Extending ASR Systems Error Measurements: Reporting LEXICAL and GRAMMATICAL Errors

Simon Gonzalez, Jason Littlefield, Tao Hoang, Maria Kim, Tim Cawley, Jennifer Biggs

Defence Science and Technology Group

simon.gonzalez@defence.gov.au, jason.littlefield@defence.gov.au,
tao.hoang@defence.gov.au, myung.kim@defence.gov.au, tim.cawley@defence.gov.au,
jennifer.biggs@defence.gov.au

Abstract

We address the limitations of the current Automatic Speech Recognition evaluation metric Word Error Rate. While used for broad assessment, it lacks the granularity to discern errors in specific linguistic categories. We offer a metric based on parts of speech and grammatical categories. Using the Whisper ASR system on English, Japanese, and Spanish, within the CommonVoice 15, we analyse GRAMMATICAL and LEXICAL error rates. Results show that GRAMMATICAL words trigger less errors than LEXICAL words, and case markers combined with LEXICAL words in Japanese, trigger higher accuracy. Our approach enhances the explanatory power of error analysis in ASR system performance.

Index Terms: Automatic Speech Recognition, error metrics, parts of speech, lexical and grammatical evaluation

1. Introduction

Automatic Speech Recognition (ASR) technologies have undergone significant advancements [1][2] and the widespread adoption of ASR systems in various industries (e.g., Healthcare, Defence and Automotive) highlight the critical role of accurate evaluation to ensure their effectiveness, reliability, and user satisfaction. However, evaluating the performance of ASR systems remains a challenging task [3]. Traditional rate measurements, such as the widely used Word Error Rate (WER), offer valuable insights into overall system performance. WER is calculated as the ratio of the total number of errors – comprising substitutions, deletions, and insertions in the transcription output – to the number of words in the audio signal input to the ASR system [4]. But WER has been reported to have limitations [5]. The primary limitation of WER lies in treating all errors uniformly, without distinguishing between those fully detrimental to the meaning of the ground truth reference and those with closer semantic or syntactic relevance. Furthermore, WER cannot gauge the relative importance of specific words in the ground truth transcription, prompting the proposal of alternative metrics that account for semantics [6], entity recognition [7], and parts of speech [8]. Previous research also indicates that WER does not consistently correlate with human judgment on ASR system performance [3][9]. These findings underscore the need for linguistic metrics offering a more detailed understanding of errors, moving beyond the holistic view currently provided. As the field of ASR matures, there is a growing recognition that a more detailed analysis of errors is imperative for refining and advancing these technologies. One limitation of existing metrics is their inability to unveil the specific nature of errors. While systems may exhibit similar overall error rates, this metric fails to elucidate whether these errors target distinct linguistic categories. For example, two ASR systems may boast comparable WERs, yet one might prove more detrimental to GRAMMATICAL words, while the other, with a seemingly identical error rate, might manifest more errors in LEXICAL or content words. To address this deficiency, there is a pressing need to delve deeper into the intricacies of errors by examining linguistic categories, thereby shedding light on the distinct areas of vulnerability within ASR systems [10][11].

Recognizing the limitations of current approaches, we propose the integration of linguistic metrics in the evaluation of ASR systems. An in-depth analysis based on linguistic categories, including parts of speech and grammatical classifications, can help in understanding the complexities of errors. By categorizing errors according to linguistic attributes, valuable insights can be gained into the nature of errors and how they behave within the context of these systems. This linguistic perspective not only brings clarity to the nature of errors but also enhances the explanatory power of error analysis in ASR, providing a more comprehensive understanding of system performance.

The purpose of our current approach is to present methodology for analyzing and reporting errors in ASR systems. Taking a multilingual approach, we analyze errors in English, Japanese, and Spanish, leveraging the Whisper ASR system [12] on the CommonVoice 15 dataset [13] described in Section 3 below. Here, we propose analyzing errors based on word classes, grouped into two major categories: GRAMMATICAL words and LEXICAL words. Utilizing Parts of Speech tagging (POS) (described in Section 3.4.2), we measure GRAMMATICAL/LEXICAL errors, breaking down errors into two distinct values: one for GRAMMATICAL or function words and another for LEXICAL or content words. This approach aims to provide a more detailed and informative perspective on the performance of ASR systems, catering to the need for specific error analysis relevant for speech recognition technologies.

2. Previous work and current approach

There have been important advancements in methodologies for measuring ASR performance, including the adoption of various error metrics. Recent efforts within ASR evaluation have shifted focus towards metrics that go beyond word counts, such as *word embeddings* [14], *sentence embeddings* [15], and *semantic proximity* [16]. Taking inspiration from machine translation, where linguistic metrics have proven instrumental in refining translation quality, this paper extends the discourse to ASR. Works such as [17] successfully integrated linguistic attributes, like parts of speech, into translation evaluation, extending beyond WER and BLEU scores. They proposed the

Position Independent Error Rate (PER) across different parts of speech, estimating the contribution of each POS class to the overall word error rate. Although their work involved errors based on POS in two languages, English and Spanish, and compared them to human assessors, the remaining question is whether these findings can be generalized to other languages with different typological characteristics. Our goal is then to build from this by developing metrics within a single widely used ASR system, facilitating a comparison between the reference (**REF**) and the hypothesized text (**HYP**) produced by the ASR system across multiple languages with major typological differences.

We aim to improve upon the work of [8], who proposed a linguistic-based error metric, offering a finer-grained analysis of errors and discrepancies. We aim to do this in two ways. Firstly, we compare three languages, English, Japanese, and Spanish, all with different levels of inflections. Linguistic inflections are changes in the form of a word to mark distinctions such as tense, person, and number. For example, verb conjugations are a type of inflections and regular plurals in English. For our second improvement, we propose making distinctions between GRAMMATICAL and LEXICAL categories, grouping POS categories into these classes since their errors have different impacts on the **HYP** text. We make this distinction because LEXICAL errors directly contribute to the misunderstanding of the message and the incorrect interpretation of the text [18]. In contrast, GRAMMATICAL errors, while potentially leading to misunderstanding, have a less disruptive impact than LEXICAL errors. This distinction enables us to compare ASR errors not only at the overall level but also how they manifest concerning LEXICAL and GRAMMATICAL words. Such an approach contributes to the understanding of ASR errors, addressing the existing literature gap in the exploration of linguistic metrics, particularly in the context of evaluating errors based on GRAMMATICAL and LEXICAL categories. In summary, while existing metrics have laid a solid foundation for ASR evaluation, our work contributes by incorporating linguistic metrics inspired by both ASR advancements and successful methodologies in machine translation. We aim to fill the current gap in detailed error analysis within the ASR domain, adapting linguistic metrics to provide a more comprehensive understanding of errors.

3. Methodology

Developing performance metrics is a substantial undertaking. Based on [9] and [19], we developed our metrics to meet four criteria deemed crucial for a metric to possess. Firstly, it should reflect some level of human judgment, aiding in the identification of how much information is effectively communicated and how much is lost. Secondly, it must be straightforward to apply, which is a crucial feature when comparing across different ASR systems. Thirdly, it should be language-independent, which helps in comparing errors across languages from different typological classifications and does not favor one language structure over another. Finally, the metric should be easy to interpret from the outputs. In the development of the metrics presented here, we adhered to these principles to align with the practicality and real-world applicability prevalent in the field.

3.1. Languages chosen

The selection of languages was driven by both data availability and the authors' expertise, resulting in the choice of English, Japanese, and Spanish. These languages serve as robust testing grounds due to their shared characteristics and notable differences. Both English and Spanish belong to the Indo-European language family, and Japanese belongs to the Japonic language family [20]. They also exhibit divergences in their levels of inflection, a factor relevant to ASR system errors. Research has indicated that word classes with higher inflection are more prone to errors compared to those with less or no inflection [21]. For instance, the English article *the* remains uninflected, while its Spanish counterparts (feminine singular: "la", masculine singular: "el", feminine plural: "las," masculine plural: "los") carry gender and number inflections. Additionally, variations in inflection levels are evident in verb paradigms. While English may have six main forms (base, infinitive, past simple, past participle, gerund, and third person singular) [22], Japanese has 12 inflections [23], and Spanish can have 52 distinct forms reflecting person, number, tense, aspect, and mood [24]. These linguistic differences in inflection levels contribute to the richness of errors observed in ASR systems.

3.2. Speech datasets

This study utilized the Common Voice 15 dataset, a publicly available collection of multilingual and open voice data provided by the *Mozilla Common Voice Project* [13]. Designed for training and validating automatic speech recognition systems, the dataset encompasses a diverse range of voices and linguistic contexts. Table 1 below displays the characteristics of the datasets per language.

Table 1:.*Dataset descriptions for each language.*

Descriptors	EN	JA	SP
Total Number of Files	*16,386*	*4,978*	*15,796*
Total Duration	*26.9 hr*	*6.6 hr*	*26.8 hr*
Average File Duration	*5.9 sec*	*4.8 sec*	*6.11 sec*
Total Characters	*890K*	*105K*	*960K*
Total Words	*153K*	*55K*	*156K*
Unique Words	*21K*	*8K*	*23K*

The dataset encompasses contributions from a substantial number of speakers, providing a rich variety of linguistic and acoustic characteristics. In our analysis, we focused on a subset consisting of recordings from the test sets for the three languages. The dataset comprises over 16,000 sentences for English, approximately 5,000 for Japanese, and more than 15,000 sentences for Spanish. This offers a comprehensive sample of spoken language for evaluating ASR systems. The inclusion of a broad range of sentences and speakers enhances the robustness and generalizability of our findings, contributing to a more comprehensive understanding of the performance of the ASR system in diverse linguistic contexts. This includes variations in syntactic, semantic, and phonetic-phonological contexts.

3.3. Speech datasets

All our experiments were conducted using *OpenAI Whisper* [12]. Whisper comprises multilingual multitask models trained on 680,000 hours of labeled and curated speech data from diverse internet sources. In this experiment, we employed *Whisper-Tiny* (**T**), *Whisper-Medium* (**M**), *Whisper Large-v2* (**LV2**) and *Whisper Large-v3* (**LV3**). Comparing these four model sizes allows us to examine whether there are relevant accuracy gains across all language models.

3.4. Analysis

3.4.1. Word Error Rate

To assess the performance of the ASR system, we utilized the WER metric, a widely accepted measure for transcription accuracy assessment [25]. WER is computed by comparing the reference transcript (ground truth) with the output generated by the ASR system. The formula for WER is given by:

$$WER = (S+D+I) / N$$

Where, *S* represents the number of substitutions, *D* represents the number of deletions, *I* represents the number of insertions, and *N* is the total number of words in the reference transcript.

The analysis was conducted in R [26] using the outputs of Whisper. Our focus lies in ASR errors when comparing the reference text (**REF**) to the hypothesis text (**HYP**). SCLITE was employed for error calculation, identifying substitutions, insertions, and deletions per sentence. SCLITE, part of the NIST SCTK Scoring Toolkit, is a tool for scoring and evaluating speech recognition system output. It compares the **HYP** to the correct **REF**. Post-comparison statistics are gathered, and various reports can be generated to summarize recognition system performance.

3.4.2. Parts of Speech and Lexical Tagging

Linguistic tagging was conducted using the UDPIPE library [27] in R to enhance the textual analysis of transcribed speech data. UDPIPE, a state-of-the-art natural language processing (NLP) library, incorporates pre-trained models for various linguistic tasks. Specifically, we employed UDPIPE's pipeline for POS tagging. The tagging process consisted of three main steps.

Firstly, in text preprocessing, raw transcripts underwent preprocessing to eliminate artifacts or noise that might impact tagging accuracy (e.g. punctuation, case sensitivity, and text normalization), which allows for accurate comparisons between **REF** and ASR-generated transcripts. Secondly, during tokenization, preprocessed transcripts were tokenized into individual words or sub-word units using UDPIPE's tokenization module. The third step involved POS Tagging, where the POS tagging module assigned grammatical categories – such as nouns, verbs, adjectives – to each token in the transcripts. This information was crucial for understanding the syntactic structure of the spoken content.

3.4.3. Linguistic Metrics Analysis

We propose a metric that categorizes errors based on whether they occur in any of the two categories within a *Word Class*: GRAMMATICAL Words or LEXICAL Words. From the UDPIPE output, each POS was grouped into either the GRAMMATICAL group (ADP, AUX, CCONJ, DET, PART, PRON, SCONJ) or the LEXICAL Group (ADJ, ADV, NOUN, NUM, PROPN, VERB). From this, we calculated errors at the POS tagging in the **REF** and **HYP** texts, defined as POS_er, following the formula below:

$$Pos\_er = (S_{POS}+D_{POS}+I_{POS}) / N_{POS}$$

Word Class errors are then calculated for each group across the entire dataset per language and language model size: LEXICAL errors (LEX_er) and GRAMMATICAL errors (GRAM_er). To ensure fair comparisons, the analysis was conducted on sentences with matching number of words, i.e. when **REF** and **HYP** have the same number of words, avoiding penalization for incorrect pairs due to deletions and insertions. After comparing the accuracy metrics for POS, we then carried on further analysis explore the role of predictable linguistic patterns with the LEXICAL and GRAMMATICAL categories.

4. Results

Table 2 provides a summary of the experiment results, highlighting observable differences in performance across the **T**, **M**, **LV2** and **LV3** models for all evaluated languages. The **T** models consistently show the highest error rates (English = 23.7%; Japanese = 24.6%; Spanish = 23.5%), while the other models (**M**, **LV2** and **LV3)** demonstrate notably lower and more uniform WERs across all languages. Among these, the **LV3** model yields the most accurate results (English = 8.3%; Japanese = 5.7%; Spanish = 4%). It is evident that the **T** models show comparable WERs for the three languages, whereas the larger models exhibit higher accuracy, with Spanish being the most accurate and English the least accurate.

Table 2. *Breakdown of error rates results.*

Lang.	Category	Language Model			
		LV3	LV2	M	T
EN	WER	8.3%	8.9%	9.9%	23.7%
	POS_er	5.2%	5.5%	6.1%	17%
	LEX_er	11%	11.4%	11.9%	22.4%
	GRAM_er	4.6%	4.8%	5.3%	15.1%
JA	WER	5.7%	6.4%	7.5%	24.6%
	POS_er	1.7%	2%	2.4%	12.7%
	LEX_er	3.6%	3.9%	5.4%	22.9%
	GRAM_er	1.9%	2.7%	3.2%	13.7%
SP	WER	4%	4.9%	5.8%	23.5%
	POS_er	1.5%	1.9%	2.2%	10.9%
	LEX_er	4.2%	4.7%	5%	12.4%
	GRAM_er	1.6%	1.3%	3.4%	8.5%

When delving into the other metrics, a more detailed understanding emerges, shedding light on the categories to which ASR systems are more susceptible for errors. POS_er results demonstrate lower error rates in comparison to WER. This is notably more distinctive for Japanese and Spanish (English = 8.3%WER vs 5.2% POS_er; Japanese = 5.7% vs 1.7%; Spanish = 4% vs 1.5%). These results indicate that errors are more generalizable at the POS level, as compared to the word level. As such, this can help better our understanding of what types of errors can be consistently expected from ASR outputs, and in what morphological contexts.

A more in-depth analysis looked at those cases where the word form was incorrect (which counts to more WER) but it still had the same POS (which did not count as error for the POS_er). In case of inflectional languages, this difference can be observed when the **HYP** text has a singular form of a noun (e.g. *cat*), but the **REF** was the same word in the plural form (e.g. *cats*). The purpose was to see how much linguistic information is not captured if we stopped at the WER level. Figure 1 below shows a breakdown by language and model size for this experiment. It shows that Japanese and Spanish have more cases where errors are explained by inflectional differences between words (i.e. words are different but not their POS), as compared to English.

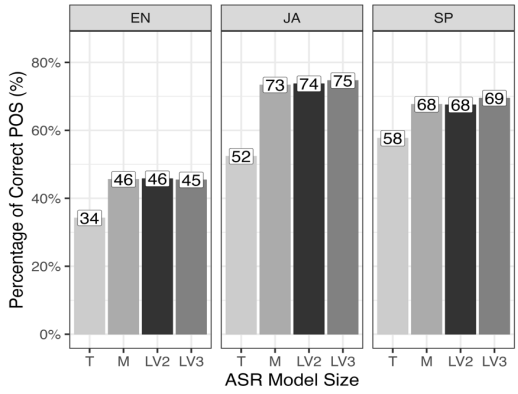

Figure 1: *Percentage of cases (and counts) of wrong words but with correct POS across all languages.*

The third layer of analysis distinguishes between LEX_er and GRAM_er, revealing patterns not captured by the other two layers (WER and POS_er). Figure 2 below presents the error rates broken down by language, model size, and word class (GRAMMATICAL or LEXICAL). The horizontal dotted line for each language represents the overall POS_er as reference.

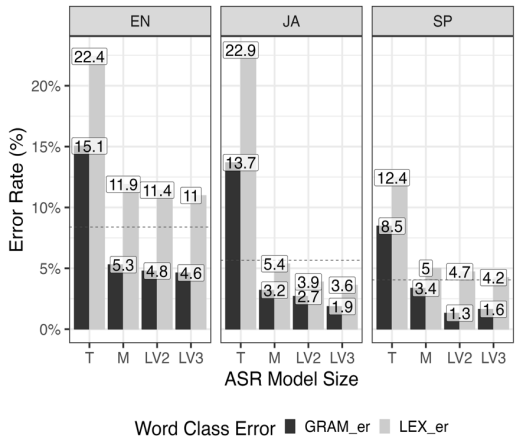

Figure 2: *Error rates across all languages and model sizes split by Word Class errors.*

Among the languages examined, Spanish consistently shows the lowest overall error rates, while English presents the highest. In the **LV3** model analysis, for LEX_er, Japanese records slightly lower rates than Spanish, while English exhibits the highest error rates (English = 11%; Japanese = 3.6%; Spanish = 4.2%). This variation can be explained linguistically by the fact that LEXICAL categories in Japanese and Spanish have higher inflections than in English, and these inflections are presented as affixes in both languages, helping the ASR system to understand the patterns of occurrence, useful to identify and predict the word form and its function in the language. This indicates that correct inflectional words significantly enhance predicting LEXICAL words. Although this finding is in contrast with [21], our results show that higher inflections are related to higher accuracy. Future research will help analyzing these results differences.

Further examination explored the extent to which predictable linguistic patterns helped in correctly identifying words. For this, we chose PROPER NOUNS (PROPN), as a subclass of the LEXICAL words. Our results show that Japanese is the language with least error rates, and English with the most errors (English = 39.6%; Japanese = 5.7%; Spanish = 18.6%). This is attributed to the use of case markers for PROPER NOUNS in Japanese, feature that is absent in English and Spanish, facilitating more accurate identification and prediction of this word class. The analysis revealed that the top six occurring words after PROPN were the case markers さん (3.4% – honorific particle), の (3.6% – possessive), に (2.4% – place), と (1.8% – joining nouns), は (1.8% – topic marking particle), and が (1.5% – grammatical subject), all accounting for approximately 15% of all words after PROPN in the Japanese dataset.

GRAM_er results show that Spanish had the lowest error rates compared to Japanese (slight difference of 0.3%) and, more prominently, to English (English = 4.6%; Japanese = 1.9%; Spanish = 1.6%). An in-depth analysis highlighted that the primary errors in English were associated with subordinating conjunctions (e.g., *if*, *that*, *while*) whereas the coordinating conjunctions were the ones driving more errors in Japanese (e.g., と *and*; も *also*) and Spanish (e.g., y *and*; o *or*). This indicates that a combination of grammatical assessment and linguistic function helps in a deeper understanding of how languages use specific words and the impact it has on the ASR accuracy. This approach is not necessarily language-dependent, but rather relies more on the typological function a word class has across multiple languages.

5. Discussion

This study discusses two interconnected error metrics to assess the accuracy of Whisper, using English, Japanese, and Spanish as test languages. Japanese and Spanish, being more inflectional than English, provided a valuable context for analysis, particularly in GRAMMATICAL words. Results reveal that relying solely on WER obscures important observations about ASR performance. In cases where WER percentages appear similar, as seen in the **T** model results, a closer examination at the GRAMMATICAL vs LEXICAL level unveils distinct accuracies. Conversely, even with differing WERs, such as in the **Large** models where Japanese and Spanish outperform English, a nuanced analysis exposes the ASR system's consistent performance on LEXICAL words but divergence in handling GRAMMATICAL words, with English being more prone to errors with the coordinating conjunctions. These metrics strike a balance between WER and individual POS errors. While reporting each POS category individually could complicate cross-language system performance comparisons, the analyzed metrics offer a middle ground. They enable users to identify the strengths and weaknesses of ASR at crucial linguistic levels, providing clarity on areas requiring attention when refining outputs. This approach enhances the interpretability and practical utility of ASR performance assessments.

6. Conclusions

The automatic processing and annotation of natural speech are complex tasks influenced by both the systems themselves and, most importantly, by the inherent characteristics of languages. Current systems have made significant progress in addressing these complexities. One notable advancement is the ability to perform automatic grammatical error comparisons across languages with different typological classifications. This advancement requires a cautious approach to understanding intrinsic language differences and variations based on the ASR system or the data used for training.

7. Acknowledgements

We would like to thank Hayden Ooi and Chloe Dean for their valuable feedback on an earlier version of this paper. We also wish to thank the anonymous reviewers for their helpful comments on the shape and content of the submission.

8. References

[1] O'Shaughnessy, D., "Trends and developments in automatic speech recognition research", Computer Speech & Language, vol. 83, 2023.

[2] Reitmaier, T. et al, "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers", Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), New York: USA, 299, 1-17, 2022.

[3] Whetten, R. and Kennigton, C., "Evaluating and Improving Automatic Speech Recognition using Severity", The 22nd Workshop on Biomedical NLP and BioNLP Shared Tasks, 79-91, 2023.

[4] Kumalija, E. and Nakamoto, Y., "Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech", Frontiers in Signal Processing, vol. 2, 1-10, 2022.

[5] He, X., Deng, L., and Acero, A., "Why word error rate is not a good metric for speech recognizer training for the speech translation task?", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5632-5635, 2011.

[6] Kafle, S. and Huenerfauth, M., "Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing", Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, 165-174, 2017.

[7] Garofolo, J. S., Voorhees, E. M., Auzanne, C. G. P., Stanford, V. M., and Lund, B. A., "1998 TREC-7 spoken document retrieval track overview and results", Proceedings of the 7th Text REtrieval Conference. NIST, 79-89, 1998.

[8] Roux, T. B., Rouvier, M., Wottawa, J., and Dufour, R., "Qualitative evaluation of language model rescoring in automatic speech recognition", Proceedings INTERSPEECH 2022 - 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 3968-3972, 2022.

[9] Morris, A. C., Maier, V., and Green, P., "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition", Proceedings INTERSPEECH 2004 - 8th Annual Conference of the International Speech Communication Association, Jeju Island, Korea, 2765-2768, 2004.

[10] Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A., and Bensaali, F., "Deep transfer learning for automatic speech recognition: Towards better generalization", Knowledge-Based Systems, vol. 277, 2023.

[11] Lee, S., Noh, H., Lee, K., and Geunbae Lee, G., "Grammatical error detection for corrective feedback provision in oral conversations", Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI'11), AAAI Press, 797-802, 2011.

[12] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust speech recognition via large-scale weak supervision", Proceedings of the 40th International Conference on Machine Learning, 28492-28518, 2023.

[13] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G., "Common Voice: A Massively-Multilingual Speech Corpus", Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 421-4215, 2020.

[14] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "Bert: Pretraining of deep bidirectional transformers for language understanding", North American Chapter of the Association for Computational Linguistics, 2019.

[15] Reimers, N. and Gurevych, I., "Sentence-bert: Sentence embeddings using siamese bert-networks", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, 3973-3983, 2019.

[16] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y., "Bertscore: Evaluating text generation with bert", 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.

[17] Popović, M. and Ney, H., "Word Error Rates: Decomposition over POS classes and Applications for Error Analysis", Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 48-55, 2007.

[18] Hemchua, S. and Schmitt, N., "An Analysis of Lexical Errors in The English Compositions of Thai Learners", Prospect: An Australian Journal of TESOL, vol. 21(3), 3-25, 2006.

[19] McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H., "On the use of information retrieval measures for speech recognition evaluation", Technical Report. IDIAP, 2004.

[20] Ethnologue: languages of the world, Dallas:Texas, SIL International, 1999.

[21] Berg, K., Hartmann, S., and Claeser, D., "Are some morphological units more prone to spelling variation than others? A case study using spontaneous handwritten data", Morphology, 2023.

[22] Lee, J. and Seneff, S., "Correcting Misuse of Verb Forms", Proceedings of ACL-08: HLT, Columbus, Ohio, USA, June 2008, Association for Computational Linguist, 174-182.

[23] Hisamitsu, T. and Nitta, Y., "An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis", International Conference on Computational Linguistics, 1994.

[24] Centeno, J. G. and Obler, L. K., "Agrammatic verb errors in Spanish speakers and their normal discourse correlates", Journal of Neurolinguistics, vol. 14, 349-363, 2001.

[25] Park, C., et al., "Fast word error rate estimation using self-supervised representations for speech and text", arXiv preprint arXiv:2310.08225, 2023.

[26] R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna:Austria, URL https://www.R-project.org/, 2021.

[27] Wijffels, J., Straka, M., and Straková, J., "Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the Udpipe Nlp Toolkit", https://CRAN.R-project.org/package=udpipe, 2021.