Morphological Analysis of Egyptian Children Corpus by KIDEVAL Program

Heba Salama
Phonetics and linguistics Department
Faculty of Arts
Alexandria *University*
Heba.salama.slp@gmail.com
https://orcid.org/0000-0002-8480-1440

Sameh Alansary
Phonetics and linguistics Department
Faculty of Arts
Alexandria University
S. alansary @alexu. org

Amany Elshazly
English Department
Faculty of Arts
Helwan University
amanyelshazly@arts.helwan.edu.eg
https://orcid.org/0000-0001-9785-8617

2022 20th International Conference on Language Engineering (ESOLEC) | 978-1-6654-5322-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ESOLEC54569.2022.10009437

Abstract— The aim of this study is to provide a morphological analysis of the Egyptian children corpus, which is a morphologically tagged and disambiguated in CHILDES. This allows the KIDEVAL program to be readily used on the corpus to address questions regarding the acquisition of Egyptian Arabic. KIDEVAL is one of the useful tools in CLAN program which has been particularly useful toolsets in the study of language acquisition in many languages. However, applications of corpus-based analyses to Egyptian children's language have not yet been conducted. This study describes how to use the KIDEVAL program for analyzing Egyptian children's language and study the development of word frequency patterns of parts of speech and order of development of grammatical morphemes in Egyptian Arabic. The output of morphological analysis enables researchers to study and answer many questions regarding the development of a grammatical morpheme in Egyptian Arabic, as well as a lot of questions that can readily be probed with KIDEVAL. The Egyptian Arabic corpus is downloaded from the Arabic part of the CHILDES database. It comprises 10transcripts from Egyptian-speaking children aged 1;7 to3;8 years, with a total of 25,645 words. The KIDEVAL program analysis profile for Egyptian Arabic children's corpus in this study reveals extensive and valuable analysis, displaying the number of occurrences of each part of speech for each child depends on his age which includes 54 categories and subcategories. The usage of the KIDEVAL tool is efficient because it reduces the time needed to label the corpus manually.

Keywords— Analyzing children language, children annotated corpus, Egyptian grammatical development, language acquisition, morphological analysis for children.

I. Introduction

The morphologically annotated corpora provide a better way for language acquisition researchers to study the course of development of Egyptian Arabic speaking child ren. The uses of corpus-based analysis for tracking the developmental stage of Egyptian children's language have not yet to be addressed. Therefore, this is one area which EA child corpus can contribute. While English and many Indo-European languages have a long history of analyzing aspects of child language production by computing various developmental indices from spontaneous language samples. CLAN (Computerized Language Analysis) programs [2], which are used with the CHILDES (Child Language Data Exchange System) database [1] include a useful tool called KIDEVAL for studying children's language in a variety of languages. KIDEVAL program is a marvelous tool that was developed from CLAN to reduce the effort involved in carrying out data analyses. KIDEVAL tool enables a researcher studying conversational interaction, language learning, or language disorders, conduct contrastive psycholinguistic studies by comparing Egyptian Arabic children's linguistic behavior to that of English and other languages using automatic morphosyntactic analysis. Moreover, address many research questions and explore many different language types. Furthermore, clinicians can use the KIDEVAL program to analyze data from delayed language children and compare that data to a large database of similar transcripts, and treat language disorders such as aphasia, particular language impairment, stuttering, dementia, and others. KIDEVAL computes the number of speech occurrences for all children in each speech part that are entered into a single analytic report automatically. These measurements are used to determine the Egyptian children's language developmental stages. The KIDEVAL's analysis helps researchers to study the development of grammatical morphemes as well as address several questions about Egyptian child language. The Egyptian Arabic analyzed corpus can be used to investigate the usual pattern of grammatical developing abilities and to diagnose language deficits in Egyptian children.

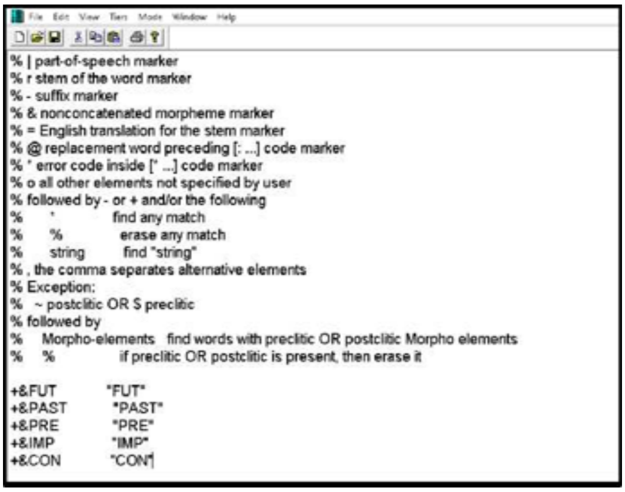
This paper starts with an overview of method as well as preparations for using the KIDEVAL program is section 2. Then section 3 demonstrates the analysis of the Egyptian Arabic children's corpus. Followed by section 4 provides the results. The final section of the paper presents the conclusion and the roadmap for future research.

II. Method Of Preparing Kideval Program

The Egyptian Arabic corpus is downloaded [3] from CHILDES database. Egyptian Arabic corpus is morphologically tagged and disambiguated version of the CHILDES (MacWhinney, 2000) consisted of 10 transcribed files based on direct spontaneous speech data collected from10 normally developing Egyptian children speak Alexandrian dialect between the age of 1;7 (1 year 7 months) to 3;7 (3 years and 7 months). The Egyptian Arabic corpus size is approximately 25,645 words. Egyptian Arabic Corpus data were elicited through playing, naming pictures, asking question, telling stories and social interaction conversation or narrating a story. All files were transcribed phonemically in the CHAT (Codes for the Human Analysis of Transcripts) format [4], which allows them to be analyzed using CLAN [2]. The transcription scheme opted to reflects morphological distinctions which have impact on the stage of analysis. Narrow phonetic transcriptions were used rather than orthographic to avoid homophonic ambiguity. The uniform and consistent coding provide by CHAT let the transcription

978-1-6654-5322-6/22/$31.00 ©2022 IEEE

of any language be readable easily. The transcripts include both the utterances of the child and of the adults, namely, the mother and the field researcher. Then, we begin the first step to prepare KIDEVAL program for analysis which is create custom subfolder called KIDEVAL folder inside ara folder in EAMOR folder downloaded [3]. The language file called ara.cut is created put into the CLAN/lib/kideval folder. The KINDEVAL has a built-in indices in ara.cut file for English. Next, ara.cut file set of headers are modified by adding or removing morphemes in ara.cut file to fit the classification of grammatical morphemes of Egyptian Arabic. Fig.1 shows how each line in the ara.cut script file defines a different type of search string. For example, in the line +&FUT,"FUT" is the label used for all instances of future tense in child's transcript.



```
% | part-of-speech marker
% r stem of the word marker
% - suffix marker
% & nonconcatenated morpheme marker
% = English translation for the stem marker
% @ replacement word preceding [: ...] code marker
% * error code inside [* ...] code marker
% o all other elements not specified by user
% followed by - or + and/or the following
%      *            find any match
%      %             erase any match
%      string         find "string"
% , the comma separates alternative elements
% Exception:
%   ~ postclitic OR $ preclitic
% followed by
%    Morpho-elements   find words with preclitic OR postclitic Morpho elements
%    %              if preclitic OR postclitic is present, then erase it

+&FUT         "FUT"
+&PAST          "PAST"
+&PRE          "PRE"
+&IMP          "IMP"
+&CON          "CON"
```

Fig. 1. *The ara.cut file inside the KIDEVAL folder*

A. Step1:running KIDEVAL in CLAN'S dialoug

KIDEVAL works on files that have been automatically annotated and analyzed with MOR. The following are the steps to efficiently run the KIDEVAL command. The Progs button is used to select KIDEVAL,and then clicks the option button and chooses 10 .cha files for Egyptian children's transcripts to analyze in one spreadsheet.

B. Step2:

Next, the dialogue box is opened and reveals KIDEVAL options. Then the speaker %CHI code which determines the child language for the input file analysis is selected and then click "done" to finish this step.

C. Step3

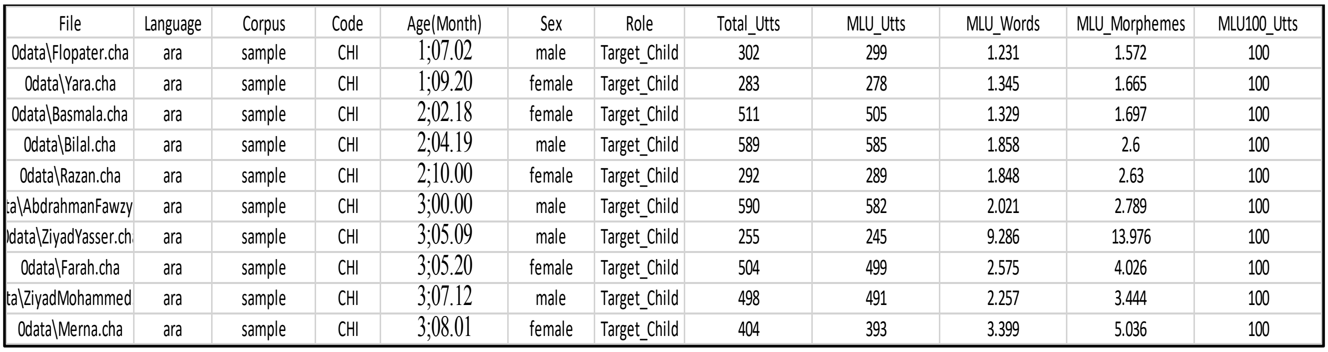
Afterwards, the language is changed to Arabic in interface window as it is not built in KIDEVAL program. Then the command will be:

```
kideval @ +lara +t*CHI
```

Where KIDEVAL is the program name, +lara specifies the language that will be analyzed. +tCHI specifies the speaker of the analysis which indicated for child

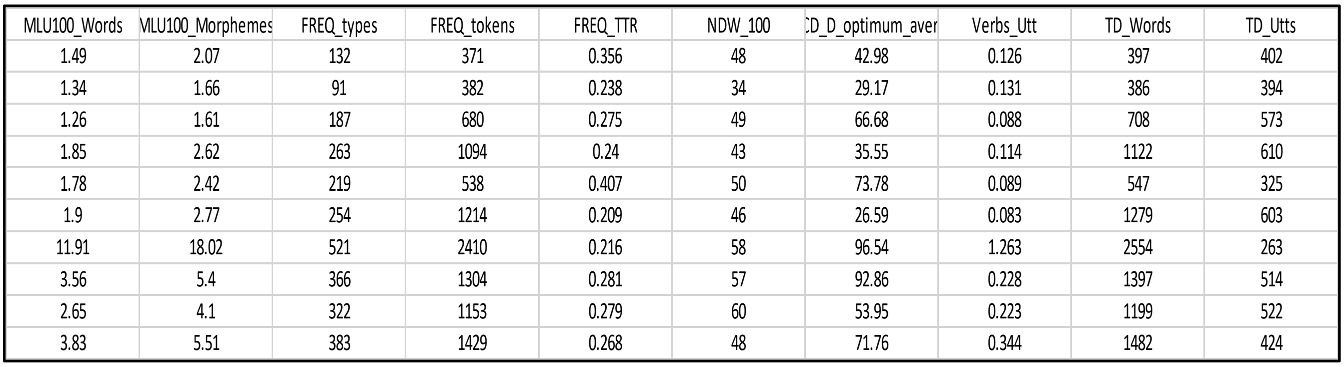
III. Analysis of The Egyptian Arabic Children's Corpus

After running KIDEVAL program morphosyntactic measures will appear, displaying all types of information that used to take several hours to compute if done by hand, such as measures of grammatical development MLU (mean length of utterance), VOCD (vocabulary diversity), and FREQ (frequency), TTR (type-token ratio) NDW (number of different words), and word errors [2]. The kideval.xls file includes 54 morphosyntactic measures for Egyptian Arabic children's corpus. Each child's data is organized in a single row. As a result, KIDEVAL was created to reduce the time and effort required to conduct data analyses in Egyptian Arabic children's corpus. CLAN's programs are specifically designed to analyze data in the CHAT format. Moreover, clinicians can use the KIDEVAL program to analyze data from delayed language children and compare that data to a large database of similar transcripts. KIDEVAL computes the number of speech occurrences for each speech part that are entered into a single analytic report automatically. These measurements are used to determine the Egyptian children's language development stage. The KIDEVAL's analysis helps researchers to study the development of grammatical morphemes as well as address several questions about Egyptian children's language. The EA analyzed corpus can be used to investigate the usual pattern of grammatical developing abilities and to diagnose language deficits in Egyptian children. The descriptive statistics produced for the word frequency patterns in the parts of speech for Egyptian Arabic children's corpusby the KIDEVAL program are demonstrated below Fig.2 to Fig.8



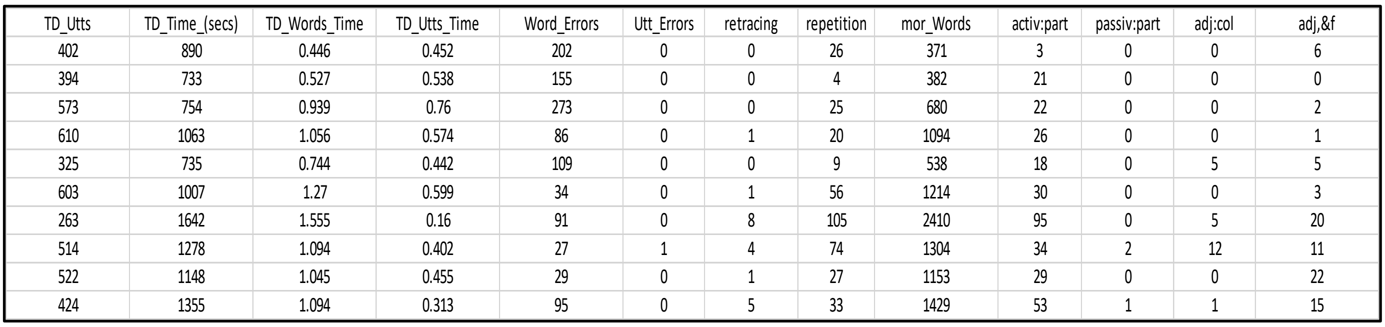
File	Language	Corpus	Code	Age(Month)	Sex	Role	Total_Utts	MLU_Utts	MLU_Words	MLU_Morphemes	MLU100_Utts
0data\Flopater.cha	ara	sample	CHI	1;07.02	male	Target_Child	302	299	1.231	1.572	100
0data\Yara.cha	ara	sample	CHI	1;09.20	female	Target_Child	283	278	1.345	1.665	100
0data\Basmala.cha	ara	sample	CHI	2;02.18	female	Target_Child	511	505	1.329	1.697	100
0data\Bilal.cha	ara	sample	CHI	2;04.19	male	Target_Child	589	585	1.858	2.6	100
0data\Razan.cha	ara	sample	CHI	2;10.00	female	Target_Child	292	289	1.848	2.63	100
ta\AbdrahmanFawzy	ara	sample	CHI	3;00.00	male	Target_Child	590	582	2.021	2.789	100
data\ZiyadYasser.ch	ara	sample	CHI	3;05.09	male	Target_Child	255	245	9.286	13.976	100
0data\Farah.cha	ara	sample	CHI	3;05.20	female	Target_Child	504	499	2.575	4.026	100
ta\ZiyadMohammed	ara	sample	CHI	3;07.12	male	Target_Child	498	491	2.257	3.444	100
0data\Merna.cha	ara	sample	CHI	3;08.01	female	Target_Child	404	393	3.399	5.036	100

Fig. 2. *KIDEVAL output for total utterances, mean length of utterance MLU, MLU for words, and MLU for first 100 utterances.*



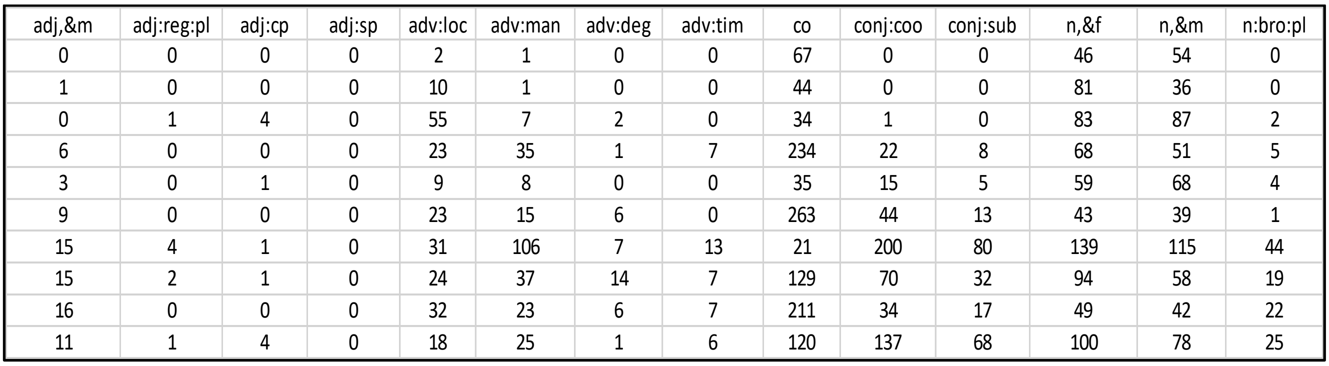
MLU100_Words	MLU100_Morphemes	FREQ_types	FREQ_tokens	FREQ_TTR	NDW_100	CD_D_optimum_aver	Verbs_Utt	TD_Words	TD_Utts
1.49	2.07	132	371	0.356	48	42.98	0.126	397	402
1.34	1.66	91	382	0.238	34	29.17	0.131	386	394
1.26	1.61	187	680	0.275	49	66.68	0.088	708	573
1.85	2.62	263	1094	0.24	43	35.55	0.114	1122	610
1.78	2.42	219	538	0.407	50	73.78	0.089	547	325
1.9	2.77	254	1214	0.209	46	26.59	0.083	1279	603
11.91	18.02	521	2410	0.216	58	96.54	1.263	2554	263
3.56	5.4	366	1304	0.281	57	92.86	0.228	1397	514
2.65	4.1	322	1153	0.279	60	53.95	0.223	1199	522
3.83	5.51	383	1429	0.268	48	71.76	0.344	1482	424

Fig. 3. *KIDEVAL output for mean length of utterance for 100 words and morphemes, frequency of types, tokens, type token ratio, number of different words, vocabulary diversity, number of verbs used in a single utterance, the total number of words spoken, and total number of utterances for each speaker.The ara.cut file inside the KIDEVAL folder*



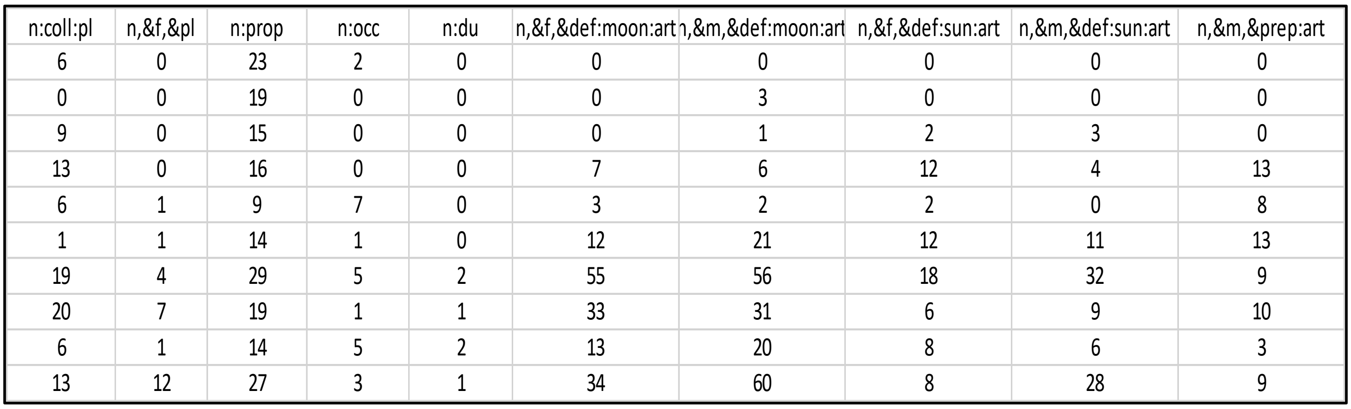
TD_Utts	TD_Time_(secs)	TD_Words_Time	TD_Utts_Time	Word_Errors	Utt_Errors	retracing	repetition	mor_Words	activ:part	passiv:part	adj:col	adj,&f
402	890	0.446	0.452	202	0	0	26	371	3	0	0	6
394	733	0.527	0.538	155	0	0	4	382	21	0	0	0
573	754	0.939	0.76	273	0	0	25	680	22	0	0	2
610	1063	1.056	0.574	86	0	1	20	1094	26	0	0	1
325	735	0.744	0.442	109	0	0	9	538	18	0	5	5
603	1007	1.27	0.599	34	0	1	56	1214	30	0	0	3
263	1642	1.555	0.16	91	0	8	105	2410	95	0	5	20
514	1278	1.094	0.402	27	1	4	74	1304	34	2	12	11
522	1148	1.045	0.455	29	0	1	27	1153	29	0	0	22
424	1355	1.094	0.313	95	0	5	33	1429	53	1	1	15

Fig. 4. *KIDEVAL output for total duration in seconds of utterances, words per second, utterances per second word errors, utterances errors, retracing, repetition, Active participle passive participle, and adjectives.*



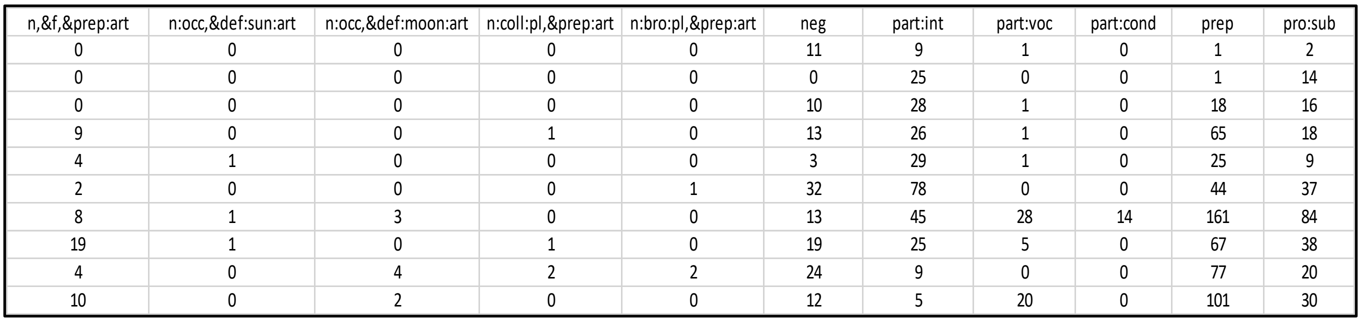
adj,&m	adj:reg:pl	adj:cp	adj:sp	adv:loc	adv:man	adv:deg	adv:tim	co	conj:coo	conj:sub	n,&f	n,&m	n:bro:pl
0	0	0	0	2	1	0	0	67	0	0	46	54	0
1	0	0	0	10	1	0	0	44	0	0	81	36	0
0	1	4	0	55	7	2	0	34	1	0	83	87	2
6	0	0	0	23	35	1	7	234	22	8	68	51	5
3	0	1	0	9	8	0	0	35	15	5	59	68	4
9	0	0	0	23	15	6	0	263	44	13	43	39	1
15	4	1	0	31	106	7	13	21	200	80	139	115	44
15	2	1	0	24	37	14	7	129	70	32	94	58	19
16	0	0	0	32	23	6	7	211	34	17	49	42	22
11	1	4	0	18	25	1	6	120	137	68	100	78	25

Fig. 5. *KIDEVAL output for masculine, regular plural. comparative, superlative adjectives, adverbs of location, manner, degree, time adverbs, communicators, conjunction coordination, conjunction subordination, feminine noun, masculine noun, broken plural, and collective nouns.*



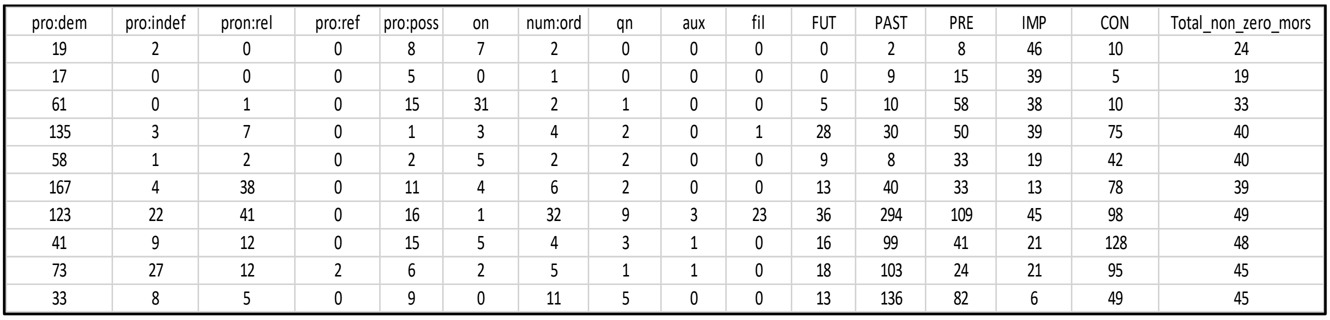
n:coll:pl	n,&f,&pl	n:prop	n:occ	n:du	n,&f,&def:moon:art	n,&m,&def:moon:art	n,&f,&def:sun:art	n,&m,&def:sun:art	n,&m,&prep:art
6	0	23	2	0	0	0	0	0	0
0	0	19	0	0	0	3	0	0	0
9	0	15	0	0	0	1	2	3	0
13	0	16	0	0	7	6	12	4	13
6	1	9	7	0	3	2	2	0	8
1	1	14	1	0	12	21	12	11	13
19	4	29	5	2	55	56	18	32	9
20	7	19	1	1	33	31	6	9	10
6	1	14	5	2	13	20	8	6	3
13	12	27	3	1	34	60	8	28	9

Fig. 6. *KIDEVAL output for noun feminine plural, proper noun. occupation noun, dual noun, noun feminine after definite moon letter, noun masculine after definite moon letter, noun feminine after definite sun letter, noun masculine after definite moon letter, noun masculine after preposition, and noun feminine after preposition.*



n,&f,&prep:art	n:occ,&def:sun:art	n:occ,&def:moon:art	n:coll:pl,&prep:art	n:bro:pl,&prep:art	neg	part:int	part:voc	part:cond	prep	pro:sub
0	0	0	0	0	11	9	1	0	1	2
0	0	0	0	0	0	25	0	0	1	14
0	0	0	0	0	10	28	1	0	18	16
9	0	0	1	0	13	26	1	0	65	18
4	1	0	0	0	3	29	1	0	25	9
2	0	0	0	1	32	78	0	0	44	37
8	1	3	0	0	13	45	28	14	161	84
19	1	0	1	0	19	25	5	0	67	38
4	0	4	2	2	24	9	0	0	77	20
10	0	2	0	0	12	5	20	0	101	30

Fig. 7. *KIDEVAL output for noun occupation after definite sun letter, noun occupation after definite moon letter, collective plural after preposition, broken plural after prep, negation, part interrogatives, part vocative, part, condition, prepositions, pronoun subject, and pronoun demonstrative.*



pro:dem	pro:indef	pron:rel	pro:ref	pro:poss	on	num:ord	qn	aux	fil	FUT	PAST	PRE	IMP	CON	Total_non_zero_mors
19	2	0	0	8	7	2	0	0	0	0	2	8	46	10	24
17	0	0	0	5	0	1	0	0	0	0	9	15	39	5	19
61	0	1	0	15	31	2	1	0	0	5	10	58	38	10	33
135	3	7	0	1	3	4	2	0	1	28	30	50	39	75	40
58	1	2	0	2	5	2	2	0	0	9	8	33	19	42	40
167	4	38	0	11	4	6	2	0	0	13	40	33	13	78	39
123	22	41	0	16	1	32	9	3	23	36	294	109	45	98	49
41	9	12	0	15	5	4	3	1	0	16	99	41	21	128	48
73	27	12	2	6	2	5	1	1	0	18	103	24	21	95	45
33	8	5	0	9	0	11	5	0	0	13	136	82	6	49	45

Fig. 8. *KIDEVAL output for pronoun indefinite, pronoun relative, pronoun reflexive, pronoun possession, onomatopoeia, number ordinal, quantifier, auxiliary, fillers, future, past, and present continues verbs.*

A. The result

The morphosyntactic analysis demonstrates not only the use of an automatic tagger tool to explore Egyptian children's language, but also has the capacity to contribute to developmental norms for Egyptian children's language. The usage of the KIDEVAL tool is efficient because it reduces the time needed to label the corpus manually and consider the linguistic components morphology and syntax more appropriately [6].

The quantitative language analysis profile produced by the KIDEVAL program can reveal extensive and valuable analysis, displaying the number of occurrences of each part of speech for each child depends on his age. The KIDEVAL analysis output for Egyptian Arabic children's corpus includes, 54 categories and subcategories. The frequency pattern of 19 parts of speech categories such as verbs, nouns, adjectives, adverbs, prepositions, negation, communicators, conjunctions, pronouns, active participles, passive participles, number ordinals, particle interrogatives, onomatopoeias, particle vocatives, particle conditions, fillers, quantifiers, and auxiliaries [5].

IV. Conclusion and Future Plan

The morphological analysis of Egyptian children can be used to build a foundation for future research into the developmental stages of Egyptian Arabic (and Arabic in general). It is important to note that more Egyptian Arabic corpora are needed in order to gain a better understanding of the hierarchical order of morphosyntactic acquisition in Egyptian Arabic. Creating a corpus for children's language is a collaborative effort rather than an individual one. The transcribing procedure took approximately 100 hours, and required 120 hours of manual disambiguation each childe's file. This is considered a huge project that needs the collaboration of different national institutions as well as institutional and government funding and support. This research project is an important step towards establishing robust developmental stages in Egyptian Arabic language acquisition. We hope that the work presented here sets the foundation for further development in this important field of child language corpus and become one of the basic steps in building a set of valid developmental measures for EA children's language that will be complemented by more research in the near future.

Acknowledgment

I would like to declare that Samantha Ghali, M.H.S, CCC-SLP, provided support and suggestions in completing this research.

References

[1] MacWhinney, B. (2000). The CHILDES database: Tools for analyzing talk, Vol 2: The database.

[2] MacWhinney, B. (2018). Tools for analyzing talk part 2: The CLAN program. Mahwah, NJ: Lawrence Erlbaum Associates. https://doi.org/10.21415/T5G10R

[3] https://childes.talkbank.org/access/Other/Arabic/Salama.html

[4] MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format. Carnegie Mellon University.

[5] Salama, H., & Alansary, S. (2016). Building a POS-Annotated Corpus for Egyptian Children. The Egyptian Journal of Language Engineering, 3(1), 12-23.

[6] Pezold, M. J., Imgrund, C. M., & Storkel, H. L. (2020). Using computer programs for language sample analysis. Language, Speech, and Hearing Services in Schools, 51(1), 103-114.