# Leveraging LEXICAL and GRAMMATICAL Errors: Extending ASR Error Measurements through NLP

**Anonymous ACL submission**


## Abstract

This paper addresses the limitations of current Automatic Speech Recognition (ASR) evaluation metrics by highlighting the inadequacies of overall error rates, particularly Word Error Rate. While this offers a broad assessment, it lacks the granularity needed to discern specific linguistic categories affected by errors. We offer an NLP-driven metric based on parts of speech and grammatical categories, to provide a more in-depth analysis of ASR errors. Using the Whisper ASR system on English, Japanese, and Spanish, within the CommonVoice 15 dataset, we analyze GRAMMATICAL and LEXICAL error rates. Results show that GRAMMATICAL words trigger less errors than LEXICAL words across all languages, and Proper Nouns in Japanese combined with case markers are related to higher accuracy. By categorizing errors based on these linguistic attributes, our methodology aims to enhance the explanatory power of error analysis in ASR, contributing to a more precise evaluation of system performance based on NLP approaches.


## 1 Introduction

Automatic Speech Recognition (ASR) technologies have undergone significant advancements (O'Shaughnessy, 2023; Reitmaier et al., 2022) 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.

Word Error Rate (WER) is a crucial metric used to evaluate the performance of ASR systems. It measures the operational accuracy of an ASR system by calculating 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 (Kumalija and Nakamoto, 2022).

While WER is used widely as a standard metric (Ali and Renals, 2004; NithyaKalyani and Jothilakshmi, 2019; Park et al., 2023), it has been reported to have some critical limitations (He et al., 2011). The primary limitation of WER lies in treating all errors equally, regardless of their impact on the meaning of the transcribed text. For instance, misrecognizing a key word might change the meaning of a sentence significantly than other non-key words, but WER weighs all the errors the same.

Additionally, WER cannot gauge the relative importance of specific words in the ground truth transcription, prompting the proposal of alternative metrics that account for semantics (Kafle and Huenerfauth, 2017), entity recognition (Garofolo et al., 1998), and parts of speech (Roux et al., 2022).

Prior studies indicate that WER does not consistently correlate with human judgment on ASR system performance (Morris et al., 2004; Whetten and Kennington, 2023). These findings highlight the necessity for refined linguistic metrics that enable a more granular analysis of errors.

Another critical limitation of the existing metric is its inability to unveil the specific characteristics of errors. Although ASR systems may exhibit similar overall error rates as measured by WER, this metric fails to distinguish between errors affecting different linguistic categories. For example, two ASR systems with equivalent WERs might impact linguistic accuracy differently: one may disproportionately affect GRAMMATICAL words, while the other could affect more frequently LEXICAL or content words. Consequently, a deeper examination of error complexities within linguistic categories is essential to identify and specify areas of vulnerability within ASR systems (Adegbegha et al., 2024; Errattahi et al., 2018; Kheddar et al., 2023; Lee et al., 2011; Li et al., 2023).

Recognizing the limitations of current methodologies, we propose the integration of linguistic metrics into the evaluation of ASR systems. An in-depth analysis based on linguistic categories, including parts of speech and grammatical classifications, enriches our understanding of error complexities. By categorizing errors based on linguistic attributes, we gain valuable insights into the nature of errors and how they behave within the context of these systems. This approach not only clarifies the types of errors but also enhances the explanatory power of error analysis, providing a more comprehensive understanding of ASR system performance.

Our proposed methodology presents an approach to analyze and report errors in ASR outputs. Adopting a multilingual perspective, we examine errors in English, Japanese, and Spanish, leveraging the Whisper ASR system (Radford et al., 2023) on the CommonVoice 15 dataset (Ardila et al., 2020). Utilizing Parts of Speech tagging (POS), we differentiate errors into two specific categories: those affecting grammatical or function words (referred as GRAMMATICAL error), and those impacting lexical or content words (referred as LEXICAL error).

The distinction between GRAMMATICAL and LEXICAL categories facilitates a layered comparison of ASR errors, examining not only their aggregate impact but also their specific manifestation across different linguistic types. This dual-level analysis enhances our understanding of ASR errors, significantly addressing the gaps identified in literature. This approach provides a more specific and informative perspective of ASR performance, addressing to the need for detailed error analysis in the advancing field of speech recognition technologies.

This work significantly advances the reporting of detailed linguistic layers in ASR systems, establishing a more consistent methodology that extends beyond the limited scope of previous research, which often confined analyses to specific databases/languages. We propose a systematic approach applicable across all languages supported by the ASR system with available universal dependencies. By developing metrics within a single widely used ASR system, we enable refined comparison between the reference text (**REF**) and the generated hypothesized text (**HYP**) across languages with varying typological characteristics.

## 2 Related Work

### 2.1 Current Progress on Metrics

ASR evaluation methodologies have undergone some refinements, incorporating diverse error metrics that surpass mere word counts, including *word embeddings* (Devlin et al., 2019), *sentence embeddings* (Reimers and Gurevych, 2019), and *semantic proximity* (Zhang et al., 2020).

Taking inspiration from machine translation, wherein linguistic metrics significantly enhance translation accuracy, this paper adapts similar methodologies to ASR. Popović and Ney (2007) effectively incorporated linguistic attributes, like parts of speech, into translation evaluation and introduced the *Position Independent Error Rate* (PER) to measure the impact of each POS class on overall word error rates. While their study analyzed POS-based in two languages, English and Spanish, and compared them with human assessments, it remains to be determined whether these findings can be generalized to other languages with different typological characteristics.

### 2.2 Main Gaps in Previous Work

Although previous studies have covered relevant aspects of the error assessment and description in ASR system outputs, there are yet three significant gaps remaining. Firstly, these studies have relied on custom-built or less standardized ASR systems, limiting the generalizability and reproducibility of their findings. Our research counters this limitation by employing the widely recognized and standardized Whisper ASR system; we ensure that our findings are more applicable to a broader range of applications and that our methodology can be more easily replicated by other researchers in the field.

Secondly, earlier studies have often focused on analyzing ASR errors in a single language or closely related languages, which limits the understanding of ASR performance across diverse linguistic parameters. Our study expands this scope by examining ASR errors in three linguistically differentiated languages – English, Japanese and Spanish – thus broadening the evaluation of our proposed error metrics and enhancing understanding of ASR systems across varied language families.

Lastly, although previous studies have measured ASR errors at various linguistic levels, they have not being consistent in proposing a direct integration of these detailed measurements into the overall reporting of ASR outputs. Our study addresses this by not only detailing these measurements but also integrating them with the general reporting of ASR errors. This integrated approach provides a more comprehensive and informative analysis of ASR performance.

## 3 Methodology

In our methodology, we build upon the linguistic-based error metrics found in Cao et al., (2023) and Roux et al. (2022), which provide a finer-grained analysis of errors and discrepancies. Our approach enhances this framework through two distinct strategies. First, we conduct a comparative analysis of three linguistically diverse languages: English, Japanese and Spanish. Each language exhibits different levels of linguistic inflections, such as changes in word form to mark distinctions such as tense, person, and number. For example, verb conjugations are a type of inflections and regular plurals in English. This comparative study allows us to assess how inflectional complexity impacts ASR accuracy across different linguistic systems.

In the second aspect of our methodology, we differentiate between Grammatical and Lexical categories, grouping POS categories into these classes since their errors have different impacts on the **HYP** text. This categorization is crucial because Lexical errors directly lead to the misunderstanding of the intended message and the incorrect interpretation of the text (Hemchua and Schmitt, 2006). In contrast, while Grammatical errors can also cause misunderstandings, their effect on the overall comprehension of the text is generally less disruptive than that of Lexical errors. Table 1 shows examples of Grammatical and Lexical errors.

In developing performance metrics for ASR systems, we adhered to four essential criteria (Morris et al., 2004; McCowan et al., 2004). First, it should reflect some level of human judgment, aiding in the identification of how much information is effectively communicated and how much is lost. Second, it must be straightforward to apply, facilitating comparisons across various ASR systems. Third, it should be language-independent to ensure unbiased error comparisons across languages from different typological classifications. Finally, the metric should be easy to interpret from the outputs. These principles ensure that our metrics are both practical and applicable in real-world settings.

| | **Words** | | | | | |
|---|---|---|---|---|---|---|
| **REF** text | **The** | **cat** | **sat** | **on** | **the** | **mat** |
| Lexical error | **The** | **dog** | **sat** | **on** | **the** | **mat** |
| Gram. error | **The** | **cat** | **sat** | **on** | **that** | **mat** |

Table 1: Examples of Differences between Grammatical and Lexical errors.

Figure 1 below summarizes the main six stages followed in this paper, and these are expanded in the following sections. The first three correspond to the ASR processing, four and five correspond to the NLP processing and categorization of errors, and the final stage corresponds to the error reporting.

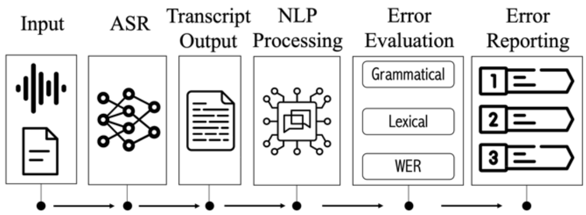



Figure 1: Data Processing and Analysis Stages.

### 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 (Ethnologue, 1999). They also exhibit divergences in their levels of inflection, a factor relevant to ASR system errors.

Some research has found that word classes with higher inflection are more prone to errors compared to those with less or no inflection (Berg et al., 2024; Smith-Lock, 1991). For instance, the English article *the* remains uninflected, while its Spanish counterparts carry gender and number inflections (feminine singular: "la", masculine singular: "el", feminine plural: "las," masculine plural: "los"). Additionally, variations in inflection levels are evident in verb paradigms. While English may have six main isolated forms (base, infinitive, past simple, past participle, gerund, and third person singular) (Lee and Seneff, 2008), Japanese has 12 inflections (Hisamitsu and Nitta, 1994), and Spanish can have 52 distinct forms reflecting person, number, tense, aspect, and mood (Centeno and Obler, 2001).

Another major difference in Japanese, unlike English and Spanish, is that it does not generally use white spaces to separate words. To identify morphemes and words in Japanese, two main approaches have been taken. The first approach is to define base unit words by first identifying syntactic words, which is done in Universal Dependencies (UD) by using *Short Unit Words* (SUW). The second approach uses *Long Unit Words* (LUW) as the base units in Japanese. Although similar results were achieved from both approaches, Omura et al. (2021) argue that lemmatization of LUW is more complex for a morphologically rich language. The pretrained model used in this paper utilized the SUW approach.

The choice of these three languages allows for typological comparisons, highlighting characteristics that are shared by all three languages, by two of them, or individually. The comparison is shown in Table 2. This summarizes the main linguistic classifications across the three languages. We present the major morphological features relevant for the current study.

| | EN | JA | SP |
|---|---|---|---|
| **Deriv. Morphology** | Prefixes and Suffixes | | |
| **Morphology** | Analytic | Synthetic | |
| **Gender** | No | | Nouns |
| **Word Order** | SVO | SOV | SVO |
| **Word Formation** | Mostly analytic | Agglutination | Inflections |
| **Inflection** | Limited | Verbs and Adjectives | Rich |
| **Case Marking** | Pronouns | Extensive | Nominative Accusative Dative |

Table 2: Main Morphological Descriptions for each Language.

These linguistic differences in inflection levels contribute to the richness of errors observed in ASR systems.

### 3.2 Speech Datasets

We utilized the Common Voice 15 dataset, a publicly available collection of multilingual and open voice data provided by the *Mozilla Common Voice Project* (Ardila et al., 2020). Designed for training and validating automatic speech recognition systems, the dataset encompasses a diverse range of voices and linguistic contexts. The data does not contain personally identifying information. Table 3 below displays the characteristics of the datasets per language.

| | **Descriptors** | **EN** | **JA** | **SP** |
|---|---|---|---|---|
| | Number of Files | *16,386* | *4,978* | *15,796* |
| audio | Duration (hr) | *26.9* | *6.6* | *26.8* |
| audio | Speech Dur. (hr) | *22.7* | *5.4* | *23.2* |
| audio | Mean Dur. (sec) | *5.9* | *4.8* | *6.11* |
| audio | Mean Speech Dur. (sec) | *4.9* | *3.9* | *5.3* |
| text | Total Characters | *890K* | *105K* | *960K* |
| text | Total Words | *153K* | *55K* | *156K* |
| text | Unique Words | *21K* | *8K* | *23K* |
| text | Characters p/Text | *54* | *21* | *61* |
| text | Words p/Text | *10* | *10* | *10* |

Table 3: Dataset Descriptions for each Language.

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 ASR System

All our experiments were conducted using *OpenAI Whisper* (Radford et al., 2023). Whisper comprises multilingual multitask models trained on 680,000 hours of labelled 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 ASR models.

### 3.4 Analysis

**Word Error Rate:** 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 (R Core Team, 2023) 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 SCT[1] 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. To assess the performance of the ASR system, we utilized the WER metric, a widely accepted measure for transcription accuracy assessment.

**Parts of Speech and Lexical Items:** Linguistic tagging was conducted using the UDPIPE (Straka and Straková, 2017) library (Wijffels et al., 2023) 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, which are based on Universal Dependencies (UD). 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. 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. Careful consideration of punctuation, case sensitivity, and text normalization procedures was carried out to ensure accurate comparisons between **REF** and ASR-generated transcripts. UDPIPE outputs have been reported to exhibit varying levels of performance. Straka and Straková (2017) reported that the automatic identification of POS has an accuracy of 93.50% for English, 88.19% for Japanese, and 98.14% for Spanish. We assess our outputs based on these reported accuracy levels.

**Linguistic Metric Analysis:** We propose a metric that categorizes errors based on whether they occur in any of the two categories within a *Word Class*: Grammatical and Lexical. 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, and we also calculated differences at the Lexical and Grammatical levels, following the formulae below:

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

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

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

*Word Class* errors are then calculated for each group across the entire dataset per language and ASR model size: Lexical errors (Lex_er) and Grammatical errors (Gram_er).

| | Total | Size | Subset | Diff. |
|---|---|---|---|---|
| **EN** | 16386 | **T** | 14459 | 12% |
| | | **M** | 14855 | 9% |
| | | **LV2** | 14886 | 9% |
| | | **LV3** | 14896 | 9% |
| **JA** | 4978 | **T** | 573 | 88% |
| | | **M** | 1544 | 69% |
| | | **LV2** | 1634 | 67% |
| | | **LV3** | 1724 | 65% |
| **SP** | 15796 | **T** | 12569 | 20% |
| | | **M** | 14833 | 6% |
| | | **LV2** | 14910 | 6% |
| | | **LV3** | 14967 | 5% |

Table 4: Number of Sentences included in the Analyses.

[1] https://github.com/usnistgov/SCT

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. Table 4 above shows the differences in the number of sentences in the original transcriptions and the ones filtered in the analyses.

# 4 Experiment and Results

## 4.1 ASR Model Size Comparison

Table 5 provides a summary of the experiment results, highlighting significant differences in performance across the **T**, **M**, **LV2** and **LV3** models for all evaluated languages.

| | Category | ASR 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 |

Table 5: Breakdown of Error Rates Results. Values are shown in Percentage of Errors (%).

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.

## 4.2 Parts of Speech Comparison

When delving into the other metrics, our results discover a more nuanced understanding, 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).

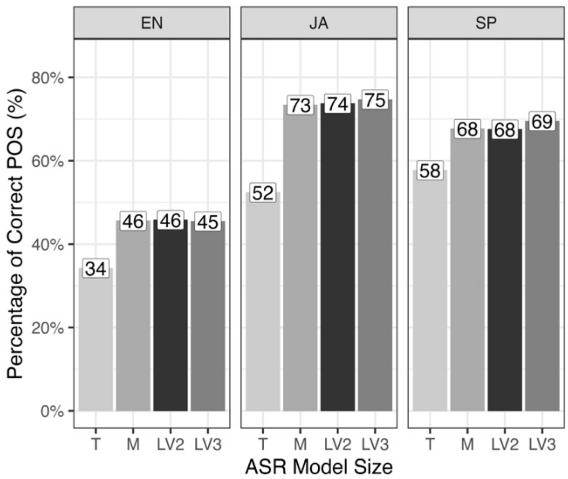



Figure 2: Percentage of Cases (and Counts) of Wrong Words but with Correct POS.

In the case of highly inflectional languages, this difference can be observed when the **HYP** text has a singular form of a noun (e.g., *cat*), but the **REF** text was the same word in the plural form (e.g., *cats*). This observation underscores the limitations of relying solely on WER, as it fails to capture subtle linguistic nuances retained in POS_er. Figure 2 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.

## 4.3 Word Class Comparison

The third layer of analysis distinguishes between LEXICAL error rate (LEX_er) and GRAMMATICAL error rate (GRAM_er), revealing patterns not captured by the previous two layers (WER and POS_er). Figure 3 presents the error rates broken down by language, model size, and word class (GRAMMATICAL or LEXICAL) with a horizontal dotted line indicating the overall POS_er as reference.

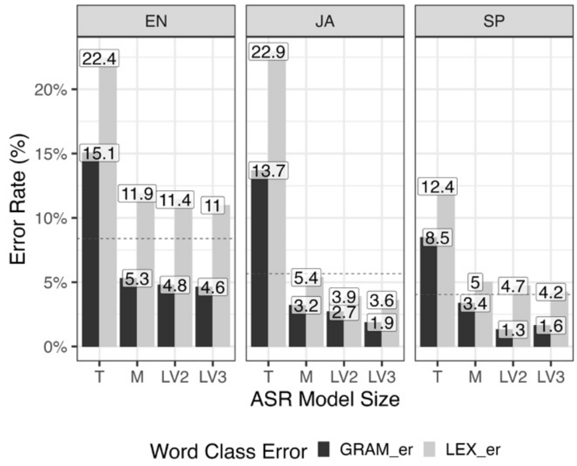



Figure 3: Error Rates across all Languages and Model Sizes split by Word Class Errors.

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 Berg et al. (2024) and Smith-Lock. (1991), our results show that higher inflections are related to higher accuracy.

Further examination explored the extent to which predictable inflections helped in correctly identifying words for the ASR system. For this, we chose PROPER NOUNS (PROPN), a subclass of the LEXICAL words (See Appendices for reference). 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 Proper Nouns.

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 significantly, 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.

### 4.4 Assessment across all Comparisons

When assessing all error rate metrics, we find that POS_er, LEX_er and GRAM_er contribute to a more robust understanding of ASR errors. The examples below are used to analyze the different layers. The first line is the transcription and below each, the POS tagging is given for each word.

(a) *English*

REF: *what is the matter with* ***a*** *thousand* ***dollars***
PRON V DET NOUN ADP **DET** NUM **N**
HYP: *what is the matter with* ***the*** *thousand* ***dollar***
PRON V DET NOUN ADP **DET** NUM **N**

(b) *Spanish*

REF: *las batallas* ***se libraron*** *primero en los territories*
DET N **PRON V** ADJ ADP DET N
HYP: *las batallas* ***el vibrado*** *primero en los territories*
DET N **DET N** ADJ ADP DET N

Sentence (a) has a WER of 25% (two wrong words over eight words in total). The errors are found in the words *a* and *dollars* in the **REF** text. The word *a*, the indefinite article, changed to definite article *the* in the **HYP** text. The change happened in a GRAMMATICAL word. The second word, *dollars*, changed to the singular form in the **HYP** text, *dollar*. In terms of classification, it is a LEXICAL word, however, the change occurred in a morpheme that conveys plurality in English. Since these errors did not change the word class, POS_er, LEX_er, and GRAM_er have 0% error rate. This also shows that the NOUN *dollar* was not changed to another NOUN, but just its plurality, which did not compromise the meaning as compared to being changed to another word, like *scholar*, for example.

Sentence (b) has the same WER as (a), 25%. However, the error patterns are different. The errors are found in the words “se” (reflexive PRON) and “libraron” (simple past on V *to fight*) in the **REF** text. Both words were substituted with different words and with different POS categories, however, they belong to the same word class (PRON > DET; V > N). POS_er is 25%, LEX_er is 12.5%, and GRAM_er is 12.5%. Compared to sentence (a), only the WER is the same, but the other error metrics are all different. These two sentences exemplify the complementary nature of the application of all metrics, rather than being competing measures.

## 5 Discussion

This study examines two NLP-driven error metrics across English, Japanese, and Spanish to assess the accuracy of the Whisper ASR system. Given the higher inflectional complexity, Japanese and Spanish provide a valuable context for analysis, particularly in GRAMMATICAL words. Our findings reveal that relying solely on WER can obscure nuanced aspect of ASR performance. For instance, while WER figures maybe comparable across models, such as 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 (**LV2**, **LV3**) where Japanese and Spanish outperform English, a detailed analysis exposes the ASR system’s consistent performance on LEXICAL words but divergence in handling GRAMMATICAL words, notably with English struggling more with subordinating conjunctions.

These metrics bridge the gap between general WER assessment and more granular POS error analyses. While reporting each POS category individually could complicate comparisons across languages, our metrics offer a balanced approach. This approach allows for a detailed identification of strengths and weaknesses in ASR systems at crucial linguistic levels, enhancing both the interpretability and practical applicability of ASR performance evaluations.

For the areas of improvement pertinent to ASR systems, we can make suggestions based on the observed patterns and the datasets used. First, the training of the ASR systems should include more accurate weighting of words based on whether they are LEXICAL vs GRAMMATICAL in function. This adjustment will enable a better-informed decision driven by word usage in context. Second, performance can also be improved if more spontaneous speech is used in the training of ASR systems. This can lead to the observation of more LEXICAL words, such as low-frequency words or those more commonly found in spoken language, in contexts that are less typical of controlled speech. Finally, the errors observed strongly suggest that errors follow specific linguistic patterns (e.g., LEXICAL vs GRAMMATICAL, or PROPN vs NOUN). In this sense, they go beyond language-dependent patterns and can be better understood under linguistic typologies.

## 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 and their typological differences. 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 necessitates a cautious approach to understanding intrinsic language differences and variations based on the ASR system or the data used for training.

Linguistically informed metrics play a crucial role in interpreting performance. This, combined with robust NLP approaches, prove to be efficient for this task. Our metric and implementation developed for assessing ASR performance help identifying areas for improvement and linguistic aspects that pose specific challenges. Additionally, these metrics assist those working on ASR systems and datasets in developing more efficient algorithms and infrastructures.

## 7 Limitations

In this study, we investigated three languages with different typological characteristics, highlighting both shared and unique features. However, the scope of our research did not extend to polysynthetic languages, presenting a limitation in the diversity of language types analyzed. Future work should include a broader range of languages to determine if the observed patterns are replicated or consistent across major language families.

Our analysis was confined to the Whisper ASR system, chosen for its broad usage and accessibility. This focus presents a limitation as it does not address a comparative evaluation across different ASR systems. Future research should examine a variety of ASR systems to ascertain whether the observed errors are influenced by specific systems or linguistic characteristics themselves.

Finally, we did not address fine-tuning of ASR models. This can help identify whether the errors are specific to the ASR model or the data used for training, and to what extent we can generalize these errors. Future research should also investigate the impact of linguistically-based fine-tuning on the performance of ASR systems.

## Acknowledgments

To be added in final version.

## References

Ahmed Ali and Steve Renals. 2018. Word Error Rate Estimation for Speech Recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Volume 2, pp. 20-24, Melbourne, Australia. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 421-4215.

Kristina Berg, Stefan Hartmann, and Daniel Claeser. 2024. Are some morphological units more prone to spelling variation than others? A case study using spontaneous handwritten data. *Morphology*, Volume 34: pp. 173-188.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. 2023. A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse. In *UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23)*, June 26--29, 2023, Limassol, Cyprus. ACM, New York, NY, USA 13 Pages.

Jose G. Centeno and Loraine K. Obler. 2001. Agrammatic verb errors in Spanish speakers and their normal discourse correlates. *Journal of Neurolinguistics,* Volume 14, pp. 349-363.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pretraining of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*.

Yetunde E. Adegbegha, Aarav Minocha, and Renu Balyan. 2024. Analyzing Multilingual Automatic Speech Recognition Systems Performance. In F. Zhao, and D. Miao, (eds.), *Communications in Computer and Information Science*, *Volume 1946*. Springer, Singapore.

Rahhal Errattahi, Asmaa E. Hannani, and Hassan Ouahmane. 2018. Automatic Speech Recognition Errors Detection and Correction: A Review. In *International Conference on Natural Language and Speech Processing*, *Volume 128*, pp. 32-37.

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

John S. Garofolo, Ellen M. Voorhees, C G. Auzanne, Vincent M. Stanford, and B A. Lund. 1998. 1998 TREC-7 spoken document retrieval track overview and results. In *Proceedings of the 7th Text REtrieval Conference*. NIST, pp. 79-89.

Xiaodong He, Li Deng, and Alex Acero. 2011. Why word error rate is not a good metric for speech recognizer training for the speech translation task?. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5632-5635.

Saengchan Hemchua and Norbert Schmitt. 2006. An Analysis of Lexical Errors in The English Compositions of Thai Learners. *Prospect : An Australian Journal of TESOL,* 21(3):3-25.

Toru Hisamitsu and Yoshihiko Nitta. 1994. An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis. In *International Conference on Computational Linguistics*, pp. 194-200.

Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM*

*SIGACCESS Conference on Computers and Accessibility*, pp. 165-174.

Hamza Kheddar, Yassine Himeur, Somaya Al-Maadeed, Abbes Amira, and Faycal Bensaali. 2023. Deep transfer learning for automatic speech recognition: Towards better generalization. *Knowledge-Based Systems, Volume 277*.

Elhard Kumalija and Yukikazu Nakamoto. 2022. Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech. *Frontiers in Signal Processing*, 2:1-10.

John Lee and Stephanie Seneff. 2008. Correcting Misuse of Verb Forms. In *Proceedings of ACL-08: HLT, Columbus, Ohio, USA, June 2008, Association for Computational Linguistics*, pp. 174-182.

Sungjin Lee, Hyungjong Noh, Kyusong Lee, and Gary G. Lee. 2011. Grammatical error detection for corrective feedback provision in oral conversations. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI'11)*, AAAI Press, pp. 797-802.

Chengzhang Li, Ming Zhang, Xuejun Zhang, and Yonghong Yan. 2023. MCRSpell: A metric learning of correct representation for Chinese spelling correction. *Expert Systems with Applications, Volume 237, Part B*.

Iain A. McCowan, Darren Moore, John Dines, Daniel Gatica-Perez, Mike Flynn, Pierre Wellner, and Herve Bourlard. 2004. *On the use of information retrieval measures for speech recognition evaluation*. Technical Report. IDIAP.

Andrew C. Morris, Viktoria Maier, and Phil D. Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings INTERSPEECH 2004 – 8th Annual Conference of the International Speech Communication Association*, Jeju Island, Korea, October 2004, pp. 2765-2768.

A. NithyaKalyani and S. Jothilakshmi. 2019. Speech Summarization for Tamil Language. In Ed. Nilanjan Dey, *Intelligent Speech Signal Processing*, Academic Press, pp. 113-138.

Douglas O'Shaughnessy. 2023. Trends and developments in automatic speech recognition research. *Computer Speech & Language, Volume 83*.

Mai Omura, Aya Wasaka, and Masayuki Asahara. 2021. Word Delimitation Issues in UD Japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021),* Sofia, Bulgaria, Association for Computational Linguistics, pp. 142-150.

Chanho Park et al. 2023. Fast word error rate estimation using self-supervised representations for speech and text. *arXiv preprint* arXiv:2310.08225.

Maja Popović and Hermann Ney. 2007. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics*, Prague, Czech Republic, pp. 48-55.

R Core Team. 2023. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna:Austria, URL https://www.R-project.org/.

Alec Radford, Jong W. Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, July 2023, pp. 28492-28518.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *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*, pp. 3973-3983.

Thomas Reitmaier, Electra Wallington, Dani K. Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22),* New York: USA, 299, pp. 1-17.

Thibault B. Roux, Mickael Rouvier, Jane Wottawa, and Richard Dufour. 2022. Qualitative evaluation of language model rescoring in automatic speech recognition. In *Proceedings INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 3968-3972.

Karen M. Smith-Lock. 1991. Errors of Inflection in the Writing of Normal and Poor Readers. *Language and Speech*, 34(4), pp. 341-350.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88-99, Vancouver, Canada. Association for Computational Linguistics.

Ryan Whetten and Casey Kennington. 2023. Evaluating and Improving Automatic Speech Recognition using Severity. In *The 22nd Workshop on*

*Biomedical Natural Language Processing and BioNLP Shared Tasks*, July 13, pp. 79-91.

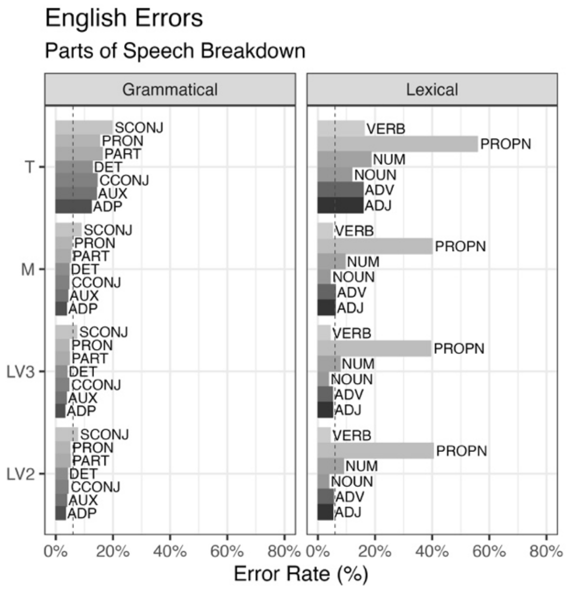
Jan Wijffels, Milan Straka, and Jana Straková. 2023. *Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the Udpipe Nlp Toolkit*. https://CRAN.R-project.org/package=udpipe.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26–30.

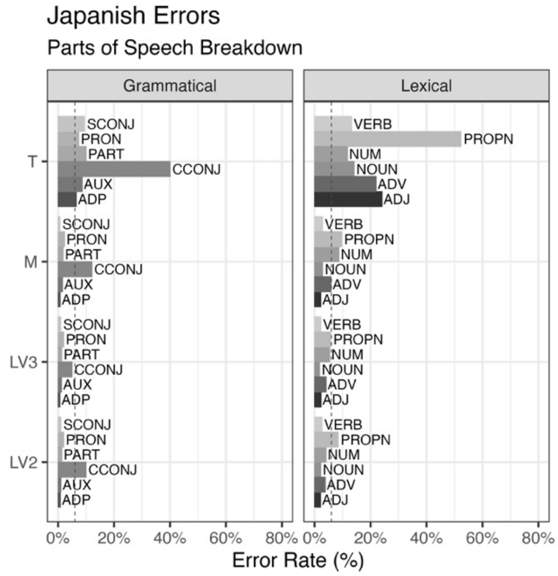
# A Appendices

This section includes the breakdown of all errors for Parts of Speech across all languages. The horizontal dotted lines indicate the overall POS_er as reference.

**English Errors Breakdown**




**Japanese Errors Breakdown**




**Spanish Errors Breakdown**