Building a POS-Annotated Corpus For Egyptian Children

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*—In this paper, we present an attempt at developing a POS annotated corpus for Egyptian children.Linguistic annotation of the corpora provides researchers with better means for exploring the development of grammatical constructions and their usage.This is an initial annotated corpus for Egyptian children. It implements part of speech tag (POS) especially a morphologically annotated corpus of spoken Arabic child language.POS are made in "%mor" 'morphology' tiers manually. Coding language transcripts for computer analysis is a daunting task. It approximately took 170 hours, and thus manual annotation focused on a particular child.The POS coding process started with a purely manually annotation of 2701words. 1380 words annotated for an adultand 1321 annotated words for the child was handled. Annotated child language proved to be challenging, and time consuming task.The MOR grammar exists in many languages, such as English, French, German, Japanese, Cantonese, Hebrew, and they are generated automatically, the CLAN has the automatic coding system "MOR program". In Egyptian Arabic, this is not applied for two reasons. First, there is no previous Egyptian Arabic work done on a constructing system for such a representation. Second, morphology of Egyptian Arabic is very rich and different from other languages. Thus, their rules cannot be applied to Arabic. In the two Arabic studies of Qatari and Emirati languages, semi-automatic and mini automatic MOR is used.Finally,certain applications of linguistic analysis commands are provided by using CLAN software. The analyses include frequency counts, word searches, co-occurrence analyses; MLU (mean length of utterance) counts and analyzes specified pairs of utterances. Transcript data provide some morphological analysis, such as mean length of utterance (MLU) counts, lexical analysis, such as frequency (FREQ) count, syntactic analysis, such as searching the data for specified combinations of words or complex string patterns (COMBO) count, as well as the discourse and interactional analysis, such as analyzes specified pairs of utterances (CHIP) count.**

Key words:POS annotated corpus, CHILDES database.

1 INTRODUCTION

A part-of-speech tagging is usually called (POS) tagging, or simply tagging, but is also known as grammatical tagging or morphosyntactic annotation [1] takes place at word level and adds morphosyntactic information next to each word in the corpus. The information added makes the grammatical category to which each word belongs explicit, by adding codes such as: adjective, comparative; noun, countable, singular; verb, simple present, third person. It increases specificity of data retrieval from a corpus, and helps in syntactic parsing, and semantic field annotation. It allows us to distinguish between the homographs.The aim of a Part of speech annotation is to assign each lexical unit in the text a code indicating its part-of-speech. Different tagsets may distinguish a different number of categories, and consequently include a different number of tags, and they may use very different codes for the same categories. POS-tagged corpora allow corpus linguists to perform advanced searches in the corpus.

Corpus annotation has become a major effort in recent years, both for linguistic research and for natural language processing applications. Linguistic annotation of the corpora provides researchers with better means for exploring the development of grammatical constructions and their usage. The main advantage of the use of a standard representation of morphosyntactic coding enable is to test the impact of universality in the development of grammatical marking and syntax in corpora from different languages. Conventions and procedures described in the present research are based on the CHAT conventions of CHILDES system. The CHAT conventions have been modified to achieve a targeted coding scheme for the Egyptian Arabic, based on the classification of [2]. The coding scheme focuses on the development of grammatical marking and syntax. This required the use of a standard representation of morphosyntactic coding.

2 PART OF SPEECH CODES

The codes for grammatical categories were from the CHAT, but with some adaptation to suit the Arabic language. More subcategories were added in Arabic that not found in English. The morphological codes on the "% mor" line begin with a part-of-speech code. The basic scheme for the part-of-speech code is a category: subcategory: subcategory. The colon character is used as the field separator. The subcategory fields contain information about syntactic features of the word that is not marked. For example, /ʔækɪl/ "ate" is a past verb and there is no single morpheme signaling past, so the part-

of-speech code is **v: past**. Information that is marked by a prefix or suffix is not incorporated into the part-of-speech code. The information is found in the right of the | delimiter.

A. Stems

The codes for the stem are found on the right hand side of the | delimiter, following any pre-clitics or prefixes. Every word on the "% mor" tier must include a "lemma" or stem as a part of the morpheme analysis. A single form is selected for each stem. Thus, the Arabic definite article is coded as **det|ʔel** with the lemma /ʔel-/ whether the actual form of the article is /ʔel-/ or /ʔe-/ if /l/ is omitted from the moon letter.

B. Affixes

The codes for affixes and clitics are in the position in which they occur in relation to the stem. CHAT conventions are used to encode the morphological structure of word forms. For example, the delimiters (-) are used for a suffix, e.g., n|qe͟s͟a͟s͟s-BROK&PL, the symbol (&) is employed to indicate inherent features (like the gender of nouns), and morphemes that are not separable. The (&) is used to mark affixes that are not realized in a clearly isolable phonological shape. For example, the form /tuffæːħ/ "apples" cannot be broken down into a part corresponding to the stem /tuffæːħ/ "apples" and a part corresponding to the plural marker. For this reason, the word is coded as n|tuffæːħ&PL. Several codes indicated with the & after the stem e.g., the form /ʔækɪl/ "ate" is coded v|ʔækɪl&PAST&1s.

3 EGYPTIAN ARABIC PARTS OF SPEECH

Languages vary considerably in morphological complexity. English, for example, has a simple morphology compared with languages, such as Arabic and Hebrew [3]. Arabic is a language of rich morphology compared to other language especially European languages. It is based on both derivational and inflectional morphology. The richness of Arabic morphology makes the analysis process difficult to deal with. On the one hand, the morphological analysis process is used in the most of the NLP (natural language processing) applications, such as information retrieval, spell checking, and machine translation. In general, morphological analysis of any word given consists of determining the values of a large number of features, such as a basic part of speech (i.e., noun, verb), gender, person, number, voice information about the clitics[1][4].

The grammar of Arabic is standardized for centuries. An initial tagset was derived from this grammatical tradition rather than from an Indo-European based tagset. Morphological tag cannot do successfully using methods developed for English because of data sparseness. Indeed, Egyptian Arabic is a very different language from Indo-European languages and should have its own tagset. In addition, Arabic linguists are based focus their studies on a traditional Arabic grammar rather than on Indo-European grammar. Arabic grammarians traditionally analyze all Arabic words into three main parts-of-speech. However, according to the present study parts-of-speech are categorized into more detailed ones, which collectively cover the whole of the Egyptian Arabic language [5]. The three main parts of-speech are:

A. Noun

A noun in Arabic is a name or a word that describes a person, thing, or idea. Traditionally the Noun class in Arabic is sub-divided into Derivatives (that is, nouns derived from verbs, nouns derived from other nouns, and nouns derived from particles) and Primitives (nouns not so derived). These nouns are sub-categorized by number, gender, and case. This class also includes what, in traditional European grammatical theory, is classified as participles, pronouns, relatives, demonstratives, and interrogatives.

B. Verb

The verb classification in Arabic is similar to that in English, although the tenses and aspects are different. The tag for the verb is sub-categorized into perfect, imperfect, and imperative. Further, sub-categorization of the verb class is possible using number, person, and gender.

C. Particle

The Particle class includes Prepositions, adverbs, conjunctions, interrogative particles, negative particle, quantifiers, communicators, determiners, and fillers.

Sometimes, it is difficult to decide to which part of speech a word belongs. Parts of speech should be clearly clarified, and the possible description of Egyptian Arabic is reviewed, as there is no previous work for part of speech in Egyptian Arabic. Thus, this applied to the possible literature dealing with more examples of Egyptian Arabic word classes to

[1]A clitic: is a morpheme that has syntactic characteristics of a word, but shows evidence of being phonologically bound to another word. For example, in Arabic the definite article, equivalent to “the” in English, appears as a two-letter proclitic at the beginning of the noun.

enable us tag words. The researcher reviewed a lot of description for Egyptian Arabic words in [6], [7], [8], [9], [10], [11], [12], and [13] as well as the whole description of Egyptian Arabic and the classification and examples of words.

4 INSIGHTS INTO EGYPTION ARABIC MORPHOLOGICAL PARADIGMS

Arabic is the most widespread member of the Semitic group of languages. The Arabic language is the most complicated and richest language. This section presents an overview of the Egyptian colloquial Arabic morphological paradigms used in POS annotated data. The following sections present the morphological paradigms of Egyptian Arabic.

A. Noun

Arabic nouns are classified according to gender and number. Arabic nouns have two genders (masculine-feminine). Gender in Arabic is animatenouns, such as those referring to people, usually have the grammatical gender corresponding to their natural gender, but for inanimate nouns the grammatical gender is largely arbitrary. Most feminine nouns end in / -a/such as cities, counties, and certain body parts. Nouns that do not fit in any of these categories are masculine.

[11] Classifies noun in Arabic into three categories: singular, dual, and plural. Singular noun is a base form, which dual or plural affixes are added to it. A dual noun is created by adding the suffix /-en/ to the stem or by adding number two before a noun. Plural nouns are sub-categorized into regular and irregular forms. Regular plurals are suffixes, /-in/ for masculine, such as /mudærris**i:n**/ 'teachers'/ and /-at/ for feminine, such as / ħæjæwæn**æ:t**/ 'animals'. Some nouns have both counted plural, such as /beḏa:ṯ/ 'eggs' and collective plural such as /be:ḏ/ 'eggs'. Irregular plural "broken plural" is predicted in some nouns, such as /ko:ra/ 'ball' , /kowwar/ 'balls', and in other nouns is unpredicted, such as /ra:ǧel/ 'man', /riggæ:læ/ 'men'. When the noun is counted except for the dual form, the cardinal number precedes the noun in the noun phrase. Numerals 3to 10 have two forms, long and short. The long form ends in /-a/ such as /**tælætæ** kilo/ 'three kilos'. The short forms end without /-a/ such as, /**tælæt**tuffæħæ:t/ 'three apples'. Numerals 11and above consist of a base which is an allomorph of numerals 1 and 2 and the suffix /-aʃar/ such as /ʔetn**aʃar**/ 'twelve'. Ordinal numbers tell the order of things in a set: first, second, third, such as /ʔettæ:ni / 'the second'.
Another type of nouns is a noun possessive. It is expressed by the word /bitæ:ʕ/ masculine 'belong', /bitæ:ʕæ/ feminine 'her', and /bitu:ʕ/ plural 'their'. It is the most common alternative to construct a phrase and indicate possession between two nouns such as /ʔekkitæ:bbitæ:ʕʔelbent/ 'the girl's book'. It is also used next to the suffix pronouns such as /bitæ:ʕu/ /ʔelʔælæmbitæ:ʕu/ 'the boy's pen', /bitæ:ʕhæ/ /ʔelʔælæmbitæ:ʕhæ/ 'the girl's pen'.

A proper noun is the special word or name that we use for a person, place, or country. A proper noun has two distinctive features: 1) it names a specific item, and 2) it begins with a capital letter. Nouns are tagged with n for common nouns, and **n:prop**for proper nouns (names of people, places, fictional characters, brand-name products). For example, **n:masc:sg|ra:gl** 'man'.

1) Occupational Nouns:

The feminine of the most occupations is formed by adding /–a/ such as /mudærres/ 'male teacher', /mudærres**æ**/ 'female teacher'. Occupational nouns are tagged **n:occu|mudærres** 'teacher'.

2) Place and Time Nouns:

Place and time nouns express the place or time of a verbal action or state. They are formed by prefixing /ma-/. For example, /**ma**ṯbax/ 'kitchen' (from /ṯabaxa/ 'to cook'), /mustæʃfæ/ 'hospital' (from the verb /istæʃfæ/ 'to cure'). Place and time nouns are tagged **n:plac|mustæʃfæ**'hospital'.

3) Instrumental Nouns:

Instrumental nouns express the instrument by which the action is performed. They are prefixed with /mi-/ and formed only by verb form I, according to the following pattern. For example, /muftæ:ħ/ 'key' from /fætæħ/ 'to open'. Instrumental nouns are tagged **n:inst|muftæ:ħ** 'key'.

B. Adjective

An adjective is a word that describes a noun. Adjectives are inflected for gender (masculine-feminine) and number (singular-plural). The masculine singular form of the adjective is the base form and is the stem to which feminine and plural affixes added as mentioned in [11]. The suffix /-æ/ is added to the stem to form a feminine adjective. Adjectives are also inflected for plural by adding /-in/ /ṣuɣajjar**i:n**/ 'small'. The adjective is inflected for comparative by adding /ʔæ-/ such as /**ʔa**kbar/ 'older', and inflected for superlative as well by adding /ʔel-/ such as /**ʔil**akbar/ 'the oldest'. Adjectives follow the noun they modify and agree with singular nouns in gender and number. An adjective is tagged **adj:masc:sg|kibi:r**'old'**.**

C. Determiner

Determiners include definite and indefinite articles. The definite article in Egyptian Arabic is /ʔel-/. It expresses the definite state of a noun of any gender and number. Definite article /ʔel-/ assimilated to a number of consonants, so the article in pronunciation is expressed only by geminating the initial consonant of the noun [8]. The gemination is expressed by putting /ʃæddæ/ on the following letters /t/, /ɵ/, /d/, /ð/, /r/, /z/, /s/, /ʃ/, /s̲/, /d̲/, /t̲/, /z̲/, /l/, /n/. The 14 letters are called "sun letters" while the remaining 14 are called "moon letters". Determiners are tagged for definitearticle **def:art:moonL|ʔel~n:sg:fem|ħæflæ**'the party'and for indefinite article**def:art:sunL|ʔeʃ~n:sg:masc|ʃæ:rʕ** 'the street'.

D. pronouns

1) Personal subject-independent pronoun:

Personal pronouns in Egyptian Arabic have singular and plural, the second and third persons differentiate gender, while the first person does not. Personal pronouns are not needed with verbs, as it is clear from the verb, but it is common to use them, especially for emphasis. They are often used with participlesas stated in [7].Personal pronouns are tagged **pron:subj:sg|ʔænæ**'I'.

2) Possessive Objective Dependent Pronoun

Dependent personal pronouns in Egyptian Arabic are affixed to various parts of speech, with varying meanings. Egyptian Arabic object pronouns are clitics. They attach to the end of a noun, verb, or preposition, with the result forming a single phonological word rather than separate words. Personal pronouns are affixed to various parts of speech, with various meanings: Dependent personal pronouns are affixed to nouns, where they have the meaning of possessive demonstratives, e.g. /be:t**i**/ 'my house' , /be:tɪ**k**/ 'your house' , /be:t**u**/ 'his house'. They are affixed to verbs, where they have the meaning of direct object pronouns, e.g. /–ni/ 'me' / ʃu:fte**ni**/ ' saw me', /–k/ 'you' /ʃu:fte**k** 'saw you', /-hum/ 'them' /ʃu:ftu**hum**/ 'saw them'. With verbs, indirect object clitic pronouns are formed using the preposition /li-/ plus a clitic. Both direct and indirect object clitic pronouns can be attached to a single verb: /ʔægi:b/ 'I bring', /ʔægíbu/ 'I bring it', /ʔægibhu:lik/ 'I bring it to you', /mægibhulki:ʃ/ 'he did not bring it to you'. They are also affixed to prepositions, where they have the meaning of objects of the prepositions, e.g. /ʕændi/ 'to me', /ʕændek /'to you', /ʕændu/ 'to him'. Dependent personal pronouns are tagged **pron:dep|hæ**. Example of possessive/objective-dependent pronoun is n:sg:masc|be:t~**pron:dep|hæ** 'her house'.

3) Pronouns with Suffixed Prepositions

A suffix pronoun is attached to prepositions, such as /fi/ 'in', /li-/ 'to', min/ 'from', /mæʕæ/ 'with', /ʕælæ/ 'on'. Pronouns with suffixed preposition are tagged **Prep|fi~Pro|hæ**. Examples of pronoun with suffixed preposition is**perp|li~pron:masc:2s|k**'for you'.

4) Demonstrative Pronouns

Demonstrative pronouns point to and identify a noun or a pronoun. Demonstrative pronouns are /dæ/ 'this, that', /di/ 'this, that ', and /do:l/ these, those'. They occur after the noun as demonstrative adjectives or before the noun as demonstrative pronouns. Another words also classified with demonstratives are /ʔæhu/ 'here is, there is', /ʔæhe:h/ 'here is, there is' , and /ʔæhum/ 'here are, there are' for dual and/or plural. They follow or precede the noun or occur in isolation. Demonstrative pronoun is tagged **dem|ʔæhu**'here'.

5) Indefinite Pronouns

In Egyptian Arabic indefinite pronouns are words like /ʔæjħædd/ 'anybody', /ħæ:gæ/ 'something'. In Egyptian, these made up of two words, but they used in exactly the same way as in English. Indefinite pronouns are tagged **Pron:indep|ħæ:gæ**'something'.

6) Relative Pronoun

The Egyptian Arabic has only one relative pronoun /ʔilli/ to represent 'that, who, and which'. There is only one relative pronoun used in reference to all nouns, regardless of gender/number. The relative pronoun is tagged **pron:rel|ʔilli** 'who'.

7) Interrogative pronouns

Egyptian Arabic pronouns indicate questions are /ʔe:hdæ?/ 'What is this?', /mi:n/ ' who', /ʔezzæj/ 'how'. Interrogative pronouns are tagged **pro:wh|ʔe:h?**'What'.

8) Reflexive Pronouns

The noun "næfs" is used as a reflexive pronoun followed by a suffix pronoun to mean that a person does an action by "himself".Egyptian Arabic reflexive pronouns are /næfsi/ 'myself ', /næfsæk/ 'yourself ', /næfsu/ 'himself'. Reflexive pronouns are used after a noun or a verb. Reflexive pronouns are tagged **Pron:ref|benefsu**'himself'.

E. Verb Tenses

[5] Classifies Egyptian Arabic into two basic tenses in Arabic. The "perfect "refers to a finished action, corresponds to the English past tense. The "imperfect" refers to an incomplete action (on going or future) and corresponds to our present, progressive, and future tenses. The imperfect is usually preceded by /bi-/ to denote present continuous and by /ħæ-/to denote the future tense. The imperative is used to give instructions or orders. There are three forms: masculine, feminine and plural. Present tense is tagged **v:PRES|be~3S:g:masc|je~ʕæmæl-INF** 'he is make', past tense is **v:PAST|ʕæmæl-INF** 'made' tagged and future tense is tagged **v:FUT|ħæ~ʔæxæd-INF** 'I will take it' .

1) Voice participle

An Egyptian Arabic participle is derived from a verb, but is used like an adjective with the verbal meaning [8]. There are two types of participles: active and passive. Active voice is the "normal" way of using a verb; it has the form of an adjective or noun. Active participles act as adjectives, and so they must agree with their subject. There are three forms: masculine, feminine, and plural. Active participles are tagged **v:activ:partic|ʕæ:rf**. Passive participles, like active participles, act as adjectives or nouns, and so they must agree with the noun they're describing. Passive participles are tagged **v:pass:partic|ʔitkætæb** 'was written'and active participles**v:activ:partic|kæ:teb** 'writer'

F. Negation

Negation in Egyptian Arabic appears in the free particles, such as /meʃ, læʔʔ, læ/ or negation bound prefix /mæ-/ and the suffix /–iʃ/. Negation is used with a verb, pronoun, adjective, and participles [11]. Negation is tagged **neg|læʔʔ**.

G. *Communicators*

Communicators are used for interactive and communicative forms, which fulfill a variety of functions in speech and conversation. Many of these are formulaic expressions, such as ba:j 'bye', bravo, ʃokran 'thank you', ʔæhlæn 'welcome', sæ:læmoʕæleko 'hello'. Words used to express emotion, as well as imitative and onomatopoeic forms, such as "ʔah, boom, mhm, wow" are included in this category [13]. Communicators are tagged **co|ʔuh** 'yes'.*Conjunctions*

H. Conjunctions

Conjunctions in Egyptian Arabic are the useful little words that join clauses together to make sentences that are more complex. Conjunctions conjoin two or more words, phrases, or sentences. A coordinating conjunction is a particle, which connects two words, phrases, or clauses together [5]. The most common conjunction is the prefixed particle /wæ/ 'and ', /fæ/ 'and so'. A coordinating conjunctions are tagged **conj:coo|wæ**. Subordinating conjunctions introduce a subordinate clause. Most subordinating conjunctions are single words, such as //bæss/ 'that's it', but, /zæj/ 'like', /bæʕd/ 'after', /ʔizæ/ 'if', /ʕæʃæ:n/ 'because', /læmmæ/'when', /jebʔæ/ 'well then' , /wælæ/ 'or' , /bardu/ 'as well'. Subordinating conjunctions are tagged **conj:sub|ʕæʃæ:n**'because'.

I. Fillers

Fillers in Egyptian Arabic are a sound or word that is spoken in conversation by one participant to signal to others that he/she has paused to think, but has not yet finished speaking /jeʕni/ 'that means' and /wallahi/ 'A word used for swearing' are common fillers[7]. Fillers are tagged **fil|jeʕni**'that means'**.**

J. Quantifier

Quantifier in Egyptian Arabic is a word or phrase, which is used before a noun to indicate the amount or quantity. Quantifier is used with bothcountable and uncountable nouns, such as /kul/ 'all'[9]. Quantifier is **tagged qn|kul**'all'.

K. Vocative Particle

The vocative particle /jæ/ is followed by a noun or proper noun for both genders [9].The vocative particle is tagged **Part:voc|jæ**.

5 METHOD

Thesecond stage in building child corpus is POS codingprocess, which is the direct result of our previous transcription process[2]. POS are made in "%mor" 'morphology' tiers manually. We hand annotated one file for a child aged 3,5,20

[2]Salama. H., Alansary, S (2014). Building a spoken Arabic corpus for Egyptian children: data collection and transcription.In*Proceedings of the Conference of language engineering*,*3*(4).Egyptian Society of Language Engineering.

years.Itapproximately took 170 hours, and thus manual annotation focused on a particular child. Hand coding of a "%mor" tier for many children would require perhaps many years of work. The POS coding process started with a purely manually annotation of 2701words. 1380 annotated words for adult and 1321 annotated words for the child washandled. ThisinitialEgyptian Arabic annotated corpus was used to run CLAN program for morphological analysis. The total number of the tagsets used in the data is 92 tags. CHAT codes were used with some adapting to fit the classification of Egyptian Arabic language. The morphological features applied to classify the words of the data were 92tagsets.The POS annotated corpus and the project is available at [14]. Following,ananalysis of the transcript as the application of CLAN program is overviewed. The commands applied in the data and analysis results are presented as well in the next section. Transcribed file after annotation process is shown in Fig .1.

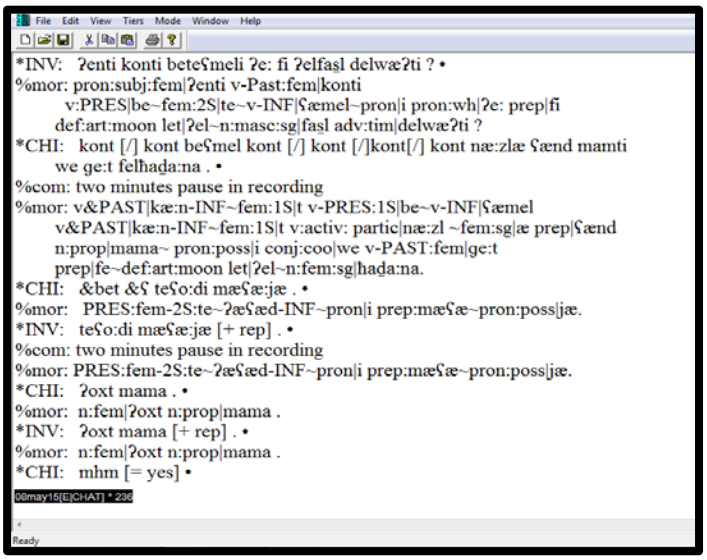

Figure 1: Transcribed File after Annotation Process.

6 SOMEFINDINGS from ANALYZING CHILD LANGUAGE TRANSCRPIT with CLAN PROGRAM

Analyzing child transcript is the final stage in building child corpus. Once a file is transcribed and annotated, the analytic work of CLAN is performed by a series of commands. These commands run from the Commands window, search for strings, and compute a variety of indices. CLAN allows the performance of a large number of analyses on transcript data; there are 29 programs inside the CLAN. The analyses include frequency counts, word searches, co-occurrence analyses, MLU counts, interaction analyses, and text changes. The CLAN programs are designed to support linguistic analysis [15]: morphological analysis, lexical analysis,syntactic analysis, discourse, and interactional analysis. The following lines review how these linguistic analyses perform in CLAN programs.

A. Morphological Analysis

Once a complete %mor tier is available, a vast range of morphological and syntactic analyses become possible.Many of the most important questions in child language require the detailed study of specific morphosyntactic features and constructions.

1) MLU

The MLU (mean length of utterance) is a command used primarily to determine the mean length of utterance of a specified speaker. It also provides the total number of utterances and of morphemes in a file. The ratio of morphemes over utterances (MLU) is derived from those two totals. [16] manifests the value of thinking of MLU in terms of morphemes, rather than words. Brown is interested in the ways in which the acquisition of grammatical morphemes reflects syntactic growth and he believes that MLU in morphemes would reflect this growth more accurately than MLU. The output of the command**mlu +t*CHI farah.cha** perform MLU analysis on the child's tier (+t*CHI) is shown in Fig.1.The MLU for investigator output is:The total number of utterances is 308 and morphemes are 2459 in a file. The ratio of morphemes over utterances (MLU) is 7.984.Where the MLU for child is:The total number of utterances is 58 and morphemes are 2374 in a file. The ratio of morphemes over utterances (MLU) is 40.931 as shown in Fig.2.

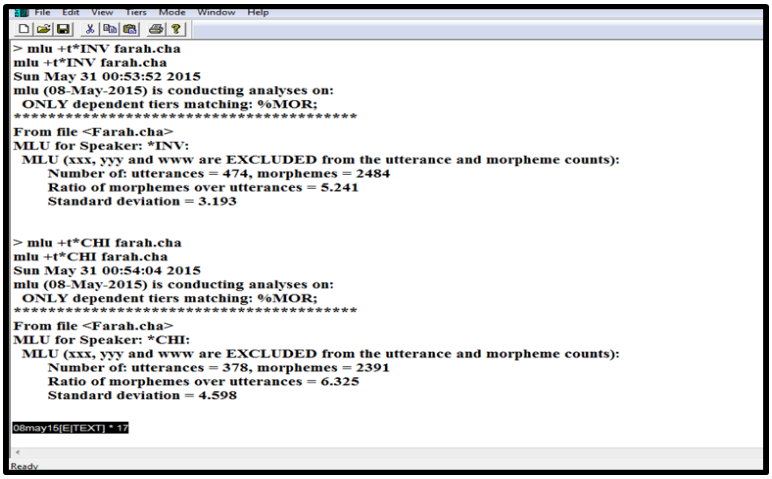

Figure 2: MLU analysis

B. Lexical Analysis

This is the easiest types of CLAN analyses, which look at the frequencies and distributions of particular word forms. The programs for lexical analysis like FREQ (frequency) and KWAL (Key Word And Line) focus on the ways of searching for particular strings. The strings to be located can be entered in a command. Many studies used these techniques to track the development of lexical fields, such as morality, kinship, gender terminology, mental states, causative verbs, and modal auxiliaries. It is also possible to track words of a given length or a given lexical frequency. An example for FREQ and KWAL is clear in the following sections.

1) FREQ:

The FREQ (frequency) command is powerful and quite flexible, permitting frequency analysis. FREQ counts the frequencies of words used in selected files. It also calculates the type–token ratio typically used as a measure of lexical diversity. It generates an alphabetical list of all the words used by all speakers in a transcript indicating frequency of each word form (morpheme) and frequency of grammatical categories.A frequency word count is the calculation of the number of times a word occurs in a file or a set of files. FREQ produces a list of all the words used in the file, along with their frequency counts, and calculates a type–token ratio. The type–token ratio found by calculating the total number of unique words used by a selected speaker (or speakers) and dividing that number by the total number of words used by the same speaker(s). It is generally used as a rough measure of lexical diversity.The output of the command **freq +t*CHI farah.cha**shows how many times a child used the word. In the last output, it is a total of 1321 words or tokens used with only five different word types. The type–token ratio is found by dividing the total of unique words by the total of words spoken. For example, the type–token ratio would be 544 divided by 1321 or a ratio of 0.412as shown in Fig. 3.

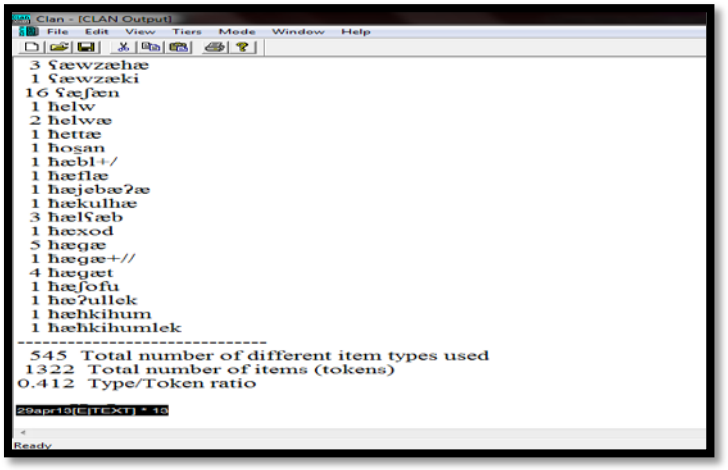

Figure 3: Frequency analysis

2) FREQPOS:

The FREQPOS (frequency position) program is a minor variant of freq. Freqpos is different in the fact that it allows us to track the frequencies of words in initial, final, and second position in an utterance. This is useful in studies of early child syntax. For example, using freqpos on the main line enables users to track the use of initial pronouns or auxiliaries. For

an open class, an item such as verbs, freqpos is useful in analyzing codes on the %mor line. For example, freqpos allows studying the appearance of verbs in second position; initial position, final position, and other positions. The frequency position command **freqpos +d farah.cha**is shown in Fig.3.

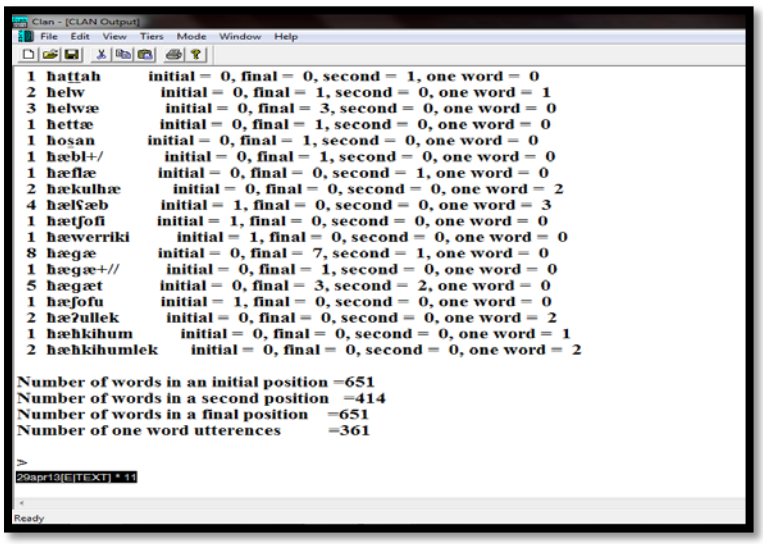

Figure 3: Freqpos analysis

3) KWAL:

KWAL is short for (Key Word And Line). It is the second major tool for conducting lexical analyses is the KWAL program. The analysis takes a word and finds the lines on which that word occurs in each transcript. This analysis is necessary to find out which lines the targets are on and in what position in the utterance each target is located. The outputs are not merely the frequencies of matching items, but also all the full context of the item. The KWAL command for the mother used the word /ʕæʃæ:n/ 'because'**kwal +sʕæʃæ:n -w2 +w2 farah.cha**is shown in Fig. 4. In this analysis, a mother used the word /ʕæʃæ:n/ 'because' nineteen times.

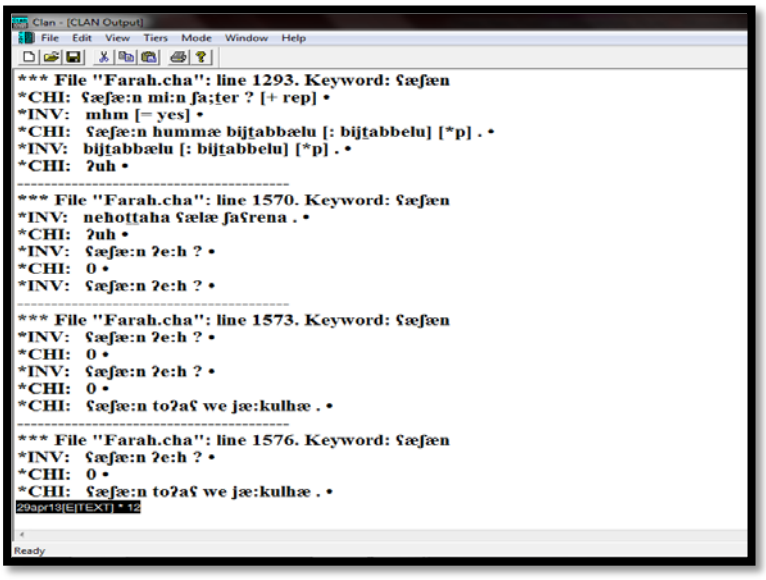

Figure 4: KWAL analysis

C. Syntactic Analysis

1) COMBO:

COMBO (combination) is a powerful program that searches the data for specified combinations of words or complex string patterns. For example, COMBO finds instances where a speaker says "beʕmelselsa:l" 'I am making dough' twice in

a row within a single utterance. The command **combo +tCHI +s"beʕmel ^s̲el̲s̲a:l" farah.cha** searches a child's tiers (+t*CHI) of the specified file 0042.cha as in Fig.5. The output in show that the combination "beʕmels̲el̲s̲a:l" 'I am making dough' is found once in the speaker's speech as in shown in Fig.5.

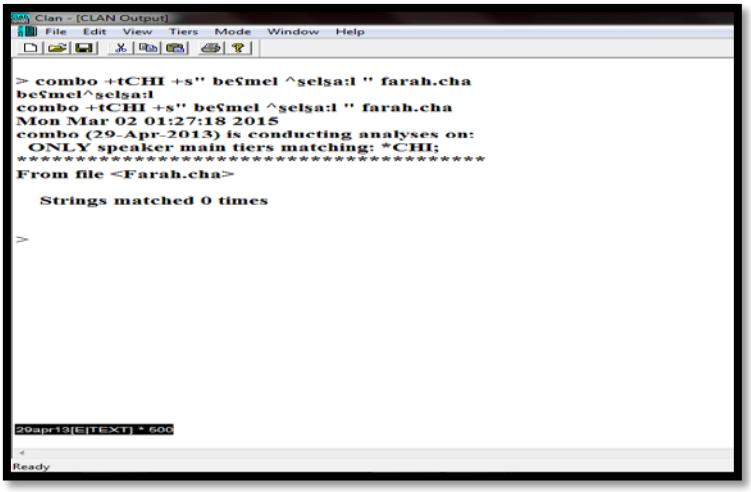

Figure 5: COMBO analysis

D. Discourse and Interactional Analysis

1) CHIP:

CHIP is useful for tracking the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. [17]Have used it successfully to demonstrate the availabilityof useful instructional feedback to a language-learning child. The program analyzes specified pairs of utterances. CHIP is used to explore parental input, the relation between speech acts and imitation, and individual differences in imitativeness in both normal and language-impaired children. CHIP compares two specified utterances and produces an analysis that then is inserted onto a new coding tier. The first utterance in the designated utterance pair is the "source" utterance and the second is the "response" utterance. The response compared to the source. An example of a minimal CHIP command **chip +bMOT +cCHIfarah.cha** is shown in Fig.6. The output in of the first ten lines shows that CHIP introduces % csr tier. This tier is an analysis of the child's self-repetitions expressed by the code $REP. Here the child is both the source and the response as shown in Fig.6.

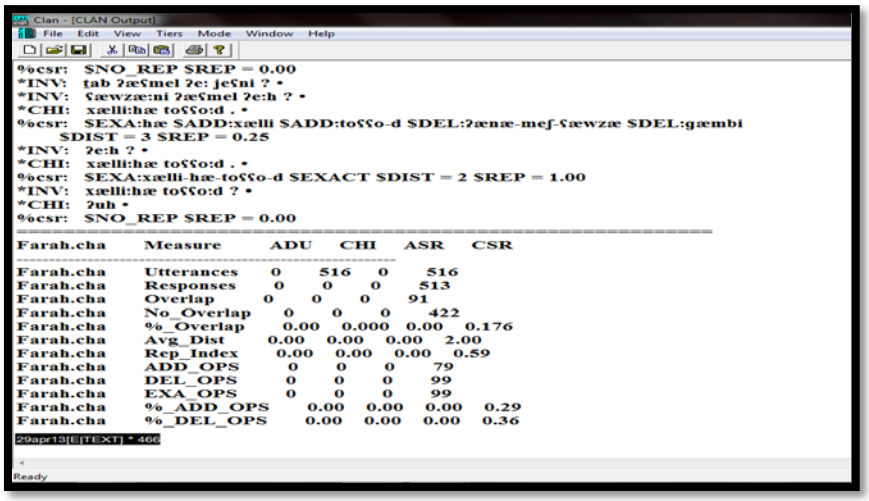

Figure 6: CHIP analysis

2) WDLEN:

The WDLEN (word length) program tabulates the lengths of words, utterances, and turns. The WDLEN program generates a histogram of maternal utterance lengths. It highlights the very high frequency of very short utterances that present language-learning children with either no or very few segmentation decisions in their efforts to locate words in the input. The command **wdlenfarah.cha** tabulates the lengths of words in child's tiers. The output shows that the investigator utterances consisted of zero single word as shown inFig.7. An additional 230 are two words long, and an additional 255 are three words long. Thus, 485 words of child directed utterances in this analysis consist of investigator turns.

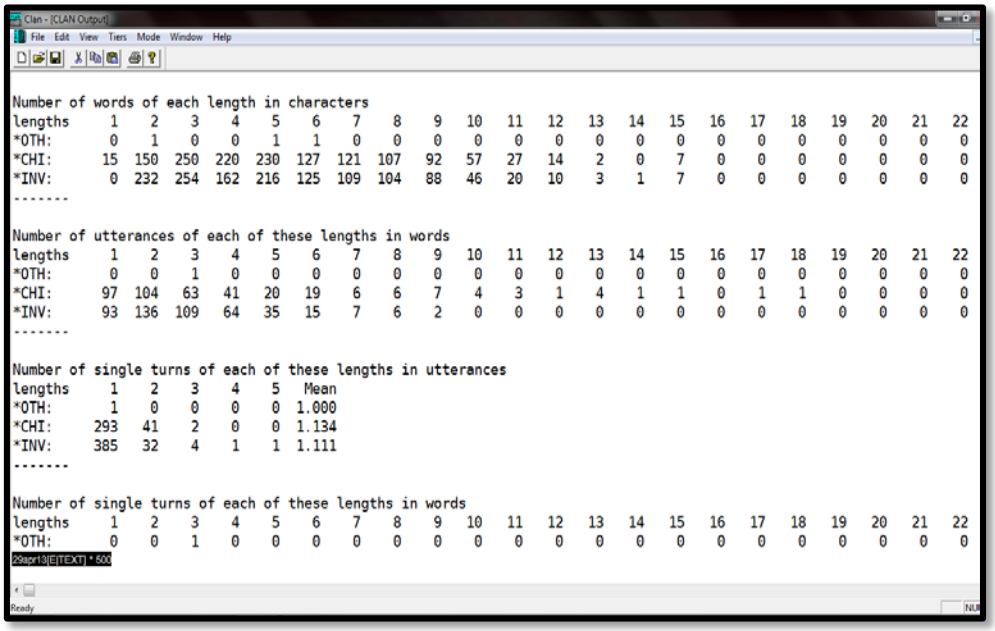

Figure 7: WDLEN analysis

7 CONCLUSIONS

We introduced POS coding and analysis by using CLAN program and CHAT format 18]. Linguistic analysis performed by using CLAN commands. Seven types of linguistic analysis were applied as an application for CLAN program. The outputs of lexical analysis, such as FREQ and KWAL help to look at the frequencies and distributions of particular word forms. The output of MLU in morphological analysis helps the researchers to investigate the grammatical development of children. The syntactic analysis, such as COMBO searches the data for specified combinations of words or character strings. Moreover, the discourse and interactional analysis, such as CHIP track the extent to which one speaker repeats, corrects, or expands upon the speech of the previous speaker. This corpus is a research tool for future investigations of Egyptian Arabic child and child-directed language, language development, language disorder, and psycholinguistics in general.

In recent years, Corpora are considered basic resources for language analysis and research. There was a major shift towards the empirical study of language rather than intuitive study. The technological advance of computers changes the area of language research. This change of trend is because of the introduction of computer and corpora in linguistic research, which, subsequently, illuminated numerous new applications of language and linguistics in the field of information exchange. Moreover, the empirical approach to language study is distinguished to be more dependable and authentic than rationalistic approach, which based on intuition. These corpora can be useful for producing many advanced automatic tools and systems, besides being good resources for language description and theory making. When child language is transcribed and compiled in a computerized database, it forms linguistic corpora. Corpora play an important role in child language research. The researchers of all theoretical persuasions make use of corpus data to investigate the development of children's linguistic knowledge. This is a high time to turn our attention towards using corpora for linguistic research. There are a lot of areas where corpora can lead to new perspectives in child language research, such as first language acquisition, second language learning, phonetic and prosodic analysis, and speech disorders.

REFERENCES

[1]McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*, Edinburgh: Edinburgh University Press.
[2]Mitchell, T. F. (1956). *An Introduction to Egyptian colloquial Arabic*. London: Oxford university press.
[3]Savoy, J. (1999). A stemming procedure and stop word list for general French corpora. *Journal of the American Society for Information Science*, 50, 944–952.
[4]Habash, N., &Rambow, O. (2005). Arabic tokenization, morphological analysis, and part-of-speech tagging in one fell swoop. In *Proceedings of the Conference of American Association for Computational Linguistics* (pp. 578-580).
[5]Haywood, J.A. and Nahmad, H.M. (1962). *A new Arabic grammar of the written language*. London: Lund Humphries.
[6]Hopkins, S. (1984). *Studies in the Grammar of Early Arabic: Based Upon Papyri Datable to Before 300 AH/912 AD*. Oxford University Press.
[7]Pipes, D. (1983). *An Arabist's Guide to Egyptian Colloquial*. Daniel Pipes.
[8]McGuirk, Russell H. (1986). *Colloquial Arabic of Egypt*.Psychology Press.
[9]Abu-Chacra, F. (2007). *Arabic: an essential grammar*. Routledge.
[10]McLoughlin, L. (2009). *Colloquial Arabic (Levantine)*.Routledge London and New York.
[11]Omar, M. (1970). The Acquisition of Egyptian Arabic as a Native Language, JanuaLinguarum, Series Practica, 160, The Hague : Mouton, 1973. (Formerly a Ph. D. dissertation at Georgetown University).
[12]Habash, N. Y. (2010). Introduction to Arabic natural language processing.*Synthesis Lectures on Human Language Technologies*, *3*(1), 1-187.
[13]Abdel-Massih, E. T. (2011). An Introduction to Egyptian Arabic.University of Michigan.
[14]project site www.eacc.ga/
[15]Crystal, D., Fletcher, P., & Garman, M. (1989). *The grammatical analysis of languagedisability*. Second Edition, London: Cole and Whurr.
[16] Brown, R. (1973). *A first language: the early stages.*Cambridge, Mass.: Harvard University Press.
[17]Sokolov, J. L., &Moreton, J. (1994). Individual differences in linguistic imitativeness.In J. Sokolov& C. Snow (Eds.), *Handbook of research in language development using CHILDES*. Hillsdale, NJ: Lawrence Erlbaum Associates.
[18]MacWhinney, B. (2012). The CHILDS project.Tool for analyzing talk Electronic Edition.part 2 : the CLAN programs. Carnegie Mellon university available on line at http://childs.psy.cmu.edu/manuals/clan.

BIOGRAPHY

HebaSalamahas a master's degree in corpus linguistics from the faculty of Arts phonetics and linguistics department Alexandria University 2015.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.Shelikescorpus linguistic because it ismore methodology that is powerful,scientific and open objective verification of results.Electronic corpora have advantages, which is unavailable to their paper based equivalents. The availability of data exchange allows the researcher to answer questions by looking for the transcript of spontaneous speech of many data, rather than single study. Sharing data make a revolution in the study of child language. She found that the most obvious advantage of using computer for language study is the speed of processing and the ease of data manipulation.E.g., searching, sorting, and formatting. Advances in computer technology enable to share child language data more readily.The 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.

Sameh Alansary is professor of computational linguistics in the Department of Phonetics and Linguistics and the head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars.
He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now.
Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland.

TRANSLATED ABSTRACT

بناء مدونة لغوية محللة علي مستوي اقسام الكلام للاطفال المصريين

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

تهدف الدراسة إلى عرض طريقة عنونه الكلمات وعرض تحليل للغة الطفل عن طريق برنامج CLAN . يعمل البحث على أقسام الكلام فيما يتعلق ببنية اللغة العربية المنطوقة لدى الأطفال. و إن الشرح اللغوى للمجموعات توفر للباحث وسائل أفضل للبحث فى التركيبات النحوية و استخدامها و تطويرها. يقوم البحث بعمل بعض التحليلات المرفولوجية مثل طول الجملة المنطوقة (MLU) و كذلك التحليلات اللفظية مثل عدد مرات التكرار (FREQ) والبحث عن كلمة معينة داخل السياق KWAL. هذا و إن بناء مدونة للأطفال قد ظهر مع وجود الثورة التكنولوجية و ثورة الحاسبات، و لقد قامت إثنتا و ثلاثون دولة حول العالم بعمل مدونة لغويه للأطفال تعتمد على قاعدة البيانات (CHILDES). أما بالنسبة للدول العربية، فلقد قامت كل من قطر و الإمارات بعمل مدونة خاصة بهما، و قامتا بعرض المدونتين على مواقع الإنترنت، إلا أن المدونة الخاصة بالعربية المصرية لم تكن متوفرة بعد. وقد قام هذا البحث بعمل اول مدونة لغوية عربية منطوقة للاطفال المصريين وعرضها علي الانترنت من اجل الاسهام في تبادل المعلومات بين الباحثين. كما يفيد ايضا فى مجال علم اللغة النفسى و البحث فى التركيبات النحوية و كذلك فى التحليل اللغوى. كما أن البحث التجريبي يمكن أن يعرفنا الكثير عن الاضطرابات اللغوية التي تحدث للاطفال ومن ثم سرعة اكتشافها وعلاجها مبكرا، كما أننا بحاجة إلى البحث فى كيفية تفاعل الطفل واستخدامه للغة فى المواقف العادية. فنحن بحاجة إلى ملاحظة و تسجيل و تحليل النماذج اللغوية التلقائية، إلا أن دراسة تلك النماذج التلقائية يتطلب وقت كبير فى جمع البيانات و الكتابة الصوتية و التحليل ، و من ثم فعمل مدونة مصرية للأطفال يسهل عملية تحليل كلام الاطفال و يساعد فى دراسة لغة الأطفال. و لقد أحدث مشروع نظام تبادل البيانات اللغوية للأطفال (CHILDES) تغيرات جذرية فى طرق البحث على العديد من المستويات (الصوتية و التركيبية و اللفظية)، فهذا المشروع مبادرة لجمع البيانات للكتابة الصوتية من مختلف الدراسات التى أجريت على لغة الأطفال وفقا لصيغة CHAT و باستخدام برنامج CLAN.