Data in Brief 11 (2017) 147–151

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems

Taha Zerrouki*, Amar Balla

The National Computer Science Engineering School (ESI), Algiers, Algeria

ARTICLE INFO

Article history:
Received 16 September 2016
Received in revised form
30 December 2016
Accepted 27 January 2017
Available online 3 February 2017

Keywords:
Natural language processing
Corpus
Arabic language
Diacritization

ABSTRACT

Arabic diacritics are often missed in Arabic scripts. This feature is a handicap for new learner to read Arabic, text to speech conversion systems, reading and semantic analysis of Arabic texts.

The automatic diacritization systems are the best solution to handle this issue. But such automation needs resources as diactritized texts to train and evaluate such systems.

In this paper, we describe our corpus of Arabic diacritized texts. This corpus is called Tashkeela. It can be used as a linguistic resource tool for natural language processing such as automatic diacritics systems, dis-ambiguity mechanism, features and data extraction.

The corpus is freely available, it contains 75 million of fully vocalized words mainly 97 books from classical and modern Arabic language.

The corpus is collected from manually vocalized texts using web crawling process.

Specifications Table

Subject Area	Computer Science
More specific subject area	Computational linguistics, natural language processing, text to speech, corpus. Arabic language.
Type of data	Text files

* Corresponding author.
E-mail addresses: t_zerrouki@esi.dz (T. Zerrouki), a_balla@esi.dz (A. Balla).

http://dx.doi.org/10.1016/j.dib.2017.01.011

How data was acquired	The data is collected from freely published texts in ancients books, these books had been rewritten and vocalized by volunteers manually, to ensure that words are vocalized.
Data format	Raw
Experimental factors	Texts are collected, filtered and converted to text format. Texts are cleaned by removing extra spaces and unnecessary stuff. Adding specific description to each book and data.
Experimental features	conduct statistical analysis about word frequencies, for vocalized, semi vocalized and un-vocalized words.
Data source location	N/A
Data accessibility	Data presented in this article is freely available at http://tashkeela.sourceforge.net

Value of the data

- This data is very helpful for the statistical training of machine learning algorithms based on natural language processing [2]; It is used by diacritization systems [1–4], and disambiguation algorithms [4–6]. It was used in training and in evaluation data as well, and it can be used for similar systems.
- It is used as a linguistic resource to extract features and linguistic data processes, i.e. building lexicons [7–9], Extraction of Arabic Modal Multiword Expressions [7].
- Furthermore, this data is integrated in many other analysis, like Morphological analysis [10], syntactical models [11], and text-to-speech rule-based extraction [12].
- Extracted texts can be used as samples in learning Arabic language for both beginners and foreigners as in Al-jazeera Learning service [13].

1. Data

Data is a collection of Arabic vocalized texts, which covers modern and classical Arabic language. The Data contains over 75 million of fully vocalized words obtained from 97 books, structured in text files.

The corpus is collected mostly from Islamic classical books [14], and using semi-automatic web crawling process.

The Modern Standard Arabic texts crawled from the Internet represent 1.15% of the corpus, about 867,913 words, while the most part is collected from Shamela Library, which represent 98.85%, with 74,762,008 words contained in 97 books (cf. Table 1).

2. Experimental design, materials and methods

The process of text vocalization is a hard task to accomplish, however, there are limited vocalized texts, mainly, in learning Arabic language for beginners, or in specific-domains texts like religious texts i.e. Quranic and Hadith scripts. For these reasons, obtaining vocalized texts is considered as very hard task to accomplish [15,16].

The only resources available to obtain vocalized texts are those religious texts [17], which are often written in classical Arabic, or as new textual scripts written by modern authors who usually use a classical language in general. The classical Arabic language is a bit different from modern standard Arabic, in terms of grammars, vocabularies and semantic [18].

This linguistic feature (language differences) can lead to obsolete evaluation and training of diacritization systems, because most of these systems are supposed to be trained on classical texts, and to be implemented in modern standard Arabic texts.

Table 1
Corpus parts.

Total words	75,629,921	Percent
Classical Arabic: – 97 Books filtered from 7079 books from Shamela Library.	74,762,008	98.85%
Modern Standard Arabic	867,913	1.15%
• 20 modern books	398,911	
• Texts crawled from Internet ◦ learning.aljazeera.net ◦ al-kalema.org ◦ enfal.de ◦ diverse …	461,283	
• Manually diacritized	7701	

However, below is a list of available vocalized resource:

- Shamila library[1]: المكتبة الشاملة is an Islamic electronic library which contains hundreds of books in many domains like Hadith (prophet citation) Fiqh (scientific dogms books), history, preaching, Islamic laws, Arabic language.
  It is freely available in many formats, like websites, desktop applications. these books are rewritten by volunteers and uploaded in suitable format to Shamila library.
  In our case, we count around 97 fully vocalized books, which represent around 75 million words, that form up the main part of Tashkeela corpus data.
- Aljazeera, learning Arabic service
  Aljazeera Network launched a new service to learn Arabic as a foreign language. Aljazeera learning Arabic site [13] provides texts, samples, exercises, courses about Arabic language with many short stories extracted from news. The texts are vocalized to ease reading and facilitate learning process. Because it is so difficult to vocalize texts, Aljazeera learning activated the manual review of their automatic diacritization system to ensure a high quality of the generated vocalized texts.
- Maqola, a citation collection: It seeks the best-citation collection in the field of Arabic and Islamic heritage for both past and present, and they display it in fully diacritical format [20].
- Diverse texts crawled from the net

There are a few and limited vocalized texts available online. The reason why, the collection of such texts is also very hard, on the other hand, most of the search engines ignore diacritics in searching process, hence, this disallow users to find vocalized texts online.

To overcome this issue, we managed to use Google verbatim search to find diacritized texts, we have used Google to find diacritics texts without significant keywords to retrieve general texts without any specific keywords, we used most frequent diacritized words [19] which are considered as stop words i.e., فِي, عَنْ ، إِلَى. However, we used vocalized stop words as they are not ignored in verbatim search, in case if the writer vocalize them, most probable that the other words in text are vocalized.

The extraction process:

– The Shamela library is basically an e-book reader software, which reads a collection of thousands of books prepared by volunteers. We use Shamela as source to extract vocalized texts.
– We search for vocalized texts in books, by looking up vocalized tags in the book index or keywords i.e., "كتاب مشكول", "فِي" (/Fi:/,in), "إِلَى" (/Ila/, to).
– After that we convert crawled texts to certain encoded file format.

[1] http://shamela.ws/

Table 2
Corpus words statistics.

Feature	Count
Total Word counted:	75 629 921
Arabic vocalized word counted	67 287 202
Punctuation and non Arabic word counted:	8 342 719
Unrepeated un-vocalized word counted:	486 524
Unrepeated vocalized word counted:	998 538
Unrepeated semi-vocalized word counted:	770 702
Estimated number of vocalizations for a word	2.05

- We extract words from text files, in order to count word number and their frequencies (cf. Table 1).
- We then truncate the last short vowel (/Haraka/) from the word, to obtain words without syntactic marks. In most case, the last mark represents the syntactic case like كتابُ (/kitab-u/ - a book in subjective case), كتابَ (/kitab-a/ - a book in objective case). In other cases, the syntactic mark in not in the end, like كتابُها (/kitab-u-ha/, her book).
  We truncate the syntactic mark, in order to count the number of semi-vocalized words and their frequencies (cf. Table 2).
- We eventually truncate all vowels (Harakat) to count the number of un-vocalized words and their frequencies (cf. Table 2).

Acknowledgments

We acknowledge the efforts made by Shamela library volunteers to write, diacritize and make free texts.

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.01.011.

References

[1] A. Chennoufi, A. Mazroui, Impact of morphological analysis and a large training corpus on the performances of Arabic diacritization, Int. J. Speech Technol. (2015).
[2] A.M. Azmi, R.S. Almajed, A survey of automatic Arabic diacritization techniques, Nat. Lang. Eng. (2013) .
[3] Y. Hifny, Higher order n-gram language models for Arabic diacritics restoration, in: Proceedings of the Twelfth Conference on Language Engineering (ESOLEC'12), pp. 1–5, 2012.
[4] M. Bebah, C. Amine, M. Azzeddine, L. Abdelhak, Hybrid approaches automatic vowelization of arabic texts 3 (4) (2014).
[5] Y. Hifny, Smoothing techniques for arabic diacritics restoration, in: Proceedings of the 12th Conference Lang. Eng. (ESOLEC '12), no. 1, pp. 6–12, 2012.
[6] S. Harrat, M. Abbas, K. Meftouh, K. Smaili, E.N.S. Bouzareah, and C. Loria, Diacritics restoration for Arabic dialect texts.
[7] R. Al-sabbagh, R. Girju, J. Diesner, Unsupervised construction of a Lexicon and a repository of variation patterns for arabic modal multiword expressions, 2014, pp. 114–123.
[8] I. Zeroual, A. Lakhouaja, Clitiques-Stemmer : nouveau stemmer pour la langue Arabe, 2014, pp. 6–9.
[9] I. Zeroual, A. Lakhouaja, Adapting a decision tree based tagger for Arabic, 2016, pp. 1–6.
[10] M. Boudchiche, A. Mazroui, M. Ould Abdallahi Ould Bebah, A. Lakhouaja, A. Boudlal, AlKhalil Morpho Sys 2: a robust Arabic morpho-syntactic analyzer, J. King Saud. Univ. - Comput. Inf. Sci. (2016).
[11] M. Achraf, B. Mohamed, S. Zrigui, A. Zouaghi, and M. Zrigui, N-Scheme Model: an approach towards reducing arabic language sparseness, in: Proceedings of the 5th International Conference on Information & Communication Technology and Accessibility (ICTA), pp. 1–5, 2015.
[12] S. Harrat, K. Meftouh, M. Abbas, K. Sma, C.S. Loria, Grapheme to phoneme conversion - an arabic dialect case to cite this version, 2014.
[13] Aljazeera Network, Aljazeera Learning Arabic service 2016. [Online]. Available: ⟨http://learning.aljazeera.net⟩.
[14] Shamela, Shamela Islamic library. [Online]. Available: ⟨http://shamela.ws⟩.

[15] W. Zaghouani, H. Bouamor, A. Hawwari, M. Diab, O. Obeid, M. Ghoneim, S. Alqahtani, K. Oflazer, Guidelines and Framework for a large scale arabic diacritized corpus, 2015, pp. 3637–3643.
[16] H. Bouamor, W. Zaghouani, M. Diab, O. Obeid, O. Kemal, M. Ghoneim, A. Hawwari, A pilot study on arabic multi-genre corpus diacritization annotation, 2015, pp. 80–88.
[17] W. Zaghouani, Critical survey of the freely available arabic corpora, 2014, pp. 1–8.
[18] A. Farghaly, K. Shaalan, Arabic natural language processing: challenges and solutions, ACM Trans. Asian Lang. Inf. Process. 8 (4) (2009) 1–22.
[19] M. Attia, P. Pecina, A. Toral, and J. Van Genabith, A lexical database for modern standard arabic interoperable with a finite state morphological transducer, in: systems and frameworks for computational morphology, in: Proceedings of the Second International Workshoppp. pp. 98–118, 2011.
[20] K. Aissa, Maqola, a collection of best arabic citations, 2016.(Online)(. Available)⟨http://maqola.org⟩.