algorithms

MDPI

Article

Automatic Simplification of Lithuanian Administrative Texts

Justina Mandravickaitė *,†, Eglė Rimkienė †, Danguolė Kotryna Kapkan †, Danguolė Kalinauskaitė † and Tomas Krilavičius *,†

Faculty of Informatics, Vytautas Magnus University, 53361 Akademija, Kaunas District, Lithuania; egle.rimkiene@vdu.lt (E.R.); danguole.kapkan@vdu.lt (D.K.K.); danguole.kalinauskaite@vdu.lt (D.K.)

* Correspondence: justina.mandravickaite@vdu.lt (J.M.); tomas.krilavicius@vdu.lt (T.K.)

† These authors contributed equally to this work.

Abstract: Text simplification reduces the complexity of text while preserving essential information, thus making it more accessible to a broad range of readers, including individuals with cognitive disorders, non-native speakers, children, and the general public. In this paper, we present experiments on text simplification for the Lithuanian language, aiming to simplify administrative texts to a Plain Language level. We fine-tuned mT5 and mBART models for this task and evaluated the effectiveness of ChatGPT as well. We assessed simplification results via both quantitative metrics and qualitative evaluation. Our findings indicated that mBART performed the best as it achieved the best scores across all evaluation metrics. The qualitative analysis further supported these findings. ChatGPT experiments showed that it responded quite well to a short and simple prompt to simplify the given text; however, it ignored most of the rules given in a more elaborate prompt. Finally, our analysis revealed that BERTScore and ROUGE aligned moderately well with human evaluations, while BLEU and readability scores indicated lower or even negative correlations.

Keywords: text simplification; Lithuanian; transformers; fine-tuning; mT5; mBART; ChatGPT

check for updates

Citation: Mandravickaitė, J.;Rimkienė, E.; Kapkan, D.K.; Kalinauskaitė, D.; Krilavičius, T. Automatic Simplification of Lithuanian Administrative Texts. *Algorithms* **2024**, *17*, 533. https://doi.org/10.3390/a17110533

Academic Editor: Frank Werner

Received: 30 September 2024
Revised: 28 October 2024
Accepted: 9 November 2024
Published: 20 November 2024

1. Introduction

Text simplification reduces the vocabulary and syntactic complexity of a text while maintaining its essential information. This is particularly important for enhancing the accessibility of information for individuals with cognitive disorders, non-native speakers, and children [1]. Additionally, text simplification, among other use cases, is vital for improving comprehension among the general public in the context of legal and administrative texts. Such texts often serve as communication means between institutions and target audiences with different reading skills [2].

In this paper, we present our experiments on text simplification for the Lithuanian language, focusing on the administrative (clerical) style. Public authorities often employ quasi-legal language to communicate with the general public, which can be ineffective in relaying information to non-experts [2]. Consequently, these texts can be challenging to comprehend for individuals without expertise in the relevant field. While the websites of public administration institutions aim to disseminate information of public interest, there is often a gap between the intended purpose of these texts and how they reach the target audience. Text simplification can bridge this gap as its task is to transform complex natural language into a simpler form, including vocabulary, sentence structure, and other relevant features, at the same time maintaining the essential content of the original text.

Governmental institutions currently make an effort in many countries to adopt the concept of Plain Language in written communications intended for the general public. Plain Language is defined as communication that is clear in wording, structure, and design so that the intended audience easily finds, understands, and uses the necessary information [3], without the need for specialized training and/or expertise in a particular field (e.g., legal, medicine, etc.) [4]. Therefore, our experiments aim to simplify Lithuanian administrative texts to a level that corresponds with the principles of Plain Language.

For our experiments, we selected mT5 and mBART as the base models, which we fine-tuned to develop text simplification models for Lithuanian texts. We also assessed the performance of the ChatGPT (GPT-4o) for this task. These models were chosen due to their support for the Lithuanian language, as not many large language models have adequate support for lower-resource languages.

In our approach to text simplification, we integrate both lexical and syntactic simplification, simultaneously simplifying sentence structures and replacing complex words or phrases with simpler, more common ones.

The rest of the paper is structured as follows: in Section 2 we briefly introduce related work, in Section 3 we describe the data, in Section 4 we describe methods we used in our experiments, in Section 5 we describe our experimental setup, in Section 6 we report the results, in Section 7 we discuss the results, and in Section 8 we end this paper with our conclusions.

2. Related Work

Text simplification has advanced significantly in recent years, evolving from rule-based methods (e.g., [5,6]) to data-driven approaches (e.g., [7,8]). These techniques encompass various strategies for simplification, including lexical and syntactic simplification, while taking into consideration different aspects of text complexity. Lexical simplification focuses on simplifying the vocabulary of the text by substituting complex words with easier-to-understand alternatives. This approach generally involves several sub-tasks, such as identifying difficult words, then selecting appropriate replacements, refining and ranking them by simplicity [9]. These tasks employ techniques that range from using simple synonym dictionaries [10] to employing machine learning [11] and deep learning models like LSBERT [12].

Meanwhile, syntactic simplification aims to improve text's readability by modifying syntactic structures, such as converting passive constructions into active ones or reducing the number of clauses. As with lexical simplification, syntactic simplification methods vary from rule-based approaches to data-driven ones. The former ones involve parsing sentences, identifying complex structures, applying rewrite rules, and post-processing [13]. Data-driven methods employ statistical or neural machine learning models, such as statistical machine translation [14].

Neural Machine Translation (NMT) has become popular for text simplification, treating text simplification task as a translation problem, i.e., translating complex text into its simpler equivalent (e.g., [15,16]). This approach has been applied for text simplification in low-resource scenarios, such as in [17–19].

Text simplification can also be framed as a text generation task, where sequence-to-sequence models have been applied for this purpose [20–23]. The Transformer architecture has been particularly successful in text simplification as it processes entire input sequences to extract essential information (e.g., [24,25]). Furthermore, large language models (LLMs) are now able to generate simplified texts that avoid complex and linked sentences [26,27].

Recent studies have highlighted the potential of these models to simplify texts through various techniques, such as specifying the desired reading grade level [28] or directly indicating necessary simplification operations [29]. To name a few, the BERT model has been employed for lexical simplification [30], text simplification via monolingual machine translation [31], and hybrid text simplification approaches [21]. Similarly, the T5 model has been used for controllable text simplification [32,33] and for simplification in low-resource scenarios [34,35]. Also, BART has been applied to the tasks of controllable text simplification [32], as well as paragraph- [36] and document-level simplification [37].

Among the various LLMs, GPT models have gained particular attention, especially in low-resource scenarios [38–40]. In such cases, ChatGPT has demonstrated a potential in simplifying texts [41,42]. However, evaluations of ChatGPT's performance indicate that while it simplifies text effectively, it may omit relevant information [43] and lack the depth and accuracy of human experts [44]. Nevertheless, when additional context is provided, ChatGPT may simplify complex texts effectively [45] and provide most of the factually correct information [26].

Some of the latest models and approaches for text simplification include SIMSUM, designed for automated document-level simplification [46], SimpleBART, which employs a specific pre-training strategy for this task [27], and KGSimple, which applies an unsupervised approach that leverages knowledge graphs to generate compressed text [47], to name a few. Also, instruction-based fine-tuning has been used for text simplification, where pre-trained models are designed to take instructions in natural language, specifying the characteristics the simplified text should enlist [48–50]. In addition to general text simplification, domain-specific models have been developed for fields such as medicine [51], law [52], and texts of specific genres [40].

Despite significant developments in text simplification, challenges remain, particularly in low-resource scenarios [53]. Using summarization data to enhance simplification models in such cases is among the proposed solutions [53]. Furthermore, domain-specific simplification may result in lower-quality outputs, as seen in medical text simplification [54,55]. Moreover, the challenge of ensuring factual accuracy remains critical [36,54]. Models fail to perform simplifications effectively by omitting important information and struggling with information addition [56]. Text simplification sometimes compromises important content [57], while sentence-level simplifications may disrupt document-level discourse structure [58].

Simplified texts can improve reading comprehension [59]; however, existing methods often overlook the complexity of individual inputs which results in simplifications that need improvements and corrections [29]. Also, existing text simplification systems tend to ignore context during simplification, which results in outcomes that are difficult to control [60]. Therefore, issues related to cultural and commonsense knowledge also persist, which highlights the need for further research [61].

In this paper, we present experiments in simplifying Lithuanian administrative texts to a Plain Language level [4], aiming to make them more accessible to the general public. We employ several automatic evaluation metrics and perform qualitative analysis to provide a more balanced assessment of automatic simplification.

3. Data

3.1. Data for Fine-Tuning

To explore the effect of data in terms of model performance, we used two datasets, see Table 1:

- **Parallel Corpus 1**—a dataset where each original sentence had one simplified equivalent;
- **Parallel Corpus 2**—a dataset where Parallel Corpus 1 was complemented with additional sentences, and part of them had more than one simplified equivalent (two to three), following such text simplification corpora as SimPA [62] and Human Simplification with Sentence Fusion Data Set (HSSF) [63].

Table 1. Basic statistics of corpora.

	Parallel Corpus 1		**Parallel Corpus 2**	
	Original Sentences	**Simplified Sentences**	**Original Sentences**	**Simplified Sentences**
Number of sentences	2142	2521	3123	2999
Number of words	36,404	34,702	64,936	52,382
Average sentence length	14.75	12.43	16.97	13.53

Simplification guidelines. We aim at simplifying Lithuanian administrative texts to a Plain Language level [4] to make them more easily understood by the general public. Therefore, lexical and syntactic simplification rules that were applied in preparing our Parallel Corpus 1 and Parallel Corpus 2 were mainly derived from cross-linguistic Plain Language principles [64,65]. Also, where applicable, we took into account text simplification rules from languages with similar grammatical structures to Lithuanian [66,67]. Additionally, we formulated Lithuanian-specific rules, particularly for handling participles. Therefore, guidelines for Plain Lithuanian feature three levels of simplification operations and can be summarized as follows:

1. **Paragraph-level simplification.** There are two main rules in this group:
   (a) Sentence shortening: sentences longer than 12 words should be divided into smaller units, preferably by converting embedded relative clauses into independent sentences.
   (b) List creation: If there are more than two coordinated elements (object or subject noun phrases, clauses, etc.) with a homogenous function in a sentence, they should be transformed into vertical lists.
2. **Lexical-level simplification.** There are three main rules:
   (a) Prioritizing the use of more frequent synonyms, as determined by the Lithuanian frequency dictionary [68], disregarding conventional formal register requirements.
   (b) Avoiding metaphors and uncommon acronyms.
   (c) Defining obscure terms in separate sentences.
3. **Syntactic-level simplification.** Key simplification strategies include:
   (a) Transforming passive voice constructions into active voice.
   (b) Replacing active participle and gerund constructions with relative clauses.
   (c) Minimizing the use of nominalizations.
   (d) Favoring affirmative sentences over negations.
   (e) Introducing demonstrative pronouns and determiners to enhance clarity where appropriate.

Parallel Corpus 1. This dataset comprises 2142 entries with two columns, where the first column contains original sentences or text fragments, equivalent to sentences, while the second column contains manually simplified versions of the corresponding original content. The original sentences were collected manually from a series of Lithuanian public institutions' websites, selected with the aim to capture a variety of topics (healthcare, security, immigration, social services, etc.). All data were simplified by four experts according to guidelines of a simplified version of the language, intended for non-specialists (general public) [11] and described above.

The data sources for this dataset were various Lithuanian governmental and non-governmental public institution websites that provide information on services such as social benefits, migration, utilities, copyright, and other issues. The data preparation process involved dividing the texts into sentences or sentence-equivalent text fragments (e.g., clauses) and simplifying them manually following the above-mentioned simplification guidelines.

Parallel Corpus 2. This dataset constitutes 3123 entries, having the same structure of two columns as Parallel Corpus 1. Also, 2142 entries of the latter corpus have been complemented with 981 additional entries. Out of them, ∼700 entries have two–three simplified sentences for one original one. All data were simplified by four experts according to defined simplification guidelines to simplify these sentences to the level of Plain Lithuanian. As with Parallel Corpus 1, all the original sentences were gathered from various Lithuanian governmental and non-governmental public institution websites.

3.2. Data for Testing

For testing, we used 554 sentence pairs not included in the Parallel Corpus 1 and 2. Governmental and non-governmental public institution websites were used as data sources. We compiled this set following diversity criteria in terms of topics covered as well as different levels of sentence complexity. This dataset has been prepared following the same guidelines and procedures as the data for fine-tuning.

4. Methods

4.1. mT5

We used the mT5 model, a multilingual variant of the T5 (Text-to-Text Transfer Transformer) model developed by Google [69], which reframes all language processing tasks as text generation problems. Key principles of the T5 model, which mT5 closely follows [70], include:

1. **Unified text-to-text framework**: T5 treats all NLP tasks as a text generation problem, so this approach simplifies the architecture and allows for flexibility;
2. **Pre-training on a diverse corpus**: T5 is pre-trained on the C4 (Colossal Clean Crawled Corpus) [71], which provides a broad understanding of language and context;
3. **Encoder–Decoder architecture**: Following the original Transformer model [72], T5 employs an encoder–decoder architecture, where the encoder is used for creating contextual representations, which the decoder then uses to generate the output;
4. **Fine-tuning for specific tasks**: While T5 is pre-trained on general corpora, it can be fine-tuned for specific tasks or languages.

To pre-train the mT5 model, the authors introduced a multilingual variant of the C4 dataset called mC4, which comprises textual data in 101 languages, among them Lithuanian as well. These data were drawn from the public Common Crawl web scrape. Being pre-trained on multilingual data made mT5 model particularly suitable for less-resourced languages [70], such as Lithuanian.

4.2. mBART

In our experiments, we employed mBART, an extension of the BART (Bidirectional and Auto-Regressive Transformers) model. mBART's architecture combines an auto-encoder and auto-regressive components to enhance language understanding and generation [73]. While mBART has been initially designed for machine translation, due to its versatility, it is well-suited for text simplification tasks as well. The main strength of mBART lies in its pre-training methodology:

1. **Denoising Auto-Encoder:** mBART functions as a denoising auto-encoder, which uses a sequence-to-sequence framework for multilingual training through full-text denoising;
2. **Multilingual Corpus:** The model was pre-trained on CC25, which is a subset of the Common Crawl (CC) corpus that comprises 25 languages from diverse language families, among them Lithuanian as well [74].

Lithuanian is less well supported in large language models, so it is not a trivial task to find a model with adequate handling of this language. Therefore, while the Lithuanian language representation was relatively small in the mBART pre-training dataset (only 1835 tokens within a 13.7-gigabyte corpus [75]), it contributes to its ability to handle this less-resourced language.

4.3. ChatGPT

ChatGPT is a variant of the GPT (Generative Pre-trained Transformer) family, which itself is part of a broader class of models using Transformer architectures [76]. This design is fundamentally built on self-attention mechanisms that allow the model to process words in context to one another across a sentence or document [72,77]. We tested ChatGPT (GPT-4o) for Lithuanian text simplification to explore a low-resource scenario.

For this purpose, we used ChatGPT in its standard, as-is configuration available via OpenAI's browser interface. Even though the model did not specifically target the Lithuanian language and its structures, ChatGPT held enough multilingual context for general text manipulation tasks in Lithuanian.

4.4. Evaluation

4.4.1. Metrics

We employed a set of evaluation metrics to assess text simplifications performed by our text simplification models. These metrics provide a quantitative evaluation of the results:

1. **BLEU (Bilingual Evaluation Understudy) Score** measures the n-gram overlap between the simplified output and reference sentences. BLEU scores range from 0 to 1, where higher scores indicate greater proximity to the reference [78].
2. **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Score** assesses n-gram overlap between simplified and reference texts, with scores between 0 (no overlap) and 1 (perfect overlap) [79]. We employed three ROUGE variants:
   (a) ROUGE-1: Unigram overlap;
   (b) ROUGE-2: Bigram overlap;
   (c) ROUGE-L: Longest Common Subsequence overlap.

   We used ROUGE to assess model performance during fine-tuning and to evaluate the automatically simplified sentences.
3. **SARI (System Aggregated Referenceless Evaluation Metric)** was developed specifically for text simplification tasks and evaluates simplified sentences by considering words kept, added, and deleted during simplification [80]. SARI scores range between 0 and 100, where higher values indicate better performance. Studies have shown that SARI correlates well with human judgments [81,82].
4. **BERTscore** leverages pre-trained contextual embeddings from BERT to identify similarities between candidate and reference phrases. BERTscore correlates well with human evaluation [83].
5. **LIX** is a readability metric that combines word and sentence factors in its formula, i.e., it considers the percentage of longer words (>6 characters) and average words per sentence. LIX scores are interpreted as follows [84,85]: 60—very difficult text, $\leq$40—universal readability, 20—very easy text.
6. **RIX** is a simplified version of LIX. For universal readability, a RIX score of 8 or below is recommended [85].
7. **SacreBLEU** is a variant of BLEU. It offers a standardized way to calculate scores, which facilitates more consistent model comparisons [86]. We utilized this metric to monitor model performance during the fine-tuning process.

4.4.2. Qualitative Evaluation

To complement quantitative metrics, we conducted a qualitative evaluation of text simplification outputs, using three popular criteria: simplicity, meaning retention, and grammaticality [87,88]:

1. **Simplicity:** determining whether the simplified text is less complex than the original one [89].
2. **Meaning retention (adequacy):** evaluating the extent the original semantics is retained after simplification.
3. **Grammaticality (fluency):** assessing whether the simplified text remains grammatical and understandable.

These criteria allow for an evaluation of simplification quality without requiring reference data.

Two experts independently assessed the simplified sentences generated by our models. They rated each sentence on a scale of 1 to 5 for each of the three criteria. To address the hierarchy of importance among these criteria (Simplicity > Meaning Retention > Grammaticality), we implemented the following rules:

1. If a simplified sentence receives a score of 1 for *simplicity*, it automatically receives a score of 1 for meaning retention and grammaticality, regardless of scores for other criteria.
2. If a simplified sentence scores higher than 1 for *simplicity* but receives a 1 for *meaning retention*, its grammaticality score is automatically set to 1.

This hierarchy of criteria was designed to prevent paradoxical situations where models might be inadvertently rewarded for merely replicating the original content or penalized for attempting simplification although with some errors.

5. Experimental Setup

This study explores text simplification for the Lithuanian language using two pre-trained models: mT5 and mBART. Both models were directly fine-tuned using a dataset of complex (original) and simplified Lithuanian sentences curated by linguists. The fine-tuning process consisted of eight epochs for the mBART model and sixteen epochs for the mT5 model, enabling us to monitor the progression and improvements in the models' performance throughout the training.

5.1. Fine-Tuning Process

The fine-tuning focused on the following key aspects:

1. **Hyperparameters.** Based on our previous research, we selected the hyperparameters that provided the best results. We explored the effects of varying batch sizes and learning rates in earlier experiments.
2. **Data Augmentation.** We assessed the impact of using the initial dataset (Parallel Corpus 1) versus the updated (augmented) dataset (Parallel Corpus 2) on the fine-tuning results for both models—mT5 and mBART.

The parameter configuration was selected based on the best performance observed in the previous fine-tuning iterations. For mBART, a batch size of 4 and a learning rate of $1e^{-4}$ were chosen, while for mT5, a batch size of 2 and a learning rate of $1e^{-4}$ were used. Given that mT5 showed better performance with more epochs, the number of epochs was increased from 8 to 16.

By systematically varying these hyperparameters, we aimed to identify the optimal settings that maximize model performance for the text simplification task. This approach allowed us to evaluate the impact of fine-tuning strategies and data augmentation on the quality and desired readability level (plain language) of the simplified texts generated by the mT5 and mBART models.

5.2. Models and Evaluation

The pre-trained mT5 and mBART models were fine-tuned on Parallel Corpus 1 and Parallel Corpus 2 separately, maintaining their encoder–decoder architecture to accommodate the language's nuances. In addition, we included ChatGPT in our study by adding simplification rules from our simplification guidelines to the prompt to test its text simplification capabilities in Lithuanian. *It should be noted that ChatGPT has not been fine-tuned for this task.*

We assessed all our models using selected metrics to compare their ability to simplify text while preserving the original meaning and intent. This extended evaluation provided insights into the effectiveness of each model in generating high-quality simplified Lithuanian text.

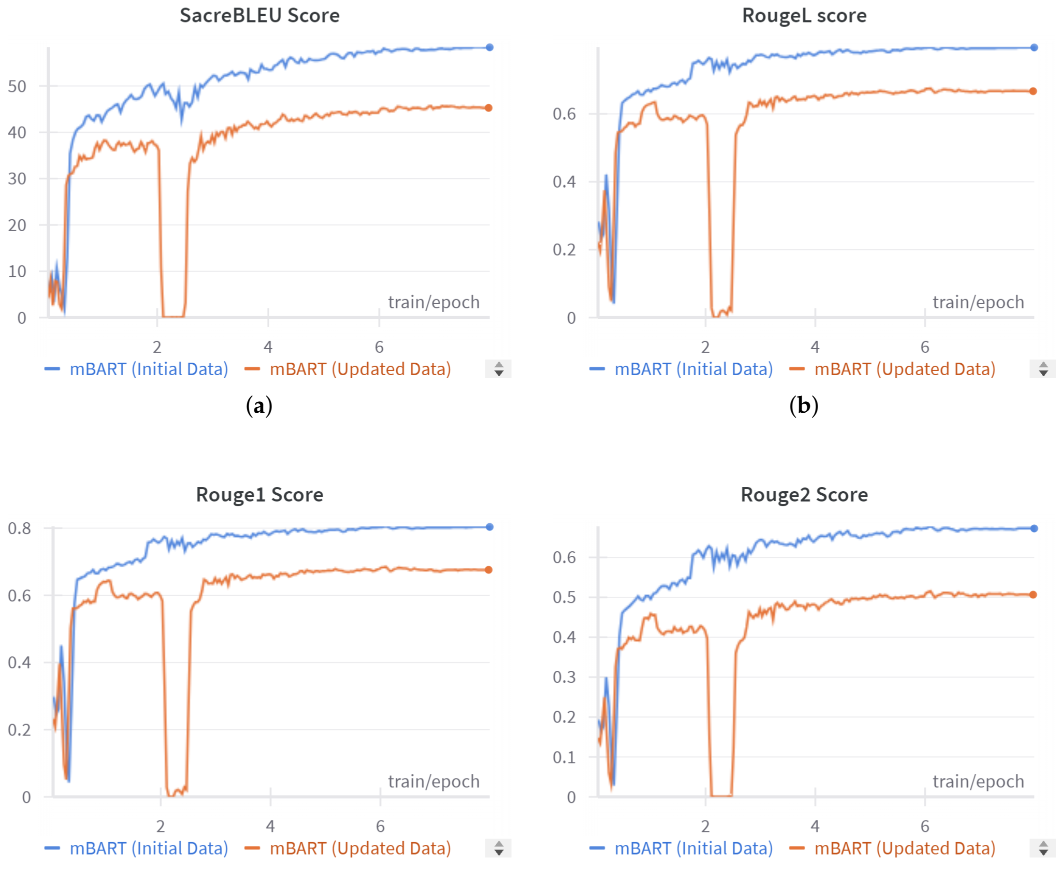
6. Results

6.1. Automatic Evaluation

6.1.1. Fine-Tuning of mBART and mT5

During the fine-tuning process of the mBART model, we evaluated the impact of two different data strategies on model performance. The initial dataset (Parallel Corpus 1) consisted of data where the original sentence had one simplified equivalent. We divided it into training and test sets. In contrast, the updated dataset (Parallel Corpus 2) included data where the original sentence had more than one simplified equivalent and was divided into training, validation, and test sets.

The performance of the model during fine-tuning was monitored using SacreBLEU, ROUGE-1, ROUGE-2, and ROUGE-L metrics across multiple epochs. The results are depicted in Figure 1. SacreBLEU scores, see Figure 1a, show that the model trained on the initial data (Parallel Corpus 1) achieved a higher and more stable performance throughout the epochs compared to the model trained on the updated data (Parallel Corpus 2). The initial data strategy led to faster convergence and higher final scores, which indicated more effective learning and better generalization on the test set.



(**a**) (**b**)

(**c**) (**d**)

Figure 1. Comparison of evaluation metrics for mBART. (**a**) SacreBLEU Score (**b**) Rougel score (**c**) Rouge1 Score (**d**) Rouge2 Score.

The ROUGE-L scores, see Figure 1b, follow a similar pattern. The strategy of fine-tuning the model on Parallel Corpus 1 resulted in higher and more consistent scores over the epochs. The augmented and restructured data (Parallel Corpus 2) caused an initial drop in performance, followed by a gradual improvement, but it did not surpass the performance achieved with Parallel Corpus 1 data. ROUGE-1 and ROUGE-2 scores, see Figure 1c and Figure 1d, respectively, also reflect this trend. The initial sharp decline in the scores of the updated data model indicates that data augmentation and the new training–validation split might have introduced complexity that the model struggled to learn initially.

Similarly, in Figure 2 we can see a similar trend for the mT5 model, which reinforces the fine-tuning pattern observed in mBART, where the model fine-tuned on Parallel Corpus 1 achieved higher and more stable performance in comparison to the model fine-tuned on the updated dataset.

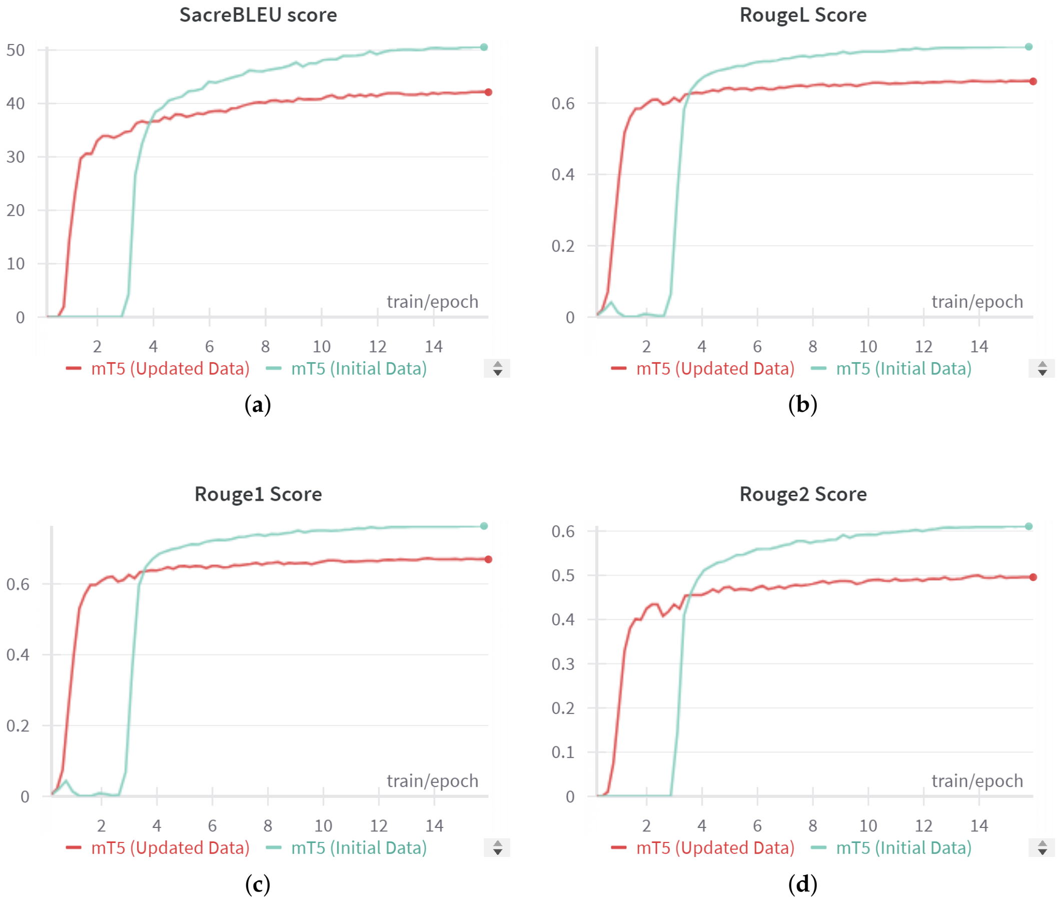


Figure 2. Comparison of evaluation metrics for mT5. (**a**) SacreBLEU Score (**b**) Rougel score (**c**) Rouge1 Score (**d**) Rouge2 Score.

In conclusion, while data augmentation and the introduction of a validation set are common strategies to improve model robustness and performance, in this case, *the initial data without augmentation provided better results*. This suggests that *the quality and distribution of the original training data were sufficient to achieve high performance*, and *the additional complexity introduced by data augmentation did not lead to improved outcomes in this scenario*.

6.1.2. Text Simplification by Fine-Tuned mT5 and mBART

The results of the simplification of Lithuanian administrative texts by mBART and mT5 models are summarized in Table 2. The performance of both models was evaluated using various metrics, including BLEU, SARI, BERTScore, ROUGE, and readability indices such as LIX and RIX. For testing our models, we used our test dataset (see Section 3.)

Table 2. Comparison of mBART and mT5 models for text simplification in Lithuanian language.

Metric	New mBART *	New mT5 *	Old mBART *	Old mT5 *
BLEU	0.2688	0.1592	**0.4196**	0.1650
SARI	57.2374	54.1182	**72.9781**	56.0943
BERTScore	0.8633	0.8342	**0.9155**	0.8498
ROUGE-1	0.6396	0.5931	**0.7797**	0.6205
ROUGE-2	0.4703	0.4323	**0.6753**	0.4652
ROUGE-L	0.5993	0.5593	**0.7555**	0.5875
Original Sentence LIX	76.3475	76.3475	76.3475	76.3475
Original Sentence RIX	9.8409	9.8409	9.8409	9.8409
Human Simplified LIX	68.2252	68.2252	68.2252	68.2252
Human Simplified RIX	7.6104	7.6104	7.6104	7.6104
Model Simplified LIX	**69.7718**	82.5804	70.2625	81.1109
Model Simplified RIX	**7.9385**	15.1725	8.1929	12.5187

* New mBART and mT5 were fine-tuned on Parallel Corpus 2, old mBART and mT5 were fine-tuned on Parallel Corpus 1.

The BLEU score, which measures the similarity between the model output and the reference text in terms of n-gram overlap, was higher for the old mBART (0.4196), fine-tuned on Parallel Corpus 1, compared to the new mBART (0.2688) and both versions of mT5, indicating that the old mBART, n-gram wise, produced outputs more closely aligned with n-grams in the reference simplifications.

SARI, another metric that we used to evaluate the quality of text simplification, was also higher for the old mBART model (72.9781) compared to the new mBART model (57.2374) and both versions of mT5. This suggests that the old mBART was more effective in simplifying the text in terms of words kept, added, and deleted during simplification. Following this tendency, old mBART achieved the highest score of BERTscore as well, outperforming the other three models. This further supports the conclusion that the old mBART generates outputs that are semantically closer to the reference.

ROUGE metrics, which measure the overlap of n-grams between the model outputs and reference texts, were consistently higher for the old mBART across all ROUGE variants: ROUGE-1 (0.7797), ROUGE-2 (0.6753), and ROUGE-L (0.7555). This demonstrates the old mBART's better performance in capturing the relevant information from the source text.

In terms of readability, the *Original Sentence* LIX and RIX scores, calculated for original sentences in both datasets (Parallel Corpus 1 and Parallel Corpus 2) separately, were identical for both models (76.3475 and 9.8409, respectively), reflecting the complexity of the input text. Meanwhile, the *Human Simplified* LIX and RIX scores, calculated for simplifications made by experts, were significantly lower (68.2252 and 7.6104), indicating that the simplified texts were easier to read and understand. Although according to the LIX score, expert simplified sentences still fell into the category of difficult text, the RIX score showed that these sentences approached the universal readability level.

The *Model Simplified* LIX and RIX scores revealed differences in readability in terms of outputs produced by our models. The new mBART achieved a LIX score of 69.7718 and a RIX score of 7.9385, which are closer to the *Human Simplified* scores, indicating better readability compared to the old mBART, which scored 70.2625 for LIX and 8.1929 for RIX. In contrast, the new mT5's scores were higher (82.5804 for LIX and 15.1725 for RIX), suggesting that its outputs were less simplified and more complex. The old mT5 had better readability scores compared to the new mT5, with scores of 81.1109 for LIX and 12.5187 for RIX.

Overall, the results indicate that *old mBART outperforms new mBART and both versions of mT5 in most metrics for simplification of Lithuanian administrative texts, providing outputs that are more accurate, semantically more similar to reference sentences, and easier to read*. However, the *new mBART shows improved readability over the old mBART in terms of LIX and RIX scores*.

6.2. Qualitative Evaluation

6.2.1. mT5 and mBART

To assess the results of the mT5 and mBART text simplification models, two linguists, familiar with Plain Lithuanian, independently evaluated $n = 554$ model-simplified sentences, comparing them to the original ones in the test set, according to three criteria, indicated in Table 4. To assess evaluators' agreement in assigning their scores, inter-rater reliability has been evaluated via Cohen's Kappa scores [90].

As Table 3 shows, while the criteria of meaning retention and grammaticality demonstrate a fair agreement between the two raters, simplicity, i.e., how well the sentence that has been simplified by mT5 and mBART models correspond to Plain Lithuanian, only shows very slight inter-rater reliability. The raters discussed all the criteria and the approximate significance of each 1 to 5 score before rating the simplified sentences; however, it seems that while it is fairly straightforward to agree on whether the sentence is grammatical, whether the meaning is retained, and whether it is simpler than the original, assigning a uniform rank from 1 to 5 for simplicity generated more discrepancy. Plain Lithuanian is a new language variety even for the experts, who are very familiar with the Plain Language rules.

Table 3. Cohen's Kappa scores.

Criterion	Score
Simplicity	0.1006
Meaning retention	0.3035
Grammaticality	0.3338

Regarding the results of the qualitative assessment, Table 4 shows that mBART scored consistently better on all three criteria. A closer qualitative analysis revealed that both models did comparatively well with shorter sentences to simplify, but differed considerably in their treatment of long sentences, which are characteristic of the Lithuanian formal and administrative language registers [91]. Notably, mT5 at times did not do well with longer passages, as it had a tendency to output nonsensical repetitions of words instead of coherent sentences or deleting whole passages from the original text, instead of transforming them into new shorter sentences. This lowered its scores considerably on all three criteria.

Table 4. Qualitative evaluation scores (averages).

	Simplicity	Meaning Retention	Grammaticality
mT5	2.67	3.16	3.33
mBART	2.80	3.72	3.88

According to evaluation results, *both models scored better on meaning retention and grammaticality than on simplification*. This resulted in sentences that were only slightly simplified, while some elements were left unaltered that were in need of simplification. Such partly performed simplifications were assigned a simplicity score of 2. These unaltered or mostly unaltered sentences were, naturally, grammatically correct and retained the full meaning of the original sentence, resulting in scores of 5.

However, the raters noticed that *both models had learned certain linguistic structures, most characteristic of Plain Lithuanian, such as adding subject and possessive pronouns as well as demonstratives and transforming the passive voice into the active voice*. The latter may pose problems with subject identification if the original sentence in passive did not include an oblique agent phrase, but it did not turn out to be a frequent issue, as in most cases, adding a first-person plural pronoun in subject position was sufficient. Finally, *mBART especially managed quite well to transform participial clauses into relative clauses with finite verbs.*

On the downside, there were very few sentences that got full scores of 5-5-5. Thus, *the models may need more and/or more diverse training data, before they can be considered reliable.*

6.2.2. ChatGPT

To test out ChatGPT (GPT-4o) for simplification of Lithuanian administrative texts, we created a prompt by providing it with a list of 18 Plain Lithuanian rules, each followed by a few example pairs of sentences (original and simplified according to the given rule).

The rules provided clear indications on what syntactic structures and lexical elements should be replaced, and with what, as well as how sentences should be split (if they exceed a certain number of words when there are embedded relative clauses). However, our experiment showed that, while ChatGPT responded quite well to a simple and short prompt to simplify the given text without any further elaboration on what the simplification entails, it ignored most of the rules given in a more elaborate prompt created for our purposes. ChatGPT struggled to follow complex text simplification rules for several reasons, including unequal amount of training data in the different languages, the fact that the main language of training was English. English is an analytical language, while Lithuanian is a synthetic language; thus, they are very different in terms of their grammatical structures and the intrinsic structural constraints, among others.

As our Plain Language guidelines provided indications on grammatical structures of Lithuanian, it soon turned out that ChatGPT ignored them completely. While the basic concepts of English grammar do not pose any problems to ChatGPT, the same notion cannot be applied to Lithuanian. For example, ChatGPT could not correctly recognize the passive voice in Lithuanian, let alone transform it into the active voice. This might be related to a general lack of resources for Lithuanian. Related studies showed that with minority languages, ChatGPT's metalinguistic awareness is doubtful [92].

Hence, *ChatGPT could be used for non-professional purposes to summarize or simplify a variety of texts, but our experiment shows that it is not reliable enough to use for the simplification of administrative texts, if these texts are to be published on public institution websites.*

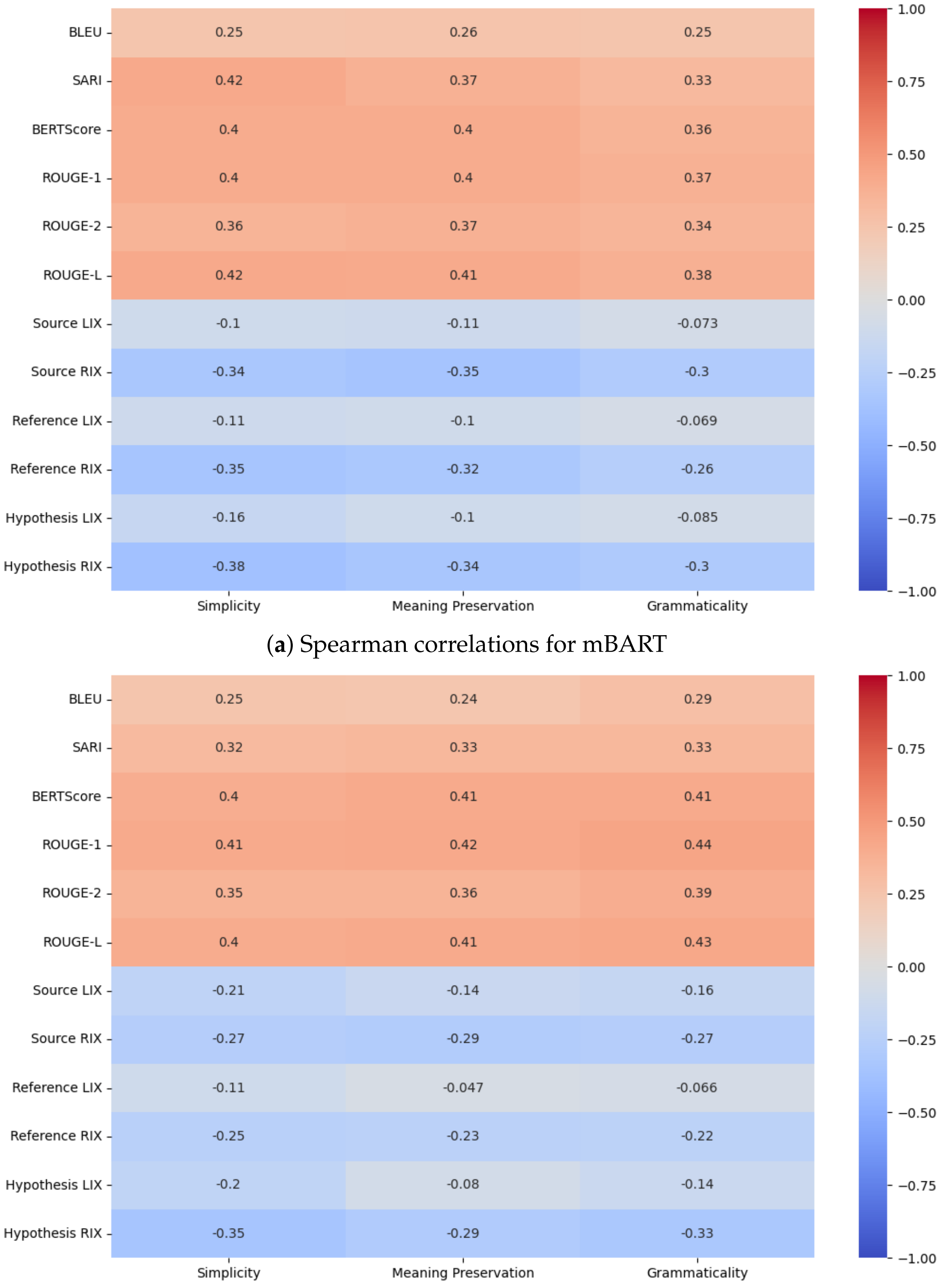
6.3. Correlation

To evaluate the effectiveness of our mBART and mT5 models in simplifying Lithuanian text, we used a combination of automatic metrics and human assessments. Spearman's rank correlation coefficient was particularly appropriate for our analysis compared to other correlation measures, such as Pearson's correlation, because it does not assume a linear relationship between variables, making it well-suited for handling our ordinal human evaluation scores. Additionally, it is less sensitive to outliers, which is beneficial given the inherent variability in human judgments [93].

Spearman correlation coefficients between automatic evaluation metrics and human scores (Figure 3) across three dimensions of text simplification (simplicity, meaning retention, and grammaticality). In both heatmaps, warmer colors indicate stronger positive correlations, while cooler colors represent negative correlations.

For the mBART model, BERTScore shows moderate positive correlations with all three human evaluation criteria, around 0.40, indicating a reasonable alignment with human judgments on simplicity, meaning preservation, and grammaticality. ROUGE metrics also displayed moderate positive correlations, with ROUGE-L having the highest correlations, up to 0.42 for simplicity. SARI also performed well, particularly in simplicity and meaning preservation, with correlations of 0.42 and 0.37, respectively.

Conversely, BLEU exhibited lower correlations, around 0.25, suggesting it might not be as reliable for this specific task. Readability indices (LIX and RIX) generally showed negative correlations, indicating that higher readability scores did not necessarily correspond to better human-judged text quality. This may have happened because readability indices focus on surface-level features like sentence length and word difficulty, which do not fully capture the nuanced aspects that human evaluators prioritize, such as meaning preservation and grammaticality. Human evaluations also consider the overall fluency of the text, an aspect that readability metrics are not designed to measure, making it difficult to directly compare these indices with human judgments.



(**a**) Spearman correlations for mBART

(**b**) Spearman correlations for mT5

Figure 3. Comparison of Spearman correlations between mBART and mT5 evaluation metrics and human scores.

The mT5 model exhibited a similar trend, with BERTScore and ROUGE metrics showing moderate to strong positive correlations with human evaluations. ROUGE-1, in particular, had the highest correlation with grammaticality at 0.44. Meanwhile, SARI showed slightly lower correlations compared to mBART, particularly in simplicity. BLEU again showed relatively low correlations. The readability indices for the mT5 model mirrored the results seen with mBART, with negative correlations, suggesting that these metrics might not effectively reflect human judgments in this context.

In summary, our analysis revealed that while automatic metrics like BERTScore and ROUGE align moderately well with human evaluations, others, like BLEU and readability indices, show lower or even negative correlations. This indicates that while automatic metrics are useful, they have limitations and should be complemented with human evaluations to ensure the nuanced aspects of text simplification are adequately captured.

7. Discussion

Our study on the simplification of Lithuanian administrative texts to plain language brings several findings in terms of model performance, data augmentation, evaluation metrics, and the capabilities of different language models.

Contrary to common expectations, data augmentation did not improve model performance in our case. The original dataset produced superior results, suggesting that the quality and distribution of the initial training data were sufficient for the current configuration of simplification experiments. This finding reinforces the importance of high-quality, representative data in text simplification tasks, particularly for less-resourced languages like Lithuanian.

While automated metrics such as BLEU and ROUGE are widely used, our findings support their limitations in capturing the nuances of text simplification. Therefore, human evaluators proved to be crucial in assessing grammaticality, semantic accuracy, and simplification. This aligns with broader discussions in the NLP community about the need for more comprehensive evaluation methods that combine automated metrics with human judgment [94–96].

However, analysis of inter-rater reliability revealed fair agreement on meaning retention and grammaticality, but only slight agreement on simplicity. This discrepancy highlights the subjective nature of assessing simplification, particularly in the context of Plain Lithuanian, which is a relatively new example of tailored language. Future research should focus on developing more standardized criteria for evaluating simplicity in plain language contexts.

In our analysis of the relationship between automatic metrics and human scores, we observed moderate positive correlations for BERTScore and ROUGE metrics, particularly for the mBART model. SARI also showed promise, especially in assessing simplicity and meaning preservation. However, BLEU exhibited lower correlations, suggesting its limitations for this specific task. Interestingly, readability indices (LIX and RIX) showed negative correlations with human evaluations, highlighting a discrepancy between surface-level textual features and human judgment of simplified text quality.

Furthermore, both fine-tuned models, especially mBART, while mT5 to a lesser extent, demonstrated adequate results in applying certain linguistic structures characteristic of Plain Lithuanian, such as adding subject and possessive pronouns and transforming passive voice to active. However, both models struggled with consistently achieving full simplification scores, indicating a need for more diverse training data. The mT5 model, in particular, struggled with longer passages, sometimes producing incoherent outputs or omitting significant portions of essential information.

Our experiments with ChatGPT revealed limitations in its ability to apply specific linguistic rules for Lithuanian text simplification. While ChatGPT performed adequately with simple prompts, it struggled to incorporate more complex, language-specific rules. This aligns with findings from related studies on ChatGPT's limited metalinguistic awareness for less-resourced languages [92,97,98], which are usually less supported in large language models.

To summarize, our findings support the need for task-specific and language-specific evaluation metrics that would better align with human judgments. Also, the importance of human evaluation in combination with automated metrics for a more comprehensive assessment of text simplification quality has been stressed. Finally, our results highlight the necessity for larger, more diverse training datasets to improve model performance on complex sentences and ensure consistent simplification.

8. Conclusions and Future Work

In this paper, we presented experiments on text simplification for the Lithuanian language, specifically targeting the simplification of administrative-style texts into plain language to improve their accessibility to the general public. We fine-tuned mT5 and mBART for this task and also explored the performance of ChatGPT. The models' outputs were evaluated using both quantitative metrics (BLEU, ROUGE, SARI, and BERTscore) and qualitative assessments (focused on simplicity, meaning retention, and grammaticality). Among the models tested, mBART proved to be the most effective, achieving the highest scores across all quantitative measures. These findings were further supported by the results of the qualitative evaluation. Our experiments with ChatGPT revealed adequate results when concise, straightforward prompts were given for text simplification. However, ChatGPT tended to disregard a significant portion of more complex prompts with simplification rules and examples. Furthermore, BERTScore and ROUGE demonstrated moderate alignment with human evaluations, whereas BLEU and readability scores exhibited lower or even negative correlations.

Our future work will focus on improving model performance via different fine-tuning techniques and more comprehensive experimentation with training parameters. We also plan to experiment with data size and diversity to enhance the models' performance and generalizability. Additionally, we aim to conduct a more thorough analysis of the models' decision-making processes, particularly concerning factual accuracy and potential biases.

Author Contributions: Conceptualization, J.M. and D.K.K.; Methodology, J.M., D.K.K. and D.K.; Software, E.R.; Validation, D.K.K. and D.K.; Writing—original draft, J.M.; Writing—review & editing, D.K. and T.K.; Supervision, T.K.; Project administration, J.M.; Funding acquisition, J.M. and T.K. All authors have contributed equally to the publication. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by THE RESEARCH COUNCIL OF LITHUANIA (LMTLT), grant agreement No. S-LIP-22-77.

Data Availability Statement: The data used for fine-tuning the models will be available by request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Štajner, S. Automatic text simplification for social good: Progress and challenges. In Proceedings of the Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, 1–6 August 2021; pp. 2637–2652.
2. François, T.; Müller, A.; Rolin, E.; Norré, M. AMesure: A Web platform to assist the clear writing of administrative texts. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations, Online Event, 4–7 December 2020; pp. 1–7.
3. Adler, M. The Plain Language Movement. In *The Oxford Handbook of Language and Law*; Solan, L.M., Tiersma, P.M., Eds.; Oxford University Press: Oxford, UK, 2012.
4. Maaß, C. *Easy Language–Plain Language–Easy Language Plus: Balancing Comprehensibility and Acceptability*; Frank & Timme: Berlin, Germany, 2020.
5. Rennes, E.; Jönsson, A. A tool for automatic simplification of Swedish texts. In Proceedings of the 20th Nordic Conference of Computational Linguistics, Linköping University Electronic Press: Linköping, Sweden, 11–13 May 2015; pp. 317–320.
6. Suter, J.; Ebling, S.; Volk, M. Rule-based Automatic Text Simplification for German. *Boch. Linguist. Arbeitsberichte* **2016**, 279. [CrossRef]
7. Štajner, S.; Saggion, H. Data-driven text simplification. In Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts, Santa Fe, NM, USA, 20–26 August 2018 pp. 19–23.
8. Srikanth, N.; Li, J.J. Elaborative simplification: Content addition and explanation generation in text simplification. *arXiv Prepr.* **2020**, arXiv:2010.10035.
9. Al-Thanyyan, S.S.; Azmi, A.M. Automated Text Simplification: A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–36. [CrossRef]

10. Swain, D.; Tambe, M.; Ballal, P.; Dolase, V.; Agrawal, K.; Rajmane, Y. Lexical text simplification using WordNet. In Proceedings of the Advances in Computing and Data Sciences: Third International Conference, ICACDS 2019, Ghaziabad, India, 12–13 April 2019; Revised Selected Papers, Part II 3; Springer: Berlin/Heidelberg, Germany, 2019; pp. 114–122.
11. Alarcon, R.; Moreno, L.; Martínez, P. Lexical simplification system to improve web accessibility. *IEEE Access* **2021**, *9*, 58755–58767. [CrossRef]
12. Qiang, J.; Li, Y.; Zhu, Y.; Yuan, Y.; Wu, X. LSBert: A simple framework for lexical simplification. *arXiv Prepr.* **2020**, arXiv:2006.14939.
13. Kumar, A.P.; Nayak, A.; Shenoy, M.; Manoj, R.J.; Priyadarshi, A. Pattern-based syntactic simplification of compound and complex sentences. *IEEE Access* **2022**, *10*, 53290–53306. [CrossRef]
14. Štajner, S.; Popović, M. Automated text simplification as a preprocessing step for machine translation into an under-resourced language. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; pp. 1141–1150.
15. Vu, T.; Hu, B.; Munkhdalai, T.; Yu, H. Sentence simplification with memory-augmented neural networks. *arXiv Prepr.* **2018**, arXiv:1804.07445.
16. Agrawal, S.; Carpuat, M. Controlling Text Complexity in Neural Machine Translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 1549–1564.
17. Mallinson, J.; Sennrich, R.; Lapata, M. Zero-shot crosslingual sentence simplification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: Kerrville, TX, USA, 2020.
18. Sakakini, T.; Lee, J.Y.; Duri, A.; Azevedo, R.F.; Sadauskas, V.; Gu, K.; Bhat, S.; Morrow, D.; Graumlich, J.; Walayat, S.; et al. Context-aware automatic text simplification of health materials in low-resource domains. In Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Online Event, 20 November 2020; pp. 115–126.
19. de Lima, T.B.; Nascimento, A.C.; Valença, G.; Miranda, P.; Mello, R.F.; Si, T. Portuguese neural text simplification using machine translation. In Proceedings of the Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, 29 November–3 December 2021; Proceedings, Part II 10; Springer: Berlin/Heidelberg, Germany, 2021; pp. 542–556.
20. Botarleanu, R.M.; Dascalu, M.; Crossley, S.A.; McNamara, D.S. Sequence-to-sequence models for automated text simplification. In Proceedings of the Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, 6–10 July 2020; Proceedings, Part II 21; Springer: Berlin/Heidelberg, Germany, 2020; pp. 31–36.
21. Maddela, M.; Alva-Manchego, F.; Xu, W. Controllable text simplification with explicit paraphrasing. *arXiv Prepr.* **2020**, arXiv:2010.11004.
22. Ulčar, M.; Robnik-Šikonja, M. Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Front. Artif. Intell.* **2023**, *6*, 932519. [CrossRef]
23. Dmitrieva, A.; Tiedemann, J. Towards Automatic Finnish Text Simplification. In Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024, Torino, Italy, 21 May 2024; pp. 39–50.
24. Zhao, S.; Meng, R.; He, D.; Andi, S.; Bambang, P. Integrating transformer and paraphrase rules for sentence simplification. *arXiv Prepr.* **2018**, arXiv:1810.11193.
25. Omelianchuk, K.; Raheja, V.; Skurzhanskyi, O. Text Simplification by Tagging. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, Online Event, 20 April 2021; pp. 11–25.
26. Jeblick, K.; Schachtner, B.; Dexl, J.; Mittermeier, A.; Stüber, A.T.; Topalis, J.; Weber, T.; Wesp, P.; Sabel, B.O.; Ricke, J.; et al. ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *Eur. Radiol.* **2023**, *34*, 2817–2825. [CrossRef] [PubMed]
27. Sun, R.; Xu, W.; Wan, X. Teaching the Pre-trained Model to Generate Simple Texts for Text Simplification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 9345–9355. [CrossRef]
28. Huang, C.Y.; Wei, J.; Huang, T.H. Generating Educational Materials with Different Levels of Readability using LLMs. *arXiv Prepr.* **2024**, arXiv:2406.12787.
29. Agrawal, S.; Carpuat, M. Controlling Pre-trained Language Models for Grade-Specific Text Simplification. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 12807–12819.
30. Qiang, J.; Li, Y.; Zhu, Y.; Yuan, Y.; Wu, X. Lexical simplification with pretrained encoders. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8649–8656.
31. Alissa, S.; Wald, M. Text simplification using transformer and BERT. *Comput. Mater. Contin.* **2023**, *75*, 3479–3495. [CrossRef]
32. Sheang, K.C.; Saggion, H. Controllable sentence simplification with a unified text-to-text transfer transformer. In Proceedings of the 14th International Conference on Natural Language Generation (INLG), Aberdeen, UK, 20–24 September 2021; Association for Computational Linguistics: Kerrville, TX, USA, 2021.
33. Seidl, T.; Vandeghinste, V. Controllable Sentence Simplification in Dutch. *Comput. Linguist. Neth. J.* **2024**, *13*, 31–61.
34. Monteiro, J.; Aguiar, M.; Araújo, S. Using a pre-trained SimpleT5 model for text simplification in a limited corpus. In Proceedings of the Working Notes of CLEF 2022, Bologna, Italy, 5–8 September 2022.

35. Schlippe, T.; Eichinger, K. Multilingual Text Simplification and its Performance on Social Sciences Coursebooks. In *Proceedings of the International Conference on Artificial Intelligence in Education Technology*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 119–136.
36. Devaraj, A.; Wallace, B.C.; Marshall, I.J.; Li, J.J. Paragraph-level simplification of medical texts. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online Event, 6–11 June 2021; Volume 2021, pp. 4972–4984.
37. Vásquez-Rodríguez, L.; Shardlow, M.; Przybyła, P.; Ananiadou, S. Document-level Text Simplification with Coherence Evaluation. In Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability associated with RANLP-2023, Varna, Bulgaria, 7 September 2023; pp. 85–101.
38. Wen, Z.; Fang, Y. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 506–516.
39. Deilen, S.; Hern'andez Garrido, S.; Lapshinova-Koltunski, E.; Maaß, C. Using ChatGPT as a CAT tool in Easy Language translation. In Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability associated with RANLP-2023, Varna, Bulgaria, 7 September 2023; Štajner, S., Saggio, H., Shardlow, M., Alva-Manchego, F., Eds.; INCOMA Ltd.: Shoumen, Bulgaria, 2023; pp. 1–10.
40. Li, Z.; Shardlow, M.; Alva-Manchego, F. Comparing Generic and Expert Models for Genre-Specific Text Simplification. In Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability, Varna, Bulgaria, 7 September 2023; Štajner, S., Saggio, H., Shardlow, M., Alva-Manchego, F., Eds.; INCOMA Ltd.: Shoumen, Bulgaria, 2023; pp. 51–67.
41. Ayre, J.; Mac, O.; McCaffery, K.; McKay, B.R.; Liu, M.; Shi, Y.; Rezwan, A.; Dunn, A.G. New frontiers in health literacy: Using ChatGPT to simplify health information for people in the community. *J. Gen. Intern. Med.* **2024**, *39*, 573–577. [CrossRef]
42. Sudharshan, R.; Shen, A.; Gupta, S.; Zhang-Nunes, S. Assessing the Utility of ChatGPT in Simplifying Text Complexity of Patient Educational Materials. *Cureus* **2024**, *16*, e55304. [CrossRef]
43. Tariq, R.; Malik, S.; Roy, M.; Islam, M.Z.; Rasheed, U.; Bian, J.; Zheng, K.; Zhang, R. Assessing ChatGPT for Text Summarization, Simplification and Extraction Tasks. In Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), Houston, TX, USA, 26–29 June 2023; pp. 746–749. [CrossRef]
44. Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; Wu, Y. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv* **2023**, arXiv:2301.07597.
45. Doshi, R.; Amin, K.S.; Khosla, P.; Bajaj, S.; Chheang, S.; Forman, H.P. Utilizing Large Language Models to Simplify Radiology Reports: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, Google Bard, and Microsoft Bing. *medRxiv* **2023**. [CrossRef]
46. Blinova, S.; Zhou, X.; Jaggi, M.; Eickhoff, C.; Bahrainian, S.A. SIMSUM: Document-level Text Simplification via Simultaneous Summarization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 9927–9944.
47. Colas, A.; Ma, H.; He, X.; Bai, Y.; Wang, D.Z. Can Knowledge Graphs Simplify Text? In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023; pp. 379–389.
48. Raheja, V.; Alikaniotis, D.; Kulkarni, V.; Alhafni, B.; Kumar, D. mEdIT: Multilingual Text Editing via Instruction Tuning. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024; pp. 979–1001.
49. Saini, A.; Chernodub, A.; Raheja, V.; Kulkarni, V. Spivavtor: An Instruction Tuned Ukrainian Text Editing Model. In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING 2024, Torino, Italy, 25 May 2024; pp. 95–108.
50. Tran, H.; Yao, Z.; Li, L.; Yu, H. ReadCtrl: Personalizing text generation with readability-controlled instruction learning. *arXiv Prepr.* **2024**, arXiv:2406.09205.
51. Basu, C.; Vasu, R.; Yasunaga, M.; Yang, Q. Med-EASi: Finely annotated dataset and models for controllable simplification of medical texts. In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, Cambridge, MA, USA, 7–14 February 2023; AAAI Press: Washington, DC, USA, 2023; AAAI'23/IAAI'23/EAAI'23. [CrossRef]
52. Kaur, P.; Kashyap, G.S.; Kumar, A.; Nafis, M.T.; Kumar, S.; Shokeen, V. From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility. *arXiv Prepr.* **2024**, arXiv:2402.16142.
53. Sun, R.; Yang, Z.; Wan, X. Exploiting Summarization Data to Help Text Simplification. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 39–51.
54. Joseph, S.; Kazanas, K.; Reina, K.; Ramanathan, V.; Xu, W.; Wallace, B.; Li, J.J. Multilingual Simplification of Medical Texts. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 16662–16692. [CrossRef]
55. Flores, L.J.Y.; Huang, H.; Shi, K.; Chheang, S.; Cohan, A. Medical Text Simplification: Optimizing for Readability with Unlikelihood Training and Reranked Beam Search Decoding. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 4859–4873.

56. Yamaguchi, D.; Miyata, R.; Shimada, S.; Sato, S. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 359–375.
57. Chatterjee, N.; Agarwal, R. Studying the Effect of Syntactic Simplification on Text Summarization. *IETE Tech. Rev.* **2022**, *40*, 155–166. [CrossRef]
58. Cripwell, L.; Legrand, J.; Gardent, C. Document-Level Planning for Text Simplification. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 993–1006.
59. Ivchenko, O.; Grabar, N. Impact of the Text Simplification on Understanding. *Stud. Health Technol. Informatics* **2022**, *294*, 634–638. [CrossRef]
60. Wang, J. Research on Text Simplification Method Based on BERT. In *2022 7th International Conference on Multimedia Communication Technologies (ICMCT), Xiamen, China, 7–9 July 2022*; IEEE Computer Society: Washington, DC, USA, 2022; pp. 78–81. [CrossRef]
61. Corti, L.; Yang, J. ARTIST: ARTificial Intelligence for Simplified Text. *arXiv Prepr.* **2023**, arXiv:2308.13458.
62. Scarton, C.; Paetzold, G.; Specia, L. Simpa: A sentence-level simplification corpus for the public administration domain. In Proceedings of the LREC 2018, Miyazaki, Japan, 7–12 May 2018.
63. Schwarzer, M.; Tanprasert, T.; Kauchak, D. Improving human text simplification with sentence fusion. In Proceedings of the 15th Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), Mexico City, Mexico, 11 June 2021; pp. 106–114.
64. Harris L.; Kleimann S.; Mowat C. Setting plain language standards. *Clarity J. Lansing* **2010**, *64*, 16–25.
65. Martinho, M. International standard for clarity—We bet this works for all languages. *Clarity J.* **2018**, *79*, 17–20.
66. Brunato, D.; Dell'Orletta, F.; Venturi, G.; Montemagni, S. Design and Annotation of the First Italian Corpus for Text Simplification. In Proceedings of the 9th Linguistic Annotation Workshop, Denver, CO, USA, 5 June 2015; pp. 31–41. [CrossRef]
67. Dębowski, Ł.; Broda, B.; Nitoń, B.; Charzyńska, E. Jasnopis—A program to compute readability of texts in Polish based on psycholinguistic research. In *Natural Language Processing and Cognitive Science*; Liberia Editrice Cafoscarina: Venezia, Italy, 2015; pp. 51–61.
68. Utka, A. *Dažninis Rašytinės Lietuvių Kalbos žOdynas*; Vytautas Magnus University Press: Kaunas, Lithuania, 2009.
69. Kale, M.; Rastogi, A. Text-to-Text Pre-Training for Data-to-Text Tasks. In Proceedings of the 13th International Conference on Natural Language Generation, Dublin, Ireland, 15–18 December 2020; pp. 97–102.
70. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online Event, 6–11 June 2021; pp. 483–498.
71. Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; Gardner, M. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv Prepr.* **2021**, arXiv:2104.08758.
72. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762
73. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**. arXiv:1910.13461.
74. Wenzek, G.; Lachaux, M.A.; Conneau, A.; Chaudhary, V.; Guzmán, F.; Joulin, A.; Grave, E. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. *arXiv* **2019**. arXiv:1911.00359.
75. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual Denoising Pre-training for Neural Machine Translation. *arXiv* **2020**, arXiv:2001.08210. [CrossRef]
76. Yenduri, G.; Ramalingam, M.; Chemmalar Selvi, G.; Supriya, Y.; Srivastava, G.; Maddikunta, P.K.R.; Deepti Raj, G.; Jhaveri, R.H.; Prabadevi, B.; Wang, W.; et al. GPT (Generative Pre-trained Transformer)—A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access* **2024**, *12*, 54608–54649. [CrossRef]
77. Rothman, D. *Transformers for Natural Language Processing: Build, Train, and Fine-Tune Deep Neural Network Architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4*; Packt Publishing Ltd.: Birmingham, UK, 2022.
78. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the ACL 2002, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
79. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 26–27 July 2004; pp. 74–81.
80. Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; Callison-Burch, C. Optimizing Statistical Machine Translation for Text Simplification. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 401–415. [CrossRef]
81. Alva-Manchego, F.; Martin, L.; Scarton, C.; Specia, L. EASSE: Easier Automatic Sentence Simplification Evaluation. In Proceedings of the EMNLP-IJCNLP 2019 (Demo Session), Hong Kong, China, 3–7 November 2019; pp. 49–54.
82. Ribeiro, V.H.A.; Cavalin, P.; Morais, E. A Dynamic Multi-criteria Multi-engine Approach for Text Simplification. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Online Event, 18–22 July 2021; pp. 1–8.
83. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
84. Ondov, B.; Attal, K.; Demner-Fushman, D. A survey of automated methods for biomedical text simplification. *J. Am. Med Inform. Assoc.* **2022**, *29*, 1976–1988. [CrossRef] [PubMed]

85. Arshad, M.; Yousaf, M.M.; Sarwar, S.M. Comprehensive readability assessment of scientific learning resources. *IEEE Access* **2023**, *11*, 53978–53994. [CrossRef]
86. Post, M. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, 31 October–1 November 2018; pp. 186–191.
87. Nisioi, S.; Štajner, S.; Ponzetto, S.P.; Dinu, L.P. Exploring neural text simplification models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 85–91.
88. Alva-Manchego, F.; Scarton, C.; Specia, L. Data-driven sentence simplification: Survey and benchmark. *Comput. Linguist.* **2020**, *46*, 135–187. [CrossRef]
89. Grabar, N.; Saggion, H. Evaluation of Automatic Text Simplification: Where are we now, where should we go from here. In Proceedings of the Traitement Automatique des Langues Naturelles. ATALA, Avignon, France, 27 June–1 July 2022; pp. 453–463.
90. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282. [CrossRef]
91. Vladarskienė, R. Lietuvių bendrinės ir administracinės kalbos santykis. *Bendrinė Kalba (Iki 2014 Metų–Kalbos KultūRa)* **2007**, *80*, 55–63.
92. Massaro, A.; Samo, G. Prompting Metalinguistic Awareness in Large Language Models: ChatGPT and Bias Effects on the Grammar of Italian and Italian Varieties. *Verbum* **2023**, *14*, 4. [CrossRef]
93. Hauke, J.; Kossowski, T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* **2011**, *30*, 87–93. [CrossRef]
94. Sai, A.B.; Dixit, T.; Sheth, D.Y.; Mohan, S.; Khapra, M.M. Perturbation CheckLists for Evaluating NLG Evaluation Metrics. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 7219–7234.
95. Zhang, S.; Bansal, M. Finding a Balanced Degree of Automation for Summary Evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6617–6632.
96. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 2511–2522.
97. Lamprinidis, S. LLM cognitive judgements differ from human. In Proceedings of the International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications, Athens, Greece, 25–26 September 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 17–23.
98. Leivada, E.; Dentella, V.; Murphy, E. The Quo Vadis of the Relationship between Language and Large Language Models. *arXiv e-prints* **2023**, arXiv:2310.11146

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.