Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels

Anonymous EMNLP submission

Abstract

While large multilingual automatic speech recognition (ASR) models achieve remarkable performance, the internal mechanisms of the end-to-end pipeline, particularly concerning fairness and efficacy across languages, remain underexplored. This paper introduces a fine-grained analysis of Whisper's multilingual decoder, examining its sub-token hypotheses during transcription across languages with various resource levels. Our method traces the beam search path, capturing sub-token guesses and their associated probabilities. Results reveal that higher resource languages benefit from markedly higher likelihood of the correct token being top-ranked in candidate guesses, higher confidence, lower predictive entropy, and more diverse alternative candidates. Lower resource languages fare worse on these metrics, but also exhibit distinct clustering patterns in sub-token usage sometimes influenced by typology in our PCA analysis. This sub-token probing uncovers systematic decoding disparities masked by aggregate error rates and points towards targeted interventions to ameliorate the imbalanced development of speech technology.

1 Introduction

Large multilingual Automatic Speech Recognition (ASR) models like Whisper (Radford et al., 2023) demonstrate impressive capabilities on high-resource languages, yet their performance often degrades significantly for low-resource languages (Javed et al., 2022), alongside persistent concerns about fairness across diverse linguistic groups (Zee et al., 2024). Aggregate metrics such as Word Error Rate (WER) can obscure the nuanced ways these models falter internally and may not capture critical issues like model hallucination (Koenecke et al., 2024). This necessitates a deeper analysis of the decoding process itself, with some prior work also highlighting the utility of evaluating models at the sub-unit level, for instance, in assessing calibration (Ballier et al., 2024b).

This paper posits that a granular, sub-token level investigation of Whisper's decoder is crucial for a more comprehensive understanding of these performance variations. In this work, we use the term 'sub-token' to refer to the sub-word units (e.g., Byte Pair Encoding (BPE) units) that models like Whisper generate; these are often broadly referred to as 'tokens' in relevant literature. Recognizing that tokenization strategies can themselves introduce biases and affect model behavior (Petrov et al., 2023; Ahia et al., 2024), we scrutinize how key characteristics of the sub-token generation process systematically differ when processing languages with varying levels of resources in training: specifically, the rank of chosen sub-tokens, model confidence in its predictions, predictive uncertainty (entropy), the diversity of the hypothesis space, and overall sub-token usage patterns.

Our analysis empirically demonstrates two primary findings: first, higher resource languages consistently benefit from more robust decoding metrics at the sub-token level, including higher prediction confidence and lower predictive entropy. Second, sub-token usage patterns as revealed through Principal Component Analysis (PCA) indicate poorer handling of tokenization for lower resource languages, but also reveal typologically coherent clusters that can transcend simple resource-level distinctions, highlighting the interplay between linguistic structure and data availability. These fine-grained insights are valuable for developing targeted interventions, such as specialized adapter fine-tuning (Song et al., 2024; Pfeiffer et al., 2021), to improve the equity and efficacy of multilingual ASR systems.

2 Background

2.1 Whisper and Tokenisation

Whisper is an influential foundation encoder-decoder Transformer model (Radford et al., 2023).

It was trained using large-scale weak supervision on approximately 680,000 hours of multilingual audio data, covering a wide array of tasks, including speech transcription and translation. This extensive pre-training enables strong zero-shot performance.

A core component of Whisper's architecture, as well as many modern large language and speech models, is its tokenization strategy. As detailed by Radford et al. (2023), Whisper utilizes two separate Byte Pair Encoding (BPE) vocabularies: one derived from the GPT-2 tokenizer (Sennrich et al., 2016; Radford et al., 2019) for English-only models, and a distinct, refitted vocabulary of the same size for multilingual models. This refitting was intended to avoid excessive fragmentation on other languages since the BPE vocabulary is English only (Radford et al., 2023). Our work focuses on the behavior of models using this multilingual BPE vocabulary, which is shared across all non-English languages the model supports. During decoding, the model generates a sequence of these sub-tokens, typically guided by special tokens like a language ID. While sub-word units allow handling large vocabularies and morphological variations more effectively than word-level tokenization, and can facilitate cross-lingual transfer (Conneau et al., 2020), Radford et al. (2023) themselves acknowledge potential limitations, particularly for languages distant from the Indo-European family which forms the bulk of the training data. They note that performance outliers could be due to a lack of transfer across languages and that the BPE tokenizer could be a poor match for these languages or variations in data quality.

However, the nature of sub-word tokenization, especially in a multilingual context, is not without its challenges. The way texts are segmented into tokens can vary significantly across languages, potentially leading to disparities in processing efficiency, context window utilization, and even model performance (Petrov et al., 2023). For instance, some languages might be systematically broken into more tokens than others for equivalent semantic content, an issue explored in the context of text-based LLMs (Petrov et al., 2023; Ács, 2019). Such tokenization artifacts can contribute to unfairness, as models might inherently find it more complex to process or learn representations for languages that result in longer token sequences (Ahia et al., 2024). While recent research also explores discrete acoustic or semantic tokens for ASR (Guo et al., 2025; Cui et al., 2024), the BPE approach as employed in Whisper remains a common paradigm, making the study of its sub-token characteristics critical.

2.2 Beam Search Decoding

For generating transcriptions, Whisper typically employs beam search decoding. Beam search maintains a set of k (the beam width) most probable partial hypotheses (sequences of tokens). At each step, it extends these hypotheses with possible next tokens and re-ranks them based on their cumulative probabilities, pruning the set back to k hypotheses. This exploration of multiple paths aims to find an overall sequence with a higher likelihood than what might be found through a purely greedy approach. Our analysis focuses on the characteristics observed along the single, final beam path chosen by the model, representing its ultimate transcription output. Probing model predictions, particularly the probability scores assigned by the decoder to chosen sub-tokens, provides insights into the model's decision-making at each step of this generation process. These output probabilities are commonly used as a direct measure of model confidence in the ASR literature (e.g., Jiang, 2005; Ballier et al., 2024b,a; Aggarwal et al., 2025). However, it is also well-established that the raw softmax probabilities from deep neural networks may not always be well-calibrated and can exhibit overconfidence (Guo et al., 2017).

2.3 Resource Disparity and Fairness in Multilingual ASR

A significant challenge in developing truly equitable multilingual ASR systems is the vast disparity in available training resources across languages. While Whisper's pre-training dataset is exceptionally large, the distribution of data per language varies by orders of magnitude (Radford et al., 2023). This imbalance directly impacts model performance, generally leading to superior results for higher resource languages (Javed et al., 2022). Scaling models and data (e.g., Tjandra et al. (2023) and Pratap et al. (2024)) can improve overall performance but does not necessarily resolve fairness issues or guarantee equitable performance across all languages and speaker groups (Zee et al., 2024). In fact, Zee et al. (2024) found that larger models can sometimes exhibit greater worst-case performance disparities.

The challenges for low resource languages are multifaceted. Beyond raw data quantity, issues in-

clude the representation of diverse scripts (Pfeiffer et al., 2021; Muller et al., 2021), the quality of sub-token vocabularies for these languages (Downey et al., 2023), and the potential for "capacity dilution" where a fixed-size model struggles to adequately represent many languages (Conneau et al., 2020). These factors can lead to higher error rates, lower model confidence, and increased susceptibility to issues like hallucination (Koenecke et al., 2024) for low resource languages. Prior work often evaluates these disparities at the word or utterance level (e.g., WER), with dedicated studies benchmarking performance on specific low-resource language sets like Pashto, Punjabi, and Urdu (Sehar et al., 2025). Our research instead asks: how do decoder-level uncertainties and hypothesis characteristics manifest differently between resource tiers even before a full word or sentence is outputted? This sub-token perspective is crucial for understanding the foundational biases and uncertainties that may contribute to downstream performance gaps and for developing targeted interventions.

3 Data

To investigate Whisper's sub-token decoding characteristics across different linguistic contexts and resource availability, we curated a diverse set of 20 languages. These languages were categorized into three tiers, High, Medium, and Low resource for better visualization, based on their representation in Whisper's own training data composition (Radford et al., 2023). It should be noted that the definition does not correspond to actual resource levels in the real world beyond Whisper's training dataset. The selected languages are detailed in Table 1. English, while being the most represented language in Whisper's training data, was intentionally excluded from our analysis. Its training data volume is disproportionately larger (over 430,000 hours) compared to the other high resource languages analyzed (e.g., German with approximately 13,000 hours), which would skew the comparative analysis across resource tiers and make it less informative for studying graduated cross-linguistic differences.

To maintain consistency in sub-token analysis and simplify cross-linguistic comparisons at the sub-token level, our analysis primarily focuses on languages that predominantly use the Latin alphabet. We made an explicit decision to exclude languages for which our initial baseline ASR performance using Whisper-large-v2 was exceptionally poor (specifically, WER higher than 60%). Preliminary qualitative analyses on languages such as Uzbek, Swahili and higher resource Danish revealed that the model frequently misrecognized the target language entirely or produced outputs with non-canonical orthography, making a meaningful analysis impractical.

Resource Tier	Language	Training Hours
High	German	13,344
	Spanish	11,100
	French	9,752
	Portuguese	8,573
	Turkish	4,333
Medium	Italian	2,585
	Swedish	2,119
	Dutch	2,077
	Catalan	1,883
	Finnish	1,066
	Indonesian	1,014
	Hungarian	379
	Romanian	356
	Norwegian	266
Low	Welsh	73
	Lithuanian	67
	Latvian	65
	Azerbaijani	47
	Estonian	41
	Basque	21

Table 1: Whisper training hours by language, categorized by resource level.

For each of the selected languages, we used approximately 10 minutes of speech data randomly sampled from the validated subset in the Common Voice 17.0 dataset (Ardila et al., 2020).

4 Methodology

4.1 Sub-token Extraction

Our methodology involves two key stages: generating hypothesis transcriptions and capturing the decoder's state at each generation step. For each audio utterance, we first obtain a transcription using Whisper-large-v2 with beam search (beam size=5, temperature=0.2). We provide the correct language ID token in the initial prompt to ensure transcription in the target language. This process yields the hypothesis sequence of sub-tokens $C = (c_1, c_2, \ldots, c_{N_H})$.

Once this hypothesis is generated, we re-trace its generation path step-by-step. For each position s in sequence C, we:

1. Provide the decoder with the audio features

(encoder output) and the prefix of already chosen sub-tokens $(c_1, c_2, \ldots, c_{s-1})$

2. Extract the decoder's full probability distribution over all possible sub-tokens for step s

3. Record the top-$K_{cand} = 50$ sub-token candidates $T_{s,K_{cand}} = (t_{s,1}, \ldots, t_{s,K_{cand}})$ along with their respective log-probabilities

This re-tracing procedure captures the decoder's internal state—its ranked candidates and their probabilities—at each decision point in the transcription process. These snapshots form the basis for calculating all our analytical metrics and provide insight into the model's decision-making across different languages.

4.2 Metrics

We compute several metrics by analyzing the beam search path of hypotheses generated by Whisper. For each utterance, we denote $C = (c_1, c_2, \ldots, c_{N_H})$ as the sequence of N_H sub-tokens in the hypothesis. At each step s in the generation process:

- c_s is the sub-token selected by beam search
- $T_{s,K_{cand}} = (t_{s,1}, t_{s,2}, \ldots, t_{s,K_{cand}})$ represents the top-K_{cand} (50 in our experiments) sub-token candidates predicted by the decoder, with corresponding probabilities $(p_{s,1}, p_{s,2}, \ldots, p_{s,K_{cand}})$

For metrics requiring comparison against ground truth, we use the reference transcription $G = (g'_1, g'_2, \ldots, g'_{N_R})$, tokenized with Whisper's tokenizer. Metrics are aggregated over all relevant items for a given language L, with N_{H_L} denoting the total sub-tokens generated across all hypotheses and N_{R_L} the total sub-tokens in all reference transcriptions for language L.

4.2.1 Average Rank of Correct Sub-token

This metric assesses how highly the ground truth sub-tokens rank among the decoder's predictions. We use Levenshtein alignment to map each reference sub-token g'_k to a hypothesis sub-token c_s (or identify it as a deletion). For each reference sub-token g'_k:

- If g'_k is aligned with hypothesis sub-token c_s: The rank of g'_k is its 1-indexed position in $T_{s,K_{cand}}$. If g'_k is not found within the top-K_{cand} list, its rank is assigned a penalty value of $K_{cand} + 1$.
- If g'_k is deleted: Its rank is also assigned the penalty value $K_{cand} + 1$.

Position	GT	Output	Operation	Rank
0	⟨|BOS|⟩	⟨|BOS|⟩	equal	1
1	Sil	S	replace	1
2	ah	el	replace	5
3	lar	am	replace	7
4	...	lar	replace	4
5	⟨|EOS|⟩	(deleted)	delete	$K+1$
6		A	insert	–
7		ley	insert	–
8		kum	insert	–
9		.	insert	–
10		⟨|EOS|⟩	equal	2

Table 2: Alignment between ground truth (GT) sub-tokens ("Silahlar..."/"Weapons...") and model output sub-tokens ("Selamun Aleykum."/"Peace be upon you.").

The average rank for a language L is the mean of these individual ranks across all reference sub-tokens:

$$\overline{\text{Rank}}(g')_L = \frac{1}{N_{R_L}} \sum_{k=1}^{N_{R_L}} R(g'_k)$$

A lower average rank indicates that the correct sub-tokens are more frequently found among the model's top predictions.

To illustrate this process, Table 2 shows the alignment between ground truth and model-generated sub-tokens for a Turkish phrase. The table demonstrates possible alignment operations: equal (where the model correctly identified the token), replace (where the model chose a different token), and delete (where a ground truth token has no corresponding model token).

For this example, the average rank is calculated as: $\overline{\text{Rank}} = \frac{1+1+5+7+4+(K+1)+2}{7} = \frac{20+(K+1)}{7} =$ 10.14 with $K = 50$.

4.2.2 Confidence

Confidence measures the average probability assigned to the sub-token c_s that was ultimately chosen at each step along the beam search path:

$$\overline{\text{Conf}}_L = \frac{1}{N_{H_L}} \sum_{s=1}^{N_{H_L}} p(c_s|\text{utterance}, c_{<s})$$

where $p(c_s|\text{utterance}, c_{<s})$ is the probability assigned to c_s given the utterance audio input and previous tokens.

4.2.3 Entropy

Token-level entropy quantifies the uncertainty in the model's predictions, calculated over the top-$K_H = 50$ predicted sub-tokens. After normalizing the probabilities of these candidates, the entropy (in bits) at step s is:

$$H_s = -\sum_{i=1}^{K_H} p'_{s,i} \log_2 p'_{s,i}$$

where $p'_{s,i}$ are the renormalized probabilities over the top K_H candidates. The average entropy $\bar{H}_L$ for a language is reported as the mean across all decoding steps.

4.2.4 Alternate-candidate Diversity

This metric evaluates the variety within the set of predicted candidates using the Type-Token Ratio (TTR). Specifically, we calculate the TTR of the non-top-1 candidates within the top-$K_D = 50$ predictions. For each language, we collect all sub-tokens that appear as candidates ranked from 2 to K_D across all decoding steps. The diversity is then computed as:

$$\text{Diversity}_L = \frac{|\text{unique non-top-1 tokens in L}|}{|\text{total non-top-1 tokens in L}|}$$

This approach explores the richness of the hypothesis space beyond the model's single best guess. Our underlying assumption is that lower-resourced languages might exhibit less diversity in these alternative candidates, potentially reflecting a more constrained or less nuanced hypothesis space learned by the model due to limited training data. A higher TTR indicates a broader range of unique alternatives being considered.

4.2.5 Sub-token PCA

To visualise cross-lingual patterns in sub-token usage, we build a frequency vector for each language from the sub-tokens that appear among the top-$K_{\text{PCA}} = 10$ candidates at every decoding step. These frequency vectors represent the distribution of sub-token IDs from Whisper's multilingual vocabulary of 51,865 tokens (Radford et al., 2023). Using a wider window (e.g. $K_{\text{PCA}} = 50$) would fold in many low-frequency alternates and blur fine-grained distinctions, so we fix K at 10 to keep language differences salient.

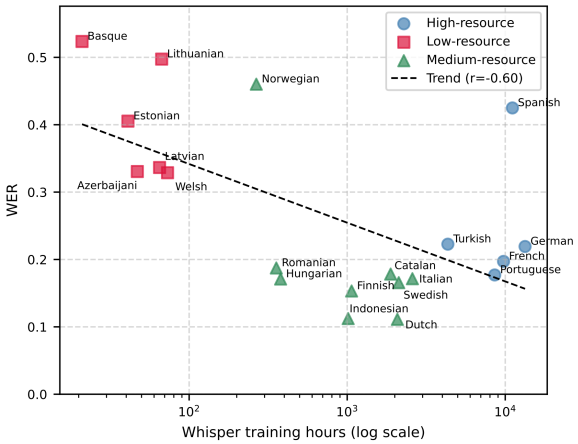

Figure 1: WER of Whisper on the Common Voice dataset versus training hours. Higher resource languages tend to have lower WER (p-value < 0.001).

After standardizing these vectors, we apply Principal Component Analysis and project the data onto the first two principal components. This allows visualization of language clusters based on sub-token usage patterns, revealing relationships that may correlate with linguistic typology or resource levels.

5 Results

Our analysis of Whisper's sub-token decoder reveals systematic variations in its behavior that correlate strongly with language resource levels, quantified by training hours. As a baseline, Figure 1 reports the Word Error Rate (WER) of the languages studied.

Consistent with previous findings we observe that WER is generally lower for languages with more training hours (Radford et al., 2023). Our subsequent sub-token level analyses aim to delve deeper into the internal decoding characteristics that might contribute to these performance differences.

5.1 Rank of Correct Sub-token

The average rank of the correct sub-token correlates with language resource levels, as shown in Figure 2. Higher resource languages have their correct sub-tokens ranked higher. This indicates that for higher resource languages, the beam search predominantly follows the locally highest-probability path. Despite the general trend, we do observe some deviations in individual languages.

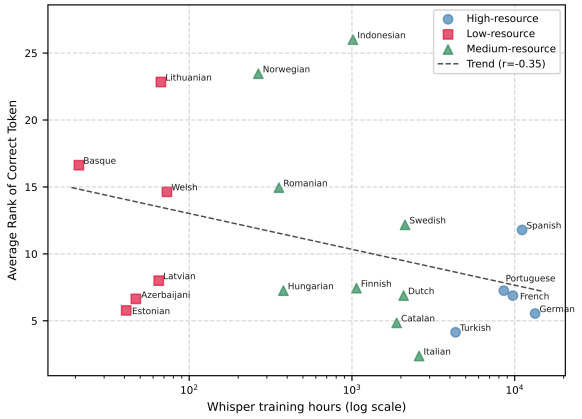

Figure 2: Average rank of the correct sub-token versus Whisper training hours. Higher resource languages tend to have correct tokens ranked higher (closer to 1, p-value < 0.001).

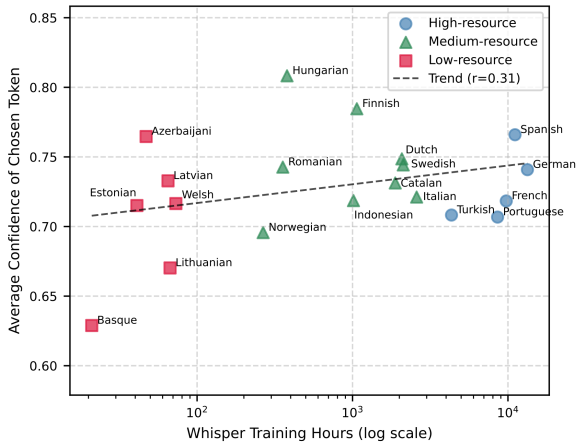

Figure 3: Average model confidence (probability of chosen sub-token) versus Whisper training hours. Higher-resource languages generally exhibit higher confidence (p-value < 0.001).

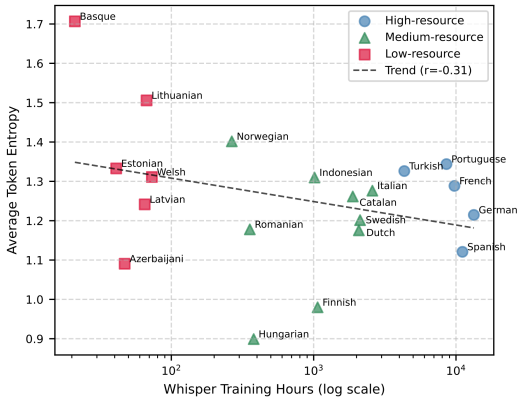

Figure 4: Average predictive token entropy of the top-50 candidates versus Whisper training hours. Higher resource languages generally exhibit lower entropy (p-value < 0.001).

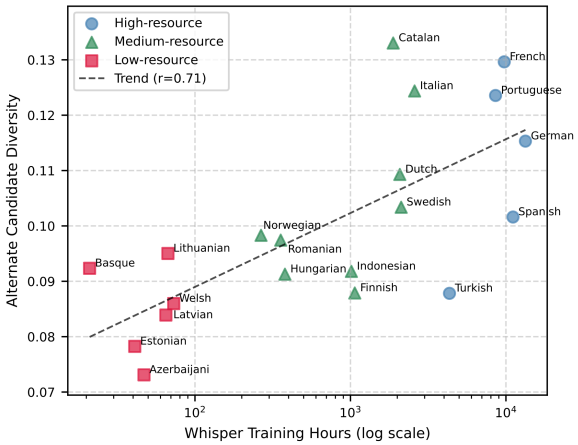

Figure 5: Alternate-candidate diversity (TTR of non-top-1 candidates in top-50 candidates) versus Whisper training hours. Higher resource languages tend to have higher alternate-candidate diversity (p-value < 0.001).

5.2 Decoder Confidence and Predictive Entropy

Decoder confidence, measured as the average probability of the chosen sub-token, shows a strong positive correlation with language training hours, as shown in Figure 3. High-resource languages tend to exhibit higher average confidence values.

Conversely, predictive entropy, calculated over the top-5 predicted sub-tokens, demonstrates a negative correlation with training hours, as shown in Figure 4. High-resource languages show lower average entropy, indicating more peaked and certain predictive distributions. This inverse relationship between confidence and entropy is expected: when the model is more confident in its chosen token, its distribution over alternatives is sharper (less entropic).

5.3 Alternate-Candidate Diversity

Alternate-candidate diversity, measured as the TTR of non-top-1 candidates within the top-5 predictions, exhibits a generally positive correlation with language resource levels, as shown in Figure 5. Higher resource languages tend to populate the upper range of diversity scores. This suggests that for higher resource languages, the model often considers a richer set of unique alternatives beyond its top choice. The overall trend indicates that increased training data may lead to a more varied set of hypotheses being considered.

5.4 PCA Clustering of Sub-token Usage

The PCA visualization of language-specific sub-token frequency vectors (from top-10 candidates) reveals distinct clustering patterns that reflect both typological relationships and resource levels, as presented in Figure 6. A prominent observation is the separation of language families. Romance languages (Spanish, French, Portuguese, Italian, Catalan, and Romanian) form a relatively cohesive cluster. Similarly, Turkic languages (Turkish and Azerbaijani), North Germanic languages (Swedish and Norwegian), Uralic languages (Finnish and Estonian, and to a lesser extent Hungarian), tend to stay clustered closer together and distinctly separated from other groups along the principal components. The clustering is also interesting as it groups languages of varying resource levels together, suggesting that shared linguistic structures (e.g., common morpho-phonological features captured in common sub-tokens) could influence sub-token usage patterns.

Notably, typologically unrelated lower resource languages, such as Welsh, Lithuanian, Latvian, and Basque, cluster together despite their significant linguistic differences. This unexpected grouping suggests these languages are handled similarly by Whisper's decoder not because of linguistic similarity, but due to their shared status as low-resource languages in the training data. Unlike high-resource languages that form distinct family-based clusters, these unrelated low-resource languages appear to share common sub-tokenization characteristics that transcend actual linguistic relationships, indicating deficiencies in the model's sub-token representations where it may be falling back to more generic decoding patterns due to insufficient exposure during training.

6 Discussion

Our sub-token probing of Whisper's decoder uncovers systematic variations in its behavior. These variations not only correlate strongly with language resource levels but are also significantly shaped by linguistic typology, offering a more nuanced understanding of the model's internal mechanisms beyond aggregate error rates.

6.1 Resource-Driven Variations in Sub-token Decoding

The analysis consistently reveals that higher resource languages benefit from more favorable decoding characteristics. Specifically, the model exhibits higher average confidence in its chosen sub-tokens for higher resource languages, and their predictive distributions are marked by lower entropy, indicating greater certainty in its predictions. Furthermore, the correct sub-token is more frequently highly ranked among the candidates considered during beam search for these languages. These advantages can be attributed to the extensive exposure to diverse linguistic phenomena afforded by larger training datasets (Radford et al., 2023). Such exposure likely enables the formation of more robust and well-calibrated sub-token representations. The generally higher alternate-candidate diversity observed for higher resource languages also suggests that increased training data allows the model to consider a richer and more varied set of plausible alternatives during the decoding process, potentially contributing to improved overall transcription accuracy.

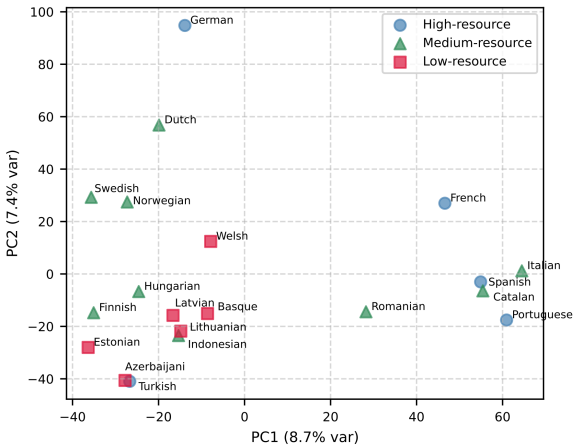

Figure 6: PCA of sub-token usage frequency vectors (top-10 candidates).

6.2 The Influence of Linguistic Typology on Sub-token Usage

Beyond the sheer volume of training data, the PCA of sub-token usage (Figure 6) illustrates the influence of linguistic structure on the model's internal representations. The distinct typological clusters, such as those observed for Romance languages and Turkic languages, underscores that shared morpho-syntactic or phonological features can drive similar sub-token utilization patterns.

The clustering is particularly interesting as it groups higher resource languages together with lower resource languages. This finding suggests that inherent linguistic properties (e.g. an agglu-

tinative morphology which might lead to a commonality in frequent, meaningful sub-tokens) can shape the model's representational space. Such structural similarities may partially mitigate the disadvantages associated with data scarcity for certain lower-resource languages when they belong to typologically related families. This suggests that typological relatedness can be a signifcant factor in conditioning sub-token representations and may influence the efficacy of cross-lingual transfer.

Notably, typologically unrelated lower resource languages, such as Welsh, Lithuanian, Latvian, and Basque, cluster together despite their significant linguistic differences. This unexpected grouping suggests these languages are handled similarly by Whisper's decoder not because of linguistic similarity, but due to their shared status as low-resource languages in the training data. Unlike high-resource languages that form distinct family-based clusters, these unrelated low-resource languages appear to share common sub-tokenization characteristics that transcend actual linguistic relationships, indicating deficiencies in the model's sub-token representations where it may be falling back to more generic decoding patterns due to insufficient exposure during training.

6.3 Sub-token Prediction Accuracy versus Global Performance

An intriguing aspect of our findings is the observed incongruity between local sub-token prediction and global Word Error Rate (WER) for particular languages. For instance, languages like Estonian and Azerbaijani, despite demonstrating remarkably low average ranks for their gold sub-tokens (signifying that the correct orthographic units are included as high-probability candidates by the acoustic model at a local, per-step level) do not invariably achieve the lowest overall WERs in our analyzed set.

This phenomenon highlights the inherent complexities of the end-to-end ASR decoding pipeline. While the model may possess robust local "knowledge" of correct sub-units, the ultimate transcription quality is a cumulative function of global sequence optimization during beam search, the effective mitigation of cascading errors arising from any single misstep, and the nuanced handling of intricate linguistic features that span multiple tokens. Consequently, for such languages, interventions aimed solely at further refining local sub-token prediction accuracy might yield diminishing returns on WER improvement. Strategies that enhance the model's capacity for global context modeling or its specific handling of overarching linguistic complexities may prove more fruitful.

6.4 Practical Implications

The above analysis offer several avenues for targeted interventions. The identification of language-specific weaknesses, such as a consistent tendency for correct sub-tokens to be ranked lower or for predictive distributions to exhibit high entropy in low resource languages, can directly inform the design of language-specific adapters or more sophisticated parameter-efficient fine-tuning strategies (Song et al., 2024; Pfeiffer et al., 2021). Furthermore, decoder-internal metrics, including dynamic measures of confidence and entropy, could potentially serve as valuable signals for adaptive adjustments to decoding algorithms. Finally, the sub-token usage patterns unveiled by PCA and the diversity metric may illuminate critical deficiencies in current vocabulary coverage and suggest novel data augmentation techniques.

7 Conclusion

This study introduces a fine-grained sub-token probing framework for Whisper's decoder, revealing systematic disparities in how the model processes languages across resource levels. Higher resource languages consistently benefit from more favorable decoding characteristics: correct tokens more frequently top-ranked, higher confidence, lower entropy, and more diverse alternative candidates. Our PCA analysis of sub-token usage further demonstrates that linguistic typology significantly influences these representations, with related languages clustering together regardless of resource tier, while unrelated low-resource languages unexpectedly cluster due to similar sub-tokenization patterns rather than linguistic similarity. These insights, often masked by aggregate metrics like WER, highlight how resource disparities manifest within the decoder's internal mechanisms and point toward targeted interventions such as language-specific adapters, dynamic decoding strategies, and focused data augmentation to promote more equitable multilingual ASR development.

Limitations

Although we tried to present a comprehensive analysis, our study has several limitations. First, we fo-

cused primarily on languages using the Latin script, excluding many writing systems and potentially missing script-specific tokenization effects. Additionally, our 10-minute samples per language provide only a snapshot that may not capture the full phonological diversity or domain variation present in natural speech.

Our methodology for extracting and evaluating sub-token performance also relies on assumptions about alignment between hypothesis and reference transcriptions that may not perfectly represent ground truth. While we analyze the chosen beam search path, we do not directly probe Whisper's internal beam search heuristics or alternate paths, which might offer additional insights into model decision-making. Finally, our resource-level categorization is based solely on Whisper's training hours rather than real-world language resources.

Furthermore, our analysis is inherently dependent on the quality and accuracy of the labels provided within the Whisper training dataset. As noted by Radford et al. (2023) themselves, there can be instances of labeling errors within this large dataset (e.g., some English audio being mislabeled as Welsh). Similar mislabelings for other languages could exist and potentially influence the model's learned representations and, consequently, our observations, particularly for languages where such noisy data might constitute a non-negligible portion of their training subset.

Future work should extend this analysis to more scripts, longer and more diverse audio samples, and explore how sub-token behavior correlates with specific linguistic features across typologically diverse languages.

References

Vaibhav Aggarwal, Shabari S Nair, Yash Verma, and Yash Jogi. 2025. Adopting Whisper for Confidence Estimation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A. Smith. 2024. MAGNET: Improving the Multilingual Fairness of Language Models with Adaptive Gradient-Based Tokenization. *Advances in Neural Information Processing Systems*, 37:47790–47814.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. 2024a. Whisper for L2 speech scoring. *International Journal of Speech Technology*, 27(4):923–934.

Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara B Ng, Richard Wright, and Jean-Baptiste Yunès. 2024b. Probing Whisper Predictions for French, English and Persian Transcriptions. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 129–138, Trento. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mingyu Cui, Daxin Tan, Yifan Yang, Dingdong Wang, Huimeng Wang, Xiao Chen, Xie Chen, and Xunying Liu. 2024. Exploring SSL Discrete Tokens for Multilingual ASR. ArXiv:2409.08805 [cs].

C.M. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. 2023. Embedding Structure Matters: Comparing Methods to Adapt Multilingual Vocabularies to New Languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 268–281, Singapore. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR. ISSN: 2640-3498.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2025. Recent Advances in Discrete Speech Tokens: A Review. ArXiv:2502.06490 [eess].

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards Building ASR Systems for the Next Billion Users. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10813–10821. Number: 10.

Hui Jiang. 2005. Confidence measures for speech recognition: A survey. *Speech Communication*, 45(4):455–470.

Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X. Mei, Hilke Schellmann, and Mona Sloane. 2024. Careless Whisper: Speech-to-Text Hallucination Harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681. ArXiv:2402.08021 [cs].

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. *Advances in Neural Information Processing Systems*, 36:36963–36990.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook 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*, pages 28492–28518. PMLR. ISSN: 2640-3498.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking Whisper for Low-Resource Speech Recognition: An N-Shot Evaluation on Pashto, Punjabi, and Urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 202–207, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Zheshu Song, Jianheng Zhuo, Yifan Yang, Ziyang Ma, Shixiong Zhang, and Xie Chen. 2024. LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR. In *Interspeech 2024*, pages 3934–3938. ISCA.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L. Seltzer. 2023. Massively Multilingual ASR on 70 Languages: Tokenization, Architecture, and Generalization Capabilities. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ISSN: 2379-190X.

Anna Zee, Marc Zee, and Anders Søgaard. 2024. Group Fairness in Multilingual Speech Recognition Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2213–2226, Mexico City, Mexico. Association for Computational Linguistics.

Judit Ács. 2019. Exploring BERT's Vocabulary.