Computational Benchmarks for Egyptian Arabic Child Directed Speech

Salam Khalifa,[1,2] Abdelrahim Qaddoumi,[1] Nizar Habash,[2] Owen Rambow[1]
Institute for Advanced Computational Science, and Department of Linguistics
[1]Stony Brook University
Computational Approaches to Modeling Language (CAMeL) Lab,
[2]New York University Abu Dhabi
{first.last}@stonybrook.edu, {first.last}@nyu.edu

Abstract

We present AraBabyTalk-Egy, an enriched release of the Egyptian Arabic CHILDES corpus, that opens the child-adult interactions genre to modern Arabic NLP research. Starting from the original CHILDES recordings and IPA transcriptions of caregiver-child sessions, we (i) map each IPA token to fully diacritized Arabic script, and (ii) add core part-of-speech tags and lemmas aligned with existing dialectal Arabic morphological resources. These layers yield ∼26K annotated tokens suitable for both text- and speech-based NLP tasks. We provide a benchmark on morphological disambiguation and Arabic ASR. We outline lexical and morphosyntactic differences between AraBabyTalk-Egy and general Egyptian Arabic resources, highlighting the value of genre-specific training data for language acquisition studies and Arabic speech technology.[1]

1 Introduction

Child-adult interactions corpora (commonly known as Child Directed Speech; CDS) are a gateway to studying child language acquisition and language development in children. Such corpora are used extensively in psycholinguistics, speech language pathology, and even theoretical linguistics. In computational linguistics and NLP, CDS corpora are, however, not as popular, either due to their relatively limited size, or due to the lack of linguistic annotation. However, such corpora are relevant for NLP as well. Children learn language efficiently given a small and sparse input.Thus, studying CDS corpora may point to new ideas for improving NLP tools for low resource languages. At the same time, core NLP technology, such as morphosyntactic tagging and dependency parsing, can facilitate the study of child language acquisition, and the development of educational technologies and policies.

[1]http://arababytalk.camel-lab.com/

B	ʔæbu- be dæ?	أَبُوا بِدَه؟
M	ʔuh nelʕæbu be dæ	أَه نِلْعَبُوا بِدَه
	jalla ʔesmu ʔe:h dæ?	يَلَّا اِسْمُه اَيه دَه؟
B	di?	دِي؟
M	ʔesmu ʔe:h?	اِسْمُه اَيه؟
B	ʔo:ʃ.	أُوش.

Table 1: A short Baby-Mother (**B**-**M**) dialog in IPA and Arabic orthography. ***B:****-lay with this?* ***M:*** *Yes we play with this ok what is it called?* ***B:*** *this?* ***M:*** *what is it called?* ***B:*** *Oosh. (rouge, lipstick)*

English, a well-studied language in NLP, has a large repository of CDS corpora with rich representations. They have facilitated advances in multiple areas such as syntactic parsing (Liu and Prud'hommeaux, 2023) or cognitively inspired language modeling (Warstadt et al., 2023). Crucially, the existing CDS corpora use exactly the same orthographic conventions as standard English, which allows the use of existing NLP resources when studying these corpora.

The situation for Arabic is very different. While a few CDS corpora exist for some dialects, they are either closed-source or they lack useful representation, such as a standard orthographic transcription and linguistic annotations.

In this work, we present AraBabyTalk-Egy, an open-source enriched CDS corpus of Egyptian Arabic (EGY). We provide orthographic transcriptions, lemmatization, and part-of-speech (POS) tagging, all aligned to an existing IPA-only transcription CDS corpus of EGY (Salama, 2015). Our work will enable the immediate use of this CDS corpus in the NLP community.

We review related work (Section 2), describe the original corpus (Section 3), and present our enriched version (Section 4). A snapshot and comparison with other resources in Sections 5 & 6. Benchmarks are presented in Section 7.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2296–2307

2 Related Work

CDS Corpora The Child Language Data Exchange System (**CHILDES**; MacWhinney, 2000) has one of the most extensive open source collections of CDS corpora in many languages; the North American dialect of English alone has 54 datasets with hundreds of recorded children and rich data representations, including orthographic and phonological transcription, morphological and syntactic annotations, and more. CHILDES is open-source under the TalkBank project (MacWhinney, 2025). Within the TalkBank project, Arabic has CHILDES datasets for only two dialects: Egyptian Arabic (Salama, 2015) and Palestinian Arabic (Nazzal, 2021). Both datasets include children of diverse ages and genders but provide only IPA phonemic transcriptions, with no additional annotations. Since Arabic NLP tools rely on orthographic input, these CDS corpora are not accessible even for the simplest NLP tasks. The only dialectal Arabic CHILDES corpus with morphological annotations is the Emirati Arabic Language Acquisition Corpus (**EMALAC**; Ntelitheos and Idrissi, 2017), but it is outside TalkBank and is neither open-source nor redistributable.

Our dataset, AraBabyTalk-Egy, extends the EGY CHILDES corpus by providing fully diacritized orthography, lemmatization, and POS tagging that are directly aligned with the existing IPA transcriptions. We chose EGY to complement its existing NLP resources.

CDS Corpora in NLP Most NLP work on CDS corpora has focused on English CHILDES, due to its rich annotations. Syntactic parsing is especially active (Sagae et al., 2007, 2010; Huang, 2016; Abend et al., 2017; Liu and Prud'hommeaux, 2023; Szubert et al., 2024; Yang et al., 2025), with related work on Dutch (Odijk et al., 2018) and Hebrew (Gretz et al., 2015). Morphological inflection and productivity have also been widely studied in English and German CHILDES (Kirov and Cotterell, 2018; McCurdy et al., 2020; Belth, 2021; Kodner and Khalifa, 2022). Cognitively-inspired language modeling is another active area (Huebner et al., 2021; Xu et al., 2021; Warstadt et al., 2023; Feng et al., 2024).

To our knowledge, no similar efforts exist for Arabic. This paper aims to address that gap by providing a richly annotated, community-accessible CDS corpus.

Name		G	Age	
Flopater	(FL^M)	M	1;07.02	(1.6)
Yara	(YR^F)	F	1;09.20	(1.8)
Basmala	(BM^F)	F	2;02.18	(2.2)
Bilal	(BL^M)	M	2;04.19	(2.4)
Razan	(RZ^F)	F	2;10.00	(2.8)
AbdrahmanFawzy	(AF^M)	M	3;00.00	(3.0)
ZiyadYasser	(ZY^M)	M	3;05.09	(3.4)
Farah	(FR^F)	F	3;05.20	(3.5)
ZiyadMohammed	(ZM^M)	M	3;07.12	(3.6)
Merna	(MR^F)	F	3;08.01	(3.7)

Table 2: **Demographic information** as provided in EgyChildes. Age is in Y;MM.DD format and in a decimal year format for short. **G**ender: **M**ale and **F**emale. Children's names are abbreviated with their gender as a superscript for easier reference.

3 Egyptian Arabic CHILDES

Demographics The original EGY CHILDES corpus (**EgyChildes**; Salama, 2015) includes audio recordings of 10 monolingual Egyptian Arabic-speaking children (5 girls, 5 boys) residing in Alexandria, Egypt, each with a single 30-minute session. Ages range from 1.6 to 3.7 years. Full demographic details are shown in Table 2.

Recording Contents Each session is an audio recording of a spontaneous, unstructured interview with an investigator, a parent, or both, while doing any combination of the following: asking questions, conversing, naming objects, naming pictures, playing, singing, or telling stories. Recordings took place in kindergartens, homes, or the investigator's home.

Data Format Each session has two files, a text-based `.cha` transcription file, and an `.mp3` audio file. Each `.cha` file has a metadata header that provides all the necessary demographic information about the child and the adult interlocutor(s). The sessions were transcribed near-phonemically in IPA, and each speaker's turn (utterance) is in a separate line. The transcriptions also include special annotations by the investigator, such as certain characteristics of the speech according to the general CHILDES transcription guidelines found in (MacWhinney, 2000). At the end of each turn, the duration of the utterance is provided in milliseconds, which provides direct alignment with the audio files. Table 3-(a) shows an example from a raw `.cha` file.

	Spk	Utterance	Duration
(a)	*CHI:	ʔæbu [: nelʕæbu] [* f:p] be dæ ?	1795_2670
	*MOT:	ʔuh nelʕæbu be dæ jalla ʔesmu ʔe:h dæ ?	2670_5007
	*CHI:	di [= lipstick] ?	5007_5879
	*MOT:	ʔesmu ʔe:h ?	5879_6791
	*CHI:	ʔo:ʃ [: ro:ʒ] [* p] .	6791_8087
(b)	*CHI:	ʔæbu-:nelʕæbu be dæ ?	1795_2670
	*MOT:	ʔuh nelʕæbu be dæ jalla ʔesmu ʔe:h dæ ?	2670_5007
	*CHI:	di ?	5007_5879
	*MOT:	ʔesmu ʔe:h ?	5879_6791
	*CHI:	ʔo:ʃ-:ro:ʒ .	6791_8087
(c)	*CHI:	نلعَبُوا بِ# دَه ؟	1795_2670
	*MOT:	اَه نِلعَبُوا بِ# دَه يِلَّا اِسمُه اِيه دَه ؟	2670_5007
	*CHI:	دِي ؟	5007_5879
	*MOT:	اِسمُه اِيه ؟	5879_6791
	*CHI:	روج	6791_8087

Table 3: A 6 second **transcription excerpt** from the session of `Basmala,Female,2;02.18`. The speaker (Spk) codes '*CHI' and '*MOT' refer to the child, Basmala, and her mother, respectively. Table (a) shows the raw transcription as stored in the `.cha` file in EGYCHILDES, Table (b) shows the preprocessed IPA, and Table (c) shows the Arabic orthography of the transcription in the final ARABABYTALK-EGY.

4 ARABABYTALK-EGY

This section outlines the creation of ARABABYTALK-EGY, an orthographically diacritized EGY child-adult interaction corpus with lemmas, POS tags, and a lexicon, built on the IPA provided in **EGYCHILDES** corpus.

Starting from the preprocessed IPA transcripts, we (a) extract a frequency list of unique word types, (b) use GPT-4o to generate an initial fully diacritized orthographic form for each type, (c) manually revise the orthographic forms and add lemma and core POS in the lexicon, and (d) map the lexicon back to all sessions followed by manual in-context validation of the orthography, lemmas, and POS, and consulting the audio as needed.

The subsequent sections, including the benchmarks uses the and lemma/POS annotations as reference; the original IPA transcripts are retained for alignment to **EGYCHILDES** only.

4.1 Preprocessing

The only preprocessing step we perform is removing the various annotations provided by the transcriber as noted in §3, as they are orthogonal to the orthography itself.[2] The only annotation we kept is the reference to children *mispronunciations*. The annotations of mispronunciations follow this format: `child_form [: adult_form]`. We replace unintelligible words or utterances transcribed as `xxx` with the keyword `NONE`. In cases where the removal of annotations results in an empty utterance, we also inserted the keyword `NONE`. Table 3-(a) shows multiple examples of such annotations. Table 3-(b) shows the preprocessed version of the example.

4.2 Automatic Orthography Transcription

Once all files are preprocessed, we extract the frequency list of all unique uttered words appearing in all of EGYCHILDES across the different speakers. This list acts as the seed for the final lexicon, which includes a lemma and a core POS for each entry. The total size of the lexicon is 4,170 words. The frequencies follow a Zipfian distribution, as expected in naturally occurring speech.

In order to make this resource compatible with mainstream dialectal Arabic resources, we opted to follow the Conventional Orthography for Dialectal Arabic (**CODA**; Habash et al., 2018) guidelines since there is no standard orthography for the dialects. CODA aims to balance between preserving dialectal uniqueness while maintaining the relationship with Modern Standard Arabic (MSA). CODA's set of guidelines, as described by its authors, is *"a consistent ad hoc convention that balances being MSA-like with being generally phonemic, and morphologically and syntactically faithful to the dialect"*. CODA is written using Arabic script, and it maintains etymological consonants spelling and vowel length. Similar to standard MSA orthography, diacritics representing short vowels and gemination are optional; however, in this work we aim for a rich representation, so we provide a fully diacritized orthography.

As an initial step, we passed all the lexicon's IPA entries through GPT-4o (OpenAI, 2024a,b) to provide an initial orthographic transcription in diacritized CODA.[3] Appendix Table 9 shows the prompt we used. We enforced a strictly structured output through the OpenAI API to account for the occasional missing tokens in the output that LLMs are known for. The prompting was done in batches

[2]These annotations are recoverable if needed, as each preprocessed sentence remains aligned with its raw counterpart.

[3]The Copyright of EGYCHILDES permits this use: https://talkbank.org/0share/rules.html

of 20 words at a time for maximal diacritized output, as pilot experiments showed that larger batches often yielded undiacritized output. At this stage, the lexicon contains IPA entries, their frequencies, and initial orthographic transcriptions for each entry.

4.3 Manual Annotation

For efficiency, we manually annotated in batches of related entries by leveraging an approximation of the consonantal root of the words, which we automatically generated by removing vowels. This "*root*" is not necessarily the canonical Arabic tri- or quad-literal root, but an approximation that helps group related entries together. Within each batch, the orthography is carefully revised and fixed if needed, then the diacritized lemma is provided along with the core POS tag. The lemma is the singular masculine (if applicable) form of the nominals, and the perfective masculine singular for verbs. We use the stem tags from the Buckwalter tagset (Buckwalter, 2004) since it is fine-grained and backwards compatible with existing resources for EGY. We follow CODA guidelines regarding word boundaries even if they don't align with phonological boundaries. For example, the negation particle /ma-/ is part of the phonological word, but is always split in CODA guidelines; for such cases, we separate them using '_' in the lexicon to signal a split in the full text. Similarly, particles written as a single letter, such as conjunctions and prepositions, e.g., the preposition /be/ 'with' بِ, must be attached to the word following it; in such cases, we mark those particles with a '#' in the lexicon to signal a merge in the full text.

The overall accuracy of the diacritized output of GPT-4o compared to the manually validated version is 27.5% at the word level, and when omitting diacritization, it is 44.4%. When looking at the most frequent entries, with 10 or more occurrences, which make up only 7.5% of the entries, the accuracy is 46% diacritized and 67.7% undiacritized. The errors made by GPT-4o varied; some were different diacritizations (extra or missing), hallucinated letters, wrong consonants, CODA non-compliant word segmentation, and very few were complete nonsensical words and characters. This indicates that generating a diacritized orthography from a given pronunciation out of context is not trivial for such models, especially the longer tail of the distribution.

4.4 Lexicon

Our intuition behind this top-down annotation approach is that since the starting point of the annotation is the transcription of the pronunciation, the out-of-context ambiguity is low, especially knowing the genre of the text where the expected vocabulary is limited. However, we found a few cases where there is real ambiguity, such as /ʔæ:lu/ where it could mean 'he said it[m.sg]' قَالُه or 'they said' قَالُوا. In this case, the orthography entry will have both options; however, they both share the same lemma قَال and POS VERB_PV. Another example is /gebnæ/, which could mean 'cheese' جِبنَة or 'we brought' جِبنَا. This entry will also have both orthographic options, and the lemmas جِبنَة and جَاب and POS for each option NOUN and VERB_PV, respectively. Out of all the entries in the lexicon, only 114 (2.7%) entries were phonologically ambiguous, corresponding to 2% of ARABABYTALK-EGY. Another type of ambiguity is that of POS, such as active and passive participles (اسم الفاعل and اسم المفعول) which can take either ADJ or NOUN, which is a result of a common semantic shift from nouns to adjective (Marzouk et al., 2025). Those cases and others make around 1.2% of the lexicon. Table 4 shows entries from the final lexicon along with the output from GPT-4o for comparison. The examples were chosen based on two consonantal roots /ʕ.r.f/ and /l.b.s/ to illustrate the annotation batches. The final lexicon comprises 4,170 IPA types which are also the key entries corresponding to 26,265 tokens, 3,113 orthographic types, 1,101 lemmas, and 29 POS tags. It is worth noting that the number of unique orthographic forms is less than the number of unique IPA entries, which is expected for the following reasons. First, some words could occur multiple times differently because of the different (mis)pronunciation of the children as we discussed in §4.1, these make up around 19.5% of the lexicon, and 0.8% of ARABABYTALK-EGY. Second, inconsistent transcriptions such as the noun بِالسِّكِّينَة 'with the knife' are found to be transcribed in two ways: /besseki:næ/, and /bessekki:næ/, those cases do not represent actual pronunciation differences, but rather transcription mistakes or inconsistencies; in this example, the latter one is the correct form. Finally, there are some subtle pronunciation differences that are phonetic rather than phonemic and therefore are not reflected in standard orthogra-

phy, such as the verb دَوَّرِي 'find [2.f.sg]' which is transcribed as /dawwari/, and /dʕawwari/, in this example, the first is the regular pronunciation. Table 3-(c) shows the final version of the example in full Arabic orthography, which is in the final AraBabyTalk-Egy.

Following the annotation of the lexicon, all 10 sessions were then mapped to their respective orthographic forms along with their lemmas and POS. We manually validated the orthography, lemmas, and POS in context and made corrections when necessary, including disambiguation of cases as mentioned earlier. Most of the phonologically ambiguous cases in the lexicon were resolved later by the context itself. We referred to the audio in cases where the IPA transcription and the context did not resolve inconsistencies we encountered. It turns out that for the majority of cases, the transcription in fact did not match the audio; these errors primarily consisted of the absence of determiners and vowel quality. We found that 96% of all the orthography remained unchanged, which indicates the feasibility of this annotation approach.

Lexicon construction and in-context annotation/validation were performed by a single expert annotator. The annotator is an NLP researcher and co-author of this paper with extensive prior experience developing and annotating Egyptian Arabic and Arabic dialect corpora. Because the annotation targets standardized orthographic normalization under CODA, the decisions are highly constrained and largely non-interpretive; therefore, and given the small corpus size, we did not compute inter-annotator agreement.

Release We release AraBabyTalk-Egy and its lexicon under CC BY-NC-SA 3.0. This is in compliance with the licensing of EgyChildes (Salama, 2015), which AraBabyTalk-Egy was built upon.

5 AraBabyTalk-Egy: A Snapshot

In this section, we present insights drawn from AraBabyTalk-Egy in terms of statistics and observations. The main focus is on the children's specific portions of the corpus. We then contrast it with the adults' portions of the corpus as well.

5.1 Children vocabulary

Tables 5-(a) shows the summary per child in terms of the number of utterances (full turns) and tokens, and the mean length of utterance (MLU). Table 5-(b) is the summary in terms of unique types

Freq	IPA	Arabic	Lex	POS	GPT-4o
19	maʕrafʃ	مَا_اَعرَفش	عَرَف	VERB_IV	مَعرَفش
19	ʕæ:rfæ	عَارِفَة	عَارِف	ADJ	عَارِف
8	bijeʕraf	بِيِعرَف	عَرَف	VERB_IV	بِيَعرَف
7	teʕrafi	تعرَفِي	عَرَف	VERB_IV	تعرِفِي
5	maʕrafhuʃ	مَا_اَعرَفهُوش	عَرَف	VERB_IV	مَاعرِفهُش
5	ʕæ:refhæ	عَارِفهَا	عَارِف	ADJ	عَارِفهَا
44	læ:bes	لَابِس	لَابِس	ADJ	لَعبِس
8	læ:bsæ	لَابِسَة	لَابِس	ADJ	لَعبَسَا
5	ħælbeshæ	حَالَبِسهَا	لِبِس	VERB_IV	هَلبِسهَا
5	ħætelbes	حَتلبِس	لِبِس	VERB_IV	حَاتلبِس
4	ʔelbeshæ	اِلبِسهَا	لِبِس	VERB_CV	البِشَا
4	læ:besu	لَابِسه	لَابِس	ADJ	لَاعِبُوَا

Table 4: A **portion of the lexicon** based on the unique types of IPA utterances in EgyChildes along with the fully diacritized Arabic orthography, lemma, and core POS tag. The last column is the initial automatic orthographic transcription generated using GPT-4o for comparison.

of IPA forms, orthographic forms, lemmas, and POS. The general trend shows that vocabulary and MLU in general increase as children grow older. One child, `ZY,M,3.4`, has relatively longer utterances and produced a richer vocabulary compared to other children around his age. This is very apparent in the number of unique forms and lemmas, which suggests a larger number of lexical items and paradigms produced.

Vocabulary Complexity Table 5-(c) shows another view of the vocabulary per child. The words per lemma ($\frac{words}{lex}$) can be seen as an approximation of the morphological complexity acquired by the children. Lemmas per POS ($\frac{lex}{pos}$) shows the diversity of lemmas per POS, i.e., the richness of the vocabulary. For both those metrics, we see a relative increase as the children grow older. On the other hand, the type-token ratio ($\frac{type}{tok}$) decreases as age increases, which indicates less repetition and more diverse usage of the vocabulary. Similarly, the phonology-orthography ratio ($\frac{ipa}{orth}$) also decreases with age, which is an indicator of diversity in unstable pronunciation. We found that the lower $\frac{ipa}{orth}$, the more the utterances are adult-like, as we found a strong correlation (0.92, $p < 0.001$) between $\frac{ipa}{orth}$ and the ratio of mispronounced words to all types.

Across the ten children, there seems to be no systematic difference between the two genders. While boys exhibited a moderately higher lemma diversity within POS categories (M = 10.74 vs. 8.75

		(a) Counts			(b) Types				(c) Complexity			
Age	**C**	**UTT**	**MLU**	**Tokens**	**IPA**	**Orth**	**Lex**	**POS**	$\frac{words}{lex}$	$\frac{lex}{pos}$	$\frac{type}{tok}$	$\frac{ipa}{orth}$
1.6	FL^M	402	1.24	498	202	140	132	19	1.1	6.9	0.3	1.4
1.8	YR^F	394	1.26	495	135	87	76	15	1.1	5.1	0.2	1.6
2.2	BM^F	574	1.35	776	292	220	182	22	1.2	8.3	0.3	1.3
2.4	BL^M	610	1.85	1,129	388	355	230	25	1.5	9.2	0.3	1.1
2.8	RZ^F	372	1.78	661	320	286	211	24	1.4	8.8	0.4	1.1
3.0	AF^M	603	2.13	1,286	375	349	225	23	1.6	9.8	0.3	1.1
3.4	ZY^M	265	9.35	2,478	854	773	404	27	1.9	15.0	0.3	1.1
3.5	FR^F	514	2.64	1,357	552	521	307	27	1.7	11.4	0.4	1.1
3.6	ZM^M	522	2.25	1,176	456	441	278	22	1.6	12.6	0.4	1.0
3.7	MR^F	424	3.43	1,453	556	527	321	27	1.6	11.9	0.4	1.1
Average		468	2.73	1,131	413	370	237	23	1.5	9.8	0.3	1.2
		±114	±2.4	±591	±206	±205	±95.7	±3.9	±0.3	±2.9	±0.1	±0.2
Adults		415	3.0	1,245	354	407	219	24	1.6	9.3	0.3	1.1
		±151	±0.3	±491	±133	±354	±64	±2.3	±0.2	±1.9	±0.1	±0.0

Table 5: Per child (C) sub-vocabulary: (a) **Counts**: number of utterances (UTT), mean length of utterance (MLU), number of tokens (Tokens). (b) **Types** unique type counts: IPA forms (IPA), orthographic forms (Orth), lemmas (Lex), and POS tags. (c) **Complexity** in terms of: $\frac{words}{lex}$ words per lemma, $\frac{lex}{pos}$ lemmas per POS, $\frac{type}{tok}$ type-token ratio, and $\frac{ipa}{orth}$ phonological forms per orthographic form. The two bottom parts of the table are the overall average across all children, followed by the average across the adults' sub-vocabulary of AraBabyTalk-Egy.

$\frac{lex}{pos}$), it is due to one outlier child `ZY,M,3.4`, who, as we saw earlier, exhibits above average vocabulary richness. We confirmed this by recomputing the average across gender without `ZY`, see Table 6 for details. It is essential to note that, given the number and size of the samples, as well as the diverse recording environments and other potential confounds, no definitive conclusions can be drawn according to gender for this corpus.

Age	**G**	$\frac{words}{lex}$	$\frac{lex}{pos}$	$\frac{type}{tok}$	$\frac{ipa}{orth}$
2.8	M	1.5	10.7	0.3	1.2
±0.8		±0.3	±3.1	±0.0	±0.2
2.8	F	1.4	9.1	0.3	1.2
±0.8		±0.3	±2.7	±0.1	±0.2
2.6	M	1.4	9.6	0.3	1.2
±0.9		±0.3	±2.3	±0.1	±0.2

Table 6: Averaged vocabulary complexity metrics across **G**ender. The top part is all the children, the second is without the outlier `ZY;M;3.4`.

Similar summary measures (e.g., MLU and lexical diversity) are available for many other CHILDES languages, and are easily accessible through childes-db (Sanchez et al., 2019), enabling future cross-linguistic comparisons beyond the scope of this work.

5.2 Comparison with Adult Vocabulary

The adult vocabulary refers to the adult interlocutor's vocabulary in AraBabyTalk-Egy. In Table 5-(**Adults**) we show the average vocabulary complexity metrics across all the adults in AraBabyTalk-Egy.

The MLU, $\frac{words}{lex}$, and $\frac{type}{tok}$ are higher than the average of the children's; the rest are lower. The contrast between the lexical diversity ($\frac{lex}{pos}$) and the morphological complexity ($\frac{words}{lex}$) seems to suggest that children get more diverse morphological input rather than lexical input, however, more investigation is needed to confirm this.

On the other hand, $\frac{type}{tok}$ is easily explainable since the adults in the sessions appeared to repeat a lot of what the children say multiple times. Similarly, the lower $\frac{ipa}{orth}$ confirms the instability in children's pronunciation; the reason it is larger than 1 is that the adults sometimes repeat the mispronunciation of the children and the general transcription inconsistencies in the annotation, as mentioned earlier.

We also computed the Jaccard Similarity index of the vocabulary between the children and the adults and found that it is 0.4 for full word forms and 0.7 for lemmas, which is another indication of the diversity in morphology between the input and the output to children.

6 AraBabyTalk-Egy and Egyptian Arabic Resources

In this section, we compare AraBabyTalk-Egy to the following existing EGY resources which are being actively used in many NLP tasks such as language modeling, morphological analysis and disambiguation, and morphophonological modeling (Pasha et al., 2014; Inoue et al., 2021, 2022; Khalifa et al., 2025) among others.

- **ECAL** The Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002) is a pronunciation dictionary primarily based on CALLHOME Egypt (Gadalla et al., 1997). Each entry in ECAL includes an orthographic (undiacritized) form, phonological form, and morphological analysis, including a phonemically transcribed lemma. We compare the coverage in terms of orthographic forms and lemmas only since the phonological forms were transcribed using different schemes. The phonemically transcribed lemma is easily mappable to orthography since it is the base form of the word in isolation, and therefore it usually does not undergo major morphophonological changes.
- **ARZTB** The Egyptian Arabic Treebank (Maamouri et al., 2014) is the primary resource in developing morphological disambiguation systems for EGY. We compare the coverage in terms of words and lemmas that appeared in the corpus.
- **CALIMA$_{EGY}$** The Egyptian Arabic Morphological Analyzer (Habash et al., 2012) is a morphological analyzer that generates a set of possible analyses for a given input token out of context. Each analysis includes a diacritized orthographic form, diacritized lemma, POS tag, and morphological features. We compare the coverage with the morphological database of CALIMA$_{EGY}$ in terms of the diacritized lemmas. We measure the coverage of the analyzer in terms of the correctly generated diacritized word, lemma, and POS given the words from AraBabyTalk-Egy as input. The analyzer ignores any diacritization in the input, therefore, their presence has no effect in the generated analysis.

Table 7 shows an overview of the coverage between the different sub-vocabularies within AraBabyTalk-Egy (in types) and the different resources described above. Of the three resources, ARZTB has the least lemma coverage with AraBabyTalk-Egy, which is expected since the source of the corpus is online discussion forums which tend to be more text-based formal discussions with MSA code-switching. ECAL, on the other hand, has more substantial coverage than ARZTB as it is based on CALLHOME Egypt which is transcribed natural conversations between adult speakers. Finally, CALIMA$_{EGY}$ has the most coverage in terms of lemmas since it has a comprehensive lexicon that is based on multiple resources as described in (Habash et al., 2012). When it comes to fully inflected words, the coverage is lower than that of lemmas, which is expected, however, consistent across the resources. For CALIMA$_{EGY}$, we measured the morphological analysis coverage by running each word through the analyzer and finding a match on the triplet of the diacritized word, diacritized lemma, and the POS. We found that words in the children, adults, and across sessions vocabularies have a full coverage of 64.9%, 62.3%, and 61.1%, respectively. A relaxed measure of matching only the diacritized form has a coverage of 73.2%, 71.0%, and 69.6% for children, adults, and session vocabularies, respectively. The full match is a proxy for the full array of morphological features, hence, a large portion of the lexicon will get rich morphological representations.

Across sub-vocabularies, adults have a consistently larger coverage than children in lemmas, but the opposite in fully inflected words. This is consistent with the vocabulary complexity metrics we discussed in §5. Adult vocabulary has a lower $\frac{lex}{pos}$ than children's, meaning fewer lemmas. On the other hand, children's vocabulary has a lower $\frac{words}{lex}$ which indicates less morphological diversity.

These observations are a clear indication of the divergence between the CDS genre and other mainstream resources.

7 Benchmarks

In this section we present results of evaluating state-of-the-art (SOTA) systems for morphological disambiguation and automatic speech recognition tasks. Since CDS corpora are underrepresented in NLP, and even more so Arabic CDS corpora, we believe it is necessary to evaluate well established

Resource	Diacritized			Undiacritized		
	Child	Adult	Session	Child	Adult	Session
Lemma Coverage						
ECAL	67.9	69.0	67.2	79.5	79.6	78.2
ARZTB	64.0	65.0	62.6	72.4	73.2	71.0
CALIMA$_{EGY}$	83.0	82.7	81.5	93.7	93.0	92.2
Word Coverage						
ECAL	—	—	—	62.6	60.7	58.0
ARZTB	40.1	38.6	35.9	49.4	47.3	44.7

Table 7: **Lexical coverage** between AraBabyTalk-Egy and existing resources for Egyptian Arabic, expressed as percentage coverage for inflected words and lemmas across the different sub-vocabularies of AraBabyTalk-Egy and the whole sessions. **ECAL** does not have readily available diacritized tokens. **CALIMA$_{EGY}$** is a morphological analyzer with a lemma-based database.

tasks within the field on this data. For each of the tasks, we evaluate a single SOTA system on complete sessions rather than sub-vocabularies, where each session represents the full conversation between a child and an adult (see §3).

7.1 Morphological Disambiguation

Morphological Disambiguation is the task of providing the morphological analysis for a given token in context. A morphological analyzer takes in one word at a time and gives all possible analyses unranked out of context. A morphological disambiguator, unlike an analyzer, takes in a whole sentence and gives one analysis per word or a list of *ranked* analyses.

We evaluate the state-of-the-art morphological disambiguator for Arabic and its dialects (Inoue et al., 2022) through the CamelTools API (Obeid et al., 2020). This specific disambiguator uses dialect-specific taggers that predict POS and other morphological and morphosyntactic tags. A morphological analyzer corresponding to the dialect generates all possible context-independent analyses for the word. The predicted tags are then used to select the correct analysis. From the chosen analysis, we also obtain the diacritized word and lemma. Since we are working with EGY, the morphological analyzer paired with the model is CALIMA$_{EGY}$.

We report the performance on the fully diacritized word (Diac), diacritized lemma (Lex), core POS tag (POS), and the full analysis of **D**iac, **L**ex, and **P**OS (DLP) as we show in Table 8. While the performance of the POS tagging is the highest among the other metrics, it is still lower than the reported (94%) on ARZTB using the same morphological disambiguator. This is then followed by the performance of Lex then Diac. Finally, the strictest metric, DLP, performs the lowest.[4] However, the average DLP (50.3%) is around 11% behind the upper bound of CALIMA$_{EGY}$ coverage as we reported earlier in §6, which shows that disambiguation for this genre is not performing optimally. However, this is not a surprise since the disambiguator was solely trained on ARZTB which, as we already showed, is a different genre and overlaps the least with AraBabyTalk-Egy in terms of lexical coverage.

Age	Child	(a) Disambiguation ↑				(b) ASR ↓	
		Diac	Lex	POS	DLP	WER	CER
1.6	FL^{M}	58.9	57.4	80.3	47.2	95.0	70.8
1.8	YR^{F}	60.0	59.6	85.1	51.8	79.7	67.4
2.2	BM^{F}	53.3	62.4	86.8	47.6	69.8	54.3
2.4	BL^{M}	62.2	67.1	88.2	55.7	84.2	66.9
2.8	RZ^{F}	55.1	63.5	87.9	49.0	91.0	79.8
3.0	AF^{M}	61.8	65.3	88.2	55.8	89.9	75.1
3.4	ZY^{M}	57.8	62.7	85.3	49.2	96.7	82.1
3.5	FR^{F}	56.9	64.0	86.2	50.1	96.8	80.9
3.6	ZM^{M}	54.4	58.4	83.4	47.8	96.4	78.6
3.7	MR^{F}	57.1	64.2	85.5	48.5	94.8	69.8
Average		57.8 ±3.0	62.5 ±3.1	85.7 ±2.4	50.3 ±3.2	89.4 ±9.0	72.6 ±8.6

Table 8: Benchmarking results per *complete session*, i.e., children and adults: (a) **Morphological disambiguation** (accuracy %) in terms of full diacritized form (Diac), full diacritized lemma (Lex), and core part-of-speech (POS), using the system described in Inoue et al. (2022). We also report the accuracy of all of them together (DLP). (b) **Automatic Speech Recognition (ASR)**, using GPT-4o-transcribe, in terms of Word Error Rate (WER %) and Character Error Rate (CER %).

[4] To investigate the effect of the different demographics, we evaluated on the sub-vocabularies of children and adults, and the trend was similar. See Tables 10 and 11 in Appendix A.

7.2 Automatic Speech Recognition

ASR is the task of transcribing a speech audio signal. The transcription produced by ASR is usually in the mainstream orthography of the given language. Since EgyChildes was transcribed in IPA only, it can not be directly evaluated for ASR. Therefore, AraBabyTalk-Egy fills this gap by providing the reference orthography. We opted to use GPT-4o-transcribe (OpenAI, 2025), which is the flagship for OpenAI speech-to-text models. Compared to their Whisper family of models, the GPT-4o based models allow for text prompting and have better performance according to OpenAI's documentation.

GTP-4o-transcribe accepts a maximum length of 1,400s of audio signal input and 16,000 tokens of text input for prompting. We experimented with different lengths of the audio, 30s, 60s, and 1,000s for input, i.e., the `.mp3` files are split into those durations. As for the prompt, we used the simple prompt in Arabic "اكتب المحادثة بالعامية المصرية" 'write the conversation in colloquial Egyptian'. We kept it simple to avoid changes due to prompt optimization. We passed "ar" for Arabic for the language parameter, and "temperature" was set to 0.

To evaluate, we performed minimal normalizations to both the reference text and the output as follows. We removed punctuation and normalized all types of *hamzated Alif* (word and stem initial glottal stop) into bare Alif. Predictions are scored using word error rate (WER %) and character error rate (CER %) using the Jiwer toolkit (Vaessen, 2025). Of the three audio lengths we tested, input of length 30s had the lowest WER and CER on average. Table 8 shows the results per session. The full results for all the setups are in Table 12 in the appendix. Salhab et al. (2025) provided some recent results on a variety of Arabic dialect corpora using a variety of systems including GPT-4o-transcribe. Their results provide WER in the 10-70% range. This shows that AraBabyTalk-Egy is significantly more challenging than existing dialectal Arabic speech corpora for ASR systems.

Upon further inspection of the output we found that the quality of the audio has an observable effect on the performance of ASR. By computing the noise for each `.mp3` file we found that the session of `BM,F,2.2` to have the least amount of noise while the sessions of `FR,F,3.5` and `ZY,M,3.4` have the poorest quality. This is apparent in the respective WER and CER scores shown in Table 8. In the samples we inspected from the audio files with the poorest quality, the model produced perfectly coherent and legible output; however, it was completely unrelated to the reference whatsoever.

8 Conclusion and Future Work

We presented AraBabyTalk-Egy, a linguistically enriched version of the Egyptian Arabic CHILDES corpus, bridging a major gap in Arabic NLP by enabling computational modeling of Child-Directed Speech (CDS). We demonstrated the utility of this resource via benchmarks in morphological disambiguation and ASR, and highlighted how Egyptian Arabic CDS differs structurally from general Egyptian Arabic.

Future work will expand annotations to syntactic structures, and explore cross-dialectal comparisons with other dialectal Arabic CDS datasets, such as Palestinian Arabic CDS (Nazzal, 2021). We also aim to investigate the cognitive and linguistic development reflected in children's speech. Integrating this corpus into pretraining pipelines and child-centric educational technologies presents opportunities for NLP and developmental linguistics.

Limitations

While AraBabyTalk-Egy offers valuable enhancements to the Egyptian Arabic CDS corpus, several limitations remain. First, the corpus is relatively small (≈26K tokens), which restricts the training of data-intensive models and limits generalizability. Second, our annotations (orthographic, morphological, and POS) are limited in scope and granularity; deeper syntactic and semantic layers are not yet included. The mapping from IPA to Arabic script, while systematic, may introduce ambiguity or errors due to dialectal variability and phonetic overlap despite being fully checked manually. Third, the corpus in its current version reflects the corrected child speech in cases where children mispronounce words. We are aware that this has consequences when evaluating ASR systems, however, we leave this empirical question to future work. Additionally, the current benchmarks focus only on morphological disambiguation and ASR, leaving other tasks like dependency parsing or lexical acquisition underexplored. Finally, the resource is specific to Egyptian Arabic, and its findings may not transfer easily to other dialects or MSA, highlighting the need for broader cross-dialectal CDS

corpora in Arabic. Addressing these limitations is critical to fully unlocking the potential of CDS data for Arabic NLP.

Acknowledgments

We thank Jordan Kodner for his helpful discussion. We also thank Rawan Bondok and Mostafa Saeed for their help with specifics of the Alexandrian variety of Egyptian Arabic. We thank the anonymous reviewers for their valuable feedback. Rambow gratefully acknowledges support from the Institute for Advanced Computational Science at Stony Brook University.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

Caleb Belth. 2021. The Greedy and Recursive Search for Morphological Productivity.

Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic transcripts LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

Shai Gretz, Alon Itai, Brian MacWhinney, Bracha Nir, and Shuly Wintner. 2015. Parsing hebrew childes transcripts. *Language Resources and Evaluation*, 49(1):107–145.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Workshop of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, pages 1–9, Montréal, Canada.

Nizar Habash, Salam Khalifa, Fadhl Eryani, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified Guidelines and Resources for Arabic Dialect Orthography. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Rui Huang. 2016. An evaluation of pos taggers for the childes corpus. Master's thesis, City University of New York (CUNY).

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1708–1719, Dublin, Ireland. Association for Computational Linguistics.

Salam Khalifa, Abdelrahim Qaddoumi, Jordan Kodner, and Owen Rambow. 2025. Learning cross-dialectal morphophonology with syllable structure constraints. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 157–167, Abu Dhabi, UAE. Association for Computational Linguistics.

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Jordan Kodner and Salam Khalifa. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Modeling inflection in language acquisition. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 157–175, Seattle, Washington. Association for Computational Linguistics.

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven parsing evaluation for child-parent interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an egyptian arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ. Supported by NICHD Grant HD082736.

Brian MacWhinney. 2025. Understanding language through talkbank. *Current Directions in Psychological Science*, 34(2):75–81.

Reham Marzouk, Sondos Krouna, and Nizar Habash. 2025. A derivational ChainBank for Modern Standard Arabic. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 78–87, Abu Dhabi, UAE. Association for Computational Linguistics.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756.

Tala Nazzal. 2021. CHILDES Palestinian Arabic Nazzal Corpus. https://doi.org/10.21415/VJYY-KA80. DOI:10.21415/VJYY-KA80.

Dimitrios Ntelitheos and Ali Idrissi. 2017. Language Growth in Child Emirati Arabic. *Perspectives on Arabic Linguistics*, 29:229–248.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Jan Odijk, Alexis Dimitriadis, Martijn van der Klis, Marjo van Koppen, Meie Otten, and Remco van der Veen. 2018. The AnnCor CHILDES treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

OpenAI. 2024a. GPT-4o. Large language model accessed via OpenAI API (model name: gpt-4o, snapshot 2025-07-06).

OpenAI. 2024b. Gpt-4o system card. Technical Report arXiv:2410.21276, OpenAI.

OpenAI. 2025. GPT4o7transcribe. https://platform.openai.com/docs/models/gpt-4o-transcribe (accessed July 25, 2025). Large language model, API documentation.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1094–1101, Reykjavik, Iceland.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. 2007. High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 25–32, Prague, Czech Republic. Association for Computational Linguistics.

Kenji Sagae, Eric Davis, Alon Lavie, Brian Macwhinney, and Shuly Wintner. 2010. Morphosyntactic annotation of childes transcripts. *Journal of Child Language*, 37(3):705–29.

Heba Salama. 2015. CHILDES Egyptian Arabic Salama Corpus. https://doi.org/10.21415/78CE-VW65. DOI:10.21415/78CE-VW65.

Mahmoud Salhab, Marwan Elghitany, Shameed Sait, Syed Sibghat Ullah, Mohammad Abusheikh, and Hasan Abusheikh. 2025. Advancing arabic speech recognition through large-scale weakly supervised learning. *Preprint*, arXiv:2504.12254.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior research methods*, 51(4):1928–1941.

Ida Szubert, Omri Abend, Nathan Schneider, Samuel Gibbon, Louis Mahon, Sharon Goldwater, and Mark Steedman. 2024. Cross-linguistically consistent semantic and syntactic annotation of child-directed speech. *Language Resources and Evaluation*, pages 1–50.

Nik Vaessen. 2025. Jiwer: Similarity measures for automatic speech recognition evaluation.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. Ud-english-childes: A collected resource of gold and silver universal dependencies trees for child language interactions. *Preprint*, arXiv:2504.20304.

A Supplementary Tables

```
Prompt "role": "system",
       "content": "You are a linguist.
       Given a list of word utterances
       in Egyptian Arabic in IPA, convert
       it into FULLY DIACRITIZED ARABIC
       text. The output should be a list
       diacritized Arabic words Proper
       names are the only utterances that
       are not in IPA and are capitalized.
       They should be converted as well.
       Make sure to have the Arabic words
       fully diacritized. 'ae' is not two
       vowel, it is one vowel same as 'a'",
       "role": "user",
       "content": f"Provide the
       diacritized Arabic orthography for
       the following list of IPA words:
       \n{list_ipa}"
```

Table 9: Prompt used to get the initial orthographic transcription

Age	Child	Diac	Lex	POS	DLP
1.6	FL^M	63.1	53.6	72.0	47.2
1.8	YR^F	59.9	58.8	75.3	53.7
2.2	BM^F	54.9	61.4	78.4	49.4
2.4	BL^M	64.5	69.5	84.9	58.7
2.8	RZ^F	59.4	65.6	87.0	54.6
3.0	AF^M	65.1	67.9	90.8	61.5
3.4	ZY^M	57.9	62.8	85.0	48.9
3.5	FR^F	58.4	66.9	86.8	52.1
3.6	ZM^M	51.7	55.4	74.6	45.8
3.7	MR^F	59.9	66.2	84.4	50.3
Average		59.5 ±4.2	62.8 ±5.4	81.9 ±6.3	52.2 ±5.0

Table 10: **Morphological disambiguation** results (accuracy %) per *child* sub-vocabulary per session, i.e., children and adults. We report here the full diacritized form (Diac), full diacritized lemma match (Lex), and core part-of-speech (POS). We also report the accuracy of all of them together (DLP).

Age	Child	Diac	Lex	POS	DLP
1.6	FL^M	57.7	58.4	82.5	47.2
1.8	YR^F	60.0	59.8	88.3	51.3
2.2	BM^F	52.8	62.7	89.4	47.0
2.4	BL^M	60.9	65.7	90.2	54.0
2.8	RZ^F	53.2	62.6	88.2	46.5
3.0	AF^M	59.6	63.6	86.5	52.2
3.4	ZY^M	57.2	62.1	86.4	50.4
3.5	FR^F	55.4	61.1	85.7	48.2
3.6	ZM^M	56.3	60.5	89.5	49.1
3.7	MR^F	53.8	61.8	86.7	46.3
Average		56.7 ±2.9	61.8 ±2.0	87.3 ±2.3	49.2 ±2.7

Table 11: **Morphological disambiguation** results (accuracy %) per *adult* sub-vocabulary per session, i.e., children and adults. We report here the full diacritized form (Diac), full diacritized lemma match (Lex), and core part-of-speech (POS). We also report the accuracy of all of them together (DLP).

		WER			CER		
Age	Child	1,000	60	30	1,000	60	30
1.6	FL^M	86.4	87.8	945	80.3	66.2	70.8
1.8	YR^F	80.5	143.6	79.7	56.5	120.4	67.4
2.2	BM^F	75.3	63.7	69.8	53.0	44.1	54.3
2.4	BL^M	90.5	120.1	84.2	71.0	88.2	66.9
2.8	RZ^F	90.7	91.3	91.0	66.3	79.6	79.8
3.0	AF^M	97.0	158.2	89.9	72.3	109.5	75.1
3.4	ZY^M	98.7	96.8	96.7	95.0	83.3	82.1
3.5	FR^F	98.6	98.5	98.6	93.1	79.0	89.9
3.6	ZM^M	98.3	96.3	96.4	91.2	78.1	78.6
3.7	MR^F	94.5	167.4	94.8	75.9	156.4	69.8
Average		91.1 ±8.1	112.4 ±33.8	89.6 ±9.1	75.46 ±14.7	90.5 ±31.3	73.5 ±9.9

Table 12: Average WER(%) and CER(%) per varying length of audio signal: 1,000s, 60s, and 30s.