Proceedings Paper:

Alrabiah, M, Al-Salman, A and Atwell, ES (2013) The design and construction of the 50 million words KSUCCA. In: Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics. The WACL'2 Second Workshop on Arabic Corpus Linguistics, 22 Jul 2013, Lancaster University, UK. The University of Leeds , 5 - 8.

The Design and Construction of the 50 Million Words KSUCCA

Maha Alrabiah
King Saud University
msrabiah@gmail.com

AbdulMalik Al-Salman
King Saud University
salman@ksu.edu.sa

Eric Atwell
Leeds University
e.s.atwell@leeds.ac.uk

1 Introduction

In this paper, we report the design and construction of King Saud University Corpus of Classical Arabic (KSUCCA) [1], which is part of ongoing research that attempts to study the meanings of words used in the holy Quran, through analysis of their distributional semantics in contemporaneous texts.

The holy Quranic text was revealed in pure Classical Arabic, which forms the basis of Arabic linguistic theory and which is well understood by the educated Arabic reader. Therefore, it is necessary to investigate the distributional lexical semantics of the Quran's words in the light of similar texts (corpus) that are written in pure Classical Arabic.

To the best of our 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)[2] and the other is the Classical Arabic Corpus (CAC) (Elewa, 2009). However, neither of the two corpora is adequate for our research; the former does not cover many genres such as: Linguistics, Literature, Science, Sociology and Biography; and it only contains 17+ million words, so it is not very large. While the latter is even smaller with only 5 million words. Therefore, we made an effort to carefully design and compose our own corpus 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 addition, we tried to make the design general enough in order to make the corpus also appropriate for other research.

2 Purpose of KSUCCA

The main purpose of KSUCCA was to be used for studying the distributional lexical semantics of words in the holy Quran. However, it was designed as a general corpus analogous to the Brown, LOB, BNC, Corpus of Contemporary Arabic (CCA) and other general corpora that can be used for a variety of Linguistics and Computational Linguistics research, such as (Atkins et al. 1992), (Dash 2008) and (Waly 2012). This corpus could be utilized for:

- Building lexicons.
- Building semantic lexicons.
- Looking for linguistic treasures including forgotten vocabulary, synonyms and syntactic structures, and help in spreading them back in current education and literature.
- Studying language (vocabulary and syntactic structures) change through history.
- Studying terminologies and idioms.
- Studying collocations.
- Studying difference between synonyms.
- Studying how certain words relate to specific syntactic structures.
- Studying the relation between words and text type.
- Studying the relation between syntactic structure and text type.
- Building parallel corpora that can be used for automatic translation.
- Studying the distribution of roots and morphological patterns in different types of texts and how this is related to their meaning.
- Studying semantic relations to build a WordNet for the Holy Quran.
- Building an ontology for the Holy Quran and Classical Arabic.
- Studying the linguistic styles of authors, which can be used in identifying unknown authors.
- Studying the linguistic style of a specific book or poem.
- Studying language difference between texts from different genres.
- Language learning for Arabs and non Arabs, and building special lexicons for language teaching.
- Building question answering systems.
- Building specialized information retrieval systems.

3 KSUCCA Representativeness and Balance

The main criterion for selecting the population of data of interest was to include only pure Classical Arabic texts, which are Arabic texts dating back to the period of the pre-Islamic era until the end of the fourth Hijri[3] century (equivalent to the period from

[1] The corpus site is available online in English (http://ksucorpus.ksu.edu.sa), and in Arabic (http://ksucorpus.ksu.edu.sa/ar)

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

[3] The Hijri calendar is the official calendar for Muslims. Its first year was the year when the Hijra, migration, of Prophet

the seventh until early eleventh century CE) (Eid, 1980). Due to the limited number of such available texts, we did not use random sampling to select the sources from the population; instead, we tried to gather as many authenticated Classical Arabic texts as we could. One of the major resources we have relied on was the Ccomprehensive library "Almaktabah Alshamilah" website[1], which is a voluntary project accomplished by the cooperation between the site's owners and Alrawdhah Cooperative Office for Call and Guidance[2]. The goal of that project was to collect and digitize a wide range of Islamic and Arabic classical and modern books.

Table 1: The classification of population into genres and subgenres

Genre	Subgenre
Religion	The Holy Quran
	Hadith
	Exegesis of The Quran
	Quranic Studies
	Hadith Studies
	Belief
	Jurisprudence
	Principles of Jurisprudence
Linguistics	Grammar and Morphology
	Language
	Lexicons
	Proverbs
Literature	Poetry
	Novels
	Literature and Eloquence
Science	History
	Geography and Travel
	Medicine
	Physics
	Astronomy
	Philosophy
	Politics
	Miscellaneous
Sociology	Ethics and Morals
	Genealogy
Biography	Prophet Muhammad Peace be upon him biography
	Other biographies

After investigating the different domains of available texts, and in order to achieve corpus representativeness, we classified the population into 6 broad genres covering most of the topics that were popular in that period of time. These are: Religion, Linguistics, Literature, Science, Sociology and Biography. We further classified these genres into 27 subgenres as shown in Table 1.

Later on, we classified texts into their appropriate genre/subgenre. The result of the classification revealed that texts were not evenly distributed between genres. However, this is consistent with our knowledge of the overall writing trends at that period of Arab history, and it is an indication of the balance of our corpus. Table 2 shows the number of texts included in each genre together with their corresponding number of words, and the portion of words in each genre compared to the whole 50 million words corpus.

Table 2: A general description of KSUCCA content

Genre	Number of texts	Number of words	Percentage
Religion	150	23645087	46.73 %
Linguistics	56	7093966	14.02 %
Literature	104	7224504	14.28 %
Science	42	6429133	12.71 %
Sociology	32	2709774	5.36 %
Biography	26	3499948	6.92 %
Total	410	50602412	100 %

4 Sampling size

Since KSUCCA was designed to study distributional lexical semantics of words in the holy Quran, and Classical Arabic in general, it was sampled as "full text", where the whole book or poem text was considered as a sample. This is more suitable for detecting the linguistic features and meanings that may be distributed evenly throughout the text as suggested by Sinclair (2005). KSUCCA was initially designed as a raw corpus containing only plain text with no tagging or lemmatization. The appendix shows a sample of the corpus text.

5 Copyright permission

The materials included in this corpus date back to more than ten centuries ago, so they do not require any copyright permissions.

6 Character Encoding

We chose to use UTF-8 for character encoding for all the corpus files since it is backward compatible with ASCII and can be work with available software (McEnery and Xiao, 2005).

Muhammad Peace be upon him from Makkah to Madinah occurred, which is equivalent to 622 CE.

[1] http://shamela.ws

[2] www.arrawdah.com

7 File Organization

The corpus is composed of text files where each file represents a single document (a book or a poem). These files are organized into six main genres, and each genre is organized into a number of subgenres. Both the genres and the subgenres are given unique alphabetical codes. Files are named with the combination of their genre and subgenre codes together with their number in the subgenre. Table 3 shows the genres and subgenres codes.

Table 3: Corpus files names

Genre (code)	Subgenre (code)	Files names
Religion (A)	The Holy Quran (A)	AA1
	Hadith (B)	AB1-44
	Exegesis of The Quran (C)	AC1-13
	Quranic Studies (D)	AD1-29
	Hadith Studies (E)	AE1-10
	Belief (F)	AF1-23
	Jurisprudence (G)	AG1-26
	Principles of Jurisprudence (H)	AH1-4
Linguistics (B)	Grammar and Morphology (A)	BA1-16
	Language (B)	BB1-6
	Lexicons (C)	BC1-27
	Proverbs (D)	BD1-7
Literature (C)	Poetry (A)	CA1-42
	Novels (B)	CB1-2
	Literature and Eloquence (C)	CC1-60
Science (D)	History (A)	DA1-19
	Geography and travel (B)	DB1-14
	Medicine (C)	DC1-3
	Physics (D)	DD1
	Astronomy (E)	DE1-2
	Philosophy (F)	DF1
	Politics (G)	DG1
	Miscellaneous (H)	DH1
Sociology (E)	Ethics and Morals (A)	EA1-23
	Genealogy (B)	EB1-9
Biography (F)	Prophet Muhammad Peace be upon him biography (A)	FA1-8
	Other biographies (B)	FB1-18

8 An example: analysing the usage of the word water

We said that KSUCCA can be used to depict how words in the Quran were used in general Classical Arabic. Let us take as an example the word "ماء", which means water in English; this word was mentioned 35 times in the Quran with three different meaning as indicated in Table 4.

Table 4: Meanings of the word water in the Quran

Meaning	Co-occurring words	Number of occurrences
rain	send down, sky	22
regular water	drink	10
semen	creation, human beings, worthless water, gushing water	3

We used Sketch Engine[1] to extract collocations of the word "ماء" in KSUCCA. The MI.log_f statistical measure showed the most significant collocations, and by analyzing the highly ranked among them, we can come out with the meanings on Table 5.

Table 5: Meanings of the word water from KSUCCA

Meaning	Collocates	MI.log_f
plants boiled in water	barley "الشعير"	63.325
drinks mixed with water	honey "العسل"	54.434
Rain	sky "السماء"	54.210
regular water	find "تجدوا"	51.729
The water from the well of Zamzam	Zamzam "زمزم"	50.467
Semen	gushing "دافق"	46.671
Juice	Pomegranate "الرمان"	46.055

Table 5 shows that the word "ماء" was used to convey the same meanings that appeared in the Quran along with other meanings depending on its co-occurring words. The high ranks of the first two collocates are due to the availability of several books on medicine in KSUCCA, and these used the word water more often than regular text.

9 Conclusion

KSUCCA is a pioneering 50+ million word corpus that captures the culture of a nation. In fact, it is a project with scientific, linguistic, cultural, social and religious aspects. It was designed and compiled

[1] http://www.sketchengine.co.uk/

with the goal of supporting research in both Linguistics and Computational Linguistics; however, it can also be used by researchers from other disciplines such as Literature and History.

KSUCCA has a markedly different selection of genres compared to other balanced corpora such as Brown, LOB, BNC, or CCA; this reflects the fact that much of the written text from that period of Arab history was inspired by the Quran and the growth of Islam. Nearly half of all the texts from this time were Religious; Linguistics texts developed largely to help non-Arabic speakers access the Quran and related Islamic texts and to preserve the Arabic language from being distorted by foreigners who embraced Islam and joined the Arabic community; many of the Biography texts focused on the prophet Muhammad and Islamic scholars; and many of the Science, Sociology and even Literature texts related to and/or drew from the Quran, Hadith and other Islamic sources.

KSUCCA will allow us to see how words in the Quran were used in Arabic at that time, as demonstrated in the analysis above. For a given word in the Quran, we can extract concordance and collocations of that word in KSUCCA, enabling a distributional semantic analysis of the word's meanings and contexts. We can also compare Quran vocabulary and language against contemporaneous Classical Arabic vocabulary and language, to highlight words and linguistic constructs which stand out.

KSUCCA will give scholars a new perspective on the analysis of the language of the Quran.

References

Atkins, S. Clear, J. and Ostler, N. (1992). "Corpus Design Criteria". Literary and Linguistic Computing, 7, 1: 1-16.

Dash, N.S. 2008. "Corpus Linguistics: An Introduction", Addison-Wesley Longman, Limited.

Eid, M. 1980. Manifestations Emerging on Arabic. A'alam Alkutub, Cairo, 20.

Elewa, A. 2009. "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.

McEnery, A. and Xiao, R. 2005. Character Encoding in Corpus Construction. In Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books: 47-58, Available online at http://ahds.ac.uk/linguistic-corpora/

Sinclair, J. 2005. Corpus and Text - Basic Principles. In Developing Linguistic Corpora: a Guide to Good Practice, ed. M. Wynne. Oxford: Oxbow Books: 1-16, Available online at http://ahds.ac.uk/linguistic-corpora/

Waly, A. 2012. "How To Build a Computational Linguistic Corpus" (In Arabic). A workshop for Al-Jazirah Research Chair for Modern Linguistics, Princess Noura Bint AbdulRahman University.

Appendix: sample from the KSUCCA corpus

الحمد لله رب العالمين الذي أنزل على عبده الكتاب ليكون للعالمين
نذيرا. والصلاة والسلام على سيدنا محمد بن عبد الله الذي أرسله الله
تعالى رحمة للناس وآتاه الحكمة وجوامع الكلم وعلمه ما لم يكن يعلم
وكان فضل الله عليه عظيما وعلى آله وصحبه ومن تبعهم بإحسان
إلى يوم الدين

أما بعد فإن السنة هي المصدر التشريعي الثاني - من المصادر
المتفق عليها لدى المسلمين - بعد كتاب الله عز وجل فهي أصل من
أصول الدين ومنهل خصيب للتشريع ودليل أساسي من أدلة الأحكام
تعرفنا حكم الله سبحانه وتعالى في كل كبير وصغير فهي جامعة
مانعة عامة شاملة لا تفوتها شاردة ولا واردة إلا وقد أعطتها حكما
شرعيا فيها بيان لما كان وما سيكون وفيها تنظيم عملي رائع لشؤون
الحياة مستوحى عن الله تعالى خالق الحياة ومن يحيا ومرتبط بمالك
الملك والملكوت الذي لا يعزب عنه مثقال ذرة في الأرض ولا في
السماء. فقلما تحدث حادثة أو تنزل نازلة إلا ونجد في السنة المطهرة
الحكم الشافي والبيان الوافي لها. وذلك أن رسول الله صلى الله عليه
وسلم هو المبلغ عن ربه {يا أيها الرسول بلغ ما أنزل إليك من ربك}