



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/125569/>

Version: Accepted Version

Proceedings Paper:

Alosaimy, A and Atwell, E (2017) Sunnah Arabic Corpus: Design and Methodology. In: Proceedings of the 5th International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2017). IMAN 2017, 26-28 Dec 2017, Semarang, Indonesia.

This is an author produced version of the paper 'Sunnah Arabic Corpus: Design and Methodology', presented at IMAN 2017.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Sunnah Arabic Corpus: Design and Methodology

Abdulrahman Alosaimy¹, Eric Atwell²

School of Computing, University of Leeds, United Kingdom

¹scama@leeds.ac.uk, ²e.s.atwell@leeds.ac.uk

ABSTRACT

Sunnah Arabic Corpus is an annotated linguistic resource that consists of 144K words/170K tokens of the Hadith narratives (an utterance attributed to prophet Mohammed) extracted from Riyāḍu Aṣṣāliḥīn book. As a first layer of annotation, the corpus has been fully diacritized. In addition, each orthographic word/token is segmented into its syntactic words. And each syntactic word is tagged with its part-of-speech in addition to multiple morphological features. Several hadith translations in different languages are provided and aligned at the narrative/paragraph level. Hadith Arabic Corpus follows the successful Quranic Arabic Corpus in its standards (corpus.quran.com). Sunnah Arabic Corpus is freely available under the Creative Commons Attribution-ShareAlike 4.0 International License

Keywords: Arabic, corpus, annotation, Hadith, Sunnah, morphology

1. Introduction and Motivation

Language resources (LRs) are recognized as key components in the development of Natural Language Processing. Annotated corpora, as one example of LR, used to perform statistical analysis, hypothesis testing, verifying grammar within a language domain and for building statistical computational models. Several scholars show the lack of freely available Arabic resources, especially gold standard annotated corpora (Albared, Omar, & Ab Aziz, 2009; Yaseen et al., 2006). In the case of Classical Arabic, Quranic Arabic Corpus (Dukes & Habash, 2010) and Al-Mus’haf Corpus (Zeroual & Lakhouaja, 2016) are two available annotated corpus but they are limited to Quranic texts. Mohamed (2012) built a small corpus of religious texts (and one of the texts is a small Hadith book) and confirmed the need of a larger classical corpus.

Sunnah Arabic Corpus (SAC) is the first corpus of Arabic Hadith (prophet sayings) that is freely available morpho-syntactically annotated corpus using a fine-grained tagset that conforms with traditional Arabic grammar. It follows the success of the Quranic Arabic Corpus (QAC).

In the rest of the paper, we will list the main features and potential uses of SAC, its collection process and content, and finally its availability and accessibility.

2. Corpus Collection Process

Sunnah Arabic Corpus currently has only one book: Riyāḍu Aṣṣāliḥīn (aka The Meadows of the Righteous) a compilation of 1896 hadith narratives written by Al-Nawawi and published on 1334. The book will henceforth be referred to as Riyad. Riyad was chosen for several reasons:

1. It has widely accepted as a valid source of prophet sayings.
2. Its codex was validated and investigated by several scholars by scientific palaeographical process.

3. A small part (42 narratives, 4479 words, ~5% of the book), famously known as Al-Arbaeen Al-Nawawyyah, has been studied linguistically in traditional Ia'rab Arabic grammar (by two books)AlOmari, 2005; Yosef, 2003)).
4. It has been translated into at least 18 languages (e.g. English).
5. It's narratives has been explained by 6 written books, at least by 11 scholar (spoken explanation).

While it is available through many websites (e.g. IslamHouse.com, Sunnah.com), we chose to download an e-book version of the book from Shamela¹, a downloadable library that contains at least 5300 Arabic books in Islamic studies , as this library become the standard library of Arabic classical books. It has been used to obtain Arabic classical text in building several corpora (Alrabia, Al-Salman, Atwell, & Alhelewh, 2014; Belinkov, Magidow, Romanov, Shmidman, & Koppel, 2016; Zerrouki & Balla, 2017).

Two versions of Riyad were available in Shamela library and we chose the version with ID# 2348 (Alfahal, 2007). This version is the one investigated by Maher Alfahal who made his investigation and commentaries open freely. Both versions has same numbering, hadith text (with some slight differences), but they both differ greatly in the commentaries.

Diacratization of both version is not full (not every letter has its short vowel); Maher's version is more thorough and accurate using a sample of five narratives randomly chosen. Quranic verses are fully diacratized in both versions (a common standard in book editing).

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
"http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xml:lang="ar" lang="ar" dir="rtl"
xmlns="http://www.w3.org/1999/xhtml"
xmlns:epub="http://www.idpf.org/2007/ops">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
<link href="../style.css" rel="stylesheet" type="text/css" />
<title>رياض الصالحين ت الفحل</title></head>
<body class="rtl"> <div dir="rtl" id="book-container"><hr/>
<a id='C159'></a><a id='C160'></a>
<span class="title">(6) - كتاب عيادة المريض وتشييع الميِّت والصلاة عليه
</span><span class="red">144 -
</span><span class="title">باب عيادة المريض</span><br /><span
class="red">894 - </span> عن البراء بن عازب رضي الله عنهما، قال: أمرنا رسولُ
الله - صلى الله عليه وسلم - بعيادة المَريض، وَاتِّبَاعِ الجَنَازَةِ، وَتَشْمِيَةِ العَاطِسِ،
وَإِبْزَارِ المُقْسِمِ، وَنَصْرِ المَظْلُومِ، وَإِجَابَةِ الدَّاعِي، وَإِفْشَاءِ السَّلَامِ. متفقٌ عليه.
1)><span class="footnote-hr">&nbsp;</span><span class="footnote">(1)
239) انظر الحديث</span>
</div><hr/>
<div class="center"> الحديث: 894 | الجزء: 1 | الصفحة:
273</div></body></html>
```

Figure 7.2.1: XML version of one page of Riyad book extracted from its EPUB version.

¹ <http://shamela.ws/index.php/book/2348>

Shamela books are available in three formats: PDF, EPUB, and BOK (used for their downloadable desktop software). PDF format is used for the scanned images of the book, and the text is not extractable easily. BOK version is not suitable as it requires their software to open. Therefore, we chose to proceed with EPUB format, an e-book file cross-platform widely-used format to view and read the book. Since EPUB format is XML-based, the extraction of the xml version of the book is easy.

We chose to use the EPUB version. However, we found that the xml version of Riyad does not tell the difference between some components of the text (like footnotes' co-reference, and page numbers). It does not separate the chain of narrators, prophet sayings, and citation. It does not either provide a table of contents (see Figure 7.2.1 for details). Therefore, we developed a custom software² to extract the narratives and verses in a structured format which identifies footnotes, chapters, and sections, remove inline annotations (for example page break of original book: “[p.34 started]”) and separate footnotes from original text and link it with co-reference. It also merge narratives that spans into multiple pages. It also import Quranic units annotations from QAC corpus. Section 7.3 describes the output of the software, i.e. corpus content.

3. Corpus Content

Riyad is a collection of 2330 units (precisely 435 Quranic verses and 1896 hadith narratives). It is classified into 20 chapters, and each chapter contains several sections, with a total of 372 sections the covers Islamic morals, acts of worship, and manners. Each section covers a specific topic, and verses and narratives that support the topic.

The corpus consist of ~119K *Arabic* words: which about 110K words compose Hadith narratives. The rest are either compose authors commentaries, his introduction, or Quranic verses. More statistics about the corpus are in Table 1. After removing short vowels and punctuations, the number of word types of hadith narratives is ~17K. The word frequency list is presented in Table 2.

Table 1: Some statistics about the Sunnah Arabic Corpus.

Counts		Counts	
Tokens	170453	Word Types	17786
Words	144106	Fully diacritized Words	102746
Sentences	7670		86.08%
Paragraphs	2075	Distinct 5-grams	90347
Documents	372	Hadith Narratives	1896

Table 2: The frequency list of Sunnah Arabic Corpus.

Word	Count	Word	Count	Word	Count	Word	Count
الله	7883	إلى	528	هذا	251	قلت	160
عليه	3476	يا	528	تعالى	243	أنس	157
قال	3182	كان	523	يوم	243	بها	157
صلى	2528	له	516	أي	237	و	145
وسلم	2470	ولا	450	عمر	218	فإذا	142
من	2068	حتى	447	فيه	213	منه	141
رسول	1990	إن	412	الذي	212	بين	139

² http://github.com/aosaimy/wasim/external_tools/

رضي	1856	أو	405	صحيح	210	وما	138
عنه	1419	إذا	373	لي	206	النار	137
في	1417	أبو	347	لم	206	عائشة	134
وعن	1406	وقال	339	الترمذي	206	أنه	132
أن	1173	ابن	337	قالت	205	فلما	130
رواه	1151	هريرة	336	رجل	202	شيء	130
ما	876	عنهما	331	فإن	200	والله	128
لا	867	حديث	319	وهو	195	إني	127
فقال	833	يقول	303	هو	194	مع	124
بن	829	حسن	297	عنها	192	فقلت	124
عن	825	عبد	282	الجنة	192	الرجل	123
على	808	رواية	275	قد	188	الصلاة	122
أبي	804	وفي	274	وعنه	185	أهل	122
متفق	747	البخاري	269	كل	177	أحد	121
النبي	650	به	266	اللهم	174	أحدكم	120
مسلم	615	ذلك	266	وإن	174	علي	118
ثم	564	الناس	258	ومن	169	عند	118
إلا	530	داود	255	سمعت	165	شينا	117

For each unit, we keep a record of its numbering, chapter and page in the original printed book, and automatically split its text into sentences: with tags to describe the purpose of that sentence using a simple rule based segmenter.

Hadith units were POS tagged and annotated semi-automatically using Wasim annotation tool, and Quranic units are matched with their annotation of QAC annotation. The format for storing annotation is CoNLL-U format v2.0, which is used by the Universal Dependencies project (Nivre & Agic, 2017). We used a slightly modified tagset of the Quranic Arabic Corpus.

4. Potential Uses

- It can help Arabic learner by understanding the interaction of sentence components, since it follows Arabic grammar of 'i' rāb (إعراب).
- It may help linguistic researchers interested in Hadith to study the stylistic and vocabulary and other linguistic studies.
- It can help researchers in translation studies to compare different translations of same Hadith.
- For researchers of Arabic Language Processing, it may help improving machine translation and Classical Arabic understanding using morphological and syntactical annotations.

5. Main features

The project aims to provide an easy-to-use online access to:

- Fully diacritized text
- A part-of-speech concordance search results organized by lemma or surface form.

- A morpheme-based part-of-speech tagged corpus with its morphological features³.
- I'rāb of a sample of hadiths in a novel visualized way⁴.
- Morphological and lemma-based search for the corpus.
- Narrator-based search for the corpus.
- A parallel text of English-Arabic aligned on the hadith level.
- A parallel text of Arabic-Arabic commentaries aligned on the hadith level.

6. A Layer of Annotation: Diacritization

Since Riyad is highly cited, we have fully diacritized the text, by automatically “borrowing” diacritization from other books. We developed a software that matches the undiacritized version of one word in Riyadh with its equivalent in other books using their word n-gram concordance. The original source of Riyad is about 47.1% fully diacritized⁵, and after borrowing diacritization, the percentage jumps to 86.08%. In terms of letters, 48.66% was diacritized, and after borrowing diacritization, the percentage jumps to 76.41% with low diacritic error rate (0.004).

7. Accessibility and Availability

Sunnah Arabic Corpus is freely available under the Creative Commons Attribution-ShareAlike 4.0 International License. This permissive license allows commercial uses and allows adaptations be shared as long as others share alike. The corpus will also be available online⁶ soon which allows an easy-to-use corpus functionalities.

8. Conclusion and Future Work

We have introduced the corpus and described its collection process, content, and its distribution and availability. We described briefly its potential uses in different fields including linguistics studies, natural languages processing, and translation studies.

We are working on the morphological and syntactical annotation of the corpus, and plan to publish its data, guidelines and evaluation soon. For future work, it will be very helpful to manually align the corpus to different languages/commentaries in sentence level. In addition, word-to-word translation proved to be helpful in Quran understanding, and we might consider automatic word alignment with other languages.

9. References

- Albared, M., Omar, N., & Ab Aziz, M. J. (2009). Arabic part of speech disambiguation: A survey. *International Review on Computers and Software*, 4(5), 517–532.
- Alfahal, M. Y. (2007). *Riyad-us-Saliheen (with commentary on Ahadith)*. Dar Ibn Katheer, Damascus, Syria.
- AlOmari, O. (2005). *إعراب الأربعين النووية - The Iaarab Of The Nawawi Forty Book*.

³ Manually verified version will be available next year.

⁴ Rest of Hadiths will be added later.

⁵ The definition of fully diacritized form is hard to compute accurately. For example, using our diacritization standard, the letter the precedes a long vowel are not diacritized. However, deciding whether Yaa letter is consonant or a vowel is ambiguous.

⁶ <http://corpus.al-osaimy.com>

- Alrabia, M., Al-Salman, A., Atwell, E., & Alhelewh, N. (2014). KSUCCA : A Key To Exploring Arabic Historical Linguistics. *International Journal of Computational Linguistics*.
- Belinkov, Y., Magidow, A., Romanov, M., Shmidman, A., & Koppel, M. (2016). Shamela: A Large-Scale Historical Arabic Corpus. *arXiv Preprint arXiv:1612.08989*.
- Nivre, J., & Agic, L. ~Zeljko. (2017). Universal dependencies 2.0 CoNLL 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied.
- Yaseen, M., Attia, M., Maegaard, B., Choukri, K., Paulsson, N., Haamid, S., ... Ragheb, A. (2006). Building Annotated Written and Spoken Arabic LR's in NEMLAR Project. *Lrec*, 533–538.
- Yosef, H. A. (2003). *إعراب الأربعين حديثًا النبوية - The Iarab of Al-Nawawi's Forty Hadith*. Cairo: AlMukhtar.
- Zeroual, I., & Lakhouaja, A. (2016). A new Quranic Corpus rich in morphosyntactical information. *International Journal of Speech Technology*, 1–8.
- Zerrouki, T., & Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11, 147–151. <http://doi.org/10.1016/j.dib.2017.01.011>

Abstract in Arabic

ذخيرة السنة العربية هي مصدر لغوي موسوم صرفياً ونحوياً يتكون من ما يقارب 144 ألف كلمة / 170 ألف رمز من الأحاديث النبوية الشريفة المستخرجة من كتاب رياض الصالحين. كطريقة أولى من الشرح، فقد تم تشكيل النص بشكل كامل. بالإضافة إلى ذلك، تم تقسيم كل كلمة إملائية إلى كلماتها النحوية. ثم تم توسيم كل كلمة نحوية بجزء الكلام المناسب بالإضافة إلى العديد من الخصائص الصرفية. بالإضافة لذلك، تحتوي الذخيرة على مجموعة من الترجمات بلغات مختلفة متوازية مع النص العربي للحديث الشريف على مستوى الحديث/الفقرة. هذه الذخيرة تتبع ذخيرة القرآن الكريم من جامعة ليدز في معايير التوسيم (corpus.quran.com). هذه الذخيرة متوفرة بشكل حر ومجاني عبر موقع الذخيرة على شبكة الإنترنت.