Generating acceptable Arabic Core Vocabularies and Symbols for AAC users

E.A. Draffan, Mike Wald, Nawar Halabi, Ouadie Sabia[1], *Wajdi Zaghouani* [2]

Amatullah Kadous, Amal Idris,[3] *Nadine Zeinoun, David Banes, Dana Lawand*[4]

[1]University of Southampton, UK
[2]Carnegie Mellon University, Qatar
[3]Hamad Medical Corporation, Qatar
[4]Mada Assistive Technology Center, Qatar

ead@ecs.soton.ac.uk, mw@ecs.soton.ac.uk, nh1g12@ecs.soton.ac.uk, o.sabia@soton.ac.uk, wajdiz@cmu.edu, tullahk@hotmail.com, aahmad2@hamad.qa, nzeinoun@mada.org.qa, dbanes@mada.org.qa, dlawand@mada.org.qa

Abstract

This paper discusses the development of an Arabic Symbol Dictionary for Augmentative and Alternative Communication (AAC) users, their families, carers, therapists and teachers as well as those who may benefit from the use of symbols to enhance literacy skills. With a requirement for a bi-lingual dictionary, a vocabulary list analyzer has been developed to evaluate similarities and differences in word frequencies from a range of word lists in order to collect suitable AAC lexical entries. An online bespoke symbol management has been created to hold the lexical entries alongside specifically designed symbols which are then accepted via a voting system using a series of criteria. Results to date have highlighted how successful these systems can be when encouraging participation along with the need for further research into the development of personalised context sensitive core vocabularies.

Index Terms: symbols, Augmentative and Alternative Communication, AAC, core vocabularies

1. Introduction

In the last few years it has become clear that many therapists and teachers working with individuals who have speech and language difficulties in the Arabic speaking Gulf area, are depending on westernized symbols and English core vocabularies. Issues around limited Arabic language knowledge and dependency on translations or working in English can cause difficulties for those who need Augmentative and Alternative forms of Communication (AAC) due to disabilities. Huer [1] reports that "observations of communication across cultures reveal that non-symbolic as well as symbolic forms of communication are culturally dependent" and her later work "suggests that consumers, families, and clinicians from some cultural backgrounds may not perceive symbols in the same way as they are perceived within the dominant European-American culture" [2].

With this in mind the Arabic Symbol Dictionary research team were determined to take a participatory approach to their project, involving AAC users and those supporting them as well as other researchers working in the field of Arabic linguistics and graphic design.

2. Background

Much has been written by speech and language therapists about the necessity for core vocabularies that have been adapted to suit symbol users who need to enhance their language skills [3], [4], [5] and [6]. Research has shown that with a few hundred of the most frequently used words 80% of one's communication needs can be accommodated [7]. More recently concept coding [8] with the idea of mapping different symbol vocabularies along with a focus on psychosocial and environmental factors [9] to improve outcomes have been added to the mix. However, there is very little research that has been undertaken to provide therapists with suitable vocabularies for Arabic AAC users [10]. In English these vocabularies tend to be lists of frequently used words from spoken and written language across all age groups and some from AAC users. Despite considerable searching there are very few of these vocabularies available in Arabic with most coming from language learning or frequently used word lists with no specified ages or Arabic AAC users.

In some areas there is also a lack of understanding regarding the complexities of Arabic spoken and written language that disproportionately affect those who may have communication and reading difficulties [11], [12] and [13]. Usziel-Karl et al [13] cite several researchers in the course of their study concerning Arabic and Hebrew linguistic frameworks and discuss the "critical importance of morphology as the main organizing principle both of the lexicon and of numerous grammatical inflections". The authors go on to point out the diglossia [two variations of a language in different social situations] nature of Arabic which means there is a 'phonological distance [in grapheme-to-phoneme mapping] that has a negative impact on the acquisition of basic literacy skills in young Arabic children…" Words or word phrases (referents) may also be presented above or below a corresponding symbol, with changing forms depending on

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 91–96,
Dresden, Germany, 11 September, 2015.

grammatical status, gender and/or number plus many letters will change their shape depending on their position within a word.

The authors of this research and others have also found there are key cultural and family values/orientations that should be considered in order to increase the effectiveness of symbol-referent vocabulary interventions [14] with individuals who use AAC within Arab communities. To this end not only has research concentrated on word frequency lists and collating an AAC user core vocabulary, but also instigating a voting system for symbol acceptance, so that words or multiword/word phrases are represented by symbols that are suitable culturally, linguistically and for the settings in which they will be used.

3. Methodology for Building a Core Vocabulary

The building of an Arabic AAC core vocabulary is ongoing, but began with the collection of word lists used by AAC users, their families, carers, speech and language therapists and teachers in Doha (Qatar) (List a). Sixty three of these individuals joined an AAC forum and these participants have continued to work with the team as symbols for the vocabularies have been developed.

The initial aim was to collect around 100 localised Arabic most frequently used words and multiwords to compare with those already in use that were in English or translated into Arabic based on English core vocabularies. Participating therapists felt a further 400 words/multiwords would be the maximum the majority of their users would have in their communication books or devices. Most English speaking three year olds use over a thousand words [15] so it was essential that the fringe vocabulary should be enlarged with words specific to the environment and personal needs including Qatari colloquial words and place names as well as to be relevant to all ages.

Surveys of core vocabularies in Arabic have revealed that few are freely available [16] and even less make good companions when thinking of basic language and literacy learning for AAC users. In order to expand the list of 500 words a comparison was carried out against five other Arabic word frequency lists. Those for general conversation included the Kelly Project [17], 101languages.net 1000 most common spoken Arabic words and Aljazeera comments often using colloquial language [18]. The Supreme Education Council (SEC) literacy lists Grade 1,2,3 and Lebanese reading lists [19] have been used for literacy skill building in Modern Standard Arabic (MSA).

3.1. Building a vocabulary list analyser

An automatic system was developed that took as an input two main pieces of information:

List a: The list to be analyzed as a basis for the new core vocabulary list: This list could optionally have frequency of each entry included. If no frequency is available then a default value should be added to all the entries before running the program. Frequency in this case equated to how often a word was used. This frequency does not have to correspond to an actual frequency of occurrence in a text somewhere.

Lists b: Lists combining existing vocabularies from a number of sources with the same structure as List a. Multiple vocabularies are used in Lists b in an attempt to weight the occurrence of individual words. These vocabularies are ideally from different sources and should be large enough so that the frequencies of the entries listed are reliable.

The system produced three lists shown in Figure 1:

List 1: Initial list containing the words in List a (the in-put list to be analyzed) that did not occur in any of Lists b. This output only contained the words with no frequency scores.

List 2: The coverage list: containing the words that occurred in List a and at least once in a source vocabulary in Lists b. This output also contained scores for each word by source vocabulary list (each word was given several scores, one for each list in Lists b). Each score equals the frequency with which each word appeared in the list from Lists b, normalized by dividing the frequencies of each word by the sum of all frequencies in that list. The score was set to 0 if the word did not occur in that list.

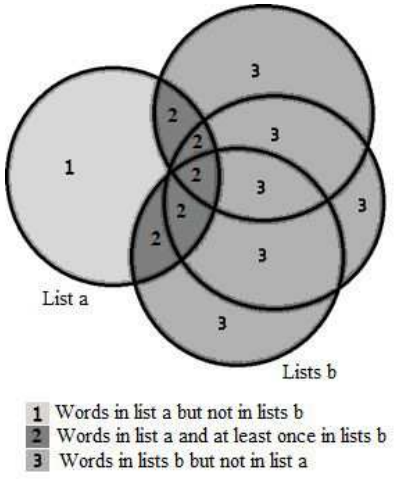

Figure 1. *Input lists (list a and lists b)*

List 3: Remaining word list: This list contained all the words that were in Lists b but were not contained in List a. This output also contained the scores for each word and is the example of the system in use (Figure 2). This is the list on which the comparison in the section 3.2 is based.

	word	lists\Kelly Project.c	lists\Aljazeera Comments	lists\Most common 1000 fro	sum
2	ان	0	0.023466719	0.001	0.024466719
3	الله	0.002	0.021198718	0.001	0.024198718
4	على	0	0.011530471	0.001	0.012530471
5	با	0	0.009577154	0.001	0.010577154
6	الى	0	0.008842202	0.001	0.009842202
7	الذي	0	0.007461319	0.001	0.008461319
8	ولا	0	0.007403351	0.001	0.008403351
9	الشعب	0	0.007820514	0	0.007820514
10	لم	0	0.006506916	0.001	0.007506916
11		0	0.007146635	0	0.007146635
12	الذي	0	0.006086648	0.001	0.007086648
13	في	0	0.005657063	0.001	0.006657063
14	حتى	0	0.005199529	0.001	0.006199529
15	او	0	0.005103261	0.001	0.006103261
16	علي	0.002	0.002888052	0.001	0.005888052
17	قط	0.002	0.002698621	0.001	0.005698621
18	فقط	0.002	0.002525752	0.001	0.005525752
19	الا	0	0.004469753	0.001	0.005469753
20	العرب	0	0.005398277	0	0.005398277
21	يجب	0.002	0.002301126	0.001	0.005301126

Figure 2. *Example Output from lists viewed in Excel*

Figure 2. shows frequencies are normalized to allow source vocabularies to be compared (column one), this process can be problematic if the list is too small as the numbers may become too high and significantly affect results. Even if there is

sufficient data, it is still imperative that an expert goes through the different output list to inspect the results, correct errors and choose the set of words to be added or removed from the input list. The scores given only act as a guide to assist the expert in the process.

In practical terms words with high scores in List 3 could be deemed suitable for inclusion in the Arabic Symbol Dictionary and added to List a. The system has been run repeatedly as lists have been added so that results become more robust.

3.2. Results of the Core Vocabulary building

When comparing the list provided by participants as examples of AAC users' vocabularies (List a), there were very small overlaps with those words most frequently found where the top words were based on very high frequency scores for those most commonly used (Lists b).

To provide an instant comparison between Output 1 and 3 the top 20 words translated from Arabic are listed below.

Output from 1 (List a) ordered by those most often used in AAC lists.

"I/me (am), go, ball, car, banana, on/to, thing/something, to, chair, clock/watch, want, in, sit, was, eat, bike, flower/rose, play, cup, door"

Output from 3 (Lists b) ordered by frequency

"the, God, about, oh, to, which (masculine), and not, people, no, which (feminine), in, even, or, on, against, only, however, Arabs, must, order"

Further analysis of the Lists b that were about spoken and colloquial language shows that nouns only made up 5% of the total list from the Kelly project, 25 to 30 % of the Aljazeera and Oweini-Hazoury lists, but 50% of the AAC lists. A concrete noun, even if it is considered part of a fringe vocabulary, is a much easier concept to illustrate with a symbol and may be seen as one of the early building blocks to language acquisition. Verbs, however are more complex and have low frequency rates; between 5 to 20 %. The Aljazeera list has the lowest and the AAC lists have the highest. The other parts of speech, equally pertinent in communication, such as adjectives, adverbs, prepositions, pronouns and conjunctions were found to be variably frequent from one list to another. The Aljazeera list has a quarter of its frequencies made up of prepositions, whereas Kelly's list, SEC and the AAC user list have only 5%. Conjunctions also show low frequencies through the lists in question; between 1% and 15%. It is worth mentioning that pronouns are totally nonexistent in Kelly's project list, either under their detached form or attached form. It should also be noted that therapists may choose nouns rather than pronouns for the purpose of symbol transparency. The other lists had less than 20% of pronouns all types combined. Arabic pronouns, and also some prepositions combine with nouns or with other parts of speech as single words, this morphological aspect could be the reason why their frequencies are rather undermined. Adverbs are also rarely listed, The Owein-Hazoury list has none; the highest adverb frequency is found in the 1000 most common Arabic words list (4%). In Arabic most adverbs of time and space are prepositional groups; typically a structure made of a preposition followed by a noun. This structural definition of adverbs explains the low number or even the lack of adverbs in some of the core vocabulary lists. The users would frame appropriate phrases to express adverbs by using existing prepositions combined with nouns.

Further confirmation for these differences in the frequency of various parts of speech was sought for the literacy skill vocabularies. The conversational based lists were replaced with reading lists forming Lists b. Arabic lists such as those used SEC and Arabic sight words [19]. It was found that in their top 100 frequently used words 30 and 38 were nouns respectively.

3.3. Discussion about the core vocabulary data collection

As can be seen from the top 20 words in List a and Lists b, both show nouns that would not be found in the top twenty frequently used words in an English core vocabulary and in reality would be considered fringe words. However, the lists do illustrate that in Arabic there are elements of the grammar that are equally as important such as conjunctions and prepositions.

There are considerable issues with the fact that root words in Arabic clearly appear within other words and this can affect the results as well as the fact that the lists collected from AAC users are based on popular use, rather than large scale frequency levels within a huge corpus. There will always be the need to improve outcomes by collecting more lists from AAC users in the future to improve the balance between words used for symbol communication and those based on frequency of use, although the latter informs vocabulary development

By using this system the combined AAC word lists from the Doha schools and clinics making up 'List a' once translated into English, could be compared to the Prenke Romich 100 Frequently Used Core Words [20], [21] (as Lists b). It was noted that the Doha Arabic AAC user list (List a) contained 38 nouns in the top 100 words compared to none appearing in the English core vocabulary. It has been said that in English the use of nouns goes from 7% in the top 100 words to 20% in the top 300 [22] whereas in MSA the corresponding frequency levels are 26% and 45% according to one of the largest frequency lists [23].

These results highlight the need for further exploration into this aspect of vocabulary building. In particular there is a need to collect more wide ranging conversations to evaluate the differences in the type of words and multiwords required to successfully build Arabic AAC personalised and context sensitive vocabularies. There is also the need to be aware of the differences in lists used for enhancing reading skills where MSA is used rather than the colloquial dialects of the area. A further distinction may be needed between adult and children's vocabularies where religious and social language requirements may impact on AAC use. The Speech and Language therapists attending meetings with the team also noted the importance of vocabularies sensitive to user's characters, interests and social setting commenting on dress and gender issues as well as being aware of the issues of using lists from AAC users of school age due to the lack of available adult AAC users in the region at the time of writing.

4. Methodology for Symbol Management

Just as it was found that there was a paucity of core AAC vocabulary lists in Arabic, the same could be said about the symbols provided for AAC devices. Some centres in Doha

were providing specifically designed symbols for the Arabic culture, environment, social and personalised linguistic needs but there were no adapted symbol sets that were freely available for sharing. Nor had any symbols been evaluated for transparency or cultural sensitivity by local AAC users, their supporting professionals and families.

A bespoke Symbol Management system was developed that allowed the team to store symbols. The system also offered participants the chance to take an active role in the decisions made around the development and evaluation of appropriate symbols as they could see and vote on uploaded symbols representing the core vocabularies previously collected.

The online database was based on a Model-View-Controller (MVC) framework using MongodB with JavaScript (NodeJS and an Express JS plugin). The code is open source and available on bitbucket. View templates which generated the html pages were built suing the Jade templating engine. The only other plugins used were for authentication and list filtering. The latter will provide the basis for browse and search features in the final Arabic Symbol Dictionary website.

4.1. Building symbol acceptance system

As part of the online management system a simple voting set up was created using the filters developed for batches of symbols. During voting sessions participants have been presented with a series of around 60-65 images of newly designed symbols, the referent in MSA, Qatari (where applicable) and English. The voting criteria are presented with large selection areas on a scale of 1 to 5 where 5 is completely acceptable (see Figure 3) so that different visual displays can be used. The four criteria are listed with a free text box for comments:

- Feelings about the symbol as a whole
- Represents the word or phrase
- Color contrast
- Cultural sensitivity

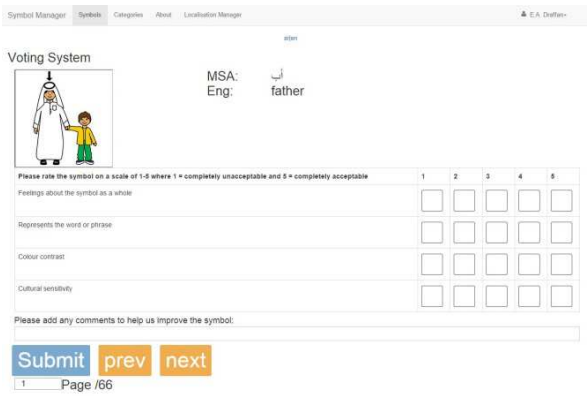

Figure 3 *Voting system with criteria for acceptance on a scale of 1-5 where 5 is completely acceptable*

4.2. Results from voting sessions

The initial batch of symbols had 63 voters logging into the Symbol Manager resulting in 2341 votes for 65 symbols. Overwhelmingly the decisions were very favourable with all mean ratings significantly greater than a rating of 3.5. The average was 4.0. (See Table 1) All voting data was anonymized and comments collated to inform the graphic designer.

Two AAC users were also able to vote on the symbols via an adapted system using their own Sensory Software Grid 2 systems with the symbols added plus a 1-5 or 1-3 'thumbs up' to 'thumbs down' scoring depending on their ability. This produced equally good results and comments were captured via recordings. More AAC users are being encouraged to join the forum and as further batches of symbols are developed it is hoped that voting sessions will continue to occur both during face to face meetings and remotely.

Table 1. *One Sample T test for Difference of Mean Ratings from 3.5*

Criteria	Number of voters	Mean rating	2 tail P Value for difference from 3.5
1	63	3.94	<0.0001
2	63	3.90	<0.0001
3	63	4.07	<0.0001
4	63	4.10	<0.0001

4.3. Discussion about the Symbol Management system

The initial development of the Symbol Management system was purely for the team to upload lexical entries and symbols with a set of filter systems based on parts of speech, gender, number and symbol descriptions. However, as the participation by AAC users, their families, therapists and teachers grew it became essential to offer a voting system that quickly produced results because specialists wanted to use the symbols as they were developed. As all the speech therapists and teachers involved had worked for several years with AAC users, but were mainly from countries other than Qatar, it was felt that there should be a method to check acceptability within the community before releasing them for download, not just depending on the team's opinions. The team had already set up a Google+ method for initially evaluating iconicity and transparency [22].

Those therapists working in the Doha area were very willing to express their opinions about symbol suitability and the links with the corresponding word lists collected. It was noted that there was a general understanding that the lexical entries in Modern Standard Arabic and those entries in Qatari colloquial Arabic may share the same symbol for similar meaning words or multiword phrases but there may need to be additional symbols and / or changes in symbol labels to represent different parts of speech, gender and number and to take into account the bilingual nature of the dictionary to aid those who were not fluent Arabic speakers.

5. Conclusion

The core vocabulary and symbol management systems have provided the research team with quick and easy ways to analyse data as well as provide a platform for user participation. Having a selection of MSA and Qatari core and fringe vocabularies has been essential for ongoing symbol development, but there is still a need to continually update the collection of local vocabularies to ensure that colloquial as

well as written language is captured. The present frequency levels of the words collected in Doha (List a) are low in comparison to global lists (Lists b). They are also subjective, based on the AAC forum input rather than a wide base of Arabic AAC users and carers. However, with support it has been shown that where suitable core vocabularies are implemented alongside appropriate symbols AAC users, who have the capacity, can enhance their communication and improve their readiness for reading [24] and already in this project AAC users have greeted the newly developed symbols with much appreciation, but there remains the need to 'focus on long-term outcomes' [9].

There remains the debate as to the differences in parts of speech seen in English core vocabulary lists compared to some Arabic lists with high levels of noun use. It is important to appreciate the limitations of the collection procedures as well as the problems of automated comparisons between lists that require normalization and have different methods for showing root words, different parts of speech and verb declensions.

There is much research still to be carried out to ensure that an appropriate vocabulary list suitable for Arabic AAC users and the development of literacy skills can be collated in a diglossia situation. But as an increasing number of words lists are provided by participants set against the further analysis of the frequency lists already gathered it is felt that this can be achieved.

6. Acknowledgements

This research was made possible by the NPRP award [NPRP 6 - 1046 - 2 - 427] from the Qatar National Research Fund (a member of The Qatar Foundation) and thanks must go to all those participants in Doha who have contributed to the work of the Arabic Symbol Dictionary team. Grateful thanks are also expressed to the ARASAAC team for allowing their symbols to be used with participants. The statements made herein are solely the responsibility of the authors.

7. References

[1] M. B. Huer, "Culturally inclusive assessments for children using augmentative and alternative communication (AAC)," *Journal of Children's Communication Development*, 19 (1), 23–34. 1997.

[2] M. B. Huer, "Examining perceptions of graphic symbols across cultures: Preliminary study of the impact of culture/ethnicity," *Augmentative and Alternative Communication* 16 (3): 180–185. 2000. doi:[10.1080/07434610012331279034]

[3] S. Balandin and T. Iacono, "A few well-chosen words," *Augmentative and Alternative Communication*, 14(September), 147–161 1998.

[4] M. Banajee, C. Dicarlo, and S. Buras Stricklin, "Core Vocabulary Determination for Toddlers," *Augmentative and Alternative Communication*, 19(2), 67–73. 2003.

[5] M. Lahey, and L. Bloom, "Planning a First Lexicon: Which Words to Teach First," *Journal of Speech and Hearing Disorders*, 340–351 1975.

[6] G. M. Van Tatenhove, "Building Language Competence With Students Using AAC Devices: Six Challenges," *Perspectives on Augmentative and Alternative Communication*, 18(2), 38–47 2009.

[7] G. C. Vanderheiden, and D. P. Kelso, "Comparative analysis of fixed-vocabulary communication acceleration techniques," *AAC Augmentative and Alternative Communication*, 3, 196-206. 1987.

[8] M. Lundälv and S. Derbring, "AAC Vocabulary Standardisation and Harmonisation," *Springer-Verlag Berlin Heidelberg*, pp.303–310. 2012.

[9] J. Light, and D. Mcnaughton, "Designing AAC Research and Intervention to Improve Outcomes for Individuals with Complex Communication Needs," *Augmentative and Alternative Communication, (ahead-of-print)*, 1-12. 2015.

[10] R. Patel and R. Dakwar-Khamis, "An AAC training program for special education teachers: A case study of Palestinian Arab teachers in Israel," *Journal of Augmentative and Alternative Communication*, 21, 3, 205-217. 2005.

[11] S. Abu-Rabia, "Learning to read in Arabic: Reading, syntactic, orthographic and working memory skills in normally achieving and poor Arabic readers," *Reading Psychology: An International Quarterly*, 16, 351–394. 1995.

[12] S. Abu Rabia, D. Share and S. M. Mansour, "Word recognition and basic cognitive processes among reading-disabled and normal readers of Arabic," *Reading and Writing: An Interdisciplinary Journal*, 16, 423-442. 2003. doi:[10.1023/A:1024237415143]

[13] S. Uziel-Karl, F. Kanaan, R. Yifat, I. Meir, N. Abugov, and D. Ravid, "Hebrew and Palestinian Arabic in Israel: Linguistic Frameworks and Speech-Language Pathology Services," *Topics in Language Disorders* Vol 34 Number 2, p 133 – 154 2014.

[14] B. Woll, and S. Barnett, "Toward a Sociolinguistic Perspective on Augmentative and Alternative Communication," *AAC Augmentative and Alternative Communication*, 14(December), pp.200–211. 1998.

[15] K.J. Hill, and C. Dollaghan, "Conversations of Three-Year Olds: Implications for AAC Outcomes," *American Speech-Language-Hearing (ASHA) Convention. San Francisco, CA.* November. 1999.

[16] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," *In the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Rejkavik, Iceland*, 26-31 May 2014.

[17] A. Kilgarriff, F. Charalabopoulou, M. Gavrilidou, J. B. Johannessen, S. Khalil, S. J. Kokkinakis and Volodina, E. "Corpus-based vocabulary lists for language learners for nine languages," *Language Resources and Evaluation*, 1-43 2013.

[18] W. Zaghouani, B. Mohit, N. Habash, O.Obeid, N. Tomeh, and K. Oflazer. "Large-scale Arabic Error Annotation: Guidelines and Framework," *In the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014). Rejkavik, Iceland*, 26-31 May 2014.

[19] A. Oweini and K. Hazoury, "Towards a list of Awards a Sight Word List in Arabic," *International Review of Education*, 56 (4), 457-478 2010.

[20] K. Hill, and B. Romich, *100 Frequently Used Core Words.* Accessed May 2015 https://aaclanguagelab.com/files/100highfrequencycorewords2.pdf

[21] K. Hill, and B. Romich, "A summary measure clinical report for characterizing AAC performance," *Proceedings of the RESNA '01 Annual Conference, Reno, NV*. pp 55-57. 2001.

[22] J. Boenisch and G. Soto, "The Oral Core Vocabulary of Typically Developing English-Speaking School-Aged Children," *Implications for AAC Practice. Augmentative and Alternative Communication*, pp.77–84. 2015.

[23] T. Buckwalter and D. Parkinson, "A frequency dictionary of Arabic: Core vocabulary for learners," Routledge. 2014.

[24] D. Evans, L. Bowick, M. Johnson and P. Blenkhorn, "Using iconicity to evaluate symbol use," *In:*

Proceedings of the 10th international conference on computers helping people. Linz, Austria, pp 874–881 2006.

[25] P. Hatch, L. Geist, and K. Erickson, "Teaching Core Vocabulary Words and Symbols to Students with Complex Communication Needs," *Presented at Assistive Technology Industry Association*, 2015. Retrieved 19/2/2015 fromhttp://www.med.unc.edu/ahs/clds/files/conference-hand-outs/atia_2015.pdf (Accessed 14 June 2015).