Audio LLM Subtokens as Encapsulated "Knowledge"

The Case of Persian Subtoken Graphemic Representations in Whisper

Behnoosh Namdarzadeh & Nicolas Ballier

Abstract. Whisper is a widely used, open-access Large Language Model (LLM) trained using a multilingual paradigm. As such, it represents an important opportunity for researchers to study how multilingual LLMs function across languages. In this paper, we probe the grapholinguistic representations of Whisper for an under-resourced language with a Perso-Arabic script: Persian.

Persian, with its abjad writing system, has a consonant-heavy inventory of letters and does not indicate vowels, making it a good candidate to test an audio LLM such as Whisper (Maleki and Ahrenberg, 2008b). Large Language Models do not deal with tokens, morphemes or graphemes as linguists know them but subtokens, fragments of words generated by an algorithm. This subtokenisation process creates fracto-morphemes, clusters of letters, that continue to represent language (this is literally "natural language processing") but the relationship of the letters to morphology, semantics and phonology is completely changed.

We want to investigate this technical paradigm shift for writing systems such as Persian, by analysing the sounds to letters mapping as mediated by acoustic representations and LLM subtokens. Using a C++ implementation (Gerganov, 2003) of Whisper (Radford et al., 2023), we have extracted the subtoken dictionary of this model and their probabilities using a customised C++ implementation of Whisper.

In this work, we probe the phonotactic and graphotactic status of the subtokens representing Persian in order to understand how "knowledge" (representation) is distributed in the Whisper subtokens used to transcribe Persian. Understanding the subtokenisation process of large language models is crucial for advancing both NLP techniques and linguistic theories. Investigating how these models represent languages such as Persian sheds light on the intricate relationship between the acoustic signal, subtoken structures, and linguistic properties.

Our experiments show the division of labour between the encoder and the decoder. We will show that the decoder is used to post-process the data, as evidenced by the fact that the internal representations of the Whisper models are only the isolated forms of graphemes and limited digraphs. We demonstrate that Whisper models tend to select the formal variant in their transcription. For example, in a read speech of 100 sentences, Whisper produces the formal variant of the object marker *rā* in the transcriptions three times more than its informal

Behnoosh Namdarzadeh[1] 0009-0000-8730-7766,
Nicolas Ballier[2] 0000-0003-2179-1043
[1,2]ALTAE, Université Paris Cité
[1]E-mail: behnooshnamdar@gmail.com, [2]E-mail: nicolas.ballier@u-paris.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 327–360. https://doi.org/10.36824/2024-graf-namd
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

variant *ro*, even though the speaker only produces the informal variant in the read speech.

1. Introduction

Whisper is an automatic speech recognition (ASR) system developed using large-scale weak supervision. It was trained to predict transcripts of audio from 680,000 hours of labeled audio data, with 117,000 hours covering 96 languages other than English. The audio was broken into 30-second segments and paired with transcripts from different sources. Audio was resampled to 16,000 Hz, and an 80-channel log-magnitude Mel spectrogram was computed. Whisper was trained for several tasks such as language identification, transcription, translation, and timestamps prediction. Whisper's training included data for 96 languages besides English, including Persian, but with only 392 hours for the translation task and 24 hours for the transcription task. The performance for transcription of Persian is reported with a word error rate of 44.8 for the large model (compared to 6.3 for English), and the impact on performance of the amount of training data is consistent with the expectations (Ballier, Burin, et al., 2024). The training data transcripts were tokenised at subword level using a byte pair encoding algorithm (Sennrich, Haddow, and Birch, 2016) that was applied to the 96 languages of the training data, resulting in a multilingual vocabulary of units we call "subtokens" in the rest of the paper.

Whisper is a neural network that learns from a training corpus made up of audio recordings (690,000 hours for the largest model), combined with their transcription or translation. The aim is to match frames (windows of the sound signal) with (sub)tokens, i.e., arbitrary sequences of characters. It is therefore a supervised learning algorithm. Whisper is therefore an auto-regressive algorithm: given an audio input and the vector representation it constructs, it predicts the tokens one after the other to generate the final output. More specifically, based on the input audio and the prefix already generated (the previously predicted tokens), Whisper's decoder determines a probability distribution for all the tokens in its vocabulary. By choosing the subtoken with the highest probability at each stage (or by sampling according to this distribution), it gradually builds the output sequence.

We wish to explore the graphophonemic status of these subtokens as applied to the transcription task (ASR) of Persian. These LLM subtokens are a paradoxical object that are reminiscent of splinters but that are not morphological entities. We want to elaborate the inventory of subtokens used by the multilingual Whisper for Persian transcription by investigating the mapping between the acoustic realisations of the Persian data used as input and the Whisper prediction. In this sense, we address the

presumed encapsulated "knowledge" in the subtoken dictionary of the speech model. An important aspect to bear in mind is the potential misleading metaphor that if LLMs have representations they "understand" speech or "recognise" words. Even though the main type of encoded representation is a dictionary of subtokens that serves as a repository for the corresponding numbers of vectors of numerical figures, that serves as a form of representation, if any, alternatively for every 30 second speech window. The different attention weights corresponding to the internal representations can also be extracted and potentially investigated (see Mohebbi, Chrupała, Zuidema, and Alishahi (2023) as an example). Linguists still need to unravel what becomes of the different languages that are processed using these multilingual models. In this respect, it is likely that the subtokenisation phase is not innocuous (Petrov, La Malfa, Torr, and Bibi, 2024). We partially discuss the phonology of the Whisper dictionary of subtokens[1], the subtokens used to represent Persian writing in the LLM. Persian offers a unique perspective for this type of investigation for three reasons. First, the scarcity of the training data (24 hours of training data in Whisper) entailed a scarcity of subtokens in the Whisper dictionary used to transcribe Persian. Second, we have a privileged window on the phonographematics of Whisper since the acoustic space is transcribed by a limited amount of subtokens, mostly Persian isolated characters, whereas Persian has graphotactic constraints for final, medial or initial letters. Third, Persian shares many characters with Arabic, which has more training data in Whisper. Therefore, Persian serves as a test case for the investigation of a language that shares some of the characters of the same Unicode group (Perso-Arabic script).

As an abjad, rather than an alphabet, the Persian written language does not note vowels; this frequently results in an error in the transcription of vowels. 28 letters of the Arabic script have been supplemented with 4 additional consonants in order to preserve consonantal distinctions that do not exist in Arabic. However, Arabic is also encoded in the Whisper dictionary of subtokens. Previous studies address the challenge of adapting multilingual language models to specific languages for different NLP tasks (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023).

The rest of the paper is structured as follows: Section 2 presents the tenets of the Persian writing system and the variety of Persian we describe. Section 3 presents the phonological features of Persian and subtokenisation process in Whisper. Section 4 details our methodology and the datasets we used in our experiments. Section 5 presents our experiments with the probing of the subtoken dictionary and discusses the distribution of graphemes encapsulated in the Whisper subtokens used to transcribe Whisper. We analyse our results in Section 6 and Section 7 discusses them. Finally, we conclude in Section 8.

1. We do not discuss the special subtokens, see Ballier, Burin, et al. (2024).

2. Background on Persian Writing System

This section provides an overview of Persian graphematics and examines the graphotactic constraints that may influence the performance of audio-based LLMs.

The history of the Persian language is divided into three periods: Old Persian (c. 525 BC–300 BC), Middle Persian (c. 300 BC–800 AD) and Modern Persian (800 AD to the present). After the advent of Islam in 642 AD, Persians adapted the Arabic alphabet to develop the modern Persian alphabet. The Arabic alphabet has 28 letters, but Iranians added four additional letters <پ>, <چ>, <ژ>, <گ> to create the current 32 Persian letters (Rastorgueva and Hill, 1964). Modern Persian emerged in the 9th century after this adoption of the alphabet and the borrowing of many Arabic words.

The Persian script, technically called "abjad", consists of 32 letters, 28 of which are shared with Arabic, while the remaining four are unique to Persian. Persian adopts a non-Latin script written from right to left (Baluch and Choudhury, 2011). It comprises six vowels: three short vowels, marked with diacritical marks, and three long vowels, marked by certain consonants (ibid.). Therefore, Persian writing can be called a "consonantal alphabet". Diacritics for short vowels are mainly used for elementary readers, disappearing from textbooks after the first grade (Rahbari and Sénéchal, 2009), and are generally absent from standard written texts. Long vowels, in contrast, are consistently represented in writing.

The relationship between graphemes and phonemes in Persian is highly consistent. Each grapheme corresponds to a single phoneme (Arab-Moghaddam and Sénéchal, 2001). Therefore, Persian is classified as having a shallow orthography when short vowels are represented or when only long vowels and consonants are present (Baluch and Choudhury, 2011). However, the omission of short vowels results in a deep and opaque orthography where the relationships between graphemes and phonemes are inconsistent.

2.1. Graphotactic Constraints for Persian

This subsection explores the graphotactic constraints of Persian[2], which may affect the performance of LLMs such as Whisper (Maleki and Ahrenberg, 2008a).

2. It should be mentioned that the dialect we are examining for the sets of our experiments is a variant of Farsi, called "Tehrani", the standard colloquial dialect spoken in Iran (Karimi, 2005). We use the term "Persian" to refer to this variant in this paper.

Ninety different forms, referred to as “allographs”, are needed to represent the 28 letters of the Persian alphabet. Persian characters can have different graphematic forms depending on their positions. The isolated form of characters (28 letters) depends on their position within a word. The isolated form of each letter of the 28 letters can vary depending on its contextual placement, taking on different forms (allographs) in the initial, medial, or final positions (ibid.). The training data may not include all the required cases to ensure an ideal distribution of these contextual character forms, specifically Segment-Initial, Segment-Medial, Segment-Final, and Isolated forms. The dictionary of Whisper subtokens (see Appendix, Table 5) will be comprehensively analysed in Section 5.2.

In Persian, a single phoneme can be represented by multiple letters, resulting in graphematic inconsistencies. This variation requires practice and familiarity to determine which specific grapheme is used in a given word. For instance, the phoneme /t/ is represented by two letters: <ت> and <ط>; the phoneme /h/ by two letters: <ه> and <ح>; the phoneme /s/ by three letters: <س>, <ص>, and <ث>; and finally, the phoneme /z/ by four letters: <ذ>, <ز>, <ض>, and <ظ>. This is particularly problematic when transliterating foreign words and selecting the appropriate grapheme for a given phoneme. For example, the graphematic representation of the French-origin word *billet* and the two variants of /t/ phoneme is controversial. This has resulted in two different graphematic forms: <بلیت> and <بلیط>. Table 3 illustrates the various letters that correspond to the same phoneme. However, in Arabic, each of these letters corresponds to a distinct phoneme.

3. Phonological System of Persian and the Speech-to-Subtoken Mapping

We first recall the phonological system of Persian before explaining how sounds are mapped to the inventory of subtokens. In this paper, we follow the standard phonological convention that indicates graphemes (letters) with angled brackets (<>), realisations in square brackets and phonemes (or targets) with slanted bars (//).

3.1. Vowels and Consonants

The standard Persian of Iran has six vowel sounds as shown in Table 1. The vowels /a, e, o/ are short vowels and /i, u/ are long vowels. Since the Persian writing system is an abjad (consonantal alphabet), short vowels do not have specific letters and are not written in written Persian, and only long vowels appear in texts. Long vowels are always marked

with consonant letters. The short vowels, as mentioned earlier, are not marked in standard written Persian, but for learners of Persian, they can be precisely marked with diacritics. Inserting diacritics is not allowed in the early years of primary school.

The three long vowels, /ɒ, i, u/, are frequently found word-finally. Less commonly occurring word-finally are the short vowels, /a, e, o/. While /e/ is present in some high-frequency words, word-final /o/ is rare (still appears in some high frequency words.)

Any vowel may begin a word, though some occur more frequently than others. Word-initial /a/ is highly prevalent, while word-initial /i/ is less common. Word-initial /u/ is rare. In summary, lax vowels such as /e/ and /a/ are the most common in word-initial positions, while word-initial /o/ is seldom encountered.

Table 1. Tehrani Persian Vowel Chart

Height	*Front*	*Back*
Close	i	u
Mid	e	o
Open	a	

Due to the complexities of inserting or omitting vowels in written Persian, significant challenges arise in accurately transcribing vowels. There are some factors such as the neighbouring graphemes and the type of syllable containing the vowel which affect the choice of grapheme (allographs) for the vowel (Maleki and Ahrenberg, 2008a). Persian script often omits short vowels, as seen in /ketɒb/ (<کتاب>), where the short vowel /e/ is not represented, reflecting the inherent complexity of vowel representation in the language. Additionally, vowel representation is influenced by neighbouring graphemes, as in /χɒhar/ (<خواهر>) where the character <و> does not denote a /v/ or /u/ sound but forms part of a <خو> digraph for /χ/ followed by /ɒ/.

This subsection explores the constraints that govern the distribution and arrangement of consonants in written Persian. Persian includes 23 consonant sounds, each represented by a dedicated letter. The table of International Phonetic Alphabet (IPA) symbols for these consonants is presented in Table 2. Notably, the representation of a letter in Persian script varies depending on its position within a word. A single letter may assume different forms when appearing in initial, medial, or final position of a word [See Table 3]. This positional variability is a defining characteristic of the Persian writing system. Word-initial consonant clusters are absent in Persian, and Persian speakers tend to insert a vowel between consonants when pronouncing borrowed words with initial clusters from the source language, thus avoiding word-initial conso-

nant clusters. Alternatively, a vowel may be inserted before the cluster, leading to a VC+C sequence (Mahootian and Gebhardt, 1997). Moreover, word-initial clusters can be broken up, resulting in a CV+C sequence. Various types of two-consonant clusters are permissible in syllable-final positions (e.g., nasal+plosive: /omq/ (<عمق>) meaning 'depth').

Table 2. Persian Consonants Organized by Manner and Place of Articulation

	Labial	*Alv.*	*Post-alv.*	*Velar*	*Uvular*	*Glottal*
Nasal	m	n				
Stop	p, b	t, d	t͡ʃ, d͡ʒ	k, g	(q)	ʔ
Fricative	f, v	s, z	ʃ, ʒ	x, ɣ		h
Tap		ɾ				
Approx		l	j			

3.2. The Semiotics of Subtokenisation

In this subsection we explain how the Persian representation between graphemes and meaning is mediated with the subtokenisation phase (Sennrich, Haddow, and Birch, 2016). Subtokens are a hybrid representation that bear testimony to the frequency of tokens in the training data of the subtokenisation algorithm (and of Whisper's training). Subword units have emerged as a crucial component in language modeling. They play a foundational role in the tokenisation processes of advanced transformer-based language models, such as those utilising Byte Pair Encoding (BPE) (ibid.). This subword-based tokenization method underpins the architecture of many LLMs (Manova, 2023). The adoption of subword tokenization strategies, like BPE, has significantly contributed to the success of pre-trained transformer models in natural language processing, enabling them to effectively handle a diverse array of languages and linguistic structures (Bostrom and Durrett, 2020). The (sub)tokens are generated automatically from the training data using tokenisation algorithms. These algorithms divide the text into units based on the frequency with which sequences appear in the corpus. The main algorithms used are Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2016), WordPiece, initially developed for BERT (Devlin, Chang, Lee, and Toutanova, 2019), and SentencePiece (Kudo, 2018). Depending on their frequency of appearance, (sub)tokens can correspond to whole words, sequences of 2 or 3 characters, or even a single character.

Subtokenisation can be challenging for Persian due to its graphematic and phonographematic constraints, as outlined in Table 2. In this context, the implementation that predicts on a sub-token basis may

differ significantly from a representation based on entire sentences. Typically, isolated tokens are expected to correspond systematically to individual letters. Figure 1 illustrates the subtokenisation in three languages—French, English, and Persian—using the Whisper tokeniser[3]. As can be seen with these examples, the frequent tokens in English are less likely to be subdivided into subwords (subtokens) than French or Persian words.

FIG. 1. Subtokenisation in French, English, and Persian sentences using Whisper tokeniser

To investigate the frequency of subdivision for Persian words, we tested the Persian FLEURS dataset (Conneau et al., 2023) for a distributional analysis to observe the frequency of the subtokens for Persian and estimate the inventory of subtokens from the Whisper dictionary used to transcribe Persian data.

3.3. Hallucinations

Another aspect of the Whisper dictionary of subtokens used for transcriptions is the fact that the Whisper multilingual dictionary of subtokens includes all the subtokens used to transcribe the 99 languages with which Whisper was trained. With Whisper smaller models (tiny, base and small)[4], hallucinations can be observed, subtokens belonging to other languages groups are used in the Whisper transcription. One way to analyse the evolution of hallucinations across the size of the models is to observe that the hallucinations in the outputs belong to different language groups, and, in that respect, they somehow follow the pecking order of the hierarchy of languages used in the training data. The Chinese characters are overwhelmingly present in the transcriptions of the tiny models and that is consistent with the role of the size of the Chinese training data.

3. https://huggingface.co/docs/transformers/model_doc/whisper
4. With 39M,74M and 244M parameters respectively (Radford et al., 2023).

4. Material and Methods

This section outlines the materials and methods used in the study.

4.1. Materials

Here are the three datasets used in our experiments:

- YouTube Audio File: A 69-minute audio recording extracted from YouTube[5] was selected for analysis. The recording features three male speakers engaging in a discussion on social issues. Due to the spontaneous and unstructured nature of the conversation, manual transcription would be both challenging and time-consuming. By utilising OpenAI’s Whisper for transcription, we gain two significant advantages: the generation of an accurate transcription and its accompanying translation. This output provides a valuable foundation for developing a bilingual corpus.
- Common Voice Mozilla: Additional Persian data for detailed experiments was sourced from the Common Voice Mozilla project[6], a diverse dataset of speech recordings designed for speech recognition tasks.
- FLEURS Dataset: The FLEURS dataset[7], which focuses on multilingual speech recognition and understanding, was also used to supplement our analysis and broaden the scope of our Persian corpus experiments.

4.2. Transcription Method With OpenAI’s Whisper

Our approach focused on leveraging a customized C++ implementation of Whisper[8], building upon the foundational work by Radford and colleagues (Radford et al., 2023). Whisper offers six model variants, ranging from tiny to large-v1, which provide valuable insights into the mapping of spoken Persian to written Persian. The Whisper.cpp implementation by Gerganov[9] facilitates efficient processing of Whisper’s parameters and includes tools for visualizing subtokens along with their associated probabilities. This functionality enables access to probability scores for each subtoken generated by Whisper, forming the basis for

5. https://www.youtube.com/watch?v=YsMDE0EH42s
6. https://commonvoice.mozilla.org/fa
7. https://huggingface.co/datasets/google/fleurs
8. https://github.com/jbyunes
9. https://github.com/ggerganov/whisper.cpp

developing a calibration model to evaluate the accuracy of the Whisper models.

4.3. Documenting Whisper's Dictionary of Subtokens Used for Persian

We compared two methods. We first compiled the prediction at subtoken levels extracted for the different models. Then we ran the tokenisation algorithms on the Persian component of the FLEURS training data to compare the distribution of subtokens. we used a subsample of the FLEURS dataset comprising 85 hours of speech. The FLEURS dataset[10] (FLores Enhanced Universal Representation for Speech) is a multilingual speech dataset, serving as the speech-focused counterpart to the FLoRes machine translation benchmark. It builds on the FLoRes dev and devtest datasets, comprising 2009 n-way parallel sentences translated into 102 languages. This extensive coverage makes FLEURS one of the most valuable resources for advancing multilingual speech research.

A key strength of FLEURS lies in its use of multilingual fine-tuning, which allows models to harness shared linguistic features across diverse languages. The dataset's results are thoughtfully organised into seven geographical regions, showcasing the dataset's cultural and linguistic breadth. Among the languages included is Persian, highlighting the dataset's inclusion of languages from the Central Asia/Middle East/North Africa region.

5. Experiments

This section describes the experiments conducted in this study on written Persian. To better understand the distribution of graphemes in Persian, we used the FLEURS training dataset (Conneau et al., 2022), which is publicly available in 102 languages and published by `Google`. We analysed the distribution of Persian graphemes across this dataset. The dataset contains a total of 280,786 graphemes, with 34 unique graphemes. Figure 2 shows the distribution of graphemes in written Persian. We hypothesize that the more frequently a subtoken appears in publicly available online corpora, the fewer challenges LLMs like Whisper encounter in accurately processing it. Frequent exposure to common subtokens during training allows the model to develop a stronger representation and mapping for these elements, reducing errors in transcription and improving overall reliability. Conversely, subtokens that

10. `https://huggingface.co/datasets/google/fleurs`

are rare or less represented in the training data are more likely to lead to inaccuracies, reflecting gaps in the model’s learned patterns.

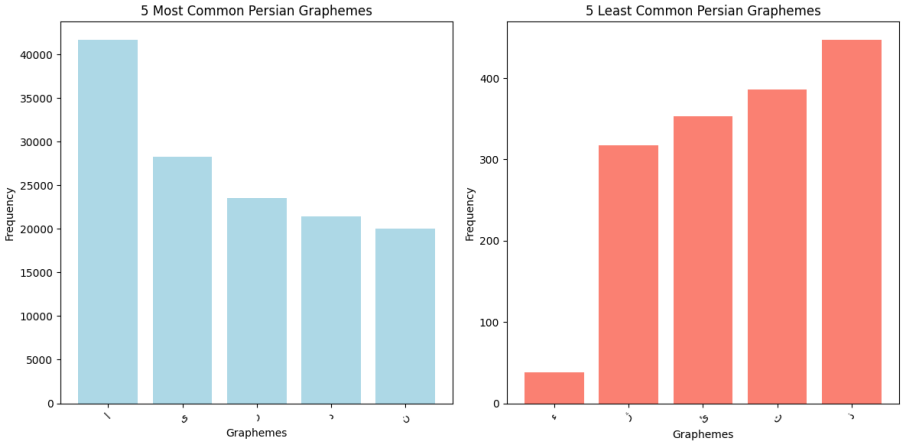


Fig. 2. Distribution of Persian Graphemes in the FLEURS dataset. As the different frequency scales indicate, the order of magnitude between most frequent and least frequent graphemes is about one hundred to one.

5.1. Distributional Frequencies in the Transcription of a Reference Dataset

To improve the performance of Natural Language Processing (NLP) tasks, the quantity of training data plays a crucial role (Ballier, Burin, et al., 2024). In general, an increase in the volume of training data directly correlates with improved task outcomes, as measured by various performance metrics. For instance, in Automatic Speech Recognition (ASR) tasks, an increase in the availability of training data for a particular language tends to lower the Word Error Rate (WER), indicating enhanced model accuracy and reliability. Figure 4 illustrates the distribution of graphemes in the FLEURS dataset.

The aim is to examine the frequency of consonants and vowels that appear at the beginning and end of words in written Persian, to be later compared to the analysis of subtokenisation outputs produced by Whisper. As shown in Figure 3, the characters <د> , <م>, <ب>, and <ک > typically appear at the beginning of words, while the consonants <ر>, <د>, <ه>, and <ن> typically appear at the end of words. Regarding vowels, the vowel <ا> frequently appears both at the beginning and at the end of

words. If we consider the FLEURS corpus as a foundational dataset for Persian, it is reasonable to assume that Whisper, when trained on similar data, has been exposed to similar distributions of graphemes. Specifically, Whisper may be more likely to accurately transcribe graphemes that are highly represented in its training data. Conversely, graphemes that are under-represented in the FLEURS dataset might pose challenges, potentially leading to transcription inaccuracies due to limited exposure during training.

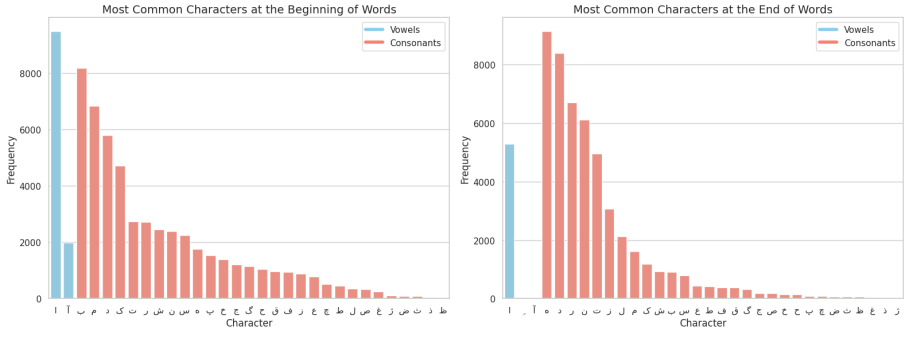


FIG. 3. Positional distribution of consonants and vowels at the beginning and end of words

5.2. Distributional Frequencies in Whisper Transcriptions of the FLEURS Speech Training Set

This subsection analyses Whisper transcription (large model) of the FLEURS speech training set. We then compare the inventory of Whisper subtokens used for transcriptions with the positional constraints of Persian graphemes.

Table 3 summarises the findings of this analysis. The first column presents the IPA representations of Persian characters. The second column lists the isolated forms of Persian graphemes, accompanied by a binary code (0–1) indicating their presence in Whisper's subtoken dictionary. A value of 1 signifies that the grapheme is included in the dictionary, while a value of 0 indicates its absence. For the isolated representation of graphemes, it was observed that <ژ>, which belongs exclusively to the Persian alphabet and not the Arabic script, is not present in the Whisper mutlilingual dictionary. The fourth column of Table 3 shows the final graphematic representations of characters. Notably, the

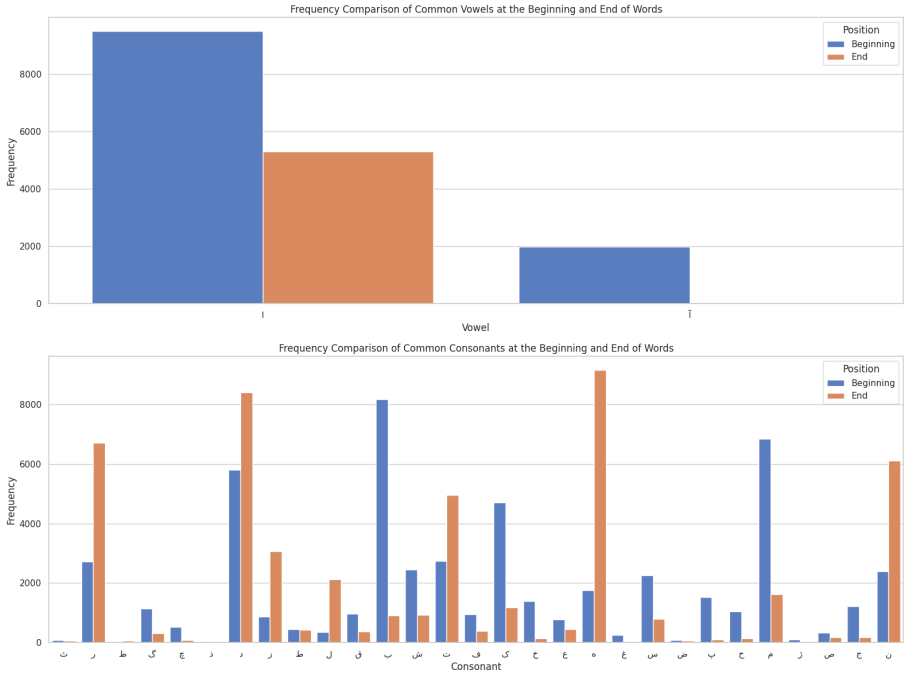


FIG. 4. Distribution of common consonants and vowels at the beginning and at the end of words in the Whisper transcriptions of the FLEURS speech training set (85 hours)

characters <پ,چ,ژ,ز,ظ,غ,گ> are absent from Whisper’s subtoken dictionary in their final forms. It is important to highlight that <پ ژ چ> are Persian-specific graphemes. The sixth column represents the medial forms of graphemes in Persian. In Whisper’s subtoken dictionary, we did not observe any Persian characters in their medial forms, except for <ت,ع,ل,ن,ی>. The eighth column reports the initial forms of characters within Persian tokens, with the following characters being absent in their initial forms: <پ,ج,چ,ض,ط,ظ,غ,گ>.

By accessing the dictionary of subtokens for all 99 languages supported by Whisper, we extracted the subtokens associated with the Perso-Arabic script. We identified four possible positions for Persian characters within a word: initial, medial, final, and isolated forms[11].

11. Some entries in the columns of graphematic representation for Persian contain a value of 0, indicating that these forms do not exist in the Persian script.

Table 3. Whisper Coding of Persian Letters in its Subtoken Dictionary

IPA	*Isolated form*	*Whisper isolated*	*Final form*	*Whisper final*	*Medial form*	*Whisper medial*	*Initial form*	*Whisper initial*
/ɒ/	ـا	1	ـل	1	0	0	0	0
/b/	ب	1	ـب	1	ـبـ	0	بـ	1
/p/	پ	1	ـپ	0	ـپـ	0	پـ	0
/t/	ت	1	ـت	1	ـتـ	1	تـ	1
/s/	ث	1	ـث	1	ـثـ	0	ثـ	1
/d/	ج	1	ـج	1	ـجـ	0	جـ	0
/ʧ/	چ	1	ـچ	0	ـچـ	0	چـ	0
/h/	ح	1	ـح	1	ـحـ	0	حـ	1
/χ/	خ	1	ـخ	1	ـخـ	0	خـ	1
/d/	د	1	ـد	1	ـد	0	د	0
/z/	ذ	1	ـذ	1	ـذ	0	ذ	0
/ɾ/	ر	1	ـر	1	ـر	0	ر	0
/z/	ز	1	ـز	0	ـز	0	ز	0
/ʒ/	ژ	0	ـژ	0	ـژ	0	ژ	0
/s/	س	1	ـس	1	ـسـ	0	سـ	1
/ʃ/	ش	1	ـش	1	ـشـ	0	شـ	1
/s/	ص	1	ـص	1	ـصـ	0	صـ	1
/z/	ض	1	ـض	0	ـضـ	0	ضـ	0
/t/	ط	1	ـط	1	ـطـ	0	طـ	0
/z/	ظ	1	ـظ	0	ـظـ	0	ظـ	0
/ʔ/	ع	1	ـع	1	ـعـ	1	عـ	1
/q/	غ	1	ـغ	0	ـغـ	0	غـ	0
/f/	ف	1	ـف	1	ـفـ	0	فـ	1
/ɢ/	ق	1	ـق	1	ـقـ	0	قـ	1
/k/	ک	1	ـک	1	ـکـ	0	کـ	1
/g/	گ	1	ـگ	0	ـگـ	0	گـ	0
/l/	ل	1	ـل	1	ـلـ	1	لـ	1
/m/	م	1	ـم	1	ـمـ	0	مـ	1
/n/	ن	1	ـن	1	ـنـ	1	نـ	1
/v/	و	1	ـو	1	ـو	0	و	0
/h/	ه	1	ـه	1	ـهـ	0	هـ	1
/j/	ی	1	ـی	1	ـیـ	1	یـ	1

This experiment based on frequency is a sort of reverse engineering that helps us understand how Persian graphemes are encoded in a multilingual model like Whisper. Beyond this architectural bias based on the distributions of the subtokens in the Whisper multilingual model, we explore how the linguistic features of Persian are encoded in Whisper when transcribing actual Persian speech.

5.3. Experiment on Spontaneous Data

To better understand the phonographemic representation of Whisper, we analysed [the decoder output] of a 69-minute spontaneous video from YouTube. The context is a studio where three young men (in their 30s) discuss various social and cultural topics. The session is mostly informal, with frequent laughter, repetition, and other linguistic and non-linguistic discourse markers, which could pose challenges for ASR systems to transcribe.

To assess the accuracy of Whisper in transcribing this audio file, we report the Word Error Rate (WER) and Character Error Rate (CER). The WER of 0.403 suggests that approximately 40% of the words in the Whisper's prediction are incorrect compared to the gold standard, indicating a significant discrepancy between the prediction and the reference. This is generally considered poor performance for many tasks, as nearly half of the words are incorrect. A CER of 0.222 suggests that 22.27% of the characters in Whisper's prediction are incorrect when compared to the gold standard, reflecting a relatively high error rate in Whisper's transcription.

Table 4. Performance Metrics for Prediction Accuracy of Whisper

Metric	*Value*
Precision	0.0004
Recall	0.0003
F1-Score	0.0003

The results for three other metrics Precision, Recall and F1-Score (see Table 4) are extremely low, indicating poor alignment between the Whisper's predictions and the gold standard at the word level. Here is what each of these metrics suggests:

Precision measures how many of the predicted words are actually correct. A precision of 0.0354% means that out of all the words Whisper predicted, almost none of them were correct. This indicates a high number of false positives (words predicted incorrectly). Recall measures how many of the actual correct words were successfully predicted. A recall of

0.029% means the system is missing nearly all the correct words, suggesting a very high number of false negatives. F1-Score combines both precision and recall into a single metric. With such a low score, it reflects that both the precision and recall are very low, indicating a significant mismatch between the prediction and the reference.

5.4. The Speech to Subtoken Mapping

The analysis of the Persian subtokens derived from the dataset shows how in the distribution of token frequencies, certain characters and sequences appear much more frequently than others, as illustrated in Figure 5. The most frequent subtokens are key characters that are integral parts of many common Persian words. The dominance of these components reflects both the linguistic characteristics of Persian and the tokenisation strategy employed by the Whisper model.

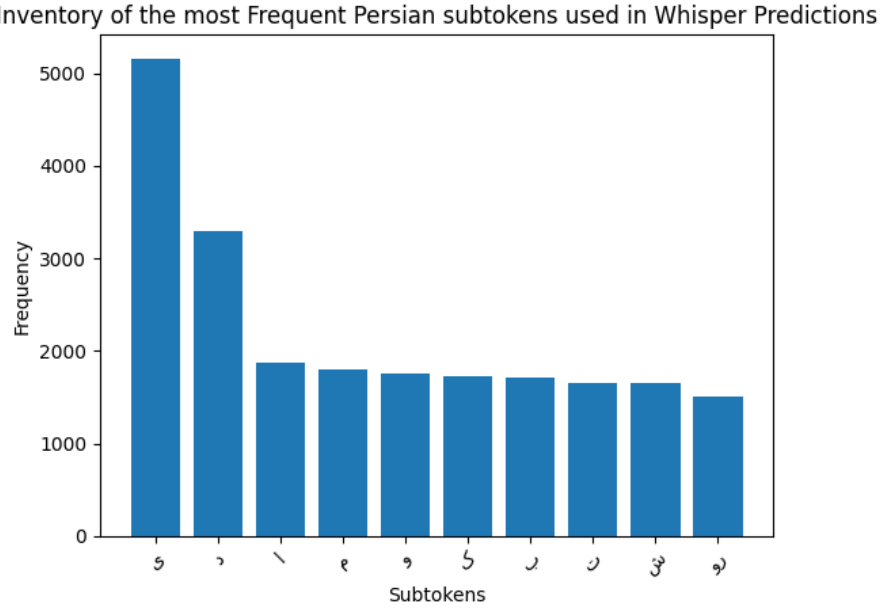


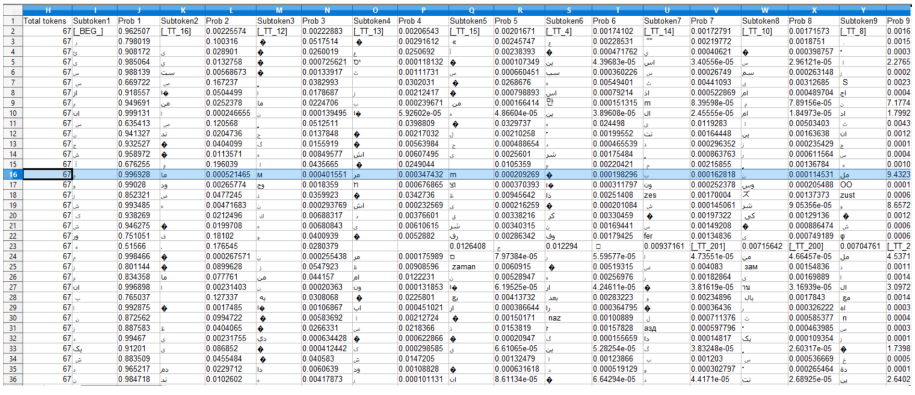
FIG. 5. Distribution of Persian Graphemes in our YouTube Dataset

Another important aspect of subtokenisation is the length of the subtokens across different languages. In the case of Persian, the tokenisation process tends to generate shorter subtokens. Based on our analysis of the subtokenisation of Persian YouTube data, out of a total of

30,544 subtokens generated by Whisper, 24,190 consist of only a single Persian character. Subtokens containing two Persian characters account for 4,325, while those with three Persian characters are limited to just 11. This observation suggests that subtokenisation for Persian, as an under-resourced language, may differ significantly from languages with larger training datasets, potentially reflecting the challenges and limitations in processing languages with less representation in model training.

5.5. Experiment on Whisper’s Alternate Predictions

We ran another experiment based on the experimentally customised implementation of C++ Whisper that extracts the 10 first predictions of the model and the corresponding probabilities[12]. Do we get subtokens reserved for Persian or character-based representations of corresponding sounds for the alternate predictions?



	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Total tokens	Subtoken1	Prob 1	Subtoken2	Prob 2	Subtoken3	Prob 3	Subtoken4	Prob 4	Subtoken5	Prob 5	Subtoken6	Prob 6	Subtoken7	Prob 7	Subtoken8	Prob 8	Subtoken9	Prob 9
2	67	[_BEG_]	0.962507	[_TT_16]	0.00225574	[_TT_12]	0.00222883	[_TT_13]	0.00206543	[_TT_15]	0.00201671	[_TT_4]	0.00174102	[_TT_14]	0.00172791	[_TT_10]	0.00171573	[_TT_8]	0.0016
3	67	ر	0.798019		0.100316	◆	0.0517514	◆	0.00291612	ه	0.00245747	ع	0.00228531	‌	0.00219772		0.0018751		0.0015
4	67	ری	0.908172	ی	0.028901	◆	0.0260019		0.0250692	ا	0.00238393	◆	0.00047162	ی	0.00040621	◆	0.000398757	ٔ	0.0003
5	67	ن	0.985064	ی	0.0132758	◆	0.000725621	ت	0.000118132	◆	0.00010749	س	4.39683e-05	است	3.40556e-05	س	2.96124e-05	ا	2.2765
6	67	س	0.988139	سب	0.00568673	◆	0.00133917	ت	0.00111731	س	0.000660451	سب	0.000360226	س	0.0002674	سم	0.000263148	ا	0.0002
7	67	س	0.669722	س	0.167237	ٔ	0.0382993		0.0302031	◆	0.0268676	ه	0.00549401	ت	0.00441093	ی	0.00312685	S	0.0023
8	67	ار	0.918557	اه	0.0504499	ا	0.0178687	ر	0.00212417	◆	0.00079893	است	0.00075214	از	0.00052869	ام	0.000489704	اج	0.0004
9	67	ه	0.949691	می	0.0252378	ها	0.0224706	س	0.000239671	می	0.000166414	◘	0.000151315	m	8.39598e-05	ر	7.89156e-05	ر	7.1774
10	67	اب	0.999131	ا	0.000246655	ر	0.000139495	اه	5.92602e-05	ه	4.86604e-05	سی	3.89608e-05	ال	2.45555e-05	ام	1.84973e-05	اد	1.7992
11	67	س	0.635413	س	0.120568	ه	0.0512511	ا	0.0398809	◆	0.0329737	ه	0.024498	ی	0.0119283	ا	0.00503403	ت	0.0043
12	67	ر	0.941327	تا	0.0204736	ه	0.0137848	◆	0.00217032	ی	0.00210258	ٔ	0.0019552	تا	0.00164448	سی	0.00163638	اب	0.0012
13	67	ر	0.932527	◆	0.0404099	ک	0.0155919	◆	0.00563984	ر	0.000488654	ه	0.000465539	ا	0.000296352	ا	0.000235429	و	0.0001
14	67	ش	0.968972	◆	0.0113571	ه	0.00849577	است	0.00607495	ی	0.0025601	شود	0.00175484	ه	0.000863763	ا	0.000611564	س	0.0004
15	67	ا	0.676255	ر	0.196039	ا	0.0435665	◆	0.0249044	ه	0.0105359	ا	0.00220421	ر	0.00215855	ٔ	0.00136784	ه	0.0010
16	67	ا	0.996928	ها	0.000521465	ما	0.000401551	می	0.000347432	m	0.000205265	◆	0.000198296	ی	0.000162818	ا	0.000114531	می	9.4323
17	67	ر	0.99028	ور	0.0026574	وج	0.0018359	تا	0.000676865	ها	0.000370393	ه◆	0.000311797	ون	0.000252378	ونی	0.000205488	OO	0.0001
18	67	ر	0.852321	می	0.0477245	ا	0.0359923	◆	0.0342736	ه	0.00945642	دا	0.00251408	zes	0.00170004	ک	0.00137373	zust	0.0006
19	67	ش	0.993485	ه	0.00471683	ی	0.000293769	است	0.000232569	ی	0.000216259	◆	0.000201084	ی	0.000145061	شود	9.05356e-05	ه	8.6572
20	67	ای	0.938269	ا	0.0212496	ای	0.00688317	ر	0.00376601	از	0.00338216	کر	0.0030459	◆	0.00197322	کرد	0.00129136	◆	0.0012
21	67	ش	0.546275	◆	0.0199708	ه	0.00680843	ی	0.00610615	شود	0.00340315	ن	0.00169441	س	0.00149208	◆	0.000886474	ش	0.0006
22	67	ق	0.751051	ق	0.18102	ه	0.0400939	◆	0.0052882	وق	0.00286342	وق	0.00179425	fer	0.00134836	ی	0.000749189	ψ	0.0006
23	67	ه	0.51566	ا	0.176545	ا	0.0280379		0.0126408	ر	0.012294	ح	0.00937161	[_TT_201]	0.00715642	[_TT_200]	0.00704761	[_TT_2	
24	67	ر	0.998466	◆	0.000267571	ر	0.000255438	ه	0.000175989	ح	7.97384e-05	ا	5.59577e-05	ا	4.73551e-05	ه	4.66457e-05	ه	4.5371
25	67	ا	0.801144	◆	0.0899628	ا	0.0547923	ا	0.00908596	zaman	0.0060915	◆	0.00519315	ه	0.004083	зам	0.00154836	ا	0.0011
26	67	ر	0.834358	ها	0.077761	ه	0.044157	اه	0.0122231	ا	0.00528947	ه	0.0025697	ا	0.00182864	ی	0.00169889	ا	0.0014
27	67	اب	0.996898	ا	0.00231403	ر	0.0002036	ور	0.000131853	اه	6.1952e-05	ار	4.24611e-05	◆	3.81619e-05	ער	3.16939e-05	ال	3.0972
28	67	ب	0.765037	ا	0.127337	به	0.0308068	◆	0.0225801	ع	0.00413732	بعد	0.00283223	ه	0.00234896	یال	0.0017843	مع	0.0014
29	67	ا	0.992875	◆	0.0017485	اه	0.00106867	اب	0.000451021	ار	0.000386644	را	0.000364795	◆	0.0003636	ا	0.000326222	اه	0.0003
30	67	ر	0.872562	ا	0.0994722	◆	0.00583692	ا	0.00212724	◆	0.00150171	naz	0.00100889	ی	0.000711376	ت	0.000585377	n	0.0004
31	67	ا	0.887583	ا	0.0404065	◆	0.0266331	می	0.0218366	ا	0.0153819	ز	0.00157828	азд	0.00059796	ٔ	0.000463985	س	0.0003
32	67	ا	0.99467	ا	0.00231755	دی	0.000634428	◆	0.000622866	◆	0.00020947	ک	0.000155659	دا	0.00014817	یک	0.000109354	ا	0.0001
33	67	یک	0.91201	ا	0.086852	◆	0.000412442	ک	0.000298585	ی	6.61065e-05	می	5.28254e-05	ک	3.83248e-05	ٔ	2.60317e-05	◆	1.7398
34	67	ش	0.883509	ا	0.0455484	◆	0.00583	ش	0.0147205	ا	0.00132479	ا	0.00123866	ب	0.001203	س	0.000536669	ر	0.0005
35	67	ر	0.965217	دم	0.0229712	دا	0.0060639	ور	0.00108828	◆	0.00063618	ر	0.00051912	و	0.000302797	ٔ	0.000265464	دی	0.0001
36	67	ن	0.984718	تا	0.0102602	ه	0.00417873	ر	0.000101131	اب	8.61134e-05	◆	6.64294e-05	ر	4.4171e-05	ات	2.68925e-05	می	2.6402

FIG. 6. Top 10 alternate predictions for the sentence 001-A.wav file from the Common Voice corpus (large model)

Figure 6 shows the prediction used in the final transcription (prediction 1) and the alternate predictions; this is similar to the beam search for machine translation, providing access to the alternative transcriptions of the model. For instance, the first line (which predicts the special token), lists the alternate candidates for the transcription, all of which correspond to timestamps close to the initial starting point of the file. In the highlighted line of predictions (l.16), the first prediction (with a high level of certainty, 0.9969) corresponds to the unique letter that

12. https://github.com/ggerganov/whisper.cpp

corresponds to the bilabial nasal <م>. Predictions 2 and 4 correspond to combinations with the vowel <ما> and the consonant <مر>, respectively, while predictions 3 and 5 correspond to the Latin letter encoding the bilabial nasal. An unknown character appears in prediction 6. The seventh and eighth generated subtokens are <ب>, a voiced bilabial plosive, and <ن>, a voiced alveolar nasal. These cases are particularly intriguing as they highlight challenges faced by audio LLMs like Whisper, in distinguishing between phonemes with similar places or manners of articulation. This comparison of the Whisper transcription (prediction 1) with the alternate predictions allows us to investigate the subtokens that are the potential graphemic candidates that correspond to a portion of the acoustic signal. This technique allows us to probe the subtokens that may correspond to the corresponding speech interval. This method allows us to answer this phonographemic question : What are the competing subtokens for a given representation of the signal for a given time slice of the signal and the competing languages or the competing subtokens for a given portion of the signal?

6. Results

This section presents the key findings, including the qualitative and quantitative analyses from the transcription task using various Whisper models (subsections 6.1 and 6.2). Additionally, subsection 6.3 highlights the findings on subtokenisation phase in Whisper for Persian.

6.1. Qualitative Analysis of the Transcriptions

This subsection evaluates the performance of various Whisper models in transcribing Persian. The base model showcases robust multilingual outputs, such as tokens observed from other languages like Serbian, Chinese, Arabic, and even mixed-language phrases, as seen in *Après (=after) getting* (a blend of French and English). Furthermore, it often produces hallucinations in the form of meaningless, repetitive phrases—particularly in the middle of transcriptions—which affect the coherence and reliability of the output (Guerreiro et al., 2023).

The tiny model exhibits a higher frequency of hallucinations than the base model, often repeating clusters of characters or sentences. These hallucinations frequently contain words from other languages, predominantly English. Hallucinations are found in multilingual models in various translation contexts. Hallucinations are most common in low-resource languages and non-English output, sometimes exposing toxic patterns associated with training data (ibid.). Toxic patterns refer to

harmful or undesirable outputs exhibited by AI systems due to problems in their training data or algorithms.

The small model reduces the presence of foreign tokens significantly, with only isolated English words occasionally appearing within Persian transcriptions. This improvement indicates better language isolation, allowing for clearer and more coherent transcriptions with fewer instances of language interference.

In the medium model, Persian transcription accuracy improves further; however, the model fails to handle the same phoneme, i.e., /z/, which has multiple graphemic representations in Persian: <ز>, <ذ>, <ض>, and <ظ>. Additionally, there are instances where the sequence of Persian characters is rendered incorrectly, resulting in absurd or invalid forms. For example, the noun <لطف> (/lotf/, meaning *grace*) is transcribed incorrectly as <لوتف> (/lutf/), which is not a valid sequence in Persian. There are two issues with this transcription. First, the short vowel /o/, which is not explicitly written in Persian, is incorrectly transcribed as the long vowel /u/ in the erroneous version of the word. Second, the transcription selects the wrong grapheme to represent the phoneme /t/ in Persian, choosing <ت> instead of <ط>. Overall, while the model achieves greater accuracy in transcribing Persian, issues with hallucinations persist.

Finally, the large models offer the highest level of transcription quality, virtually eliminating hallucinations and foreign language tokens. Transcriptions are consistently accurate and free from unrelated phrases. Named Entity Recognition (NER) remains an issue. For example, the proper noun <رضا> (Reza) is written with the grapheme <ض>, which represents the /z/ sound. However, even though we would expect such a frequent first name to be present in the training data, the model consistently uses the grapheme <ز>, which is not used for the spelling of this proper noun.

Another significant factor influencing transcription, whatever model is used, is the presence of loanwords from languages such as Arabic and Urdu. These borrowed words, which are integrated into Persian vocabulary, may exhibit unique phonetic and morphological patterns that affect how the model processes and transcribes them. The handling of loanwords can highlight the model's adaptability to cross-linguistic influences, but it also raises questions about the comprehensiveness of its training data in capturing these nuances. For example, <شکر> (/ʃokr/) means "gratitude" in both Arabic and Persian.

Moreover, the language frequency distribution within the Whisper dictionary plays a crucial role in shaping transcription outputs. The frequency of subtokens directly impacts the likelihood of their selection during the transcription process. If a subtoken appears more frequently in the training data, the model is more inclined to produce it, even in

contexts where a less frequent variant might be more appropriate. This distribution can influence both the selection of formal versus colloquial variants and the representation of loanwords, underscoring the interplay between token frequency and linguistic accuracy in Whisper's transcriptions.

Last, a noteworthy observation across Whisper models is the addition of a letter referred to as the silent /he/ to represent the *ezafe* construction, which corresponds to an unstressed /e/ sound. In standard Persian orthography, this [e] sound is not explicitly transcribed. For example, Whisper transcribes <دوسته من> (/dusteh man/, meaning "my friend"), whereas the gold standard transcription in Persian is <دوستِ من> (/duste man/, meaning "my friend"). This discrepancy reflects a common challenge encountered by Persian learners and individuals with limited formal education, highlighting the difficulty of aligning phonetic and orthographic conventions in Persian transcription.

6.2. Quantitative Analysis of the Transcriptions

This subsection presents a statistical analysis of the transcriptions generated by Whisper, focusing on Word Error Rate (WER) and Character Error Rate (CER) as key metrics to evaluate transcription accuracy (Morris, Maier, and Green, 2004). For Persian, CER is particularly valuable, as it accounts for variations in spacing, which are common in the language and can impact transcription evaluation. Unlike WER, CER effectively ignores inconsistencies in how spaces are placed between letters, providing a more precise measure of accuracy for Persian text. For Persian, CER is very close to a subtoken error rate.

Table 5. WER and CER results for different models. The best-performing model (Large) is shown in bold characters

Model	*WER*	*CER*
Base	0.97	0.84
Tiny	2.66	1.22
Small	0.89	0.58
Medium	0.83	0.56
Large	**0.42**	**0.20**
Large-v2	0.57	0.24

The large Whisper model performed best, with the lowest WER (0.42) and CER (0.20). It captures both word-level and character-level accuracy effectively. The medium model also performed well, with WER (0.83) and CER (0.56), though it is not as accurate as the large model.

Base and small models showed higher WER and CER compared to both medium and large models. Tiny model performed the worst, with extremely high WER (2.66) and CER (1.22), indicating significant transcription errors. Model large-v2 strikes a balance but does not outperform the large model.

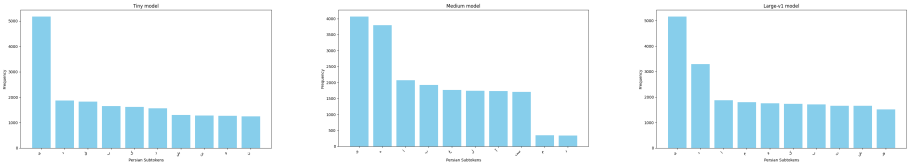
6.3. Distribution of Subtokens in Transcription Across Models

Our comparative analysis between model variants reveals hierarchical performance patterns aligned with model size. The larger models, particularly large and large-v2, demonstrate marginally superior vocabulary acquisition capabilities, especially beyond the 10^3 word threshold. The medium and small variants maintain remarkably similar growth trajectories through the middle ranges, while the tiny and base models show distinct divergence patterns at various intervals. This divergence becomes more pronounced in the upper word count ranges, suggesting that model capacity plays an increasingly significant role in vocabulary acquisition as the lexical complexity increases. Notably, several models exhibit plateau phases at different points in their growth curves, indicating potential capacity limitations or optimization boundaries specific to each model architecture.

In this subsection, we report our findings on the transcription of the data. Figure 7 displays the frequency of different Persian subtokens produced by the Whisper tiny, medium, and large-v1 models. The x-axis represents different Persian subtokens. The y-axis shows the frequency values, ranging up to 5000. As for the tiny model, the subtoken on the far left <ی> has the highest frequency, exceeding 5000. The remaining subtokens have lower frequencies, ranging approximately between 1000 and 2000. The pattern shows an interesting distribution, with one dominant token and several others at lower frequencies. The bar chart for the Medium Whisper model's Persian subtoken distribution shows that the highest frequency is 4100 occurrences for the subtoken <ی>. The second most frequent subtoken <ه> has approximately 3750 occurrences. There is a significant drop after the second token, with the next group clustering around 1500-2000 occurrences. The lowest frequencies, <م> and <ه>, are seen in the last two subtokens, with around 300-400 occurrences. In sum, the graph illustrates a clear non-uniform distribution of Persian subtokens. This distribution may affect how well the model handles different types of Persian text.

Figure 7 also shows the frequency of different Persian subtokens produced by the large-v1 model of Whisper. There appears to be a descending pattern in frequency from left to right like the other models. The most frequent subtoken <ی> has approximately 5000 occurrences,

and the second most common subtoken <د> has around 3300 occurrences. The remaining subtokens show relatively consistent frequencies between 1500-2000 occurrences. The distribution pattern is not uniform, showing a sharp decline between the two first frequent subtokens. After the second subtoken, the frequency stabilises with a more gradual decline.



(a) (b) (c)

FIG. 7. Comparison of the distribution of subtokens across models

By analysing the distribution of subtokens in Whisper outputs across different Whisper models, we can compare these distributions to those of Persian subtokens within the FLEURS dataset, as illustrated in Figure 2. All Whisper models, with the exception of smaller models (e.g., tiny), demonstrate consistency with the distribution of graphemes in the FLEURS dataset. The graphemes <ی> and <ا> are the most frequent in the FLEURS dataset, a pattern also reflected in the outputs of the larger Whisper models, namely medium and large-v1.

6.4. Probing of the Ten Best Predictions

In this experiment, we analysed the ten best predictions for the large model in 15 minutes of speech from the Mozilla Foundation. First, we show the validity of the results of our method, where we probe the alternative predictions for the special subtoken used to indicate the beginning of transcriptions [_BEG_]. We used the density probability representation (Figure 8), which is why the probabilities of the unique occurrences of [_TT_30] [_TT_6] are not represented. The most frequent second alternate predictions are [_TT_1] (5 occurrences) and [_TT_2] (23), where these special subtokens correspond to time stamps (Ballier, Burin, et al., 2024). Even if the time alignment is far from perfect with Whisper (ibid.), the results are consistent with the corresponding portion of the signal: these time stamps point to the very beginning of the signal (the bigger the [_TT_] value, the further from the beginning).

We similarly illustrate the consistency of the Whisper alternate predictions for the voiced alveolar fricative and for the bilabial nasal as case

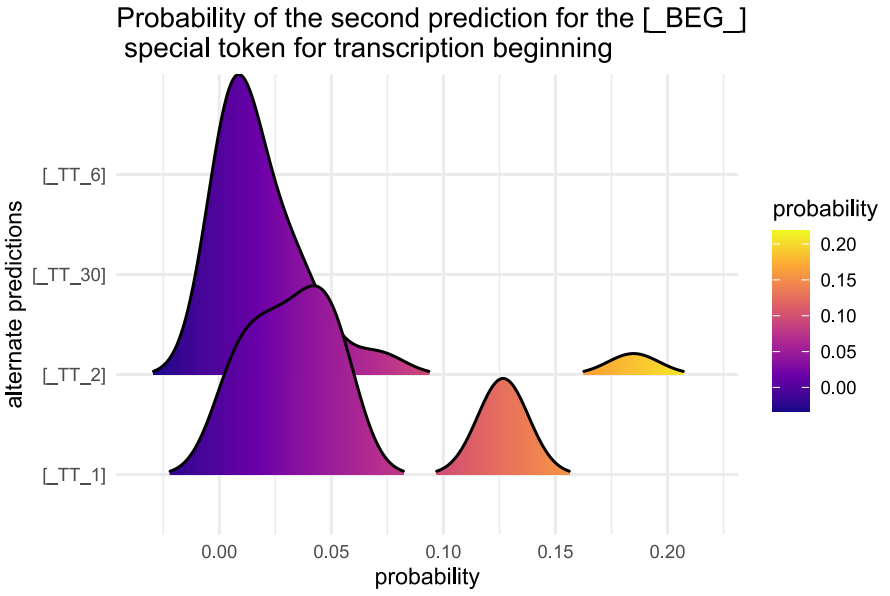


Fig. 8. The second best predictions for the [_BEG_] Whisper special token for transcription beginning (large model)

studies. For the transcription of the voiced fricative /z/ for the word <همزمان>, the expected letter would be <ز> . However, alternate predictions include subtokens belonging to other languages, like зам /zam/, which is a Russian word possibly meaning “deputy” or “assistant” or the subtoken in Latin characters *zaman* (see Figure 6). The first prediction corresponds to the correct letter in Persian for the voiced alveolar fricative. Nevertheless, the alternative letters are used in the predictions three and five and six and eight, but they are mostly used in Arabic. Alternatively, the full subtoken *zaman* is used in the fifth prediction and corresponds to the full realization of the following sounds. The alternate candidates for the same portion of the speech signal are nevertheless consistent rendering and transcriptions of the sibilant, exposing the competition of the subtokens of the different languages in the prediction. This probing of the alternate predictions confirms the correspondence between the token predicted and the acoustic properties of the corresponding speech signal.

A similar phenomenon can be observed with nasals. They all correspond to subtokens in different languages that correspond to the same phonological sounds like <m> and <M> with the Latin script. Figure 9 illustrates the alternate predictions for the bilabial nasal *min*, where the following vowel can be observed as well as the alveolar nasal <ن>.

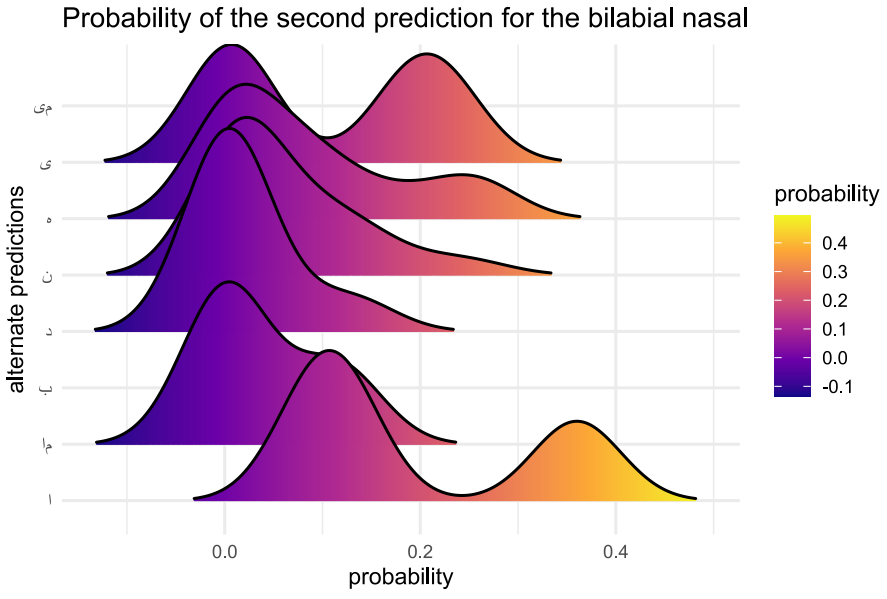


FIG. 9. The second best predictions for the bilabial nasal (large model)

Probing Whisper at the prediction phase does not capture the graphotactics of Whisper, but rather is a vantage point for the characterisation of the phonographematics of an audio LLM. It also efficiently summarises the dictionary biases of the subtokenisation: languages with a limited number of subtokens in the Whisper dictionary are more likely to be transcribed with subtokens used to represent other languages. Since all the Whisper models were trained with the same subtoken dictionary, this probably explains the hallucinations observed with the lower models: the transcriptions use 'foreign' subtokens that are deemed to be more probable, because the association of the foreign subtoken with the acoustic signal is probably more present in the training data of the other languages.

7. Discussion

7.1. Distribution of Subtokens in Whisper Dictionary Across the Perso-Arabic Script

We assume the byte pair encoding model (the subtokeniser) was trained on the full dataset of the Whisper training data, where Arabic

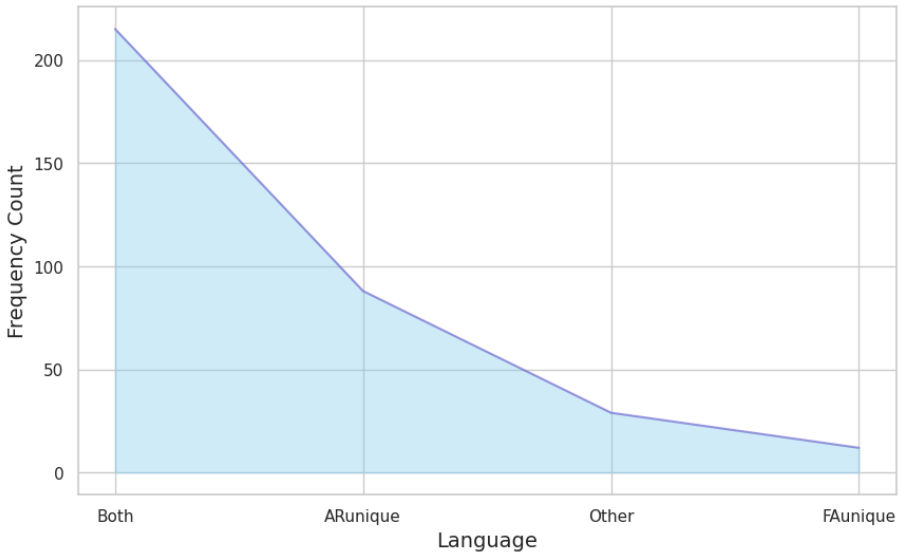


FIG. 10. Language Frequency Distribution in the Whisper Dictionary (large model)

has 739 hours and Persian 24 for the transcription task. To analyse the distribution of Perso-Arabic written characters in the Whisper multilingual dictionary, we extracted the corresponding subtokens [of the large model]. Figure 10 shows that 345 subtokens are associated with Perso-Arabic characters. Among these, 215 subtokens correspond to both Persian and Arabic (labeled as "Both"), 88 subtokens are specific to Arabic (labeled as "ARunique"), and only 12 subtokens are exclusive to Persian, representing characters that are uncommon in Arabic but unique to Persian (labeled as "FAunique"). Tadjiki and Urdu subtokens are labeled as "Other".

Figure 10 highlights the over-representation of Arabic, which shares the same script as Persian, in Whisper's subtoken dictionary. This imbalance explains the predominance of Arabic outputs, particularly in Whisper's smaller models (e.g., base, tiny, and small). Other languages are also present in the dictionary, notably Chinese, which reflects the challenges of multilingual models with uneven language distributions. This issue has been studied, and prior research suggests that fine-tuning can significantly improve performance for under-represented languages by allocating more subtokens to them (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023).

7.2. Sensitivity to Gender, Social Level, Accent, Register

Understanding how speaker attributes—such as gender, social background, accent, and register—affect ASR system performance is a critical area of investigation. Graham and Roll (2024a) and Patman and Chodroff (2024) focused on English, evaluates the impact of these variables on OpenAI's Whisper ASR system. The results reveal Whisper's strong recognition capabilities for North American English accents but highlight challenges with British, Australian, and certain non-native accents, particularly Vietnamese and Thai. Speaker-specific characteristics, including English proficiency, vowel inventory, and language typology, significantly influence recognition accuracy. Notably, male speakers generally experience higher error rates. Whisper also demonstrates superior performance with read speech compared to conversational speech, though the gender-related accuracy gap diminishes in conversational contexts. These findings underscore the importance of developing robust and inclusive ASR systems that can effectively handle diverse accents and speech patterns. Future research should prioritise reducing biases, utilizing more diverse datasets, and advancing personalized, context-aware ASR technologies to enhance real-world applicability. For Persian, building on previous experiments (Namdarzadeh et al., 2023), we report our findings on the variability of Whisper outputs for the same text read by two different native Tehrani speakers. Based on this experiment, Despite both female speakers employing the Tehrani dialect, commonly spoken in Iran's capital, variations in the articulation of certain consonants and vowels were observed when the same word was read. Specifically, phonemic confusions were noted, including between /m/ and /b/, as well as /ʃ/ and /tʃ/. Additionally, inconsistencies arose in the differentiation between short and long vowels, such as /e/ (close-mid front unrounded) and /a/ (open front unrounded). These findings suggest that subtle differences in speakers' manner of articulation can influence the transcription outputs, even for the large model.

7.3. The Subtokenisation Architectural Bias

An important factor to further investigate for the phono-graphematics of Whisper is the role of the distribution of languages and of the subtokenisation algorithm. Assuming the transcriptions of 680,000 hours of English are mixed all the other languages for the training of the byte pair encoding algorithm, under-resourced languages seem to be doomed to a character-based representation : whereas languages like English are represented by several thousand subtokens, some of them being actual words, less represented language only get a small fraction of the subtoken dictionary. The unequal distribution of the languages in the training

data lead to an architectural bias in the Whisper models (Ballier, Burin, et al., 2024). Previous research has pointed out some architectural biases of subtokenisation for NLP tasks. For example, for translation, Wisniewski, Zhu, Ballier, and Yvon (2021) have shown that the more subtokens are needed for occupational nouns, the more gender bias errors are likely to be found in translations. Therefore, it is likely that subtokenisation raises issues that are still to be uncovered, whether in the representation of the written language (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023) or for the spoken language. Our paper is an initial investigation of these issues using a language that is under-represented in the lists of languages used for the training data, sharing letters from the same Unicode group, suggesting some of the dominant effects that can also be pointed out when investigating the transcriptions from Whisper. Clearly, the model size, therefore the number of representations, plays a crucial role, as well as, predictably enough, the size of the training data. Compared to the larger model where very few representations were spotted when transcribing 24 hours of speech of Persian, errors of that type could be spotted in the different models.

It may be the case that alternate (sub)tokenisation algorithms may be preferable, a point already made for the morphological tokenisation of Soranî Kurdish (Ahmadi, 2020), which is written in a modified Perso-Arabic script. Because of its specific orthographic and structural properties, the Perso-Arabic script presents specific challenges and opportunities for subtokenization. Subtokenisation models need to account for these differences to achieve effective morphological segmentation, and it would seem that the Unigram model tend to outperform BPE for in this task. It should be noted that this issue has been raised for LLMs (Petrov, La Malfa, Torr, and Bibi, 2024) and mitigating strategies (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023; Han and Zou, 2024) have been proposed for written models, and comparable experiments should be carried out with audio LLMs.

8. Conclusion and Future Research

We have probed the graphemic outputs of Whisper and some of its internal representations for the transcriptions of spontaneous speech. Our reverse engineering perspective has enabled us to document how graphemes of Persian are encoded in a multilingual model like Whisper. Our methodology could be replicated for other languages that have a writing system potentially biased in the (sub)tokenisation phase. We have been able to discuss architectural bias and properties of subtokens and explore how linguistic properties (like the *ezafe*) are encoded in LLMs.

One of the questions is the generalisability of our reverse engineering method. We can suggest that it can be extended to all the languages which are treated by Whisper as a character-based model, namely languages for which the Whisper dictionary of subtokens mostly comprises their alphabet. As a result of which, the decoder has to apply most of the graphotactic constraints of the said language. This is what we showed with the automatic substitution of initial and final letters which were required in the final transcript as opposed to the grapheme in isolation that are predicted internally, see Table 3 that summarizes this point. The list of languages likely to be treated on a character-based system is likely to be deducted from the small number of hours in the training data as published in the (Radford et al., 2023) appendix.

The demographic of the speakers in the Whisper training data is only partially documented in Radford et al. (ibid.) since the training data is not really known, even for the small number of hours (24) of the Persian training data. As we suggested, we can nevertheless surmise that similar effects that have been diagnosed for English (Graham and Roll, 2024a; Sanabria et al., 2023) can be observed, for example, a different sensitivity to accents and gender has been documented for English (Graham and Roll, 2024a). Future research will need to see if the fine-tuned model for Persian using the Mozilla Foundation Common Voice data[13] has similar gender bias for speech recognition.

Our strategy has been to treat the Whisper dictionary of subtokens as a repository of graphemic units mapped to acoustic representations of varying scope (from phonemic, syllabic to word-based representation for languages like English) correspondences.[14] We have shown that Persian corresponds the character-based representation : a limited number of subtokens in the dictionary was used in the encoder for the prediction phase and graphotactic rules are applied by the decoder.

More generally, our roadmap for investigating the phonological "knowledge" of audio LLMs consists of investigating the existence of phonographemic mapping rules for subtokens. For this phonology of subtokens, if we assume that subtokens are associated to similar pronunciations, what is the acoustic granularity of this similarity? In the name of robustness, we have to assume some plasticity for the acoustic input matching the subtoken, a unique subtoken corresponding to several phonetic realisations. A systematic analysis of the top ten alternate predictions (see Figure 6) for the transcription task should provide insight into the phonetic representations of the different models: the subtokens

13. https://huggingface.co/speechbrain/asr-whisper-large-v2-commonvoice-fa

14. We do analyse the dictionary of the English only models, based on GPT2, nor do we detail here the fraction of special subtokens used to represent tasks, task position of time stamps, see for instance Ballier, Burin, et al. (2024).

predicted reflect the categorisation of phonetic neighbours as encoded in the Whisper subtoken representations.

9. Acknowledgements

The authors would like to express their gratitude to Prof. Richard Wright for his assistance with our revised phonological analysis of Persian and the phonetic correlates potentially associated with Persian graphemes. Additionally, they are grateful to Dr. Jean-Baptiste Yunès for providing the C++ implementation of OpenAI’s Whisper, which made this work possible. The authors would like to extend their appreciation to the reviewers for their insightful comments and constructive feedback, which have significantly enhanced the quality of this work. We also thank Guillaume Wisniewski and Alice Henderson for comments on an earlier version of this paper.

A. Appendix

Table 6. Perso-Arabic Subtokens in the Multilingual Whisper Dictionary

ا	ل	ي	ن	ر	م	و	ت	د	ه	ب	ع	ة	س	ك
ق	ح	ف	أ	ال	ع	ي	ج	ى	ه	ش	ن	ذ	في	ك
ص	ط	ار	الم	خ	ين	ية	ج	ام	عل	ض	ها	ح	ون	ز
د	إ	ق	ر	ث	ور	ير	الل	وا	أن	ست	لا	ما	ذا	على
هم	الأ	دي	ان	خ	ري	الت	اب	گ	غ	اد	الع	عن	ما	كن
لم	آ	ظ	ئ	لي	مع	وم	بال	الس	بي	الح	الله	عد	ني	ط
ى	لك	ود	را	هذا		چ	ذه	حد	اس	كم	لل	الق	الش	رب
يا	يد	رة	الج	علي										

Among the 344 Perso-Arabic characters extracted from the Whisper dictionary, 240 are Persian but not exclusive to Persian, as they are also used in Arabic. Punctuation marks are recorded in the Perso-Arabic format, following the right-to-left writing system.

The remaining characters belong to Arabic transcripts, some of which have been phased out of Persian due to the language’s evolution and domestication.

References

Ahmadi, Sina (2020). "A tokenization system for the Kurdish language". In: *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 114–127.

Arab-Moghaddam, Narges and Monique Sénéchal (2001). "Orthographic and phonological processing skills in reading and spelling in Persian/English bilinguals". In: *International Journal of Behavioral Development* 25, pp. 140–147. DOI: 10.1080/01650250042000320.

Ardila, R. et al. (2020). "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215.

Asebriy, Zahra et al. (2014). "Comparative systems of handwriting Arabic character recognition". In: *2014 Second World Conference on Complex Systems (WCCS)*, pp. 90–93. DOI: 10.1109/ICoCS.2014.7060923.

Ballier, Nicolas, Léa Burin, et al. (2024). "Probing Whisper Predictions for French, English and Persian Transcriptions". In: *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*. Ed. by Mourad Abbas and Abed Alhakim Freihat. Trento: Association for Computational Linguistics, pp. 129–138. URL: https://aclanthology.org/2024.icnlsp-1.15.

Ballier, Nicolas et al. (2023). "Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech, a Case Study with French Learners of English". In: *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*. Ed. by Mourad Abbas and Abed Alhakim Freihat. Online: Association for Computational Linguistics, pp. 282–292. URL: https://aclanthology.org/2023.icnlsp-1.30.

Baluch, Bahman and S. Choudhury (2011). "Spelling transparency and its impact on short-term memory: Evidence from persian and english". In: *Spelling Skills: Acquisition, Abilities, and Reading Connection*, pp. 93–106.

Bostrom, Kaj and Greg Durrett (2020). "Byte Pair Encoding is Suboptimal for Language Model Pretraining". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4617–4624.

Conneau, Alexis et al. (2022). "FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech". preprint arXiv:2205.12446.

———— (2023). "Fleurs: Few-shot learning evaluation of universal representations of speech". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 798–805.

Devlin, Jacob et al. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186.

Downey, C.m. et al. (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)*. Ed. by Duygu Ataman. Singapore: Association for Computational Linguistics, pp. 268–281. DOI: 10.18653/v1/2023.mrl-1.20.

Freihat, Abed Alhakim and Mourad Abbas, eds. (2021). *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Trento, Italy: Association for Computational Linguistics. URL: https://aclanthology.org/2021.nsurl-1.0.

Gerganov, Georgi (2003). "whisper.cpp : A High-performance inference of OpenAI's Whisper automatic speech recognition (ASR) model". URL: https://github.com/ggerganov/whisper.cpp.

Graham, Calbert and Nathan Roll (2024a). "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits". In: *JASA Express Letters* 4.2, p. 025206.

——— (2024b). "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits". In: *JASA Express Letters* 4.2, p. 025206. ISSN: 2691-1191. DOI: 10.1121/10.0024876.

Guerreiro, Nuno M. et al. (2023). "Hallucinations in Large Multilingual Translation Models". In: *Transactions of the Association for Computational Linguistics* 11, pp. 1500–1517. DOI: 10.1162/tacl_a_00615.

Halbout, Dominique and Muhammad-hossein Karimi (2012). *Le persan*. Assimil.

Han, Yujin and Difan Zou (2024). "Understanding and Mitigating Tokenization Bias in Language Models". In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 17480–17504. URL: https://proceedings.mlr.press/v235/han24g.html.

Hosseini, Fatemeh Sadat et al. (2021). "IDPL-PFOD: An Image Dataset of Printed Farsi Text for OCR Research". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 22–31. URL: https://aclanthology.org/2021.nsurl-1.4.

Jafari, Mohammad Mahdi et al. (2021). "Improving pre-trained Language Model for Relation Extraction Using Syntactic Information in Persian". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 38–44. URL: https://aclanthology.org/2021.nsurl-1.6.

Karimi, Sarvnaz, Falk Scholer, and Andrew Turpin (2007). "Collapsed Consonant and Vowel Models: New Approaches for English-Persian Transliteration and Back-Transliteration". In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Annie Zae-

nen and Antal van den Bosch. Prague, Czech Republic: Association for Computational Linguistics, pp. 648–655. URL: https://aclanthology.org/P07-1082.

Karimi, Simin (2005). *A Minimalist Approach to Scrambling. Evidence from Persian*. Berlin, New York: De Gruyter Mouton. DOI: doi:10.1515/9783110199796.

Khalilia, Hadi, Abed Alhakim Freihat, and Fausto Giunchiglia (2021). "The Dimensions of Lexical Semantic Resource Quality". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 15–21. URL: https://aclanthology.org/2021.nsurl-1.3.

Kudo, T. (2018). "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing". preprint arXiv:1808.06226.

Kummervold, Per E et al. (2024). "Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges". preprint arXiv:2402.01917.

Mahootian, Shahrzad and Lewis Gebhardt (1997). *Persian*. Descriptive grammars. Includes index. London: Routledge.

Maleki, Jalal and Lars Ahrenberg (2008a). "Converting Romanized Persian to the Arabic Writing Systems". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Ed. by Nicoletta Calzolari et al. Marrakech, Morocco: European Language Resources Association (ELRA). URL: https://aclanthology.org/L08-1315/.

——— (2008b). "Converting Romanized Persian to the Arabic Writing Systems". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/738_paper.pdf.

Manova, Stela (2023). "ChatGPT, *n*-grams and the power of subword units: The future of research in morphology". In: *Fourth International Workshop on Resources and Tools for Derivational Morphology*. Vol. 5, p. 1.

Manova, Stela et al. (2020). "What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology". In: *Word Structure* 13.1, pp. 1–21.

"Frontmatter" (2023). In: *Persian Computational Linguistics and NLP*. Ed. by Katarzyna Marszałek-Kowalewska. Berlin, Boston: De Gruyter Mouton, pp. I–IV. DOI: doi:10.1515/9783110619225-fm.

Mohebbi, Hosein et al. (2023). "Homophone Disambiguation Reveals Patterns of Context Mixing in Speech Transformers". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Asso-

ciation for Computational Linguistics, pp. 8249–8260. DOI: 10.18653/v1/2023.emnlp-main.513.

Morris, Andrew Cameron, Viktoria Maier, and Phil D Green (2004). "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition". In: *Interspeech*, pp. 2765–2768.

Namdarzadeh, Behnoosh et al. (2023). "Fine-tuning MBART-50 with French and Farsi data to improve the translation of Farsi dislocations into English and French". In: *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*. Ed. by Masaru Yamada and Felix do Carmo. Macau SAR, China: Asia-Pacific Association for Machine Translation, pp. 152–161. URL: https://aclanthology.org/2023.mtsummit-users.14/.

Oji, Romina, Nasrin Taghizadeh, and Heshaam Faili (2021). "PerSpellData: An Exhaustive Parallel Spell Dataset For Persian". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 8–14. URL: https://aclanthology.org/2021.nsurl-1.2.

Patman, Chloe and Eleanor Chodroff (2024). "Speech recognition in adverse conditions by humans and machines". In: *JASA Express Letters* 4.11.

Petrov, Aleksandar et al. (2024). "Language model tokenizers introduce unfairness between languages". In: *Advances in Neural Information Processing Systems* 36.

Radford, Alec et al. (2023). "Robust Speech Recognition via Large-Scale Weak Supervision". In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

Rahbari, Noriyeh and Monique Sénéchal (2009). "Lexical and nonlexical processes in the skilled reading and spelling of Persian". In: *Reading and Writing* 22.5, pp. 511–530. DOI: 10.1007/s11145-008-9122-1.

Rastorgueva, Vera Sergeevna and Steven P. Hill (1964). *A Short sketch of the grammar of Persian*. Indiana University Research Center in Anthropology, Folklore and Linguistics Publication. "Also : Part II of the International Journal of Ammerican Linguistics". Bloomington: Indiana University.

Salehi, Ali and Cassandra L. Jacobs (2024). "The Effect of Model Capacity and Script Diversity on Subword Tokenization for Sorani Kurdish". In: *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Ed. by Garrett Nicolai et al. Mexico City, Mexico: Association for Computational Linguistics, pp. 51–56. DOI: 10.18653/v1/2024.sigmorphon-1.6.

Sanabria, Ramon et al. (2023). "The Edinburgh international accents of English corpus: Towards the democratization of English ASR". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.

Sartakhti, Moein Salimi, Romina Etezadi, and Mehrnoush Shamsfard (2021). "Improving Persian Relation Extraction Models By Data Augmentation". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 32–37. URL: https://aclanthology.org/2021.nsurl-1.5.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural Machine Translation of Rare Words with Subword Units". In: *CoRR* abs/1508.07909. URL: http://arxiv.org/abs/1508.07909.

——— (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)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162.

Shakeri, N. et al. (2015). "Investigating Phonological Awareness in Persian-Speaking Children With Phonological Disorders". In: *Middle East Journal of Rehabilitation and Health* 2.4, e32200. DOI: 10.17795/mejrh-32200. URL: https://brieflands.com/articles/mejrh-21516.

Song, Xinying et al. (2021). "Fast wordpiece tokenization". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2089–2103.

Sproat, Richard and Alexander Gutkin (2021). "The taxonomy of writing systems: How to measure how logographic a system is". In: *Computational Linguistics* 47.3, pp. 477–528.

Taghizadeh, Nasrin, Ali Ebrahimi, and Heshaam Faili (2021). "NSURL-2021 Shared Task 1: Semantic Relation Extraction in Persian". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 1–7. URL: https://aclanthology.org/2021.nsurl-1.1.

Taguchi, Chihiro and David Chiang (2024). "Language Complexity and Speech Recognition Accuracy: Orthographic Complexity Hurts, Phonological Complexity Doesn't". preprint arXiv:2406.09202.

Wisniewski, Guillaume et al. (2021). "Screening Gender Transfer in Neural Machine Translation". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 311–321.

Yousef, Saeed and Hayedeh Torabi (2018). *Persian*. Routledge comprehensive grammars. Abingdon, Oxon: Routledge.