KSUCCA: A Key To Exploring Arabic Historical Linguistics

Maha Alrabia msrabiah@gmail.com
Department of Computer Science
King Saud University
Riyadh, Saudi Arabia

AbdulMalik Al-Salman salman@ksu.edu.sa
Department of Computer Science
King Saud University
Riyadh, Saudi Arabia

Eric Atwell e.s.atwell@leeds.ac.uk
Faculty of Engineering
Leeds University
Leeds, United Kingdom

Nawal Alhelewh drnawalh@gmail.com
Department of Arabic
Princess Nora bint Abdul Rahman University
Riyadh, Saudi Arabia

Abstract

Classical Arabic forms the basis of Arabic linguistic theory and it is well understood by the educated Arabic reader. It is different in many ways from Modern Standard Arabic which is more simplified in its lexical, syntactic, morphological, phraseological and semantic structure. King Saud University Corpus of Classical Arabic is a pioneering corpus of around 50 million words of Classical Arabic. It is initially constructed for the purpose of studying distributional lexical semantics of the Quran and Classical Arabic, however, it is designed in a general way making it also appropriate for other researches in Linguistics and Computational Linguistics. In this paper, we will briefly describe the structure of our corpus, and then we will demonstrate how it can be used to depict some aspect of Arabic language change between the classical and the modern periods.

Keywords: Historical Linguistics, Corpus Linguistics, Classical Arabic, Modern Standard Arabic, Lexical change, Syntactic Change, Morphological Change, Phraseological Change, Semantic Change.

1. INTRODUCTION

Historical linguistics, also called diachronic linguistics, is the study of why and how languages change or maintain their structure during time [1]. Nowadays, much research in historical linguistics is based on corpora containing texts from earlier periods of the target language allowing linguists to conduct more systematic studies on the evolution of languages and how various linguistic aspects are effected during the course of time [2]. This type of corpora that contain texts from past periods are usually referred to as historical corpora; there exist many historical corpora for English and many other languages. For example, the Helsinki Corpus of English Texts: Diachronic and Dialectal is a well-known historical corpus of English, which consists of two parts; a diachronic part containing 1.5 million words texts from the period between 750AD and 1700AD, and a dialect part consisting of transcripts of interviews with speakers of British rural dialects from the 1970's [3]. Another example is the Corpus of Historical American

English (COHA), which contains 400 million words of American English texts covering the period from 1810 until 2009 [4].

On the other hand, most of research in Arabic historical linguistics are not corpus based [5], which is a consequence of the lack of available appropriate corpora that can be used in such studies. Therefore, in order to study changes in Arabic language, very large corpora of Classical Arabic, which forms the basis of Arabic linguistic theory, should be made available to linguists so that they can compare them to existing Modern Standard Arabic (MSA) corpora in order to observe how and why did Arabic language change.

There exist a decent amount of MSA corpora with different types. For example, the Tim Buckwalter corpus for Modern Standard Arabic is the first corpus developed for Arabic. It was constructed in 1986 from an Arabic newspaper. The corpus was initially around 40000 words, and then it was expanded to more that 2.5 million words when the electronic Arabic content was available in the web. This corpus was designed for lexicographical purposes, and is not freely available for the public[1] [6]. On the other hand, Al-Sulaiti and Atwell [7] constructed the Corpus of Contemporary Arabic (*CCA*) of around one million words for the purpose of teaching Arabic as Foreign Language (*TAFL*). The corpus contains Arabic text from various categories and it is freely available online[2].

Moreover, Alansary et al., [8] compiled the International Corpus of Arabic (*ICA*) of about 100 million words of written MSA collected from a wide range of Arabic regions to insure diversity of writing styles, which makes it a good candidate for linguistic researchers who are interested in studying the influence of nationality on the speakers of MSA. They relied on machine readable sources to compile the corpus; containing newspaper articles, magazines, novels, books, web articles and other academic articles. ICA includes the following main genres: strategic sciences, social sciences, sports, religions, literature, humanities, natural sciences, applied sciences, arts and biographies. In addition, Sawalha and Atwell [9] constructed a large broad-coverage lexical resource for Arabic, which is a corpus of 23 machine-readable lexicons organized into roots, words formed from these roots and the meanings of those words. The authors evaluated the coverage of their corpus using three available Arabic corpora [9]; it scored 65-68% when using exact word matches and 82-85% when a lemmatizer was used to remove clitics. Moreover, Alfaifi and Atwell [10] developed the Arabic Learner Corpus (ALC), which is a 31272 words corpus consisting of texts written by learners of Arabic in Saudi Arabia. The corpus covers both native Arabic students who are learning to improve their Arabic language abilities and foreign students who are learning Arabic as a second language. For other examples of MSA corpora, I refer the reader to [6].

On the other hand, and to the best of the authors knowledge, there exist only two corpora of Classical Arabic; one is part of the King Abdulaziz City for Science and Technology Arabic Corpus (KACST Arabic Corpus)[3] and the other is the Classical Arabic Corpus (CAC) [11]. However, neither of the two corpora is adequate for research in distributional semantics; the former has a limited number of genres and it only contains 17+ million words, which is not very sufficient. While the latter is even smaller with only 5 million words. Therefore, it was essential to design and compose a new corpus of Classical Arabic bearing in mind that it should be large enough, balanced, and representative so that any result obtained from it can be generalized for Classical Arabic. In this paper, we will give a brief description of the design and construction of King Saud University Corpus of Classical Arabic (KSUCCA), which is a very large corpus of Classical Arabic that can be used in various corpus linguistic studies. In addition, we will demonstrate how KSUCCA corpus can be used in historical linguistics.

[1] http://www.qamus.org

[2] http://www.comp.leeds.ac.uk/eric/latifa/research.htm

[3] http://www.kacstac.org.sa/Pages/Default.aspx

The paper is structured as follows. Section 2 provides a brief description of the corpus. Section 3 demonstrates and discusses some aspects of change in Arabic language using KSUCCA. Finally, Section 4 discusses the conclusions of the work presented.

2. King Saud University Corpus of Classical Arabic (KSUCCA)

Texts included in KSUCCA are Arabic texts dating back to the period of the pre-Islamic era until the end of the fourth Hijri[4] century (equivalent to the period from the seventh until early eleventh century CE) [12]. The corpus is classified into 6 broad genres (Religion, Linguistics, Literature, Science, Sociology and Biography) covering most of the topics that were popular in that period of time, which is a strong indication of the corpus representativeness. These genres are further classified into 27 subgenres as shown in Table 1. It can be noticed that the number of texts and tokens are not evenly distributed between genres. However, this is consistent with the knowledge of the overall writing trends at that period of Arab history, and it is an indication of the balance of the corpus [13].

Genre	Subgenre	No. of documents	No. of tokens	Percentage
Religion	The Holy Quran	1	78245	0.15 %
	Hadith	44	5784326	11.43 %
	Exegesis of The Quran	13	7061862	13.96 %
	Quranic Studies	29	3665288	7.24 %
	Hadith Studies	10	643144	1.27 %
	Belief	23	486801	0.96 %
	Jurisprudence	26	5567407	11.00 %
	Principles of Jurisprudence	4	358014	0.71 %
Literature	Poetry	42	1265696	2.50 %
	Novels	2	172695	0.34 %
	Literature and Eloquence	60	5786113	11.43 %
Linguistics	Grammar and Morphology	16	1400951	2.77 %
	Language	6	401308	0.79 %
	Lexicons	27	4855732	9.60 %
	Proverbs	7	435975	0.86 %
Science	History	19	3750498	7.41 %
	Geography and Travel	14	609979	1.21 %
	Medicine	3	1837452	3.63 %
	Physics	1	61347	0.12 %
	Astronomy	2	112695	0.22 %
	Philosophy	1	24760	0.05 %
	Politics	1	4674	0.01 %
	Miscellaneous	1	27728	0.05 %
Biography	Prophet Muhammad Peace be	8	1163795	2.30 %
	Other biographies	18	2336153	4.62 %
Sociology	Ethics and Morals	23	1081566	2.14 %
	Genealogy	9	1628208	3.22 %
Total		410	50602412	100 %

TABLE 1: Classification of KSUCCA Texts.

KSUCCA is designed as a general corpus analogous to the Brown [14], LOB [15], BNC [16], Corpus of Contemporary Arabic (CCA) [6] and other general corpora that can be used for a variety of Linguistics and Computational Linguistics research. In the next section, we will demonstrate how KSUCCA can be used to detect various aspects of language change between Classical Arabic and MSA.

[4] The *Hijri* calendar is the official calendar for Muslims. Its first year was the year when the *Hijra*, migration, of Prophet Muhammad from *Makkah* to *Madinah* occurred, which is equivalent to 622 CE.

3. KSUCCA AND HISTORICAL LINGUISTICS

A key factor in understanding how language change is to look at the change in frequencies of the linguistic phenomenon under study. In fact, the change of frequency of a given word in time varying corpora can be an indication of historical, cultural or social changes [4]. Many Arabic words that were popular in the Classical Arabic period are used rarely in MSA. On the other hand, many new words have evolved in MSA Arabic due to cultural and social changes. In addition, noticeable drifts in the meanings of some words between the classical and the modern periods of Arabic have occurred. In this section, we will demonstrate how KSUCCA can be used to depict some lexical, Phraseological, vocabulary and semantic changes between Classical Arabic and MSA.

3.1 Lexical Change

One example of lexical change is the usage of the word (الغيث) (*Alghaith*), which is one of the synonyms of the word *rain*; this word witnessed a major drop in frequency in MSA. Table 2 and Figure 1 show the frequency rate (3/100,000) of this word in classical literature, taken from KSUCCA, compared to its frequency rate (0.58/100,000) in modern literature[5]. This decrease in frequency for the word *Alghaith* in modern Arabic literature was accompanied by an increase in the frequency rate (11.5/100,000) of another synonym of the word rain (المطر) (*Almatar*), as in Figure 1.

	Alghaith	***Almatar***
Classical Literature	3	5.62
Modern Literature	0.58	11.5

TABLE 2: Frequency rates of the words *Alghaith* and *Almatar* in classical and modern literature

These figures are of strong indication of a cultural and linguistic crisis of Arabic. This is due to the fact that the two synynoms *Alghaith* and *Almatar* are not absolute synonyms. In fact, it is belived by many ancient and contemporary Arabic linguists that there are no absolute synonyms in Arabic, and that there exists, definitely, a differnce in meaning between every pair of synonyms. This thoery applies to the two synonyms *Alghaith* and *Almatar*; the word *Alghaith* refers to the rain that falls when people and crops are of great need and thurst, and also to the rain that does not cause any damage to people, cattle, crops, property, etc. On the other hand, the word Almatar can be used to describe the rain that causes damages or the rain that does not [17]. A look at the figures in Table 2 shows a drastic decrease in the usage of the word *Alghaith* in modern literature. The use of that word decreases even further, as expected, in common daily language use as in newspapers articles where it reaches a frequency rate of (0.27/100,000)[6]; it is barely used.

This may indicate that Arabic speakers no longer taste language and their linguistic background does not allow them to use proper synonyms in their proper contexts, which results in a linguistic phenomenon known as semantic generalization. Semantic generalization is a strong sign of language decay, which has a shrinking effect on contemporary lexicons reducing the amount of their vocabulary tremendously.

[5] http://arabicorpus.byu.edu.

[6] The All Newspapers corpus is a 135,360,804 word sub corpus of arbiCorpus (http://arabicorpus.byu.edu), which contains newspapers from the period between 1996 and 2010.

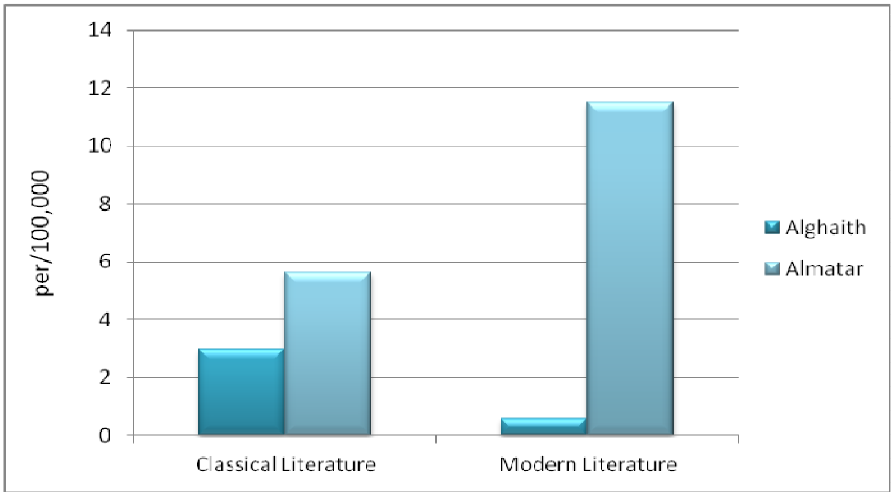


FIGURE 1: Frequency rates of the words *Alghaith* and *Almatar* in classical and modern literature.

Another example of lexical change is the emergence of the use of the word *Azya'a*, which means *fashion* in English. The frequency of this word in KSUCCA is only 1, and it was not used to mean fashion, as we know it today, it is merely the sum of the word *Zay*, which means *clothes* or *costume*. On the other hand, the word *Azya'a* is used very frequently (0.9/100,000) in contemporary newspapers with the meaning of fashion. Figure 2 shows the frequency rate of the word *Azya'a* in KSUCCA compared to its frequency rate in the All Newspapers corpus[7]. This indicates severe cultural and social changes caused by the influence of the western culture on Arabic societies, and the way that Arabic language tries to adapt and cope with these changes.

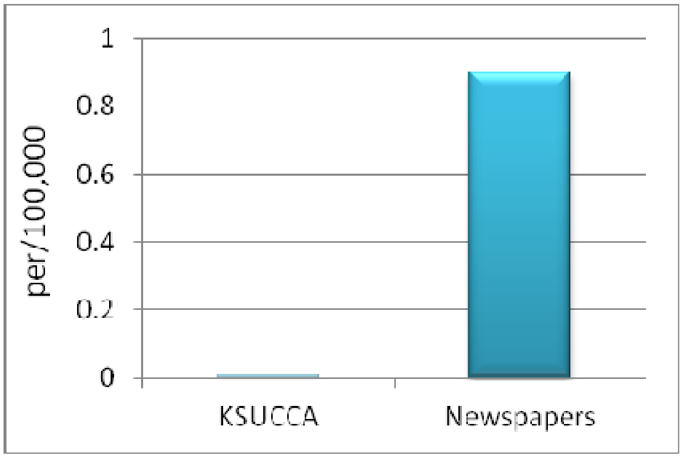


FIGURE 2: Frequency rates of the word *Azya'a* in KSUCCA and All Newspapers corpus.

3.2 Phraseological Change

Phraseology is the study of multi-word units in language, which have a range of subtypes [18]. In this section, we will discuss how KSUCCA can be used to detect phraseological changes between Classical Arabic and MSA. We will focus on collocations, which are considered one of the subtypes of the multi-word units included in phraseology. A collocation is the tendency of two or more words to appear together conveying a meaning by their association [19]. The Arabic word *Raghaba aan* (□غب عن) is an example of a collocation; it means *abstain from* in English. This

[7] http://arabicorpus.byu.edu.

collocation was common in Classical Arabic; Table 3 shows the frequency rate of its usage in KSUCCA (0.334/100,000). However, it is no longer used very frequently in MSA; it only appeared 24 times in the whole 135,360,804 words All Newspapers corpus, with a frequency rate of (0.018/100,000)[8].

	KSUCCA	Newspapers
Ragheba aan	0.334	0.018

TABLE 3: Frequency rates of the collocation *Raghiba aan.*

Figure 3 visualizes the severe difference in frequency rates of the collocation *Raghiba aan* in both corpora.

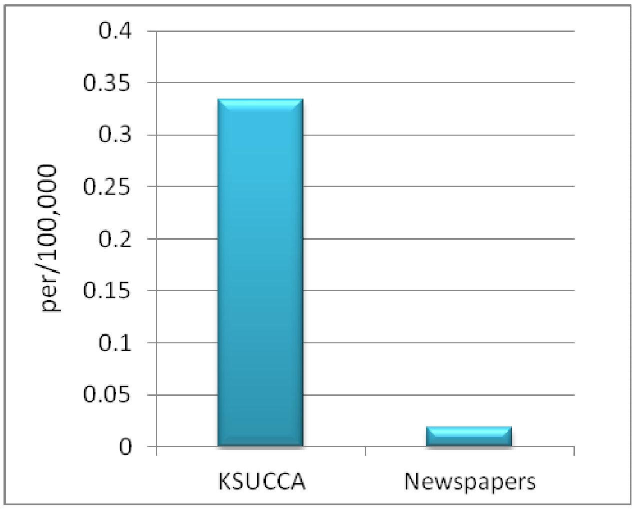


FIGURE 3: A comparison of the frequency rates of *Raghiba aan.*

There are many other collocates that were common in the classical period of Arabic and are no longer used much; they are being replaced by new ones. One of them is the collocation referring to the name of the holy mosque in *Almadinah Almonaurah* city; this mosque is commonly referred to, nowadays, as *Almasjid Alnabawy*, which means *The Prophetic Mosque*. However, a simple search for this collocation in KSUCCA would reveal nothing, because that mosque was only referred to at that period of time as *The Mosque of Allah's Apostle* or, in Arabic, as *Masjid Rasool Allah*. Table 4 shows the frequency rates of the two collocates in both KSUCCA and the All Newspapers corpus.

	Newspapers	KSUCCA
The mosque of Allah's apostle	0.01	0.47
The prophetic mosque	0.115	0

TABLE 4: The Frequency Rates of The Two Collocates.

It is clear from Figure 4 that the new collocation *Almasjid Alnabawy* is the common and most used collocation, nowadays, to refer to the holy mosque, and that the original name, *Masjid*

[8] http://arabicorpus.byu.edu.

Rasool Allah, is barely used. These two examples can be seen as a consequence of the principle of least effort, also known as Zipf's law [19], where people try to exchange long terms by shorter ones, and which is considered to be responsible for many linguistics changes [20].

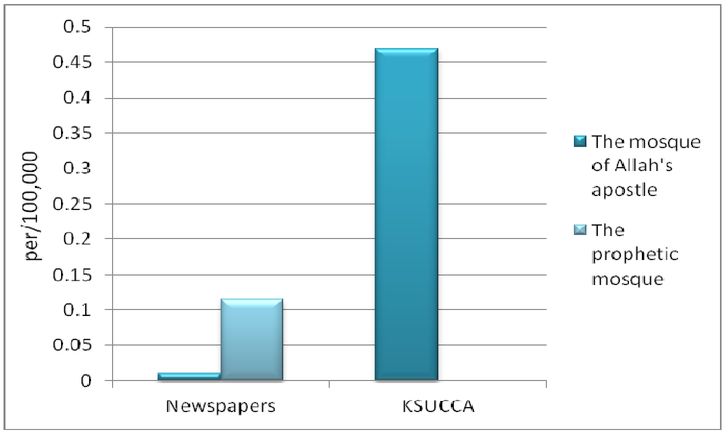


FIGURE 4: A Comparison Between The Frequency Rates of The Two Collocates.

3.3 Vocabulary Change

The Hadith, classical poems and other classical writtings are very rich in vocabulary, which is an indication of the very solid linguistic base that pople had at that era. Unfortunately, this is not the case with MSA; any regular Arabic reader can sense the decline in vocabulary wealth in MSA writings compared to Classical Arabic. One way to prove this assumption, is to study the number of roots used in samples of Classical Arabic writings from different genres and compare them to other samples from MSA with the same genres; where a root is the three letters word that form the basic source of all the forms of a given word.

We have chosen three samples, 100 word each, from KSUCCA covering the following genres: Quranic studies, philosophy and ethics and morals. To represent the MSA, we have also chosen three samples, 100 word each, falling under the same genres from the the Comprehensive library "*Almaktabah Alshamilah*" site[9] . Then we used Alkhalil morphological analyzer [21] to extract the roots of the words in each sample. Table 5 shows the numbers of unique roots extracted from each sample.

Genre	No. of roots in Classical Arabic samples	No. of roots in MSA samples
Quranic studies	58	51
Philosophy	53	49
Ethics and morals	47	31

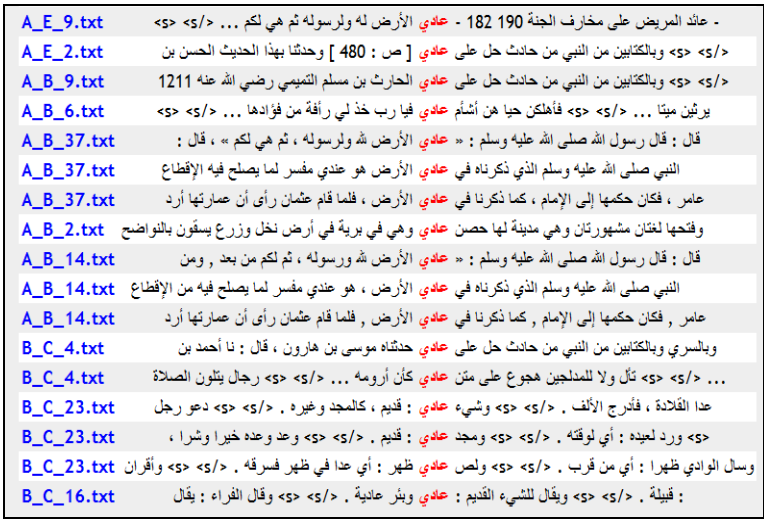
TABLE 5: Number of Unique Roots In Each Sample.

It is obvious that the number of roots in the Classical Arbic samples are larger than their equivalent samples from MSA. This can be considered as an evidence of the decline of the average vocabulary in MSA writings, which is another sign of language decay.

[9] http://shamela.ws

3.4 Semantic Change

Many Arabic words went through a series of semantic changes from the classical period until now, and sometimes their meanings were completely altered. One of these words is the word *Aady* (عادي), which was originally used to mean *old* or *an aggressor.* This word was not common in Classical Arabic, which is confirmed by looking at the concordance[10] of the word *Aady* in KSUCCA in Figure 5. The word is used with law frequency rate as in Table 6.



A_E_9.txt	- عائد المريض على مخارف الجنة 190 182 - عادي الأرض له ولرسوله ثم هي لكم ... </s> <s>
A_E_2.txt	</s> <s> وبالكتابين من النبي من حادث حل على عادي [ص : 480] وحدثنا بهذا الحديث الحسن بن
A_B_9.txt	</s> <s> وبالكتابين من النبي من حادث حل على عادي الحارث بن مسلم التميمي رضي الله عنه 1211
A_B_6.txt	يرثين ميتا ... </s> <s> فأهلكن حيا هن أشأم عادي فيا رب خذ لي رأفة من فؤادها ... </s> <s>
A_B_37.txt	قال : قال رسول الله صلى الله عليه وسلم : « عادي الأرض لله ولرسوله ، ثم هي لكم » ، قال :
A_B_37.txt	النبي صلى الله عليه وسلم الذي ذكرناه في عادي الأرض هو عندي مفسر لما يصلح فيه الإقطاع
A_B_37.txt	عامر ، فكان حكمها إلى الإمام ، كما ذكرنا في عادي الأرض ، فلما قام عثمان رأى أن عمارتها أرد
A_B_2.txt	وفتحها لغتان مشهورتان وهي مدينة لها حصن عادي وهي في برية في أرض نخل وزرع يسقون بالنواضح
A_B_14.txt	قال : قال رسول الله صلى الله عليه وسلم : « عادي الأرض لله ورسوله ، ثم لكم من بعد , ومن
A_B_14.txt	النبي صلى الله عليه وسلم الذي ذكرناه في عادي الأرض ، هو عندي مفسر لما يصلح فيه من الإقطاع
A_B_14.txt	عامر , فكان حكمها إلى الإمام , كما ذكرنا في عادي الأرض , فلما قام عثمان رأى أن عمارتها أرد
B_C_4.txt	وبالسري وبالكتابين من النبي من حادث حل على عادي حدثناه موسى بن هارون ، قال : نا أحمد بن
B_C_4.txt	... </s> <s> تأل ولا للمدلجين هجوع على متن عادي كأن أرومه ... </s> <s> رجال يتلون الصلاة
B_C_23.txt	عدا القَلادة ، فأدرج الألف . </s> <s> وشيء عادي : قديم ، كالمجد وغيره . </s> <s> دعو رجل
B_C_23.txt	<s> ورد لعيده : أي لوقته . </s> <s> ومجد عادي : قديم . </s> <s> وعد وعده خيرا وشرا ،
B_C_23.txt	وسال الوادي ظهرا : أي من قرب . </s> <s> ولص عادي ظهر : أي عدا في ظهر فسرقه . </s> <s> وأقران
B_C_16.txt	: قبيلة . </s> <s> ويقَال للشيء القَديم : عادي وبئر عادية . </s> <s> وقال الفراء : يقَال

FIGURE 3: A Concordance of The Word Aady From KSUCCA.

However, the meaning of this word have drift from time to time until it is used now in MSA to mean *normal* or *ordinary*. This drift in its meaning was accompanied, of course, by an increase in its frequency rate nowadays compared to its limited use in Classical Arabic, as can be seen in Table 5.

	KSUCCA	Newspapers
Aady	0.9	6.25

TABLE 6: The Frequency Rates of The Word *Aady.*

4. CONCLUSION

Most of the previous work on Arabic historical linguistics was not corpus-based; one major reason for that is the fact that there are no available corpora of Classical Arabic that are large enough to conduct such studies. In addition to the fact that most traditional Arabic linguists are not familiar with the use of computerized corpora in language researches.

In this paper, we present KSUCCA a pioneering 50 million words corpus of Classical Arabic with various genres and sub genres that can be used in various types of Linguistic and Computational Linguistic research. We also showed how can KSUCCA be used to detect some interesting lexical, Phraseological, vocabulary and semantic changes in Arabic language. We believe that the construction of KSUCCA and the work presented in this paper will encourage Arab linguists to take steps forward in exploring corpus-based historical linguistics and discovering other interesting aspects of language change.

[10] Using Sketch Engine (http://www.sketchengine.co.uk).

5. REFERENCES

[1] T. Bynon. Historical Linguistics. Cambridge University Press, 1977.

[2] C. F. Meyer. English corpus linguistics: An introduction. Cambridge University Press, 2002.

[3] M. Kytö. "Manual to the diachronic part of the Helsinki corpus of English texts, 3rd ed." University of Helsinki, 1996.

[4] M. Davies. "Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English." Corpora, 2012.

[5] M. Mansour. "The absence of Arabic corpus linguistics: a call for creating an Arabic national corpus." International Journal of Humanities and Social Science, vol. 3, no. 12, 2013.

[6] L. Al-Sulaiti and E. Atwell. "The design of a corpus of contemporary Arabic." International Journal of Corpus Linguistics, vol. 11, pp. 135-171, 2006.

[7] L. Al-Sulaiti and E. Atwell. "Extending the Corpus of Contemporary Arabic." In Proceedings of Corpus Linguistics conference, University of Birmingham, UK, 2005.

[8] S. Alansary, N. Magdi and N. Adly. "Building an international corpus of Arabic (ICA): Progress of Compilation Stage." In 7th Int. Conference on Language Engineering, Cairo, Egypt, pp.1-30, 2007.

[9] M. Sawalha and E. Atwell. "Constructing and Using Broad-coverage Lexical Resource for Enhancing Morphological Analysis of Arabic." In proceeding of: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, 2010.

[10] A. Alfaifi and E. Atwell. "Arabic Learner Corpus v1: A New Resource for Arabic Language Research." In proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster University, UK, 2013.

[11] A. Elewa. "Did they translate the Qur'an or its exegesis?." 3rd Languages and Translation Conference and Exhibition on Translation and Arbization in Saudi Arabia, Riyadh, Saudi Arabia, 2009.

[12] M. Eid. Manifestations Emerging on Arabic. A'alam Alkutub, Cairo, pp. 20, 1980.

[13] M. Alrabiah, A. Al-Salman and E. Atwell. “The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic." In Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster University, UK, Monday 22nd July 2013.

[14] W. N. Francis and H. Kucera. "Brown Corpus Manual: Manual Of Information To Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers." Internet: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM, 1964 [March. 2, 2014].

[15] S. Johansson, E. Atwell, R. Garside and G. Leech. "The Tagged LOB Corpus: Users' manual." ICAME, The Norwegian Computing Centre for the Humanities, Bergen University, Norway, 1986.

[16] L. Burnard. "British National Corpus: User's reference guide for the British National Corpus." Oxford, Oxford University Computing Service, 1995.

[17] A. Alaskari, Linguistic Differences. (in Arabic), Dar Alkutub Alelmiah, 2010.

[18] S. Granger and F. Meunier. Phraseology: An Interdisciplinary Perspective. Amsterdam: John Benjamins, 2008.

[19] C.D. Manning and H. Schuetze. Foundations of Statistical Natural Language Processing, 1st ed., The MIT Press, 1999.

[20] C.M. Millward. A Biography of the English Language. 2nd ed. Harcourt Brace, 1996.

[21] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. Ould Abdallahi Ould Bebah and M. Shoul. "Alkhalil MorphoSys: A Morphosyntactic analysis system for non vocalized Arabic." Seventh International Computing Conference in Arabic (ICCA 2011), Riyadh, 2011.