



Linguistic sources of performance disparities in Dutch Automatic Speech Recognition for non-native adults

Kiarash Karimi¹

Supervisor(s): Odette Scharenborg¹, YuanYuan Zhang¹

¹EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Kiarash Karimi
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, Catharine Oertel

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Abstract

Automatic Speech Recognition (ASR) systems are increasingly used in real-world applications, but their performance is not equally reliable across speaker groups. Previous work has shown state-of-the-art ASR systems to perform worse for non-native speakers; the linguistic factors that contribute to this, however, are yet to be understood. To address this gap, this study investigates how utterance length and word category affect ASR performance for Dutch speech from non-native adult speakers. In addition, it investigates how these effects differ between read speech and human-machine interaction (HMI) speech. The study uses ASR outputs from Google Chirp 2 and Whisper large-v3 on non-native adult speech from the JASMIN corpus. The analysis examines how error patterns relate to utterance length and part-of-speech category, with insertion, deletion, and substitution errors analyzed separately for each factor. The results show that ASR performance varies systematically with both factors: very short utterances are especially error-prone, length effects differ between read and HMI speech, and word-category patterns are most visible in substitution and insertion errors. These findings highlight the need for more fine-grained ASR evaluation methods that account for how linguistic context influences recognition errors in non-native speech.

1 Introduction

Automatic Speech Recognition (ASR) systems are widely used in applications such as speech translation, voice search, personal digital assistants [1], and are increasingly deployed in real-world and high-stakes settings, such as emergency call centers [2] and medical reporting [3]. As a result, ensuring that ASR systems perform reliably and equally well for speakers with different sociolinguistic backgrounds is crucial for both accessibility and fairness [4, 5]. Although recent systems such as OpenAI’s Whisper have reached accuracies comparable to human performance in English transcription [6], and systems such as Google Chirp 2 achieve word error rates (WERs) as low as 12% for typical Dutch read speech [7], significant performance disparities remain across speaker groups. Previous studies have shown that ASR systems exhibit biases related to gender [4, 8], nativeness [9], and speech impairments [7]. These disparities are often attributed, at least in part, to imbalances in training data and variability in speech characteristics across different speaker populations [5]. Understanding the sources of these disparities is an important step toward developing more inclusive ASR systems.

For Dutch, previous work has shown that ASR performance is affected by speaker background and speech variation. Accents that deviate from standard Dutch, regional dialects, and non-nativeness have been associated with lower recognition accuracy [9]. Zhang et al. [7] found that 12 state-of-the-art ASR systems consistently perform worse on non-native speech compared to native speech for both Dutch read and human-machine interaction (HMI) speech. This pattern is not limited to Dutch, as similar results have been reported for other languages, such as English [10] and Swedish [11]. These disparities have been linked to pronunciation differences and phonemic variation [5, 9], and the speaker’s mother tongue [12].

In addition to phonetic factors, previous research has examined the influence of word-level and utterance-level linguistic factors. Mansfield et al. [13] investigated the effect of word categories, such as function, content, and conversational words, on ASR performance, finding that conversational words are associated with higher error rates. Similarly, Li et al. [14] found that nouns are among the most difficult word categories to recognize, and that WER decreases as utterance length increases. Zhang et al. [7] also reported that error

rates are lowest for utterances containing 5 to 15 words, further suggesting a relationship between utterance length and ASR accuracy. However, existing work on the linguistic factors underlying ASR performance disparities remains limited, especially with respect to how utterance length and word category relate to ASR errors across different speech styles. Understanding these patterns can help identify the linguistic conditions under which ASR systems are most error-prone.

The present study investigates the role of linguistic factors at the word and utterance level in ASR performance for non-native speakers of Dutch. The analysis focuses on two factors: *utterance length* and *word category*. To avoid introducing age-related effects, the study is restricted to adult speakers. Age has been found to influence both speech characteristics and ASR performance [5, 7], so including both children and adults could add variation unrelated to the linguistic factors under investigation. Because non-native speakers may differ in Dutch proficiency, an additional exploratory analysis is conducted to examine whether the observed utterance length and word category patterns differ across proficiency groups.

Furthermore, this study compares the effects of utterance length and word category separately for read and HMI speech, as these speech styles differ in their linguistic characteristics. Read speech is typically based on planned written prompts and therefore tends to follow more prescriptive linguistic norms. In contrast, HMI speech is closer to spontaneous or conversational speech and may contain more syntactic variation, fillers, self-corrections, repetitions, and variation in speech rate [15]. Such spontaneous speech characteristics have historically made speech more difficult for ASR systems to recognize than read speech, even for strong-performing systems [15].

The main research question of this study is as follows:

To what extent do utterance length and word category affect the performance of state-of-the-art ASR systems for Dutch speech from non-native adult speakers, and how do these effects differ between read speech and HMI speech?

2 Methodology

2.1 Dataset

The data used in this study are taken from the JASMIN speech corpus, a collection of approximately 115 hours of Dutch speech produced by adolescents, non-native speakers, and older adults from Flanders and the Netherlands [16]. The corpus contains both read speech and HMI dialogues. JASMIN was developed as a complement to the Corpus Gesproken Nederlands (CGN) [17]. The present analysis focuses specifically on speech from non-native adult speakers. For these speakers, the corpus includes proficiency information based on the Common European Framework of Reference for Languages (CEFR) [18]. In addition, JASMIN provides speaker metadata such as age, place of birth, and mother tongue.

The ASR transcriptions analyzed here come from Zhang et al. [7], who investigated the recognition performance of various ASR systems and models on Dutch diverse speech. As part of their evaluation, the ASR systems transcribed speech chunks segmented using the JASMIN annotations, and the resulting ASR outputs were compared with the reference transcripts provided by JASMIN. Before calculating errors, both the reference transcripts and ASR outputs were normalized: punctuation was removed with the exception of apostrophes, all characters were converted to lowercase, and Arabic numerals were transcribed

into their Dutch textual forms. The present study uses these normalized transcripts and the corresponding result files, which align the ASR outputs with the reference utterances and annotate the resulting insertion, deletion, and substitution errors. No additional filtering was applied to the normalized transcripts.

2.2 ASR Models

The ASR outputs analyzed in the present study were selected from the systems evaluated by Zhang et al. [7]. That study evaluated 12 ASR systems and models on Dutch diverse speech: nine company-developed systems and three custom Dutch ASR models trained from scratch. The company-developed systems included API-based systems, such as Google [19] and Microsoft [20] systems, as well as models downloaded from Hugging Face [21], including Meta’s Massive Multilingual Speech system [22], NVIDIA’s NeMo-nl [23] system, and three versions of OpenAI’s Whisper [24, 25, 26].

The present study focuses on Google Chirp 2 [27] and Whisper large-v3 [25], which were selected because they were among the strongest-performing systems in Zhang et al. [7]. Google Chirp 2 achieved the lowest WERs across the four JASMIN test sets of non-dysarthric speech, while Whisper large-v3 was also among the strongest-performing systems. Google Chirp 2 is a multilingual ASR model developed by Google and applied through the Google Cloud API, and Whisper large-v3 is a multilingual ASR model developed by OpenAI and downloaded from Hugging Face in the original evaluation.

2.3 Linguistic Analysis Tools

The analysis of word categories was performed using the `nl_core_news_lg` model from the spaCy [28] library version 3.8.14 in Python 3.12.4. spaCy was used to assign part-of-speech (PoS) tags to words, allowing word categories such as nouns, verbs, and pronouns to be analyzed. spaCy is a production-ready natural language processing (NLP) library with support for more than 75 languages. It was chosen over alternatives such as NLTK [29] and Stanza [30] because of its support for Dutch and its high computational efficiency [31], while maintaining strong accuracy for PoS tagging [32], with reported accuracy of up to 97% for Dutch models [33].

The Universal PoS tagset [34] was used for the word-category analysis. Table 1 lists the PoS tags assigned by spaCy and their definitions, based on the Universal PoS tagset. The Universal PoS tagset groups the tags into three categories: open-class words, closed-class words, and other. Open-class words, such as nouns, verbs, adjectives, and adverbs, are categories that can easily gain new words and lose old ones over time. Closed-class words, such as determiners, pronouns, adpositions, and conjunctions, mainly serve grammatical functions and change much less frequently [35]. The X tag is used for words that could not be assigned a standard PoS category [34].

2.4 Experimental Setup

To answer the research question, two complementary analyses were conducted. First, the effect of utterance length was examined by relating ASR error rates to the number of words in the reference utterance. This analysis was used to determine whether shorter and longer utterances were associated with different recognition performance. Second, the effect of word category was examined by assigning PoS tags to the reference and error words, making it possible to determine which word categories were associated with higher rates of deletion,

Table 1: Universal PoS tags used by spaCy, grouped into open-class words, closed-class words, and other categories.

Open-class words		Closed-class words		Other	
Tag	Definition	Tag	Definition	Tag	Definition
ADJ	Adjective	ADP	Adposition	PUNCT	Punctuation
ADV	Adverb	AUX	Auxiliary	SYM	Symbol
INTJ	Interjection	CCONJ	Coordinating conjunction	X	Other
NOUN	Noun	DET	Determiner		
PROPN	Proper noun	NUM	Numeral		
VERB	Verb	PART	Particle		
		PRON	Pronoun		
		SCONJ	Subordinating conjunction		

substitution, and insertion errors. All analyses were carried out separately for read and HMI speech and for each ASR model, allowing the effects of utterance length and word category to be compared across speech styles and systems.

2.4.1 Utterance Length

Utterance length was calculated as the number of words in the ground-truth transcription of each utterance. The reference utterances were then grouped according to their word count, and the number of utterances in each group was calculated. To examine the relationship between utterance length and ASR performance, the mean WER was calculated for each utterance length. In addition, insertion, deletion, and substitution rates were analyzed separately as a function of utterance length. For each rate, a 95% confidence interval was computed, the exact calculation of which is given in Section 2.5.

2.4.2 Word Category

Using spaCy, the reference transcripts were PoS-tagged to obtain the distribution of word categories in the corpus. This distribution was used to determine how frequently each category occurred in the reference data and to identify categories with too few tokens for reliable category-level analysis.

The words involved in ASR errors were also PoS-tagged. To characterize the errors, each error was described in terms of a *source* and a *destination*. The source corresponds to the reference word that was misrecognized, and the destination corresponds to the word produced by the ASR model. Under this definition, substitutions have both a source and a destination, insertions have only a destination, and deletions have only a source.

Since raw error counts are influenced by the overall frequency of each word category in the corpus, deletion and substitution source counts were normalized by the total number of words in the corresponding category in the reference corpus, following the methodology of Li et al. [14]. These normalized counts are referred to as *category error rates* in this study. For each category error rate, a 95% confidence interval was calculated. The exact calculation of the category error rate and the confidence interval is given in Section 2.5. Insertions were analyzed separately using the PoS category of the inserted word and reported as percentages of all insertions, as they have no corresponding source word in the reference transcripts.

For substitutions, the most common source–destination PoS-tag pairs were examined

and ranked by category error rate. Source PoS-tag categories accounting for less than 0.5% of words in the reference transcripts were excluded to avoid inflated ranks from very small counts. Thus, X, punctuation, and interjections were excluded for read speech, while X and punctuation were excluded for HMI speech.

Subsequently, all of the PoS tags were grouped according to Table 1 into three *aggregated categories*: open-class words, closed-class words, and other. The same analysis as done for PoS tags was repeated for these aggregated categories.

To test whether open-class and closed-class words differed significantly in their deletion and substitution rates, chi-squared tests of independence were performed separately for each ASR model, speech type, and error type. For each test, a 2×2 contingency table was constructed using the number of open-class and closed-class source words involved in the error and the corresponding number of open-class and closed-class source words not involved in the error. The aggregated “other” category was excluded from these tests.

For substitutions, category error rates were also computed for each source–destination combination of aggregated categories, such as open-class to closed-class, closed-class to open-class, and open-class to other substitutions.

2.4.3 CEFR Level

To examine whether the observed patterns differed by Dutch proficiency level, the utterance-length and word-category analyses were repeated after grouping the data by CEFR level: A1, A2, and B1. For utterance length, the error-rate patterns and confidence intervals for the three CEFR groups were compared within each ASR model, speech style, and error type. For word category, the PoS-level and aggregated category patterns for A1, A2, and B1 were compared directly for each ASR model, speech style, and error type. This CEFR analysis was treated as exploratory and descriptive, and no additional statistical tests were performed for the CEFR comparisons.

2.5 Evaluation Metrics

For the utterance-length analysis, the main evaluation metric was WER. For each utterance, WER was calculated as the total number of insertion, deletion, and substitution errors divided by the utterance length. Insertion, deletion, and substitution rates (indicated below as IR, DR, and SR) were also calculated separately by dividing the number of errors of each type by the utterance length. All rates were multiplied by 100 and are therefore reported as percentages:

$$\text{WER} = \frac{I + D + S}{N} \times 100\%, \quad \text{IR} = \frac{I}{N} \times 100\%, \quad \text{DR} = \frac{D}{N} \times 100\%, \quad \text{SR} = \frac{S}{N} \times 100\%. \quad (1)$$

where I , D , and S denote the number of insertions, deletions, and substitutions, respectively, and N denotes the utterance length, defined as number of words in the reference transcription.

To indicate the probabilistic uncertainty around each mean error rate estimate, 95% confidence intervals were calculated. Specifically, the sample standard deviation of the error rate was divided by the square root of the number of utterances to obtain the standard error, and the interval was computed as the mean $\pm$ 1.96 standard errors [36]. For length groups containing only one utterance, the confidence interval was set equal to the mean.

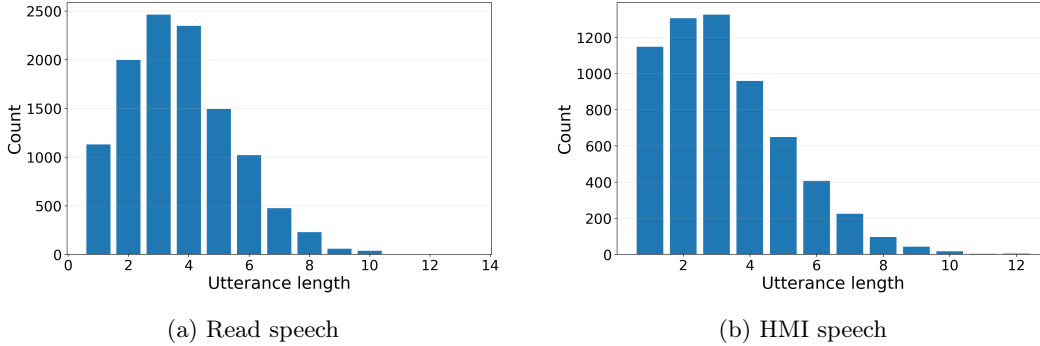


Figure 1: Utterance-length distributions of the reference utterances for read and HMI speech.

For the word category analysis, the *category error rate* was used to account for differences in the overall frequency of word categories in the reference corpus. For each word category, this metric was calculated as the number of deletion or substitution source errors in that category divided by the total number of reference words belonging to that category:

$$\text{Category error rate}_c = \frac{E_c}{N_c}, \quad (2)$$

where E_c denotes the number of deletion or substitution source errors for category c , and N_c denotes the total number of reference words belonging to category c .

For each category error rate, a 95% confidence interval was computed by treating the rate as a binomial proportion, following standard methods for categorical data analysis [37]. Since category error rates are proportions, confidence interval bounds were clipped to the interval $[0, 1]$.

For the chi-squared tests, each analysis used a 2×2 contingency table crossing word category with error type. The rows represented open-class and closed-class words, and the columns represented whether a reference token was involved in the relevant error type or not. This allowed the test to evaluate whether deletion or substitution error occurrence was independent of word category. The tests were conducted using standard chi-squared tests of independence for contingency tables [37]. A significance threshold of $p < 0.05$ was used.

3 Results

3.1 Utterance Length

Figure 1 shows the resulting utterance-length distributions separately for read and HMI speech. There were in total 11,259 read and 6,187 HMI utterances. Utterance lengths were generally short and right-skewed in both conditions. Read speech contained slightly longer utterances overall, with a median of 4 words compared with 3 words for HMI speech. The interquartile range was 2–5 words for read speech and 2–4 words for HMI speech, and the maximum utterance length was 13 words for read speech and 12 words for HMI speech. Although both distributions peaked at 3 words, very short utterances were more common in HMI speech: 61.1% of HMI utterances contained 3 words or fewer, compared with 49.7% of read utterances.

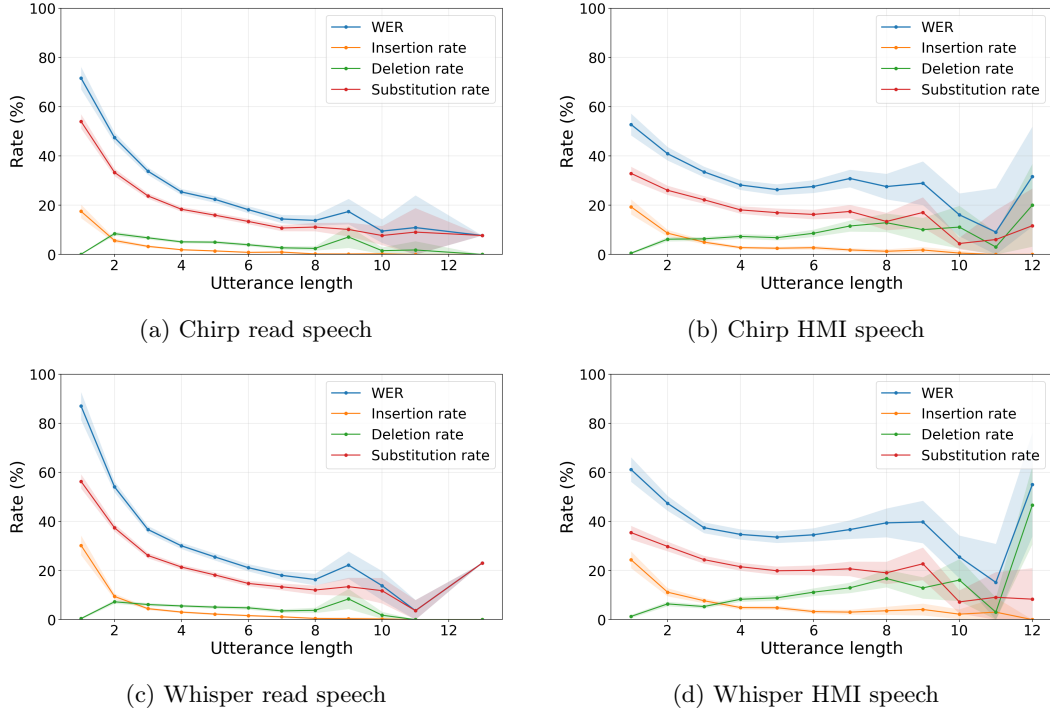


Figure 2: Mean WER, insertion rate, deletion rate, and substitution rate as a function of utterance length. Shaded regions indicate the 95% confidence intervals.

Figure 2 shows the mean WER, insertion rate, deletion rate, and substitution rate as a function of utterance length. For both ASR models and both speech types, WER was highest for one-word utterances and decreased for short utterances. However, the point at which this decrease stopped differed by speech type: for HMI speech, WER began to rise again from approximately five words onward, whereas for read speech it continued to decrease up to approximately eight words. At longer utterance lengths, the WER estimates fluctuated more strongly; the confidence intervals also became wider because fewer utterances were available at these lengths, as shown in Figure 1, indicating that the estimates are less stable. Read speech showed higher WER than HMI speech for utterances shorter than three words, but from three words onward, read speech generally showed lower WER.

The error-type patterns were highly similar for Google Chirp and Whisper. Overall, substitutions were the dominant error type across utterance lengths, and their pattern closely followed the overall WER trend: they were highest for short utterances and decreased more clearly for read speech than for HMI speech. Insertions also decreased rapidly as utterance length increased and remained relatively low for longer utterances. Deletions showed a different pattern across the two speech types: after the initial increase from one-word utterances, deletion rates decreased slightly for read speech, whereas they continued to increase for HMI speech.

The separate analysis for CEFR proficiency groups showed no clear systematic patterns. Although some differences between A1, A2, and B1 were observed for particular utterance lengths, speech styles, ASR models, or error types, these differences were not consistent.

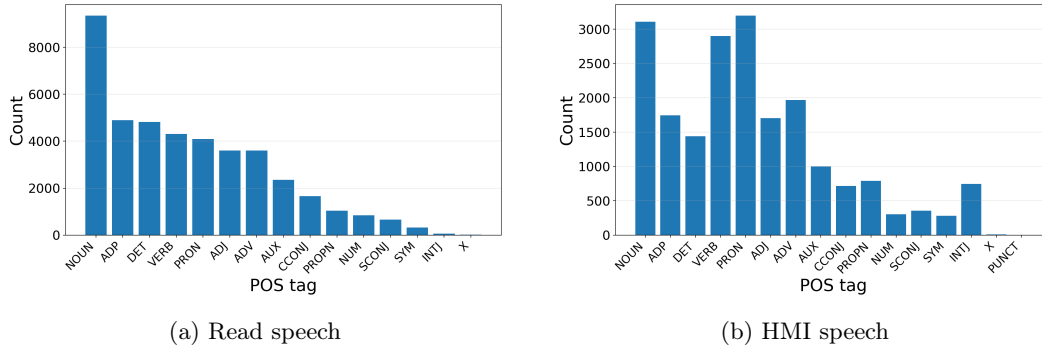


Figure 3: Distribution of PoS tags in the reference transcripts for read and HMI speech.

No proficiency group consistently showed higher or lower error rates than the others, and the confidence intervals largely overlapped. The figures comparing the error types for the different CEFR groups can be found in Appendix B.

3.2 Word Category

Figure 3 shows the PoS-tag distributions in the reference transcripts, which contained 41,623 words for read speech and 20,276 words for HMI speech. Read speech was more noun-heavy: nouns were the most frequent PoS category, accounting for more than 22% of all words, followed by adpositions, determiners, and verbs at approximately 10–12% each. In contrast, the HMI distribution was more evenly spread across categories. Pronouns were the most frequent category at approximately 16%, followed closely by nouns and verbs at approximately 15% and 14%, respectively.

Figure 4 shows the normalized deletion and substitution source category error rates, together with the insertion destination percentages, for each PoS category. Insertions are shown as bar charts because they represent a distribution over inserted-word categories, unlike deletion and substitution category error rates. Overall, substitutions showed the clearest category-specific pattern, deletion rates were more affected by low-frequency categories, and insertions were more skewed toward particular destination categories, with some differences between read and HMI speech. The raw error counts corresponding to these category error rates and insertion percentages can be found in Appendix A.

For deletions, the highest category error rates were observed for low-frequency categories such as symbols and X. These values should be interpreted cautiously, because normalization can inflate rates for categories with very few reference tokens. Excluding these low-frequency categories, proper nouns and numerals showed the highest deletion rates. Some differences between speech types were also observed: interjections had higher deletion rates in read speech, while determiners, coordinating and subordinating conjunctions, and auxiliaries had higher deletion rates in HMI speech.

For substitutions, numerals showed the highest category error rates, but inspection of the substitution pairs showed that this was largely driven by substitutions from the word *één* to *een*. This pattern was more prominent in read speech than in HMI speech. Among the remaining categories, proper nouns, nouns, adjectives, and verbs showed relatively high substitution rates. HMI speech showed higher substitution rates than read speech for several categories, including determiners, subordinating conjunctions, adverbs, and adpositions,

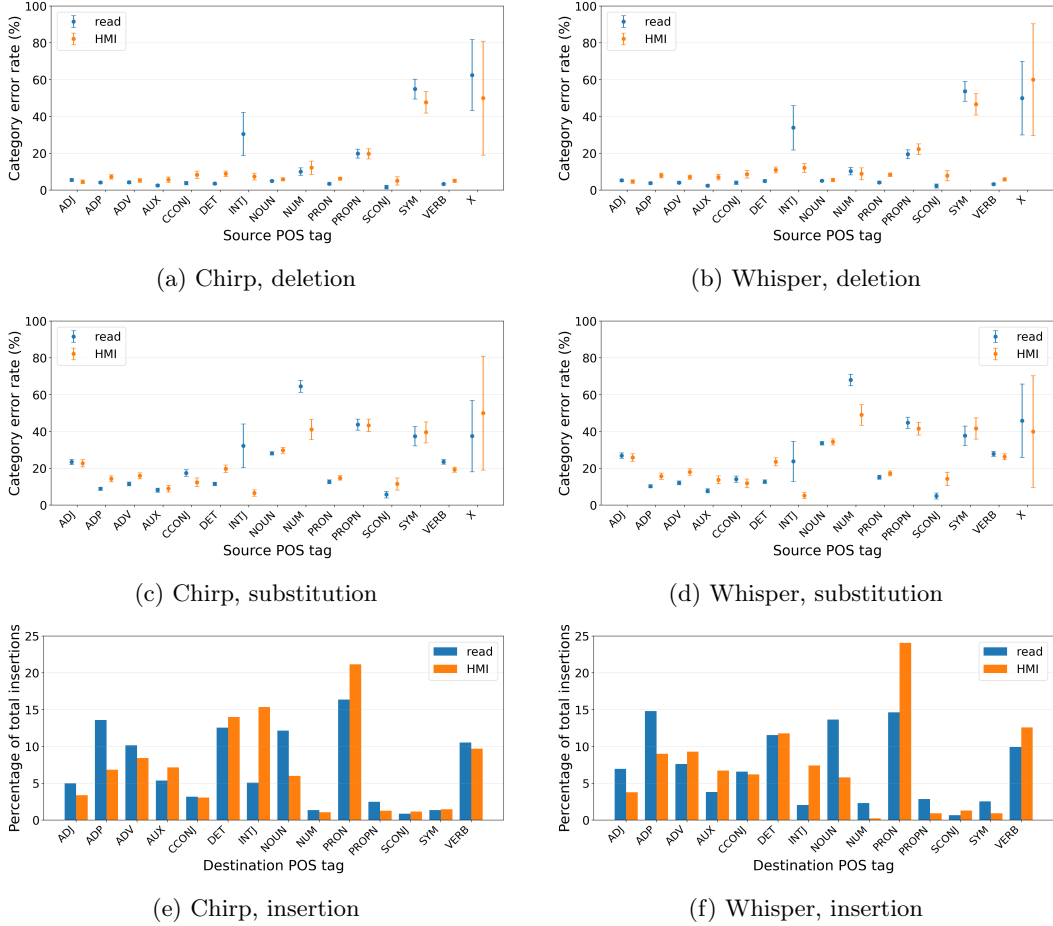


Figure 4: Deletion source category error rates (%), substitution source category error rates (%), and insertion destination percentages by PoS category. In each subplot, read and HMI speech are displayed side by side.

whereas interjections showed higher substitution rates in read speech than in HMI speech.

For insertions, pronouns, determiners, and verbs were among the most frequent inserted-word categories in both speech types. Pronoun insertions were especially prominent in HMI speech, while nouns and adpositions accounted for a larger share of insertions in read speech. Interjections were more common in HMI insertions.

Table 2 shows the top 10 substitution source–destination PoS-tag pairs ranked by category error rate. Numeral-to-determiner substitutions had the highest category error rate for both ASR models and speech styles. As noted above, this was largely driven by cases where *één* was replaced by *een*, which changes the PoS tag from numeral to determiner. Several within-category substitutions, including noun-to-noun, verb-to-verb, and adjective-to-adjective substitutions, also appeared in the top ten for both models and speech styles. This shows that many high-ranked substitutions preserve the PoS category. Proper-noun substitutions were also common, especially proper-noun-to-noun and proper-noun-to-proper-noun

Table 2: Top 10 substitution source–destination PoS-tag pairs ranked by category error rate (%), separately for each ASR model and speech type.

Rank	Google Chirp						Whisper					
	Read			HMI			Read			HMI		
	Src.	Dest.	Rate	Src.	Dest.	Rate	Src.	Dest.	Rate	Src.	Dest.	Rate
1	NUM	DET	39.0	NUM	DET	21.4	NUM	DET	40.8	NUM	DET	24.7
2	NOUN	NOUN	14.4	NOUN	NOUN	13.3	NOUN	NOUN	17.3	NOUN	NOUN	16.0
3	PROPN	NOUN	13.2	PROPN	PROPN	11.8	PROPN	NOUN	13.5	PROPN	NOUN	11.2
4	VERB	VERB	11.6	SYM	PRON	10.2	VERB	VERB	11.2	ADJ	ADJ	10.2
5	NUM	ADP	8.4	ADJ	ADJ	8.9	ADJ	ADJ	8.6	SYM	PRON	9.9
6	ADJ	ADJ	7.9	PROPN	NOUN	8.5	SYM	NOUN	8.3	PRON	PRON	8.6
7	SYM	NOUN	7.7	PROPN	PRON	8.0	ADJ	NOUN	8.1	VERB	VERB	8.4
8	CCONJ	DET	7.5	PRON	PRON	7.7	PROPN	PROPN	8.0	PROPN	PROPN	8.2
9	CCONJ	ADP	6.9	VERB	VERB	6.4	VERB	NOUN	7.6	DET	DET	7.1
10	ADJ	NOUN	5.8	DET	DET	5.4	NUM	ADP	6.9	SYM	NOUN	6.0

Table 3: Chi-squared tests comparing open-class and closed-class deletion and substitution rates (%).

Model	Speech	Error type	Open-class	Closed-class	$\chi^2(1)$	p
Chirp	Read	Deletion	5.4	3.8	58.06	< .001
Chirp	Read	Substitution	24.5	13.3	823.39	< .001
Chirp	HMI	Deletion	6.4	7.1	3.87	.049
Chirp	HMI	Substitution	23.0	15.4	177.51	< .001
Whisper	Read	Deletion	5.3	4.3	25.83	< .001
Whisper	Read	Substitution	28.4	14.3	1198.46	< .001
Whisper	HMI	Deletion	7.4	8.6	9.97	.002
Whisper	HMI	Substitution	26.7	18.1	206.04	< .001

substitutions. In HMI speech, pronoun-to-pronoun and determiner-to-determiner substitutions also appeared among the highest-ranked pairs.

When the PoS tags were aggregated into open-class, closed-class, and “other” categories, open-class words were the largest group in both speech types. The distributions were similar overall: read speech contained 52.8% open-class words, 46.4% closed-class words, and 0.8% other tokens, while HMI speech contained 55.3%, 43.2%, and 1.5%, respectively.

Figure 5 shows the deletion rates, substitution rates, and insertion percentages for the aggregated word categories. The *other* category had the highest deletion and substitution rates, but this category represented only a small proportion of the reference corpus and is therefore not treated as a central pattern. For insertions, closed-class words accounted for the largest share of inserted words for both models and speech styles.

Table 3 shows the chi-squared tests comparing open-class and closed-class deletion and substitution rates. The χ^2 statistic shows how much the observed open-class and closed-class error counts deviate from the counts expected if word category and error occurrence were unrelated. The clearest difference was found for substitutions: for both ASR models and speech styles, open-class words had significantly higher substitution rates than closed-class words. Deletion rates also differed significantly between open- and closed-class words, but these differences were smaller and varied by speech style, with higher open-class deletion rates in read speech and higher closed-class deletion rates in HMI speech.

Table 4 shows the substitution source–destination pairs for each aggregated word category. Substitutions mostly preserved the aggregated category: open-class words were more

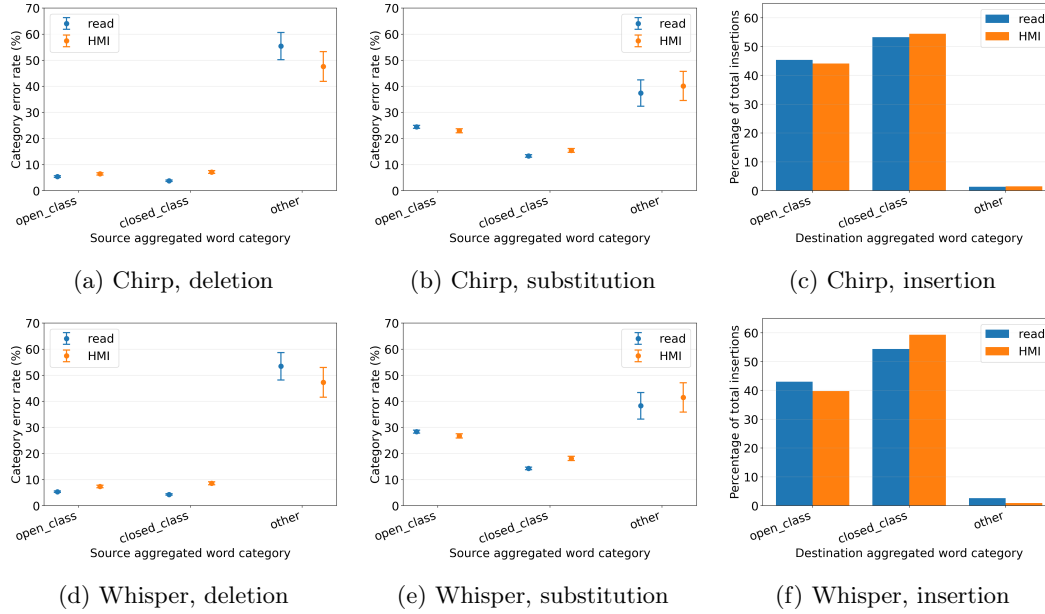


Figure 5: Deletion category error rates (%), substitution category error rates (%), and insertion percentages for aggregated word categories. In each subplot, read and HMI speech are displayed side by side.

often substituted by open-class words, and closed-class words were more often substituted by closed-class words. Cross-class substitutions were lower overall, but both open-to-closed and closed-to-open substitutions were higher in HMI speech than in read speech.

CEFR-related differences were generally limited, with deletion and substitution patterns appearing broadly similar across proficiency groups. The main exception was found for HMI deletions, where A1 showed higher deletion rates for proper nouns and symbols. However, many of these involved single-letter or incomplete tokens, as observed during inspection of the errors. The clearest differences were observed for insertions, particularly for Chirp. In HMI speech, A1 showed a different insertion distribution from A2 and B1, with a larger share of open-class insertions at the aggregated level and more interjection and verb insertions at the PoS level. For both read and HMI speech, A1 showed lower pronoun insertions than A2 and B1. Whisper showed more similar insertion distributions across CEFR groups. The figures comparing the errors for these CEFR groups can be found in Appendix C.

4 Discussion

The findings show that both utterance length and word category are associated with differences in ASR performance for non-native adult speakers of Dutch, but that these effects depend on speech style. For utterance length, read speech showed a clearer decrease in WER as utterances became longer, while HMI speech showed a weaker pattern and began to increase again from around five words. For word category, substitutions showed the strongest category-specific pattern: open-class words had higher substitution rates than closed-class words for both ASR models and both speech styles. Insertions, in contrast, were mainly

Table 4: Substitution source–destination pairs for aggregated word categories, with category error rates (%), separately for each ASR model and speech type.

Source	Destination	Chirp		Whisper	
		Read	HMI	Read	HMI
Open-class	Open-class	19.7	16.2	23.2	19.6
Open-class	Closed-class	4.3	6.2	4.6	6.7
Open-class	Other	0.5	0.6	0.5	0.4
Closed-class	Open-class	3.7	4.2	4.3	5.9
Closed-class	Closed-class	9.2	10.5	9.8	12.1
Closed-class	Other	0.4	0.7	0.2	0.2
Other	Open-class	20.6	16.7	21.7	15.6
Other	Closed-class	15.7	21.4	14.9	24.8
Other	Other	1.1	2.0	1.7	1.0

closed-class words. Overall, these patterns show that linguistic properties of the utterances themselves provide useful information about the errors produced by ASR systems.

The utterance-length analysis largely reproduces the pattern reported by Zhang et al. [7], but specifically for non-native adult speakers. In line with that study and with Li et al. [14], WER was highest for the shortest utterances and generally decreased as utterance length increased. This pattern was clearest for read speech, while HMI speech showed an initial decrease followed by an increase from around five words onward, again matching the general trend reported by Zhang et al. [7]. However, the dataset of non-native adults contained mostly short utterances, so the lowest-error range reported by Zhang et al. [7] for utterances of 5–15 words is only partly represented here. Moreover, the wider confidence intervals at longer utterance lengths indicate that these estimates should be interpreted cautiously.

The error-type analysis shows that substitutions and insertions largely followed the overall WER pattern, while deletions differed by speech style: after the shortest utterances, deletion rates decreased slightly for read speech but continued to increase for HMI speech. One possible explanation for both the increase in WER for longer HMI utterances and the continued increase in HMI deletions is that longer utterances may differ not only in length, but also in linguistic structure. Read speech is based on planned prompts and is therefore likely to remain relatively controlled as utterances become longer. HMI speech, in contrast, is closer to spontaneous interaction and may contain more variation in structure, fillers, self-corrections, repetitions, incomplete words, and speech rate, all of which can make recognition more difficult for ASR systems [15]. Previous work has also shown that ASR systems can fail to capture conversational and incomplete words [38], and that conversational words are associated with higher error rates [13]. Since recent work suggests that speech models can encode linguistic structure in their internal representations [39], differences in utterance structure may affect recognition performance. The HMI pattern may therefore indicate that additional context is not always beneficial when it is accompanied by more spontaneous or less predictable structure.

Turning to word-category patterns, the substitution results are consistent with previous analyses of ASR errors. Open-class words had higher substitution rates than closed-class words, which aligns with Mansfield et al. [13], who found that content words undergo more substitution errors. This comparison is appropriate because open-class and closed-class words largely correspond to content and function words, respectively [35]. The result also partly aligns with Li et al. [14], who reported high error rates for nouns and lower rates for function words. One possible explanation is that open-class categories are less stable over

time, as new words can be introduced and older words can become obsolete [35], requiring ASR systems to adapt to a changing vocabulary. In contrast, function words are generally more frequent and predictable than content words [40], which may make them easier for ASR systems to recognize.

A further pattern in the substitution analysis is that many high-ranked substitutions preserved the word category. Examples from the present data, such as *reis* to *reisje* and *wilt* to *wil*, suggest that words from the same category may sometimes share acoustic or morphological similarities, such as inflectional variants or closely related noun forms. Another possible explanation is that the surrounding utterance context constrains which word categories are plausible in a given position, making within-category substitutions more likely.

Inserted words were most often closed-class words. This may again be related to the frequency and predictability of function words [40]. While these properties may make function words easier to recognize when they occur in the reference utterance, they may also make ASR systems more likely to insert them when producing a fluent transcription. The insertion distribution also differed by speech style. This suggests that insertions may be influenced by the surrounding context in which the insertion occurs, which might be different for read and HMI speech.

The exploratory CEFR analysis did not show strong or systematic differences between proficiency groups. This only partially aligns with Feng et al. [5], who reported that WER generally decreases as CEFR level increases, although their results were also not fully consistent across all speaker groups. In the present study, CEFR-related differences appeared only for particular utterance lengths, speech styles, models, or error types. These differences were not stable enough to suggest a clear proficiency-level effect. Word-category patterns were also broadly similar across CEFR levels. Chirp showed somewhat more variation than Whisper, especially for insertion distributions, but these differences were limited and should be treated as descriptive rather than as main findings.

Overall, these results show that aggregate WER alone does not fully capture the linguistic sources of ASR errors for non-native Dutch speech. Utterance length affected read and HMI speech differently, suggesting that the structure and conversational nature of the utterance may shape how useful additional context is. At the word level, open-class words were more vulnerable to substitutions, while closed-class words were more prominent in insertions. Together, these findings indicate that ASR performance disparities are linked not only to who is speaking, but also to what kind of linguistic material is being recognized.

4.1 Limitations

Several limitations should be considered when interpreting these findings. First, the word-category analysis depended on automatic PoS tagging. Although this made it possible to analyze a large number of reference and error words, some transcripts contained incomplete words, single-letter tokens, and other forms that may have received unreliable tags. This is especially relevant for low-frequency categories, where a small number of tagging errors can strongly affect the estimated category error rates.

Second, the dataset contained mostly short utterances. As a result, the estimates for longer utterance lengths were based on fewer examples and had wider confidence intervals. The observed increase in error rates for longer HMI utterances should therefore be interpreted cautiously until it can be tested on data with more medium and long utterances.

Third, utterance length and word category were analyzed mostly separately. This means that the present study can show that both factors are associated with ASR performance,

but it does not directly test how they interact. For example, it remains unclear whether certain word categories are more error-prone specifically in shorter or longer utterances.

Finally, several interpretations in the discussion rely on utterance-internal structure. This includes the possibility that longer HMI utterances provide less predictable context, that surrounding context makes within-category substitutions more likely, and that insertion patterns depend on the local context in which words occur. These explanations are plausible, but they were not tested directly in the present analysis.

5 Conclusions and Future Work

This study investigated how utterance length and word category affect the performance of state-of-the-art automatic speech recognition systems for Dutch speech from non-native adult speakers. It also examined how these effects differ between read speech and HMI speech. The results show that both utterance length and word category are associated with ASR performance, but that their effects depend on speech style. Utterance length mainly affected the overall error patterns across complete utterances, while word category revealed systematic differences in the types of words that were more likely to be deleted, substituted, or inserted.

Utterance length affected ASR performance, with very short utterances being especially error-prone. This suggests that limited utterance context can make recognition more difficult. Although WER generally decreased as utterances became longer, this pattern was more consistent for read speech than for HMI speech, where error rates increased again for longer utterances. This difference indicates that additional context is not always beneficial. The error-type analysis showed that while substitutions decreased with utterance length in both speech styles, deletions showed a clearer speech-style difference, decreasing slightly in read speech but increasing in HMI speech.

Word category also played an important role in ASR error patterns. Open-class words had higher substitution rates than closed-class words, showing that the type of word being recognized affects the likelihood of substitution errors. Insertions followed a different pattern, as they mainly consisted of closed-class words, suggesting that ASR systems may be more likely to add frequent grammatical words when producing a fluent transcription. The word-category analysis also showed that many substitutions preserved the original word category, indicating that ASR systems often replace words with linguistically similar alternatives rather than with completely unrelated word types.

Future work could extend this analysis in several directions. First, PoS-based ASR error analysis could be improved through additional preprocessing or manual validation of incomplete and ambiguous tokens. Second, datasets with more medium and long utterances could be used to test whether the observed utterance length patterns remain stable when longer utterances are better represented. Third, future research could examine the interaction between utterance length and word category more directly, for example by testing whether certain word categories become more or less error-prone in shorter or longer utterances. Finally, utterance structure could be analyzed in more detail using features such as syntactic complexity, word order, dependency relations, and local word context. This would help determine whether the patterns observed here are driven not only by utterance length and word category separately, but also by the internal structure of the utterance.

6 Responsible Research

The dataset used in this study, JASMIN-CGN, is publicly available, and no new data were collected as part of this project. The analysis also used existing ASR output files from the previous study on which this research is based. Since no new interaction with human subjects took place, this project did not involve new data collection from participants. The dataset and related sources were cited, and the data were used only for the research purposes described in this paper.

To support reproducibility, the methodology was described, including the dataset, ASR systems, preprocessing steps, linguistic analysis tools, error metrics, and statistical procedures. Furthermore, where relevant, the versions of the tools were mentioned. Python scripts were used to process the data files and generate the graphs and tables. These scripts are made available together with the analysis results (plots and CSV files) so that the analysis process can be inspected and reproduced ¹.

AI tools were used during the project to help with writing and coding. ChatGPT was used to assist with writing by improving grammar, spelling, clarity, and structure. The content, arguments, references, and interpretation of the results were provided by the author. All AI-assisted text was read critically and checked against the cited literature, methodology, and results to ensure that it did not introduce unsupported claims or incorrect information. Codex was used to assist with writing Python scripts for data analysis and for generating graphs and tables. The resulting scripts were manually reviewed, and their outputs were checked against the expected input files and result patterns. In this way, AI assistance was used as a support tool, while responsibility for the final text, code, analysis, and conclusions remained with the author. The prompts and project files that were accessible by Codex did not contain any personal information.

Finally, the broader impact of the research was considered. The findings may contribute to a better understanding of how ASR systems behave for non-native Dutch speech and under which linguistic conditions errors are more likely to occur. This can support more transparent evaluation of speech technologies and may help future work identify where ASR systems require improvement. At the same time, the results should be interpreted within the scope of this study: they concern specific ASR systems, a specific dataset, and specific linguistic analyses. They should therefore not be generalized to all non-native speakers or all ASR systems without further evidence.

¹https://github.com/KiarashK2-Work/Research_Project

References

- [1] Dong Yu and Lin Deng. *Automatic speech recognition*. Vol. 1. Springer, 2016.
- [2] Alakbar Valizada, Natavan Akhundova, and Samir Rustamov. “Development of speech recognition systems in emergency call centers”. In: *Symmetry* 13.4 (2021), p. 634.
- [3] Tobias Hodgson and Enrico Coiera. “Risks and benefits of speech recognition for clinical documentation: a systematic review”. In: *Journal of the american medical informatics association* 23.e1 (2016), e169–e179.
- [4] Allison Koencke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Touns, John R Rickford, Dan Jurafsky, and Sharad Goel. “Racial disparities in automated speech recognition”. In: *Proceedings of the national academy of sciences* 117.14 (2020), pp. 7684–7689.
- [5] Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. “Towards inclusive automatic speech recognition”. In: *Computer Speech & Language* 84 (2024), p. 101567. ISSN: 0885-2308. DOI: <https://doi.org/10.1016/j.csl.2023.101567>. URL: <https://www.sciencedirect.com/science/article/pii/S0885230823000864>.
- [6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. “Robust speech recognition via large-scale weak supervision”. In: *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.
- [7] Yuanyuan Zhang, Thomas De Valck, and Odette Scharenborg. “Speech recognition performance disparities between Dutch diverse speaker groups”. In: *Phonetica* (2026). DOI: [doi:10.1515/phon-2025-0061](https://doi.org/10.1515/phon-2025-0061).
- [8] Mahault Garnerin, Solange Rossato, and Laurent Besacier. “Gender representation in French broadcast corpora and its impact on ASR performance”. In: *Proceedings of the 1st international workshop on AI for smart TV content production, access and delivery*. 2019, pp. 3–9.
- [9] Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iska Janssen. “Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers”. In: *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpED)*. IEEE, 2023, pp. 146–151.
- [10] Nathan Roll and Calbert Graham. “Scaling conformation bias in automatic speech recognition”. In: *Proc. SLaTE* (2025), pp. 133–137.
- [11] Ronald Cumbal, Birger Moell, José Lopes, and Olof Engwall. “You don’t understand me!: Comparing ASR results for L1 and L2 speakers of Swedish”. In: *arXiv preprint arXiv:2405.13379* (2024).
- [12] Albina Sarymsakova and Patricia Martín Rodilla. “Assessment of L2 Spanish pronunciation accuracy via Automatic Speech Recognition”. In: *Phonica* 21 (2025), pp. 1–23.
- [13] Courtney Mansfield, Sara Ng, Gina-Anne Levow, Richard A Wright, and Mari Ostendorf. “Revisiting Parity of Human vs. Machine Conversational Speech Transcription”. In: *Proc. Interspeech 2021* (2021), pp. 1997–2001.
- [14] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. “ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition”. In: *arXiv preprint arXiv:2305.16065* (2023).

- [15] Philipp Gabler, Bernhard C. Geiger, Barbara Schuppler, and Roman Kern. “Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition”. In: *Information* 14.2 (2023). ISSN: 2078-2489. DOI: 10.3390/info14020137. URL: <https://www.mdpi.com/2078-2489/14/2/137>.
- [16] Instituut voor de Nederlandse Taal. *TSTC JASMIN-spraakcorpus*. URL: <https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus/> (visited on 04/09/2026).
- [17] Instituut voor de Nederlandse Taal. *TSTC Corpus Gesproken Nederlands*. URL: <https://taalmaterialen.ivdnt.org/download/tstc-corpus-gesproken-nederlands/> (visited on 04/09/2026).
- [18] Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. URL: <https://www.coe.int/en/web/common-european-framework-reference-languages> (visited on 06/09/2026).
- [19] Google Cloud. *Compare transcription models*. 2026. URL: <https://docs.cloud.google.com/speech-to-text/docs/transcription-model> (visited on 06/13/2026).
- [20] Microsoft. *azure.cognitiveservices.speech package*. Microsoft Speech SDK for Python documentation. 2026. URL: <https://learn.microsoft.com/en-us/python/api/azure-cognitiveservices-speech/azure.cognitiveservices.speech?view=azure-python> (visited on 06/13/2026).
- [21] Hugging Face. *Hugging Face*. 2026. URL: <https://huggingface.co/> (visited on 06/13/2026).
- [22] Meta AI. *Massively Multilingual Speech (MMS) – Finetuned ASR – ALL*. Hugging Face model card. 2023. URL: <https://huggingface.co/facebook/mms-1b-all> (visited on 06/13/2026).
- [23] NVIDIA. *NVIDIA FastConformer-Hybrid Large (nl)*. Hugging Face model card. 2023. URL: https://huggingface.co/nvidia/stt_nl_fastconformer_hybrid_large_pc (visited on 06/13/2026).
- [24] OpenAI. *Whisper large-v2*. Hugging Face model card. 2022. URL: <https://huggingface.co/openai/whisper-large-v2> (visited on 06/13/2026).
- [25] OpenAI. *Whisper large-v3*. URL: <https://huggingface.co/openai/whisper-large-v3> (visited on 04/09/2026).
- [26] OpenAI. *‘turbo’ model release*. GitHub discussion. 2024. URL: <https://github.com/openai/whisper/discussions/2363> (visited on 06/13/2026).
- [27] Google Cloud. *Chirp 2: Enhanced multilingual accuracy*. URL: <https://docs.cloud.google.com/speech-to-text/docs/models/chirp-2> (visited on 04/09/2026).
- [28] Explosion. *spaCy: Industrial-strength Natural Language Processing in Python*. URL: <https://spacy.io/> (visited on 05/22/2026).
- [29] Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O’Reilly Media, 2009.
- [30] Stanford NLP Group. *Stanza: A Python NLP Package for Many Human Languages*. URL: <https://stanfordnlp.github.io/stanza/index.html> (visited on 05/22/2026).
- [31] Explosion. *spaCy Facts & Figures*. URL: <https://spacy.io/usage/facts-figures> (visited on 05/22/2026).

- [32] Explosion. *Linguistic Features: Part-of-speech tags and dependencies*. URL: <https://spacy.io/usage/spacy-101%5C#annotations-pos-deps> (visited on 05/22/2026).
- [33] Explosion. *Dutch spaCy Models Documentation*. URL: <https://spacy.io/models/nl> (visited on 05/22/2026).
- [34] Universal Dependencies contributors. *Universal POS Tags*. URL: <https://universaldependencies.org/u/pos/> (visited on 05/22/2026).
- [35] Allison Burkette, Maura D. Wood, William A. Jr. Kretzschmar, and John Nerbonne. “Parts of Speech”. In: *Exploring Linguistic Science: Language Use, Complexity, and Interaction*. Cambridge University Press, 2018. Chap. 9, pp. 100–111.
- [36] Avijit Hazra. “Using the confidence interval confidently”. In: *Journal of thoracic disease* 9.10 (2017), p. 4125.
- [37] Alan Agresti. *Categorical Data Analysis*. 2nd ed. John Wiley & Sons, 2002.
- [38] Alianda Lopez, Andreas Liesenfeld, and Mark Dingemanse. “Evaluation of automatic speech recognition for conversational speech in Dutch, English and German: What goes missing?” In: *Proceedings of the 18th conference on natural language processing (KONVENS 2022)*. 2022, pp. 135–143.
- [39] Marianne de Heer Kloots, Martijn Bentum, Hosein Mohebbi, Charlotte Pouw, Gaofei Shen, and Willem Zuidema. “Tracking the emergence of linguistic structure in self-supervised models learning from speech”. In: *arXiv preprint arXiv:2604.02043* (2026).
- [40] Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. “Predictability effects on durations of content and function words in conversational English”. In: *Journal of Memory and Language* 60.1 (2009), pp. 92–111.

A PoS error distribution

Table 5: Counts of deletion sources, substitution sources, substitution destinations, and insertion destinations by PoS category, separately for each ASR model and speech type.

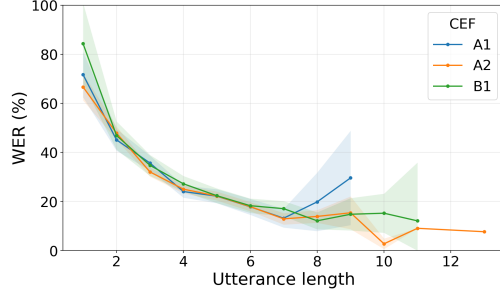
Deletion sources																	
Model	Speech	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
Chirp	Read	198	205	154	60	63	171	18	465	85	140	206	0	11	179	142	15
	HMI	77	126	104	57	60	128	55	183	37	199	156	0	18	135	146	5
Whisper	Read	191	186	146	57	67	241	20	473	87	171	203	0	15	175	140	12
	HMI	80	138	138	71	62	157	90	172	27	268	176	1	28	132	169	6

Substitution sources																	
Model	Speech	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
Chirp	Read	846	434	415	192	290	557	19	2628	545	519	455	0	38	122	1016	9
	HMI	387	250	313	90	89	284	49	927	125	473	342	1	41	112	559	5
Whisper	Read	972	503	436	182	234	613	14	3142	574	618	465	0	33	123	1200	11
	HMI	441	274	355	138	85	339	39	1069	149	551	328	0	51	118	764	4

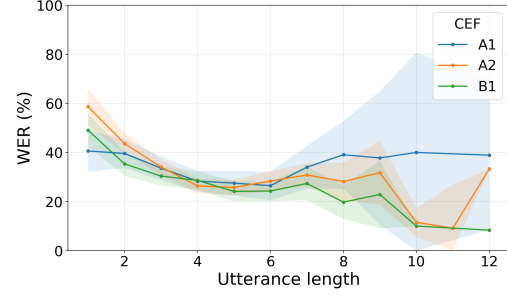
Substitution destinations																	
Model	Speech	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
Chirp	Read	820	558	376	256	162	939	52	2272	199	599	375	0	57	193	1227	0
	HMI	334	254	279	198	90	396	103	834	44	660	229	1	38	132	455	0
Whisper	Read	888	539	405	261	329	1023	37	2771	104	632	602	0	61	162	1302	4
	HMI	436	299	411	180	169	423	92	990	31	715	237	0	64	62	596	0

Insertion destinations																	
Model	Speech	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
Chirp	Read	52	142	106	56	33	131	53	127	14	171	26	0	9	14	110	0
	HMI	32	65	80	68	29	133	146	57	10	201	12	0	11	14	92	0
Whisper	Read	115	245	126	63	109	191	34	226	38	242	47	0	11	42	164	1
	HMI	53	126	130	94	87	165	104	81	3	337	13	0	18	13	176	0

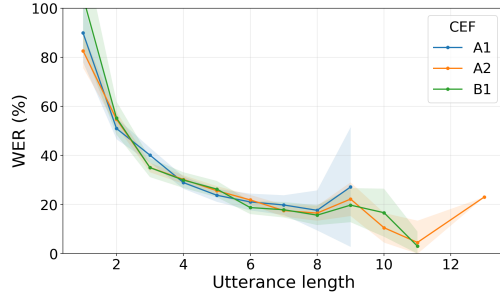
B CEFR: Utterance Length



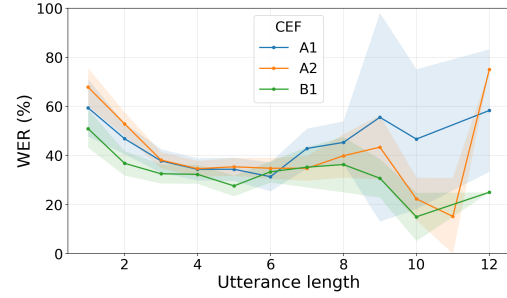
(a) Chirp read speech



(b) Chirp HMI speech

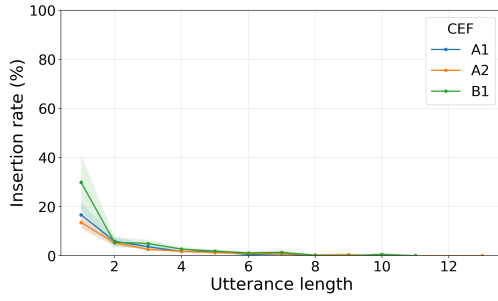


(c) Whisper read speech

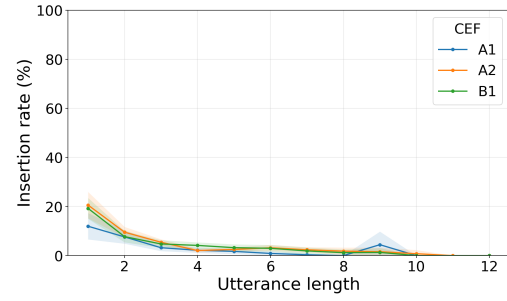


(d) Whisper HMI speech

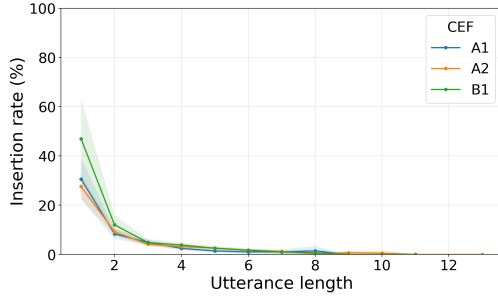
Figure 6: CEFR-specific mean WER as a function of utterance length. Shaded regions indicate the 95% confidence intervals.



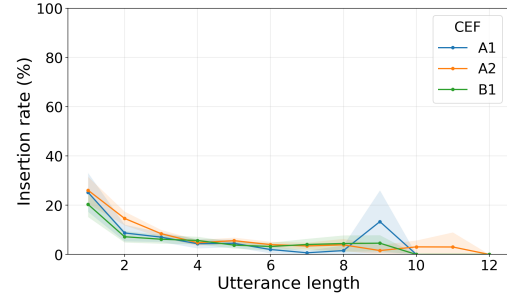
(a) Chirp read speech



(b) Chirp HMI speech

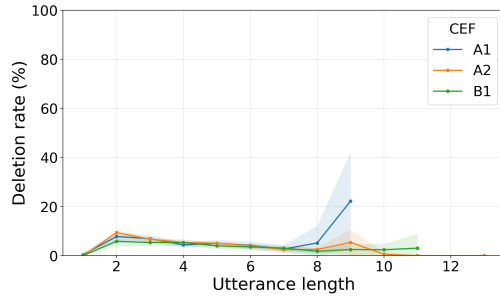


(c) Whisper read speech

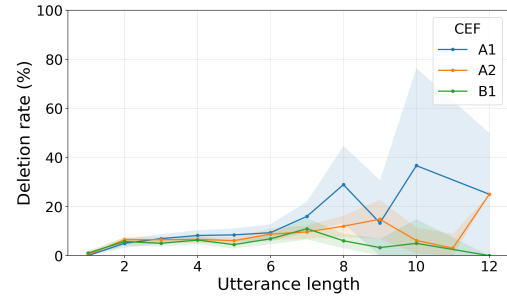


(d) Whisper HMI speech

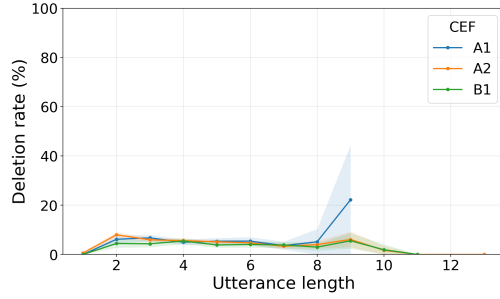
Figure 7: CEFR-specific insertion rate as a function of utterance length. Shaded regions indicate the 95% confidence intervals.



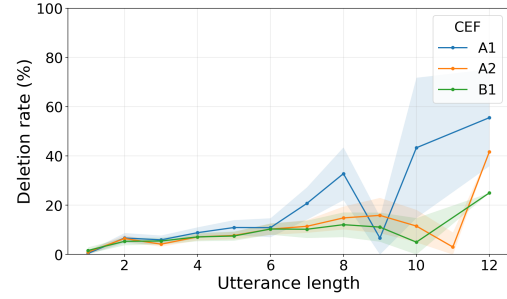
(a) Chirp read speech



(b) Chirp HMI speech

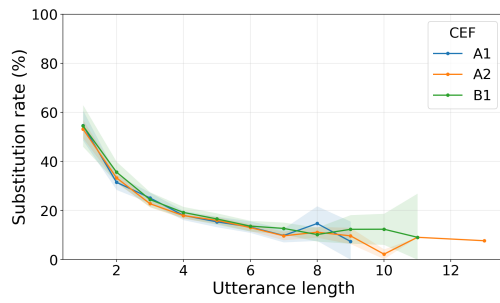


(c) Whisper read speech

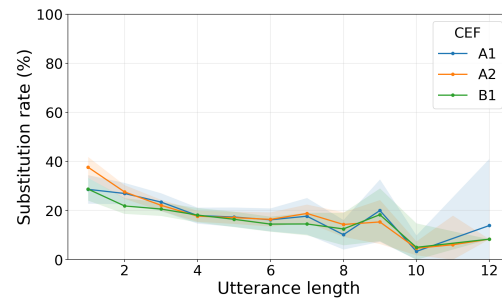


(d) Whisper HMI speech

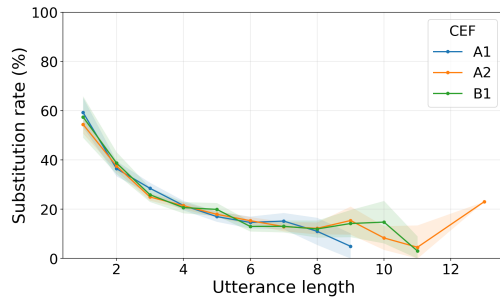
Figure 8: CEFR-specific deletion rate as a function of utterance length. Shaded regions indicate the 95% confidence intervals.



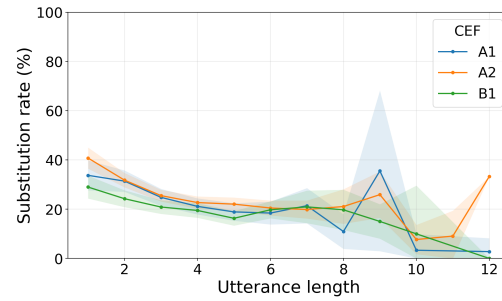
(a) Chirp read speech



(b) Chirp HMI speech



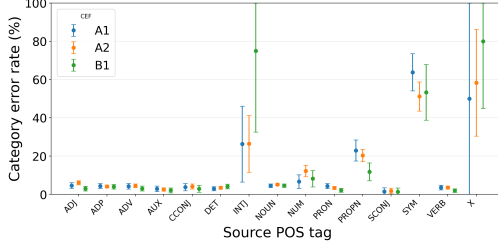
(c) Whisper read speech



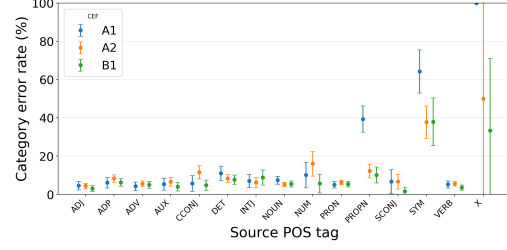
(d) Whisper HMI speech

Figure 9: CEFR-specific substitution rate as a function of utterance length. Shaded regions indicate the 95% confidence intervals.

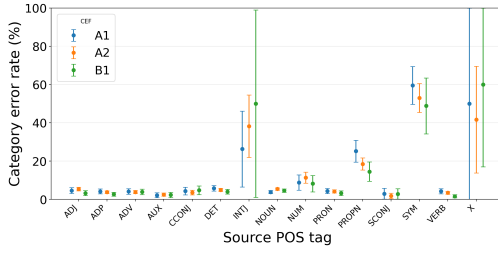
C CEFR: Word Category



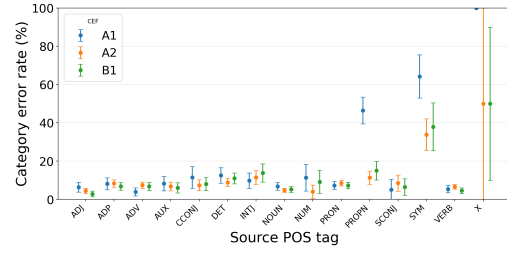
(a) Chirp read speech



(b) Chirp HMI speech

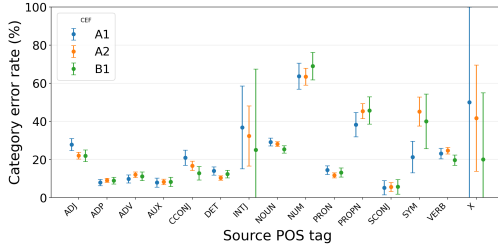


(c) Whisper read speech

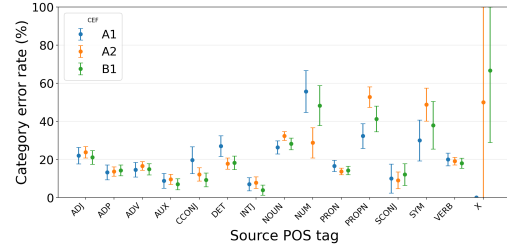


(d) Whisper HMI speech

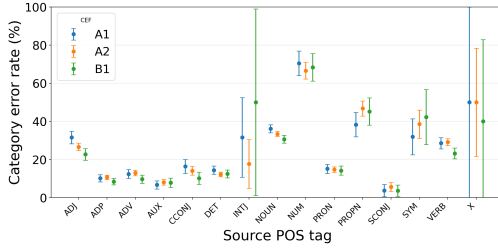
Figure 10: CEFR-specific deletion source category error rates by PoS category.



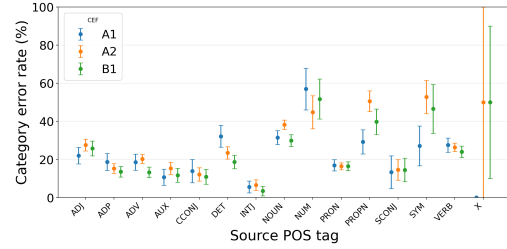
(a) Chirp read speech



(b) Chirp HMI speech

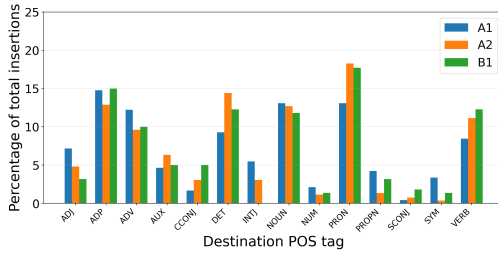


(c) Whisper read speech

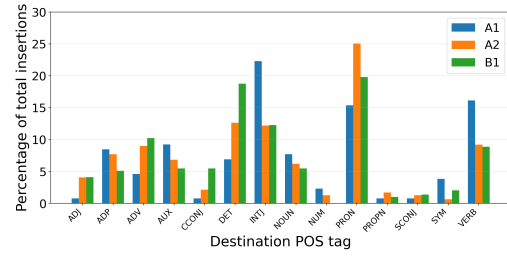


(d) Whisper HMI speech

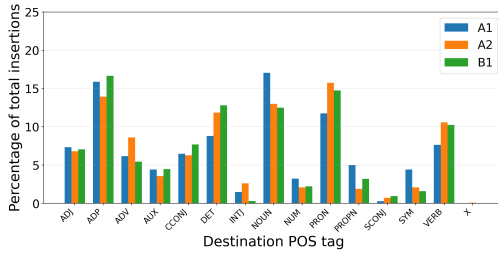
Figure 11: CEFR-specific substitution source category error rates by PoS category.



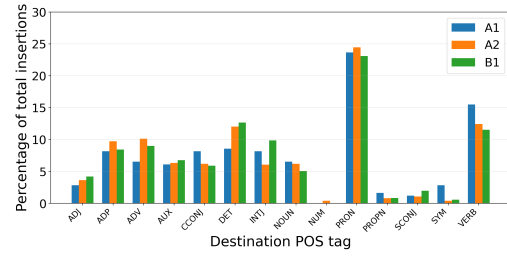
(a) Chirp read speech



(b) Chirp HMI speech

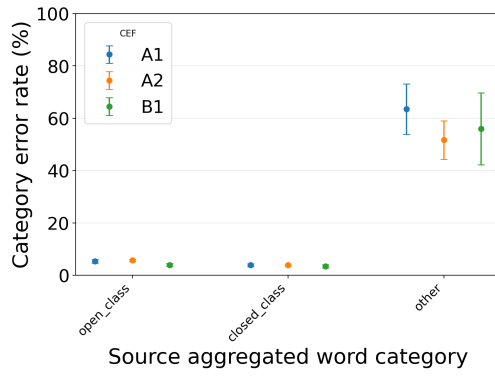


(c) Whisper read speech

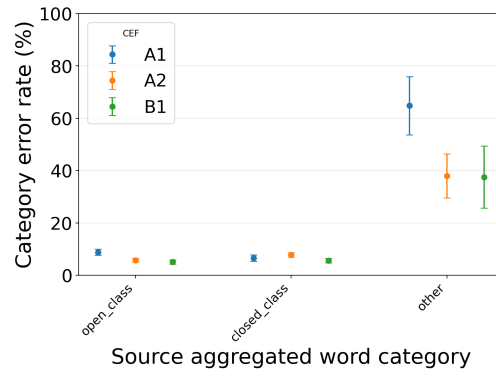


(d) Whisper HMI speech

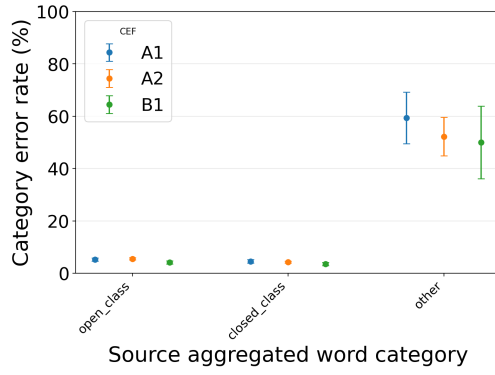
Figure 12: CEFR-specific insertion destination percentages by PoS category.



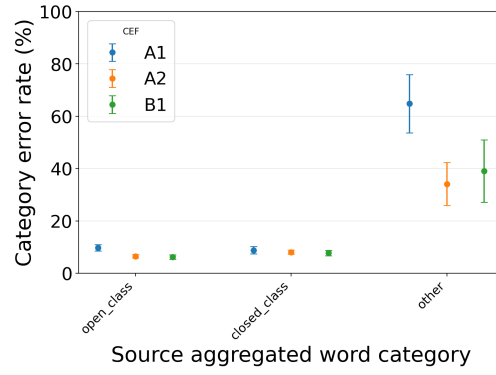
(a) Chirp read speech



(b) Chirp HMI speech

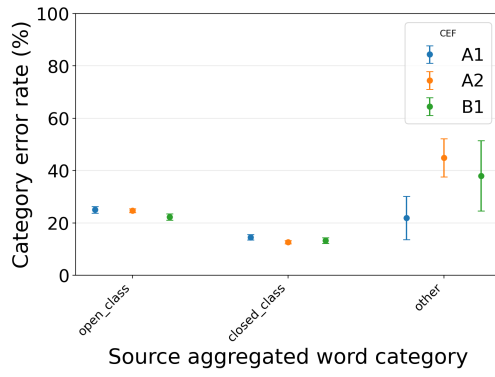


(c) Whisper read speech

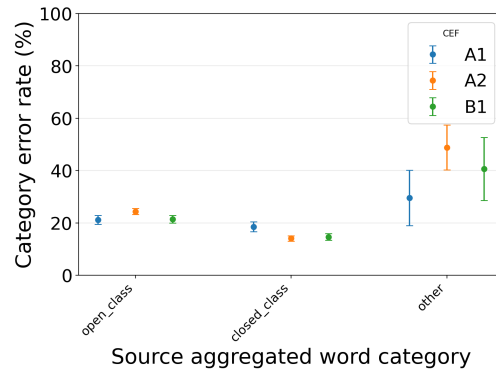


(d) Whisper HMI speech

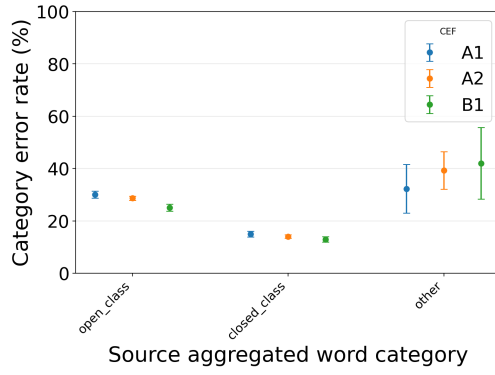
Figure 13: CEFR-specific deletion source category error rates for aggregated word categories.



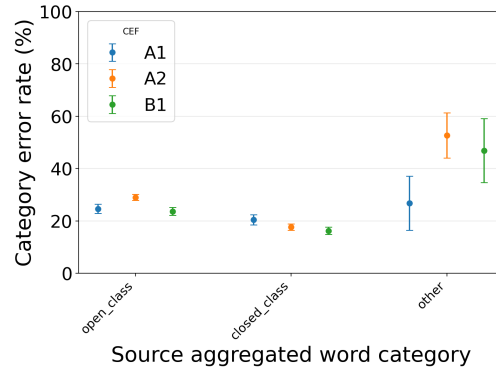
(a) Chirp read speech



(b) Chirp HMI speech

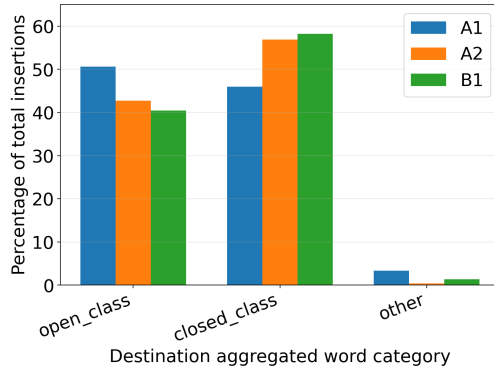


(c) Whisper read speech

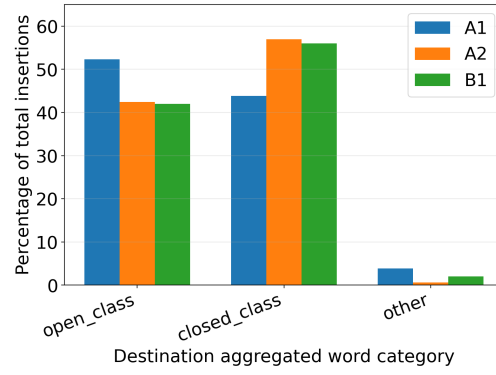


(d) Whisper HMI speech

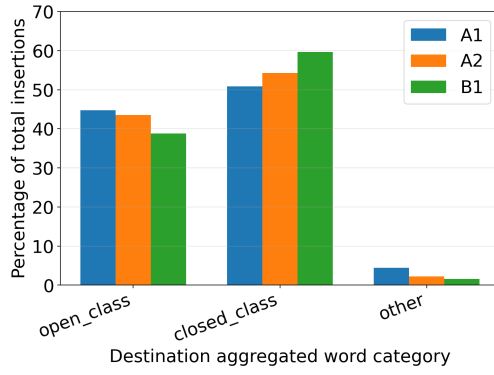
Figure 14: CEFR-specific substitution source category error rates for aggregated word categories.



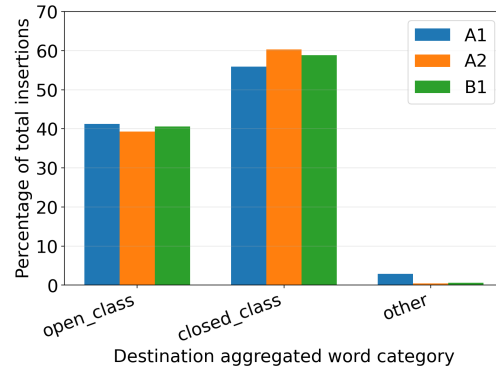
(a) Chirp read speech



(b) Chirp HMI speech



(c) Whisper read speech



(d) Whisper HMI speech

Figure 15: CEFR-specific insertion destination percentages for aggregated word categories.