56. Frank & Timme.

Vivek Miglani, Aobo Yang, Aram Markosyan, Diego Garcia-Olano, and Narine Kokhlikyan. 2023. Using Captum to explain generative language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 165–173, Singapore. Association for Computational Linguistics.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation techniques revisited: A dynamic and functionalist approach. *Meta: Journal des traducteurs*, 47(4):498–512.

Lucía Molina and Amparo Hurtado Albir. 2002. Translating techniques revisited: A dynamic and functionalist approach. *Meta*, 47(4):498–512.

Josélia Neves. 2005. *Audiovisual Translation: Subtitling for the Deaf and Hard-of-Hearing*. University of Surrey.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall, New York.

Eugene A. Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill Archive.

Charles Kay Ogden. 1932. *Basic English: A General Introduction with Rules and Grammar*. Kegan Paul, Trench, Trübner & Company Limited, London.

Horacio Saggion. 2017. *Automatic text simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Horacio Saggion and Lucia Specia. 2015. Lexical simplification: Graph-based unsupervised learning. *Natural Language Engineering*, 21(3):389–435.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328.

Jean-Paul Vinay and Jean Darbelnet. 1971. *Stylistique comparée du français et de l'anglais*. Didier. Translated into English as *Comparative Stylistics of French and English*, 1995.

Patrick Zabalbeascoa. 2000. From techniques to types of solutions. In Allison Beeby, Doris Ensinger, and Marisa Presas, editors, *Investigating Translation*, pages 117–127. John Benjamins, Amsterdam and Philadelphia.

Wei Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.