A Morphological Analyzed Corpus for Egyptian Child Language

Heba Salama, Sameh Alansary

Phonetics and linguistics Department, Faculty of Arts Alexandria University
Heba.salama.slp@gmail.com

Phonetics and linguistics Department, Faculty of Arts Alexandria University
Sameh.Alansary@bibalex.org

Abstract: ***The main focus of the present paper is on the construction of a MOR grammar for Egyptian children. We also introduce a morphological analyzer that was specifically developed for Egyptian children corpus. Morphological analysis is so crucial for child language studies. The corpus will be an instrument for future investigations of language development, child language acquisition theory and psycholinguistics in general. Egyptian children hand tagging of child transcript is time-consuming and error-prone, the MOR program for automatic analysis and part of speech tagging was introduced into the CHILDES system to address these problems. To date, MOR analysis programs have been constructed for 11 languages. Once a child language corpus has been automatically tagged by MOR, it is then possible to automate various systems for assessment and diagnosis of child language. The annotated corpora will support three different kinds of investigation: 1) Basic child language developmental research that examines the sequence of acquisition of grammatical morphemes and the various morphosyntactic processes of the language , 2) Computational modeling of child language acquisition , and 3) Diagnosis of language differences and disorders through methods such as automation of IPSYN or DSS scores. The corpus data contained 26,700 words, all the data are transcribed in CHAT. The paper discusses methods for preparing data for MOR analysis and developing MOR grammars. Describe the shape of rules for controlling Allomorphy and morpheme concatenation. The emergence of this new line of work represents a major step forward for Egyptian child language research***.

Key words: ***Morphosyntax, part of speech tagging, child language, MOR grammar***

1 INTRODUCTION

The current paper shows the construction of a MOR system for Egyptian child language, a part-of-speech (POS) tagging system for Egyptian Arabic children. The morphological (MOR) tagging relies on the application of MOR program for automatic analysis and part of speech tagging in CHILDES database. The morphological tagging of corpora is important not only for research on morphosyntactic development but also for the development of automatic ways of evaluating children's level of grammatical development. Egyptian Arabic is rich of allomorphic patterns and complex morphology. Therefore, it demonstrates a particular interesting challenge to any system for automatic morphological analysis. Explaining child language acquisition is one of the most challenges facing cognitive science, linguistics and psycholinguistics. Acquisition of grammar is expressed through the measurement of morphosyntactic competence which is important in fields such as developmental language disorders, schooling and literacy, and second language acquisition. To test and validate theoretical predictions quantitatively, researchers have increasingly come to rely on large corpora of transcript data of verbal interactions between children and parents to examine the development of morphosyntax. CHILDES database [1] is a standard source of child language corpus data to investigate the development of morphosyntax.

To date, MOR analysis programs have been constructed for 11 languages that do not include Arabic; the Arabic language do not have yet a MOR lexicons. Thus, one must create a system of part of speech using the automatic tools provided by CLAN program [2]. The CLAN software includes a language for expressing morphological grammars, implemented as a system, MOR, for the construction of morphological analyzers. There is a great need to develop such an automatic tools as a solution for the huge time required for hand annotation for child language transcript it took about 170 hours and it is a very time-consuming task. The Arabic language is characterized by much inflection; therefore, this system presents a challenge to any system for automatic MOR analysis. Part-of-speech (POS) tagging is a core natural language processing task that can benefit a wide range of processing applications. Tagging a child language has various outcomes to acquisition researchers for exploring instances, finding frequencies of particular constructions and searching specific usage. This tool serves language researchers by providing them with an automatic interface that enables quick and accurate analysis of child's language. The current paper focuses on the previously transcribed data on CHAT format from 10 children between the ages of 1;6 and 4;00 [3], cross sectional corpus. Methods for preparing data for MOR analysis, the shape of rules for controlling allomorphs, and morpheme concatenation to develop MOR grammars are generated. The project has several contributions: it presents a morphologically annotated corpus of Egyptian Arabic, with a set of tools that can be applicable to new corpora. Moreover, this is a research tool for future investigations of new perspectives

in Egyptian Arabic language research, which will influence some areas such as first language acquisition, second language learning, and speech disorders.
We discuss the MORphological device and describe the shape of rules for controlling Allomorphy and morpheme concatenation. Method for lexicon building; ongoing effort on morphological disambiguation is the topic of Section 3. Finally, Section 4 outlines the ways in which the annotated database is already being used, and concludes with future work.

2 MOR DEVICE

We developed a morphological analyzer for the Egyptian child language in complying with MORphological grammars [4] for 11 languages. The CLAN software includes a language for expressing morphological grammars, implemented as MOR program by applying a MOR grammar to a corpus, a new tier below each main tier is created, %mor tier. The morphological information for each item in the main tier is listed gives the surface representation of concatenated linear morphemes (stems + affixes) and lexical information attributed to the surface token, as shown in Fig. 1.

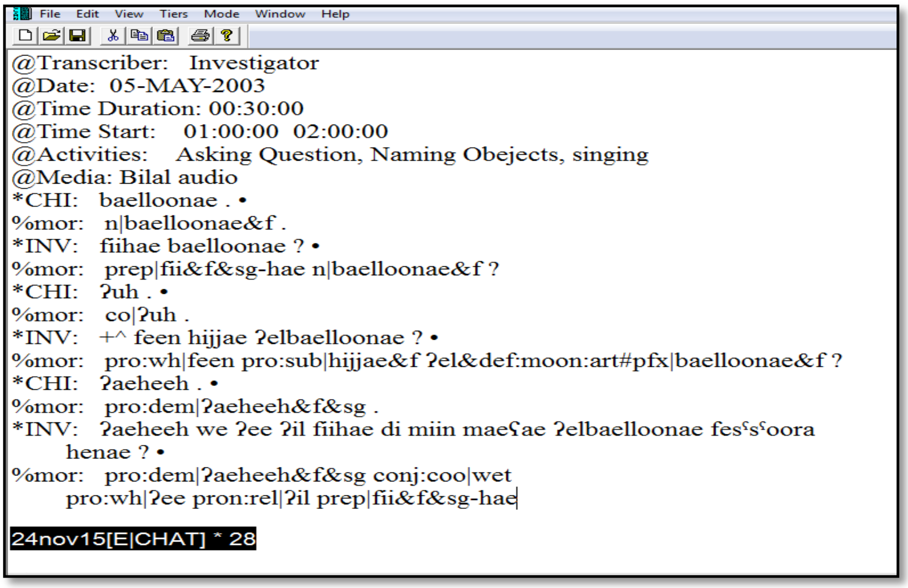

Figure 1: Word length distribution in Arabic and English

The Egyptian analyzer (EgyMor) consist of three major components: first an **a**(llophone)-rules file where it is specifies various forms a root or stem can take, second a **c**(oncatenation)-rules file which consists of higher level rules, allowing concatenation of the different categories and finally a set of lexicon files is organized by grammatical category (e.g., adverbs, function words, nouns, pronouns and verbs) and divided into several files that contain roots, stems, or whole words will described in the following sections. The use of MOR device varies in Different languages according to their requirements and needs. The Dutch grammar incorporate a short list of possible allomorphic changes and several concatenation rules; the grammar for Italian includes a more detailed A-rules file and a relatively short C-rules file; the grammars for Cantonese and Mandarin rely on lexicon files only; while Hebrew grammar as well as Egyptian Arabic extensively uses all of these devices (bases and affixes) because of the changes of the stem allomorphs is handled within a set of A-Rules; and affixation possibilities are allowed (or restricted) via the C-rules [5].

A. Lexicon

The current lexicon includes over 4,000 entries, distributed across the different parts of speech as listed in table1. Creating a lexicon is an open-ended task; we focused on adding the entries that occur in the corpora we had, and will continue to extend the lexicon as needed, given more corpora. Lexical entries are very simple. Words are entered into text files one word on each line in alphabetical order. The surface form comes first on the line, followed by the scat or syntactic category, some possible morphological features. For example, ʔutˤtˤa {[scat n][gen f]}"cat". The most complex entries are usually for verbs, as in this example, "ʔaeael" "said"{[scat v][ptn a][tense PAST][pers 3][num sg][gen m]}.

TABLE I
NUMBER OF LEXICON ENTRIES PER PART OF SPEECH

POS	Number of entries
NOUN	1730
Verb	1550
Adjective	500
Adverb	125
Preposition	25
pronoun	78
Total	4,008

B. A-rule

The first step in building MOR grammars is to determine the basic allomorphy types of Egyptian Arabic. Egyptian Arabic is characterized by rich and complex morphology (inflectional and derivational), therefore several systematic text books of inflectional and derivational patterns used. We follow the same in languages with more complex paradigm such as Spanish and Hebrew, allomorphy types is created based on the formal segments of the nominal and verbal paradigm. For example, the stems can target past, imperative. The construction of morphological analyses depends first on the generation of a runtime lexicon combined through the operation of rules of allomorphy (arules), as they operate on the items listed in the lexicon. The function of the arules is to increase the entries in the disk lexicon into a larger number of entries in the on-line lexicon. Subsequently Words that undergo regular phonological or orthographic changes when combined with an affix just need one disk lexicon entry. The arules are used to create on-line lexicon entries for all inflectional variants. These variants are called allos. The full set of generated allomorphs is stored in a trie structure [6]. Example of the allomorphy rule for noun suffix as the following:

```
RULENAME: n-sfx
LEX-ENTRY:
LEXSURF= $X$B
LEXCAT = [scat nsfx]
ALLO:
ALLOSURF = LEXSURF
ALLOCAT = LEXCAT, ADD [allo n0]
```

An arule consists of a header statement, which contains the rule name, followed by one or more condition-action clauses. Each clause has a series of zero or more conditions on the input, and one or more sets of actions. LEXSURF matches the surface form of the word in the lexical entry to an abstract pattern. The string XB is composed of variable declarations that characterize the noun suffixes to produce form as beet-u "his home" where X and B are a variable declaration indicate any string of consonant or vowels. Next line in the rule is ALLOSURF which used to produce an output surface. For example, a lexical entry surface form such as "beet" "Home" is converted to "beetu" "his home"to serve as the stem of the noun. where ALLOCAT determines the category of the output allos. From the above example, The first allo(morph) produced by the rule is "beet" "home" will produce words like "beetu" "his home or "beetnae" "our home". The use of this lock-and-key allomorphy pattern matching mechanism is controlled by the Crules. The application of Arules must be ordered from specific to general rule pattern to control the lock-and-key matching system for feature-value pairs through careful documentation and control of the [allo] features and other grammatical feature-value pairs such as [gen].for example, n|beet&m&sg-u . 347 Arule is applied in Egyptian Arabic is not too complex where in English with a simple morphology is 391 lines and Spanish is 3172 [7]. After finishing building arule it should end with a default rule to copy over all remaining lexical entries that have not yet been matched by some rule. This default rule must have this shape.

```
% default rule- copy input to output
RULENAME: default
LEX-ENTRY:
ALLO:
```

C. C-rule

The purpose of the crules is to allow stems to combine with affixes. In these rules, sets of conditions and actions are grouped together into if then clauses, performed by two input START and NEXT. The word matches are determined by the START rules that only require that a morpheme match the syntactic category [scat] of the rule. After the first rule match creates a candidate for the first few letters, MOR continues to take in letters looking for another morpheme match. Once a new morpheme fires, there can be lock-and-key process in which the STARTCAT and the NEXTCAT must match in terms of their allo features which called the MATCHCAT process. Crules depend on direction from

RULEPACKAGES statement. START and END rule should be for each part of speech. Here is an example of crule analyzed plural noun

RULENAME: n-start
CTYPE: START
if
NEXTCAT = [scat nsfx]
then
RESULTCAT = NEXTCAT
RULEPACKAGE = {n-reg-plural}

In the above example The STARTCAT is not defined, The NEXTSURF is the noun suffix that is attached to that stem then RESULTCAT = NEXTCAT lead all category information from the start input to be copied over to the result. The RULEPACKAGE identifies which rules may be applied to rule, when that result is the input to another rule. RULEPACKAGE are not tried until after another morpheme has been found. For example, in parsing the input “baellonaeaet” "ballons" , the parser first finds the morpheme "ballonae" "ballon" and applies the start rules thus the rule for noun will be fired. This rule includes a RULEPACKAGES statement specifying that the rule which handles noun suffix may later be fired. When the parser has further identified the morpheme "aet" the verb conjugation rule will apply, where "ballonae" is the start input, and "aet" is the next input. Applying the rule the analyzer associate this analysis n|baellonae&f&pl-aet.this analysis is read as the follows N is the main category (noun) and f&pl indicate feminine plural The crules on the contrary, of arule which are strictly ordered from top to bottom, crule is determined by their CTYPE and the way in which the RULEPACKAGES statement track words from one rule to the next. 456 crule lines are used in Egyptian grammar. Generally, not huge set of rules is required for producing a relatively extensive number of outputs (above 26,500 items are analyzed by the current version of the analyzer). Currently, 95.4% of all adult word tokens and 87% of all child tokens are analyzed. Out of these analyzed tokens, approximately 29.4% of all adult forms and 15.4% of all child forms are still ambiguous.

3 LEXICON BUILDING

The first step in building a grammar is to take a tour of the analysis of Egyptian morphological system [8][9] to have a good understanding of affixes, the classes of the stem Allomorphy variation and the conditions of choosing allomorphs. After that, we create a small lexicon of the most frequent words and choose to begin with adjectives because it is better to relay on one part-of-speech at a time and contains no morpheme. Then we move to another speech parts. Therefore, a related language rules from another languages with complex morphology like Hebrew and English rules files is studied well. Consequently, we create Arules and crules files for Egyptian language. We move from build up a lexicon of uninflected stems to inflected stem so affix.cut file is created which include categories and allos for the affixes it matched with the stem during crule fire. For example, f {[scat npfx] [pcat OR n pfx]} "f|prep" in this example the preposition attach to a noun as a prefix. In the process of building to check up the application of ar.cut and cr.cut we type mor +xi*.cha in the interactive mode to analyze the word in MOR, if it is analyzed the rules is ok if not we changed the rules. The unrecognized words identified by running the command mor +xl*.cha. Example of MOR analysis as follows in Fig. 2 and Fig. 3.

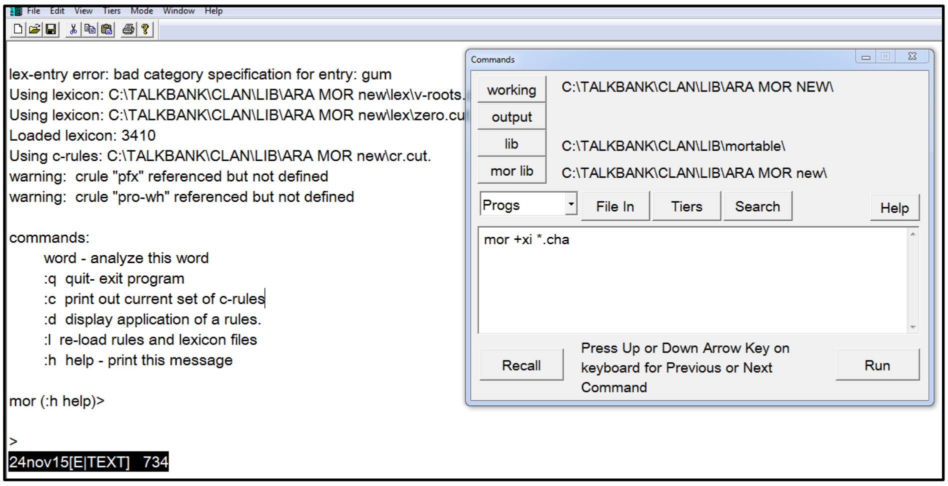

Figure 1: MOR command in interactive mode

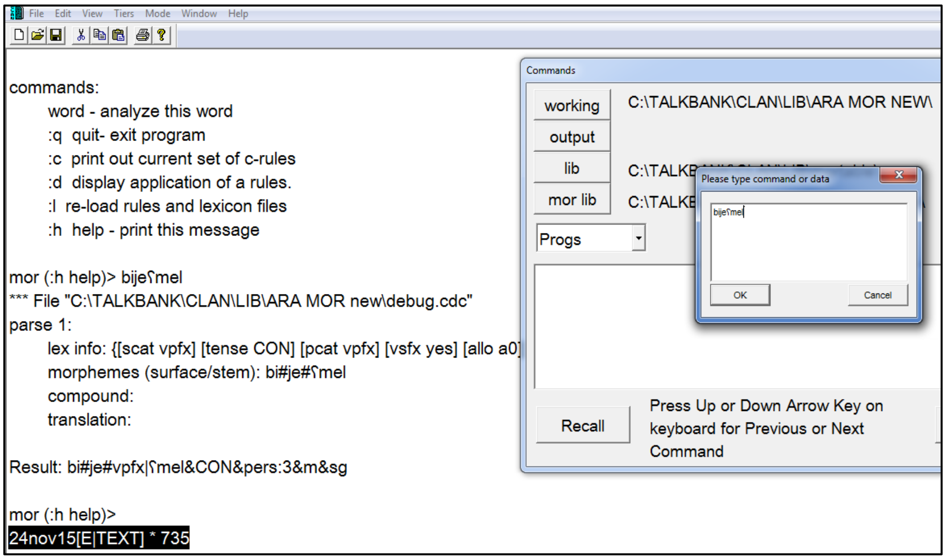

Figure 3: MOR analysis

There are 55 separate files for each of the 55 part of speech that break out the possible words of Egyptian Arabic (and additionally, 190 affixes), distributed across the different parts of speech. After all rules is built mor*.cha command is running to insert morphological analysis layer in child transcript. The output of the MOR program is a new tier or line called the %mor line in which tags stand in one-to-one correspondence with words on the main transcript line. These files are not disambiguated. Words can receive as many as 2 different analyses or more, all concatenated with the caret (^) symbol. For example, analysis output for "ʕaemaelt "I did" is v|ʕaemael&PAST&pers:2&m&sg-t^v|ʕaemael&PAST&pers:1&sg-t "maeʕae" "with" prep|maeʕae^mae#neg:bound|ʕa part of one word is part from another.

To achieve disambiguation of such combinations, the CLAN program offers two solutions 1) manual disambiguation tool called Disambiguator Mode in mode menu within CLAN. In this mode each ambiguous interpretation on %mor tier is produced in its alternative possibilities. The user choose the correct option this is called manual disambiguation. This very time consuming process that involves many online decisions 2) Automatic disambiguation POST and POSTTRAIN programs where these automatic module requires a lot of setting such as database of disambiguation rules called post.db file that we will consider in the future work. We notice that ambiguity of our data is low for example in one file o child transcript there are 41 ambiguous words from 2620 words. Ambiguity arises especially with items that have the same orthographic like first and second person in past "kaetaebt" "wrote". A lot of morphological analysis such as frequencies of morphological categories and mean length of utterance can be performed which is valuable and import to study child language acquisition.

4 DISCUSSION

This paper describes the construction and usage of first computational systems for morphosyntactic analysis for Egyptian child language. The MOR grammar approximately covers our current corpus data. The main advantage of our building system is it can analyze words contain 8 morphemes. For example, the result of analysis for "maebijeaebulinaeʃ" "they did not give us" is mae#bi#je#vpfx|aeb&CON&pers:3&m&sg-u-linae-ʃ , 7 morphemes, Where the result of the word "maebijiktibuhaelhaeʃ" " they did not wrote to her" is 8 morphemes. mae#bi#ji#vpfx|ktib&CON&pers:3&m&sg-uhae-lha-eʃ . We face a lot of challenges:

1- Short studies deal with the description of Egyptian Arabic where this work should be guided by a good descriptive grammar of the morphology of the language therefore, we try to collect possible description from different books.
2- Error and replacement in child speech
3- Ambiguity solving

Future plans

Several issues still need to be resolved with respect to our Egyptian child corpus. First, the way to handle disambiguation is to train a part-of-speech (POS) tagger using automatic tools provided by CLAN. Such a tagger will be able to automatically select the most suitable analysis for every multiple-choice output for a given lexeme and thus save the time. We are currently preparing training material for this task, and will use the POST program that is part of CLAN since it has worked to markedly reduce ambiguity for languages such as English, Spanish, and Chinese. The output of the POS tagger will serve as a basis for developing a statistical parser for Egyptian data of the kind that was developed for the English section of CHILDES and other languages[10].

REFERENCES

[1]MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.
[2] (Sagae, K., Lavie, A., & MacWhinney, B. (2005, June). Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 197-204). Association for Computational Linguistics.)
[3] Salama, H., & Alansary, S (2015). Building a POS-Annotated Corpus for Egyptian Children. *The Fifteenth Conference on Language Engineering (ESOLEC'2015)* (Doctoral dissertation, Ain Shams University Cairo, Egypt).
[4]Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, third edition, 2000.
[5]Brian MacWhinney. Enriching CHILDES for morphosyntactic analysis. In Heike Behrens, editor, *Corpora in Language Acquisition Research: History, methods, perspectives*, volume 6 of *Trends in Language Acquisition Research*. Benjamins, Amsterdam, 2008.
[6]Fredkin, E. (1960). Trie memory. Communications of the ACM, 3, 490-499.
[7]Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. Journal of Child Language, 37, 705-729
[8] Abdel-Massih, E. T. (2011). An Introduction to Egyptian Arabic.
[9]Omar, M. K. (2017). *The acquisition of Egyptian Arabic as a native language* (Vol. 160). Walter de Gruyter GmbH & Co KG.
[10]Bracha Nir, Brian MacWhinney, and Shuly Wintner. A morphologically analyzed CHILDES corpus of Hebrew. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1487–1490. European Language Resources Association (ELRA), May 2010. ISBN 2-9517408-6-7.

BIOGRAPHY

Heba Salama has a master's degree in corpus linguistics from the faculty of Arts phonetics and linguistics department Alexandria University 2015. A PhD student is building a morphologically analysed corpus for Egyptian children. She is interested in child language research. Her main interest is to collect corpus data to study child language development. She is searching for standard criteria to collect and transcribe data. She likes corpus linguistic field because it is more methodology that is powerful, scientific and open objective verification of results. The construction of database is very important in helping the researcher to manage the problem they faced and wishes to test a detailed theoretical prediction on naturalistic samples.

بناء مدونة لغوية محللة صرفيا في عربية الاطفال المصريين

هبه سلامة - سامح الانصاري
كلية الاداب- قسم الصوتيات واللغويات- جامعة اسكندرية

تهدف الدراسة الي بناء مدونة محللة صرفيا للأطفال المصريين،حيث أن التحليل المورفولوجي مهم للغاية في دراسة لغة الأطفال وستكون هذه الاداة في عمل أبحاث مستقبليةعن نظرية اكتساب الطفل للغة وعلم اللغة النفسي. يستغرق عمل المحلل الصرفي الكثيرمن الوقت والجهد والأخطاء لذلك فان انشاء محلل صرفي أليMOR الذي سوف يساعد في التغلب علي هذه المشاكل. حتي الان تم بناء برنامج MOR ل 11 لغة لا تتضمن العربية المصرية.ويجب انشاء محلل صرفي باستخدام الادوات الموجودة ببرنامج. بمجرد أن يتم عمل محلل صرفي تقوم العديد من البرامج التحليلية التي تعمل علي تقييم وتحليل لغة الأطفال.تساعد المدونات المحللة صرفيا 3 تحليلات اساسية :1) تحليل تطور النحو الصرفي للاطفال.2) النمذجة الحاسوبية لإكتساب الطفل للغة. 3) تشخيص الإختلافات والإضطرابات اللغوية من خلال أدوات برنامج CLAN. تحتوي المدونة علي 26,700 كلمة تم إعدادها بإستخدام CHAT. وتتناول الدراسة طرق إعداد المحلل الصرفي والقواعد المستخدمة ويمثل ظهور هذا الاتجاه من البرنامج خطوة كبيرة للأمام في بحث اللغة المصرية للأطفال.