Lexical Growth in Egyptian Arabic Speaking Children: A corpus Based Study

Heba Salama[1], SamehAlansary[2]

Phonetics and linguistics Department, Faculty of Arts Alexandria University

[1]Heba.salama.slp@gmail.com

[2]Sameh.Alansary@bibalex.org

Abstract:***This paper calculates developmental index of language growth in Egyptian Arabic based on a corpus consists of spontaneous speech samples from10 children 5 boys and 5 girls from 1.7 to 4 years. Depending on 30 minutes transcripts of spontaneous speech production the following properties of collected data were analysed: size of vocabulary and frequency of word use in relation to age (development of types and tokens) and individual differences in vocabulary size. The contribution of the current study lies in the use of vocabulary profile results as a measure of potential indicators of developmental language delay. The results provide a new measurement tool for lexical growth at different developmental stages.***

Keywords: Lexical growth; developmental delay; vocabulary profile

1 INTRODUCTION

This paper is a corpus-based study rooted in the so-called CHILDES (Child Language Data Exchange System) [1]. The objective of the present study is to describe the size of vocabulary in relation to age (development of types and tokens), identify individual differences in vocabulary size. The acquisition of the lexicon is a central and complex component of child language development. It has received growing interest within psycholinguistic research and is becoming a field of study in its own right. At the same time, lexical development also interacts with acquisition processes in other linguistic domains. Consequently, early lexical abilities might be indicative of following language skills. It is a question of crucial importance, whether certain abilities belong to the normal range of individual variation, or whether these abilities should rather be considered as an indication of a delayed or disturbed language acquisition process. To answer this question, empirically based knowledge about the normal course of vocabulary development within a particular language has first to be established. In this paper, a cross sectional study on early lexical development in Egyptian Arabic children is presented to analyze vocabulary growth. The results of this study provide an important new measure to help in assessment of language delay of Arabic language for children.

2 LEXICAL DEVELOPMENT

A. Vocabulary Growth

Normal development for young children's vocabularies has a number of applications in research design, assessment, and intervention that is very difficult to obtain before. While English and most Indo-European languages have a long tradition of examining aspects of child language production by computing different developmental indices from spontaneous language samples while, these aspects are lacking in this valuable area of research in Egyptian Arabic language. This may be partially because of the lack of longitudinal corpora of Arabic child language, large corpora or the appropriate set-up for experimental studies. Bridging this gap, this paper provides the first systematic cross-sectional study of the validity of TTR (type token ratio) developmental index in Egyptian Arabic language to measure vocabulary growth.

Lexical development is usually measured on the number of new words entering a child's vocabulary as they acquire a language. Statistical information is usually computed from spontaneous language samples of children in conversation or narrating a story. One of the first measures used in this context is the type-token ratio (TTR) or the ratio of new words (types) over the total number of words (tokens) in a speech sample. The categorical terms "types" and "tokens" are two important concepts in lexical analysis. If a text is 100 words long, we say it has 100 tokens. However within that text, there may be many words that are repeated, and as a result, there could be only 30 different words in the text in which case, we say there are 30 word types. The type vs. token ratio (TTR) is an important criterion that linguists use to evaluate lexical diversity in lexical analysis[2], introduced the index to child corpora and found a consistent ratio of around one different word for every two words uttered, independently of variables such as age range and gender. According to [3], "Templin's data are applicable only with children between the ages of 3 and 8 years". However, later work has shown that TTR depends on the size of the input transcript. That is, language samples which contain larger numbers of tokens give lower values for TTR and vice versa. As the children start producing longer utterances and language samples, a greater part of their acquired lexicon emerges and, as a result, the number of available new word types that could potentially be introduced decreases. Type and token frequency data, a major variable in psycholinguistic research, can be derived from corpora only. Language researchers and applied linguists working on a wide range of topics frequently need indices that quantify the range or number of different words in a text or conversation. Such measures are variously conceptualized as reflecting for example, lexical range and balance [4], total vocabulary size [5]. from a more negative perspective, particularly in the investigation of language disorders, measures are often seen as an index of repetitiveness manifested as perseveration in dementia and schizophrenia; speech automatisms in aphasia; echolalia in autism and mental deficiency. [3] cited [6] and admit that the values compared when determining TTR, the total number of words in a specified language sample and the total number of different words in the same language sample, are most valuable for evaluating the appropriateness of the child's vocabulary development. According to [3], "reductions in the total number of different words and the total number of words have been implicated as potential indicators of developmental language delays or disorders.

In this study the quantitative measures are used rather than qualitative to give general insight into the number of words known, but do not distinguish them from one another based on their category or frequency in language use. They have developed to make up for the widely applied measure type-token-ratio (TTR).This paper will proceed as follows: we will review a related work in Section 3. Method and procedures will be explained in Section 4. Next, we will present the results in Section 5, discuss them, and conclude with a summary of the importance of our findings in Section 6 and 7.

3 RELATED WORK

The following section presents a brief outline of some relevant findings in vocabulary development of children literatures.[7], as cited in [3]reported that at 18 months, normally-developing children had the ability to produce 22 meaningful and different words[8]and[9], as cited in[3], argued that the normally-developing 18 month-old child could produce nearly 50 different words[10], as cited by [3] asserted that the mean age at which children typically were capable of producing 50 different words was 19.75 months[11], [10] as cited by [7], as cited by [3]reported that the typically-developing 21 month-old toddler produced roughly 118 different words. [12] found that a control group of children, who were of mean age of 23 months, produced an average of 189.5 words. According to [13], toddlers between the ages of 23-25 months received total vocabulary scores of 196.24 words[14], as cited by [3]reported that the expressive vocabulary size of typically developing 2 year-old children was, at least 150 words.[15]found that at 30 months, children produced an average of 264.50 words. [16], as cited by [3] conducted a study obtaining TTR values for children under 3 years of age; however she only reported the TTR values, and not the total number of words or total number of different words necessary to compute TTR. There are a number of references showing typical TTR values for children across early development. [2], as cited in[3] norms for children between the ages of 3; 0-3; 5 shows total average of words 204.9 and total average of different words 92.5 TTR 0.45 while the mean score of the data collected from children between the ages of 2; 4 and 2; 9 is 140 and total average of different words 62.2 with TTR 0.447. In Emirati Arabic[17] found that there is no correlation between age and TTR was found in six children.

4 METHOD

The pre-compiled corpora include cross-sectional studies investigating the speech of 10 children in certain activity contexts (e.g. toy-playing-asking question, describe pictures). Five boys and five girls were selected randomly with no language delay from a nursery in Alexandria ranged from 1.6 to 4 years with a mean age 2.77 transcribed in chat format [1]. The children were split into five age groups each group contain (one boy-one girl) as shown in Table1. All children were normal and their first language is Arabic. The corpora contain 25,645 transcribed words from all 10 transcribed chat files. The summary of statistics of the corpus data is shown in Table2.The command in CLAN program that was used in the current research is FREQ (frequency) command. This command stands for Frequency Analysis. It is powerful and quite flexible, permitting frequency analysis. FREQ counts the frequencies of words used in selected 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 development. The following command **freq +t*CHI farah.cha** looks specifically at a child's tier. The output printed in the CLAN window comes in alphabetical order. Using this command, researchers can count the number of words appearing in selected files; in addition, the ratio of different words (Types) to the total number of words (Tokens) Type-Token ratio (TTR) of words can be reported.

TABLE I
AGE RANGE GROUPS

No	Age Range	Number of Children
1	1.6 -2	One boy and one girl
2	2 – 2. 6	One boy and one girl
3	2.6 – 3	One boy and one girl
4	3 – 3.6	One boy and one girl
5	3.6 – 4	One boy and one girl
Total	10	
mean	2.77	

TABLE 2
THE SUMMARY OF STATISTICS OF THE CORPUS DATA

Age	Number of investigator items	Number of child items	Total
19	1070 Mot, Inv311	376	1765
21	1122	385	1507
26	1448 Mot, Inv 760	699	2914
28	1882	1109	2987
34	1351	627	1978
36	1949	1269	3226
41	657	2407	3059
42	1380	1323	2701
43	1686	1162	2844
44	1252	1414	2664
Total	14,868	10,777	25,645

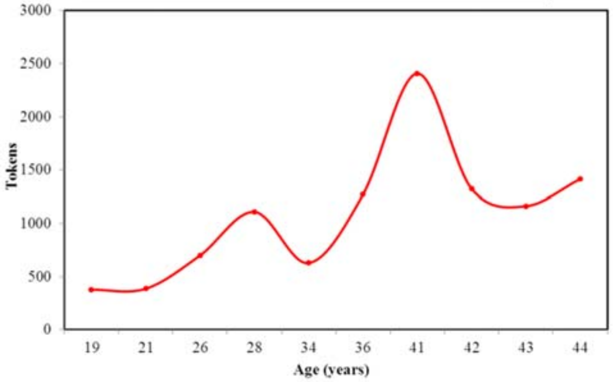

Figure 1: Growth patterns of age and word (tokens)

TABLE 3
FONT SIZES FOR PAPERS

Age	Types	Tokens	TTR
19	165	376	0.43
21	104	385	0.27
26	251	699	0.35
28	378	1109	0.34
34	333	627	0.53
36	373	1269	0.29
41	847	2407	0.35
42	547	1323	0.41
43	455	1162	0.39

44	546	1414	0.38

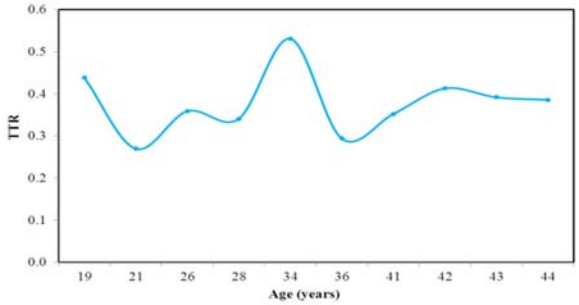

Figure 2: Growth patterns of age and TTR

5 DISCUSSION OF RESULTS

The results presented in Table 3 on spontaneous use of words in 10 Egyptian children contribute to show a picture of the process of lexical development in Egyptian children from the 19 months to four years. The results of vocabulary size in the groups can be illustrated by the language profiles of single cases. The rate at which children continue to accumulate new words in the second and third year of life differs individually. Looking at single cases as well as larger samples, varying growth patterns have been demonstrated. The vocabulary size increased from19 to 26 months gradually. A sudden acceleration in the rate of vocabulary growth has been reported around the age of 28 months (1109). This widespread phenomenon of a rapid and sudden growth is referred to as vocabulary spurt. The observed exponential increase in spontaneous word production in the 28 months, followed by a further deceleration, described as vocabulary growth within a 'region of acceleration' as [18] explained. Although the findings clearly support a general trend of a vocabulary spurt phase in the second year. The vocabularies continue increase from 36 to 41months and decrease from 42to43 months and continue to increase by the age of 44 months. The individual patterns are showing linear phases of vocabularies growth or a spurt in the third and half year. Supporting the findings of [19] has found that children vary a great deal in the rate at which they acquire vocabularies. Finally, the development of vocabularies in this research is independent of variables such as gender and parental influence.

By comparing our results with the previous work we found that 21 months produce 384 words while [12] found 23 months, produced an average of 189.5 words. The children between the ages of 21-26 months show mean vocabulary score 738 words where According to [13], toddlers between the ages of 23-25 months received total vocabulary scores of 196.24 words. The 26 months produce 706 words where [14], as cited by [3] reported that the expressive vocabulary size of typically developing 2 year-old children was, at least 150 words. The overall findings show the vocabulary increase gradually from 1,7 to 3,8 years. It appears that increase in early lexical abilities considered an indicator for later grammatical complexity. Accordingly, the study shows that lexical limitations successfully serve as a reliable, early predictor of potential language acquisition problems and sometimes, of severe and continuing disorders. This evaluation supports the results of other studies in which lexical development is taken to be a valid predictor of further language acquisition. A satisfactory level of lexical development is a prerequisite for grammatical development. The lexical limitation is an indicator of a problem in the other linguistic areas such as syntax and morphology.

The results of TTR counts of all children shows that TTR size of different files in the corpus is not constant as shown in Fig. 2. The correlation between tokens and TTR is non-linear. The vocabulary spurt affect TTR result and thus for each child, 19 months child with 318 words show a higher TTR count 0.43score where, 44 month child with 1323 words show TTR count 0.41score. There is a correlation between tokens and TTR, the large sample size give small TTR. For example, the 41 months with 2407 words give 0.35 score TTR and 44 months with 1414 words give 0.38 score.TTR is declined with increasing sample size. Therefore, any single value of TTR lacks reliability as it will depend on the length in words of the language sample used. A graph of TTR against tokens for a transcript will lie in a curve beginning at the point (1,1) and falling with a negative slope that becomes progressively less steep. (TTR) developmental index is not alone a valid measure for vocabulary growth in Egyptian Arabic language. Further measure such as VOCD (measurement of vocabulary diversity) should incorporate with TTR.

6 CONCLUSIONS

The obtained results in this research can form the base on which further research on figuring out the developmental stages of Egyptian Arabic can be based. This is an important research project because of the fundamental lack of work in this area of Arabic linguistics. It is obvious that further work needs to be done on Large-scale corpus-based to allow us shed light on the mechanisms of early lexical development. This is an important step towards establishing robust developmental stages in Egyptian Arabic language.

REFERENCES

[1] B. MacWhinney 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(2012).

[2] M. C.Templin Certain language skills in children. Minneapolis: University of Minnesota Press (1957).

[3] K. Retherford, Guide to analysis of language transcripts (3rd ed.). Eau Claire, WI: Thinking Publication (2000).

[4] D. Crystal, Profiling linguistic disability. London: Edward Arnol (1982).

[5] G. H., & Thompson, J. R. Thomson Outlines of a method for the quantitative analysis of writing vocabularies. British Journal of Psychology, 1904-1920, 8(1), 52-69. (1915).

[6] J. F. Miller, Assessing language production in children: experimental procedures. London: Edward Arnold (1981).

[7] P. S. Dale, Language Development: Structure and Function. New York: Holt Rinehart and Winston (1976).

[8] R.E. Owens, Language disorders: A functional approach to assessment and intervention. New York: Merrill/Macmillan (1991).

[9] H. Benedict, Early lexical development: Comprehension and production. Journal of child language, 6(02), 183-200. (1979).

[10] K. Nelson, Structure and strategy in learning to talk. Monographs of the society for research in child development, 1-135. (1973).

[11] M. E. Smith, An investigation of the development of the sentence and the extent of vocabulary in young children. University of Iowa Studies: Child Welfare. (1926).

[12] D., &Tobias, S.Thal, Relationships between language and gesture in normally developing and late-talking toddlers. Journal of Speech and Hearing Research, 37, 157-170 f (1994).

[13] L. Rescorla, N. B. Ratner, P. Jusczyk, & A. M. Jusczyk Concurrent Validity of the Language Development Survey: Associations with the MacArthur—Bates Communicative Development Inventories Words and Sentences. American Journal of Speech-Language Pathology, 14(2), 156-163, (2005).

[14] A. Mehrabian, The development and validation of measures of affiliative tendency and sensitivity to rejection. Educational and psychological measurement (1970).

[15] J. Heilmann, S. E. Weismer, J. Evans & C. Hollar, Utility of the MacArthur—Bates Communicative Development Inventory in Identifying Language Abilities of Late-Talking and Typically Developing Toddlers. American Journal of Speech-Language Pathology, 14(1), 40-51. (2005).

[16] J. R Phillips Syntax and vocabulary of mothers' speech to young children: age and sex comparisons. Child Development 44. 182. (1973).

[17] D. Ntelitheos & A. Idrissi. Language Growth in Child Emirati Arabic. In 29th Annual Symposium on Arabic Linguistics (pp. 9-11) (2015).

[18] E. Bates, P.S. Dale & D. Thal. Individual Differences and their Implications for Theories of Language Development. In P. Fletcher, & B. MacWhinney, (eds.), The Handbook of Child Language (pp. 96-152). Cambridge: Basil Blackwell. (1995).

[19] L. Fenson, E. Bates, P.S. Dale, S.J. Pethick, J.S. Reznick & D. Thal. Variability in Early Communicative Development. (Monographs of the society for research in child development, 59/4). Chicago: The University of Chicago Press. (1994).

BIOGRAPHY

Dr. Sameh Alansary: *Director of Arabic Computational Linguistic Center at Bibliotheca Alexandrina, Alexandria, Egypt.*

He 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.

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.

نمو المفردات في عربية الاطفال المصريين: دراسة مؤسسة علي مدونة لغوية

[1]هبه سلامة ـ [2]سامح الانصاري
كلية الاداب- قسم الصوتيات واللغويات- جامعة اسكندرية
[1]Heba.salama.slp@gmail.com
[2]Sameh.Alansary@bibalex.org

ملخص

تهدف الدراسة الي وصف تطور المفردات وعلاقتها بالعمر لدي الأطفال المصريين ومعرفة الأختلافات الفردية لدي الأطفال في نمو المفردات وحساب نسبة كلمات الطفل علي الكلمات الكليةTTR. يعد اكتساب الأطفال للمفردات محور رئيسي ومعقد في تطور اللغة عند الأطفال وقد لاقي اهتماما كبيرا في الأبحاث اللغوية النفسية وأصبح مجال للدراسة في حد ذاته. وتقوم هذه الدراسة علي مدونة لغوية للأطفال مسجلة من الكلام التلقائي لمدة 30 دقيقة لعشرة أطفال 5اولاد-5 بنات من عمر1,7 الي 4 سنوات وتحليل معدل تطور الكلمات مع تطور العمر. يساهم هذا البحث في معرفة المعدل الطبيعي لنمو المفردات لتميزومعرفة مدي تأخر اللغة عند الأطفال، ويعد من الأسهامات الأولي في اللغة المصرية للأطفال.