Arabic Corpus Processing Tools for Corpus Linguistics and Language Teaching

Sultan Almujaiwel
Arabic Language Department
College of Arts, King Saud University, Saudi Arabia
salmujaiwel@ksu.edu.sa

Abdulmohsen Al-Thubaity
The National Center for Computer Technology and Applied Math
King Abdulaziz City for Science and Technology, Saudi Arabia
aalthubaity@kacstac.edu.sa

……..

ABSTRACT

This paper illustrates a new language processing tool (Arabic Corpus Processing Tools ACPTs 4.6 Version) that is a stand-alone open/free source for analyzing both Arabic and English large-scale texts, exceeding 50 million words depending on over 8 gigabytes of personal PC RAM space. The new ACPT has sophisticated cutting-edge functions used for the literature of corpus linguistics and corpus-linguistic analysis, especially statistical packages. A comparison with other tools suitable for corpus-linguistic analysis is important. This gives merit for ACPTs to be considered the most effective tool for corpus analysis. As this paper focuses on language teaching, the most prominent functions of the intended tools that can be exploited for enhancing the progress of language teaching/learning are illustrated, arguing some key elements that should be understood before using the tool.

Keywords: Arabic corpus processing tools, corpus search and analysis tools, language teaching, statistical corpus linguistics.

1. INTRODUCTION

The Arabic Corpus Processing Tools ("ACPTs") (version 4.6, "ghawwas" given in Arabic, "diver" in English) have become more advanced (cf. Almujaiwel 2016; regarding version 3.0). This stand-alone processing tool was designed by a research team in February 2013 led by Al-Thubaity et al. (2014b). It is an open/free source system[1], and it has arguably become the most reliable tool for processing Arabic texts due to its ability to read Arabic characters. The interface of ACPTs comes with English and Arabic options. It gives the frequency and relative frequency of types/tokens and documents—if selected during the process—in the selected folders. Moreover, it supports TXT. DOC. DOCX. and HTML formats, and ANSI or UTF-8 encoding.

As far as corpus linguistics and language teaching are concerned, it is not only English or Arabic that can be processed with this tool for more practice in language learning/teaching, but it also can be used for French as well (Al-Thubaity et al. 2014a).

The ACPTs are considered to be one of the three best Arabic processing tools: Sketch Engine (Kilgarriff et al. 2004) and aConCorde (Roberts 2014; in Alfaifi and Atwell 2016. See also Roberts et al. 2006). The other tools are AntConc (Anthony 2005), WordSmith Tools (Scott 2012), and IntelliText Corpus Queries (Sharoff 2011). All of these tools are stand-alone open corpus search software tools except Sketch Engine and IntelliText, which are web-based corpus queries. In the work of Alfaifi and Atwell (2016), the eight criteria (cf. Al-Thubaity et al. 2014c) posed for such an evaluation are briefly as follows: reading Arabic text files in UTF-8 and UTF-16 formats, displaying diacritics, the accuracy of displaying right-to-left Arabic characters,

[1]See http://sourceforge.net/projects/kacst-acptool/

normalizing diacritics or Hamza, providing Arabic user interface and enabling users to upload or open their Arabic personal corpora.

The question is whether these criteria are enough for an evaluation. Since diacritics are mostly neglected in Arabic typesetting, and since files saved in UTF-16 can be resaved in other respective formats, these two criteria are less important in searching Arabic corpora. In addition, the Sketch Engine launched an Arabic interface in 2015. However, the criteria that might be more essential in a corpus search and analysis are frequency, concordancing, collocations and corpus linguistics statistics (e.g. Biber et al. 2009, Evert 2009a, Evert 2009b). It would be more useful for a user who is proficient in Arabic and English to use WordSmith Tools instead of aConCorde due to the importance of statistics adopted in corpus linguistics since the beginning of the last decade of the twentieth century (Kenneth et al. 1991, Baayen 2008, Baroni and Evert 2009). The KWIC concordance tools are also as useful as aConCorde in terms of displaying Arabic characters.

The work of Alfaifi and Atwell (2016) is helpful because of their proposed criteria. Other essential criteria need to be discussed and analyzed, especially statistical distributions of keyness, collocation, colligation and collostruction strength/weakness on one hand, and attraction/repulsion on the other hand. The latter is processed in R and RStudio programs and in the programming syntactic inputs used for statistical corpus linguistics. This might be done less in ACPTs regarding how distributions of corpus-related statistics can be tabulated. This matter needs to be studied, in order to examine its feasibility.

2. ACPTs FUNCTIONS

The parameter of the JAVA Runtime Environment is best between -xms512m -xmx4048m. This guarantees better and faster processing, and enables the teacher/analyzer to process a file that contains more than 50 million words. There are three main tabs in the ACPTs interface: Add Corpus, Preprocessing and Comparison (Figure 1). The first main tab opens an interface where the teacher/analyzer uploads text files as either a primary or reference corpus. The second tab has four functions: n-grams, the corpus query engine, the options of holding or removing diacritics and of modifying some characters, and the features selection where the user can search the whole file(s), uploading a Stop List or Include List. The third tab shows the statistics that are adopted in corpus linguistics, and which the user can choose from. The functions of ACPTs are multi-faceted. The first is the role of primary corpus and reference corpus.

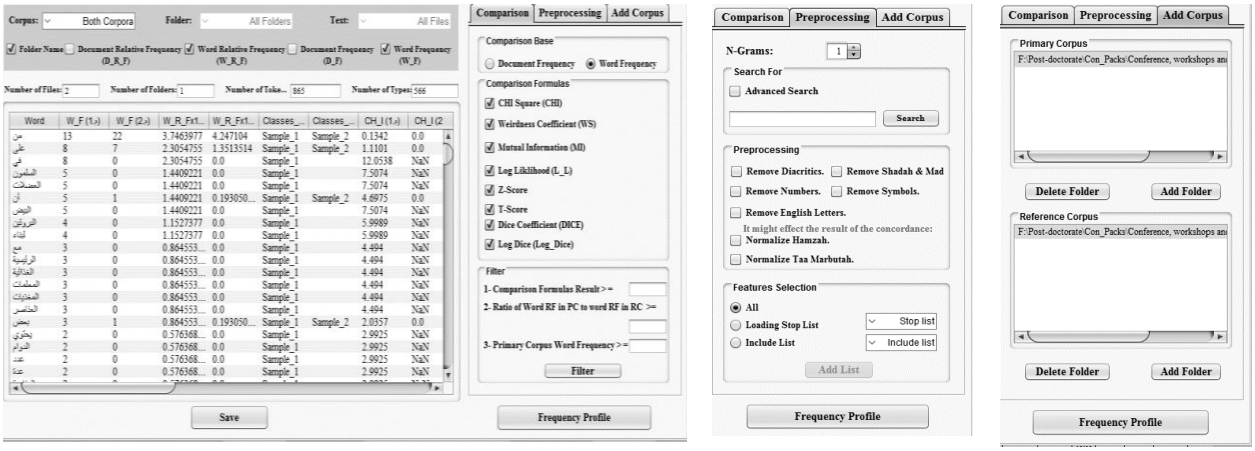

Figure 1. the three sub-interfaces of ACPTs

Uploading folders including plain text files must be treated carefully and in different ways. Not only are raw texts, sought by teachers, said to be essential in teaching. If the teacher endeavours to provide learners with significant collocations and colligations of authentic texts, he/she might have to reorganize the texts by the sort-out and/or throw-out function of duplicates and punctuation, or, for example, cleaning up extra spacing between words. He/she might also have to rearrange texts or include/exclude texts in order to achieve the objectives of linguistic skills in teaching/learning.

N-grams simply refer to the number of words (collocates) occurring on either side of the node. The nodal item is the core lexical unit/item for which the processing of the collocational span is intended. The Stop List and Include List are of great importance when the teacher/analyzer has a tendency to calculate particular resulting concentrations. This can involve, for instance, excluding the function words from the results or including terms for which the teacher wants to detect the linguistic and contextual behaviors in the text files.

3. ACPTs STATISTICS

Statistical corpus linguistics can be conducted and evaluated by any natural language toolkit software. However, the degree to which each statistical significance in ACPTs is retrieved and the extent to which each is useful are important points in language teaching and education (Table 1).

Chi-square $\chi 2$, and similarly, log-likelihood LL, are used to simply judge the values and their goodness of fit test according to their position between the (probability) P-values (usually between 0.05 or 0.999). This is to evaluate the text files in the sense that the distribution of items in them are significantly represented by chance. The LL is a comparison that evaluates the probability values that estimate the coefficient parameters. It is simply indicated by high/low scores on a standardized test (Nagao and Mori 1994).

Table 1. All built-in ACPTs statistics results between randomly sampled PC (File1) and RF (File2)

من	WF1	WF2	WRF1	WRF2	File1	File2	$\chi 2$ 1	$\chi 2$ 2	MI_1	MI_2	z- 1
	13	22	3.74	4.24	S1	S2	0.13	0.0	-0.11	0.06	-0.27
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	0.22	-0.28	0.2218	0.06	0.07	10.12	10.34	0.13	0.13	0.88	1.0
على	WF1	WF2	WRF1	WRF2	File1	File2	$\chi 2$ 1	$\chi 2$ 2	MI_1	MI_2	z- 1
	8	7	2.30	1.35	S1	S2	1.11	0.0	0.41	-0.35	0.80
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	1.0	-0.66	0.701	-0.74	0.04	0.02	9.50	8.74	1.08	1.08	1.70
في	WF1	WF2	WRF1	WRF2	File1	File2	$\chi 2$ 1	$\chi 2$ 2	MI_1	MI_2	z- 1
	8	0	2.30	0.0	S1		12.05	NAN	1.31	0.0	2.67
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	NAN	-2.18	1.69	Infinity	0.04	0.0	9.52	Infinity	14.7	14.72	Infinity
السلمون	WF1	WF2	WRF1	WRF2	File1	File2	$\chi 2$ 1	$\chi 2$ 2	MI_1	MI_2	z- 1
	5	0	1.44	0.0	S1		7.50	NAN	1.31	0.0	2.11
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	NAN	-1.73	1.33	Infinity	0.02	0.0	8.86	Infinity	9.17	9.17	Infinity
العضلات	WF1	WF2	WRF1	WRF2	File1	File2	$\chi 2$ 1	$\chi 2$ 2	MI_1	MI_2	z- 1
	5	0	1.44	0.0	S1		7.50	NAN	1.31	0.0	2.11
	z- 2	t- 1	t- 2	Dice1	Dice2	LogD1	LogD2	LL1	LL2	WC1	WC2
	NAN	-1.73	1.33	Infinity	0.02	0.0	8.86	Infinity	9.17	9.17	Infinity

The Weirdness Coefficient WC is important in education in that keywords or peculiar words/collocates are intended between files (between primary corpus PC and reference corpus RC, at a precise application). There are four values of WC. First, they are at the level of 0. Secondly, the value is more than 1 when words are more frequent in PC than in RC. Thirdly, the value is less than 1 when words are more frequent in RC than in PC. Fourthly, the value is INFINITY when the words occur only in PC.

Mutual Information MI is useful in detecting the association strength between collocates. The higher values, the stronger the collocates are associated. A value that is less than 3 usually indicates no significance in such an association. This is like the z-score, but the insignificance is at the value of 2 or less. Similarly, t-score values are dealt with like the z-score.

Dice coefficient and logDice signify the associative strength and weakness among words/documents. The former indicates very small decimal numbers, and the smaller the value becomes, the stronger the association is. While logDice produces values between 0–14, it shows when the associative collocates occur. The word/collocation occurs 16,000 times when the value is zero, and the closer the value is to 14, the higher the formulated strength.

4. ACPTs AND LANGUAGE TEACHING

This section will describe the important practices of language teaching that can take advantage of the respective tools. The most adopted practices are quoted from Timmis (2015), but with a discussion of the practical elements limited by the ACPT.

The effectiveness of using this tool in language education depends on building or gathering the corpus/files. This should be taken into consideration before exploiting frequency, concordancing and collocations.

Lexis teaching is essential when its strategies rely on texts and parts of sentences. Phraseology is the term that is preferred in language acquisition literature, instead of concordancing (citations) in corpus linguistics. This term is the focal point for vocabulary teaching and its varying dynamic use in different linguistic situations. It might be possible to say, in contrast to concordancing, that the more one reads, the more he or she learns phrasal and structural patterns and lexical-filling. However, the text selection is a vital process for further language immersion because variations of authentic language use can be easily taught with a productive method.

The psychological and sociocultural perspectives of applied linguistics and language teaching can be added along with the functional techniques of ACPTs. One group stands out—communicative competence (Brown 2000). If the teacher, or especially the learner, practiced in using ACPTs, aims at tackling some communicative elements in long texts, the function of n-grams is of great help. This function can show the sequences (whatever the span of nodal items and the number of collocates in either side of them) of the associative and semantic priming (the latter refers to the words occurring in real use of the language, whereas the former refers to the expected words in the communicative elements) in different situations. These elements can be the sender, the receiver, the message and the medium by which the texts/discourses are produced. For instance, there is the matter of fossilization, which students experience when they have only a few texts and words in a given contextual communication. The development of communicative competence might easily speed up when further naturally occurring elements do not come to a standstill.

Collocations of co-occurrences and the availability of sorting out lists that show their frequencies and contexts in an expanded concordance have played an effective role in reshaping natural language, from the last decade of the twentieth century (Shimohata and Nagata 1997). More experimental investigations into the examples of collostructions of clausal and phrasal

constituents that alters in different naturally-occurring examples are taken into account in corpus-linguistic studies (Wiechmann 2008). Further techniques to rely on reliable, learnable, acceptable, and changeable data for instruction and learning are pivotal in teaching/learning developments.

The importance of a primary corpus and reference corpus in ACPTs, which is not facilitated by or does not exist in other corpus analysis tools, increases when keyword is used. For instance, if keyword lists are achieved by gathering long primary texts from a particular domain (e.g. business management) and using reference texts from multiple domain(s), it will give a long list of the keywords, collocations and extended citations related only to business management.

5. Conclusion

This paper concludes with the advantages of ACPTs in language teaching and developments. It seems that it is best suited for teachers to exploit in the teaching process. It also seems best suited for advanced learners to direct themselves with more effective open sources than those limited in textbooks. The user needs to become familiar with all of the ACPT's functions explained in this paper. It would not be an exaggeration to conclude that ACPTs help to process and analyze texts much more—perhaps twice as much—than other natural language toolkits.

References

Alfaifi, A. & Atwell, E. (2016). Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology*, 9(2): 347-357.

Almujaiwel, S. (2016). "Free/Open KACSTAC and its Processing tools: Lexical Resources for Arabic Lexicogrammatical Microstructures Based on Collocational Indicators." In Alonso Almeida, F., I. Ortega, E. Quintana and M. E. Sánchez. (eds.), *Input a Word, Analyze the World: Selected Approaches to Corpus Linguistics*. Newcastle Upon Tyne: Cambridge Scholars Publishing, 153-170.

Al-Thubaity, A. Al-Khalifa, H. Alqifari, R. and Almazrua. M. (2014c). Proposed Framework for the Evaluation of Standalone Corpora Processing Systems: An Application to Arabic Corpora. *The Scientific World Journal* (open access journal). Retrieved July 2, 2016, from file:///C:/Users/Lenovo/Downloads/602745%20(1).pdf.

Al-Thubaity, A., & Al-Mazrua, M. (2014a). *Khawas: Arabic Corpora Processing Tool USER GUIDE*. Retrieved July 2, 2016, from http://www.sourceforge.net/projects/kacst-acptool/files/?source=navbar.

Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Almoussa, M. (2014b). KACST Arabic Corpora Processing Tool ''Khawas'' [Computer Software]. Retrieved July 2, 2016, from https://sourceforge.net/projects/ghawwasv4/.

Anthony, L. (2014). *AntConc*, (Version 3.4.2) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved July 2, 2016, from http://www.antlab.sci.waseda.ac.jp/.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baroni, M. & S. Evert, (2009). "Statistical Methods for Corpus Exploitation." In Lüdeling, A. and M. Kytö (eds.) *Corpus Linguistics: An International Handbook* Vol. 2. Berlin, New York: Mouton de Gruyter, 777-803.

Biber, D. & Jones, J. K. (2009). "Quantitative Methods in Corpus Linguistics." In Lüdeling, A. and M. Kytö (eds.) *Corpus Linguistics: An International Handbook* Vol. 2. Berlin, New York: Mouton de Gruyter, 1286–1304.

Brown, D. H. (200). *Principles of Language learning and Teaching.* 4th edition. New York: Pearson Education.

Evert, S. (2009a). "Rethinking Corpus Frequencies." Paper presented at the *ICAME 30 Conference*, Lancaster, May, 27-31.

Evert, S. (2009b). "Corpora and Collocations." In Lüdeling, A. and M. Kytö (eds.) *Corpus Linguistics: An International Handbook.* Vol. 2. Berlin, New York: Mouton de Gruyter, 1212–1248.

Kenneth Ch., W. Gale, P. Hanks, & D. Hindle, (1991). "Using Statistics in Lexical Analysis". In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Uri Zernik (ed.). N.J.: Lawrence Erlbaum Associates, 115–164.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The sketch engine. In *Proceedings of the Euralex, 6–10 July 2004,* pp. 105–116, Lorient, France. Retrieved June 28, 2016, from http://www.sketchengine.co.uk.

Nagao, M. & Mori, S. (1994). "A New Method of *N*-gram Statistics for Large Number of *N* and Automatic Extraction of Words and Phrases from Large Text Data of Japanese." In *Proceedings of the 15th Conference on Computational Linguistics*, 5 August, 611-615. Stroudsburg, PA (USA): Association for Computational Linguistics.

Roberts, A. (2014) *aConCorde*. Retrieved June 28, 2016, from http://www.andy-roberts.net/coding/aconcorde.

Roberts, A., Al-Sulaiti, L., & Atwell, E. (2006). aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora* 1: 39–60.

Scott, M. (2004). *WordSmith Tools*. Version 4.0. Oxford: Oxford University Press. Retrieved June 28, 2016, from http://www.lexically.net/wordsmith/version4/index.html.

Sharoff, S. (2014). *IntelliText Corpus Queries* [Computer Software]. Retrieved June 28, 2016, from http://www.corpus.leeds.ac.uk/itweb/htdocs/Query.html.

Shimohata, S. T. S., & Nagata. J. (1997). "Retrieving Collocations by Co-occurrences and Word order Constraints." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* in Madrid, 7-10 July. Stroudsburg: Association for Computational Linguistics, 476-481.

Timmis, I. (2015). *Corpus Linguistics for ELT*. New York: Routledge.

Tsukamoto, Satoru. *KWIC Concordance for Windows*. Retrieved June 28, 2016, from http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/index_e.html.

Wiechmann, D. (2008). "On the Computation of Collostruction Strength: Testing Measures of Association as Expressions of Lexical Bias." *Corpus Linguistics and Linguistic Theory* 4(2): 253-290.