Indian Journal of Science and Technology, Vol 9(46), DOI: 10.17485/ijst/2016/v9i46/107110, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Building Arabic Corpus Applied to Part-of-Speech Tagging

Rabab Ali Abumalloh[1]*, Hassan Maudi Al-Sarhan[2] and Waheeb Abu-Ulbeh[3]

[1]University of Dammam, Department of Computer Science, Dammam – 31451, Saudi Arabia; aliyuchikaji2@gmail.com
[2]Ajloun National University, Information Technology, Ajloun – 26810, Jordan.
[3]Universiti Teknologi Malaysia, UTM Johor Bahru, 81310 Johor, Malaysia.

Abstract

Objective: This paper aimed to review corpus linguistics sources related to part-of-speech tagging and to build up a sufficient annotated corpus for the Arabic language that contains Arabic words and their grammatical tags. **Methods/ Statistical Analysis:** An in-depth survey conducted by the author's showed that there is a need for free tagged Arabic corpus that can be used in natural language processing researches. A corpus of 25,000 words collected manually from different web sources which ware written in Modern Standard Arabic. The collected words were tagged using Arabic language grammar books. **Findings:** The developed corpus can help the researchers in natural language processing applications. **Applications/Improvements:** This corpus needed to be expanded to include more words and their grammatical tags.

Keywords: Arabic Language, Corpus, Linguistics, Part of Speech Tagging

1. Introduction

The last fifty years has seen a huge development in the field of computer science. In recent years, computer scientists have tried to develop a machine that has the ability to communicate with people and understand their languages. Thus, there has been huge progress in such related fields as: speech recognition and part of speech tagging.

The study described in this paper belongs to the area of computational linguistics. The first time the computational linguistics term was used in 1963 in a preface to an English translation by[1]. Recently, Computational linguistics and natural language processing have attracted attention from researchers around the word, with a huge shift from only studying simple models to extensive systems that can process, test and learn from large corpora (plural of corpus)[2].

There are many areas that may be regarded as properly included within the discipline of computational linguistics[3]. One of these areas is Part-Of-Speech Tagging (POS). POS-tagging is a wide research area in computational linguistics[4].

Part-of speech tagger processes the words in the corpus in order to assign a part of speech to each word. Part-of-speech refers to the syntactic class of the word such as: noun, verb, and adjective[5]. In the study of linguistics, corpus is defined as a large amount of natural language material that can be stored in machines in a way that is easily accessed and manipulated. Corpus provides part-of-speech tagger systems with the needed linguistic knowledge that helps resolve the ambiguity in the language without the need of strong linguistic skills.

For English language, a huge effort has been made to create hundreds of corpora that were extracted from written or even spoken texts. This evolutionary work began with building the Brown corpus in 1961 with one million English language words, leading to several corpora that have played an important role in compilation of English language dictionaries[6].

Arabic language differs than Indo-European- languages and is considered more complex than English language in different aspects. Although software systems and computer programs are wide spread in the Arabic

*Author for correspondence

world, most of these programs are written in English language and are used in English data processing systems. Lack of data sources for Arabic languages is one of the major problems that face researchers in the area of building Arabic part-of-speech taggers. The significance of this research comes from two perspectives. First, the lack of free resources for natural language processing applications in Arabic language and second, there is a need for tagged Arabic corpus that can be used in training and testing Arabic part of speech tagger.

At the beginning of this research that was motivated by the lack of free resources that can be used in developing part of speech tagger for Arabic language, there was a need to determine the research design that will be followed. Determining the research design is about determining the procedures that must be followed to lead the researchers in their study from the beginning to the end. This study is comprised of four main stages including: define research problem (Phase 1), literature review (Phase 2), develop the annotated corpus (Phase 3) and using the developed corpus in the training and testing process of the Arabic part of speech tagger (Phase 4). Figure 1 presents the research design phases.

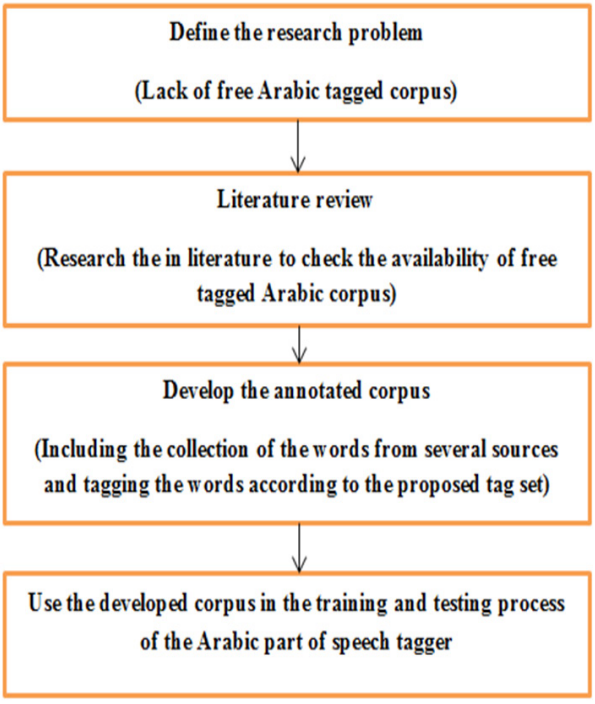

Figure1. Research design.

2. Introduction to Arabic Language

Arabic is the official language in 22 Arab countries, and it is the language of all Muslims regardless of their origin. AL-Qur'an, "The holy book of Islam", is written in Arabic language. Arabic language has given the Arab world its unique identity. The advent of Islam and AL-Qur'an in the seventh century gave the Arabic language a great religious role. This is a result of the holy book which was revealed in Arabic language to the prophet Muhammad and it must be read in Arabic language in order to understand the message of AL-Islam[7]. The presence of AL-Islam had extended the importance of Arabic to include over one billion Muslims around the world[7]. Arabic language is required by all Moslems. Arabic is needed when they are reading AL-Qur'an, in their praying and even in their understanding of the religion. AL-Qur'an reflects the beauty and the perfection of Arabic language and has become the Arabic language standard[7].

Arabic language belongs to the family of Semitic languages which includes Tigrinya, Tigre, Amharic, Modern Hebrew, Syriac, some Aramaic dialects and Maltese[7]. Semitic languages have a history dating back to thousands of years[8]. Moreover, most of these languages have died out, but Arabic language is still popular among all Arabs and Muslims.

Since the seventh century, the written Arabic has changed little comparing to the spoken Arabic which has many variations[7]. In general, however, there are many forms of Arabic language. These forms can be categorized into three main variations:

1. *Classical Arabic:* This is the language of "AL-Qur'an", "The holy book of Islam", and is also the language that was used in the old Islamic empire[7]. This type is restricted now to religious and formal texts[7].These days no one speaks this language as a native.
2. *Modern Standard Arabic (MSA):* This language is referred to as: "Al-fusha" in Arabic, which is used in formal communications, television, radio, news, education, and printed media[7]. Arab do not speak MSA in their ordinary life except in formal schools during the education process[7]. MSA development started in the late eighteenth century because of the huge spread of Al- Islam, many non- Arab people believe in Islam and through contact with Arab Moslems.

3. *Colloquial Arabic*: This referred to as: "Alaameya" that represents the locally spoken dialect, which differs from country to country. People use "Alaameya" in their ordinary life, when they are shopping, chatting or in their homes[7].

Arabic alphabet consists of 28 letters. Many letters in Arabic have no familiar letters in English language such as the letters "ayn", "gayn", and "daad". The letter "Al daad" is a special character that distinguishes the Arabic language from other languages. For this reason, Arabic language is called "daad" language "logataldaad" according to the "daad" letter.

Arabic words are written as a series of letters, in which the letters of a single word are strung together. Unlike the English letters which are written from left to right, Arabic Alphabets are written from right to left. In Arabic letters, there is no upper or lower case like the English language. Arabic letters have different shapes according to their positions in the word. The letter at the beginning of the word is written in different form than the letter written in the middle or at the end of the word[9].

Diacritics in Arabic language are short vowels which are defined as special marks that are put below or above characters. It helps in determining the correct pronunciation, especially for identical spelling words. Many words in Arabic have the same letters, but they differ in the pronunciation. The need of the diacritics is to distinguish between these words. Arabic writing is categorized into three categorizations according to the presence or absence of the diacritics. These categorizations are: vocalized, partially vocalized or completely not vocalized writing. In Arabic language, four main diacritics give short sounds. These diacritics are: Fatha, which is used at the top of a letter to give the "a" sound; Dhamma, which is put on the top of a letter to give the "u" sound; Kasra, which is put under a letter to give the "i" sound; and Sukuun, which is represented by a small circle that is put on the top of the letter to indicate the absence of a vowel.

Syntax of a given language is concerned with studying the grammatical arrangement of the words in the sentences[10]. In Arabic syntax system, there are two types of sentences, verbal sentences and nominal sentences. The verbal sentence begins with a verb that is followed by the subject of the verb. The nominal sentence begins with a noun or a subject that is followed by a verb and the object.

A word in Arabic language is defined as a collection of letters strung together as a single unit having a specific meaning. Arabic grammarians categorized Arabic words into three main part-of-speech classes which are: noun, verb and particle. A noun is a word used to describe a thing, a person or an essence[11]. Nouns could be distinguished from the verbs by the following signs:

1. *Definition Article*: If the word in Arabic language begins with the definite article "Al-Altareef", then it is a definite noun[11].
2. *Vocative*: If the word follows the vocative particle, then it is a noun[11].
3. *Nunation mark*: If the word ends by nunation mark, then it is anoun[11].
4. *Kasra mark*: When the noun in the genitive case, "kasra" appears at the end of the noun. Verbs in Arabic never receive a "kasra" mark at the end[11].

A particle in Arabic language is called "harf" It does not have a meaning without being attached to a noun or a verb. The main aim of the particle in Arabic language is to assist other words in their semantic action in the sentence[12].

The structure of Arabic language differs from Indo-European languages in many ways. The regularity of form is one of the main characteristics of the Arabic language. Usually Arabic words contain prefixes, suffixes and infixes in addition to the root. In Arabic language, while words are derived from roots, words could also be built from recurring patterns. Root in Arabic language is a collection of consonants, usually three and rarely two or four. The essential and the core meaning of the word are represented by the root of the word. Words in Arabic language are built by adding affixes or vowels to the roots using a specific pattern. In Arabic language there are two types of morphology:

1. *Inflectional Morphology*: This type of morphology is concerned with the inflectional proprieties of the words without modifying their core meaning[13].
2. *Derivation Morphology*: This type of morphology is concerned with the process of forming the word using other words, roots or stems with patterns[13].

Patterns are a collection of consonants and vowels that are combined from the root to form particular templates[14]. These templates are used to form the nouns and the verbs in Arabic language.

Table 1. Summary of the developed Arabic Corpora.

Name of Corpus	Source	Medium	Size	Availability	Tagged
CALLFRIEND Corpus	[22]	Spoken conversations	60 telephone conversations	Not available for free	Not tagged
Nijmegen Corpus	[23]	Written magazines and fictions books	Two million words	Not available for free	Not tagged
CLARA Corpus	Charles University (1997)	Written books and internet sources	37 million words	Not available for free	Not tagged
Egypt Corpus	John Hopkins University (1999)	Written corpus of Al-Qur'an	Not mentioned	Not available for free	Not tagged
An-Nahar Newspaper Text Corpus	European Language Resources Association (1995-2000)	Written newspaper	140 million words	Not available for free	Not tagged
Broadcast News Speech	University of Pennsylvania LDC (2000)	Spoken from the radio	110 broadcasts	Not available for free	Not tagged
DINAR Corpus	Nijmegen University (2000)	Written	10 million words	Not available for free	Not tagged
Arabic Newswire Corpus	(2001)[19]	Written articles from Agence France Presse (AFP)	80 million words	Not available for free	Not tagged
Buckwalter Arabic Corpus	[20]	Written resources on the web	2.5 - 3 billion words	Not available for free	Not tagged
Leuven Corpus	Mark van (1990-2004) at the Catholic University Leuven	Written and spoken web resources	3 million words	Not available for free	Not tagged
Corpus of Contemporary Arabic	[16,21]	Written and spoken resources from websites.	One million	Available online	Not tagged
Arabic Blogs (2009)and a corpus builder application	[8,17]	Written 37 blogs around the death of a Saudi female journalist.	131,836	Not available online	-
Essex Arabic Summaries Corpus	[18]	Written articles and summaries of these articles	Not mentioned	Available for free	Not tagged
Open Source Arabic Corpus (OSAC)	[25]	Written CA and MSA	18 million words	Available for free	Not tagged
2012 Arabic newspapers corpus	[26,28]	Written MSA	2.5 million word	Available for free	Not tagged
Event Corps	[27]	WEB pages	5393 documents	Not mentioned	Not tagged
KACST	[26,28]	Written	700 million words	Available for free	Not tagged
The developed corpus	The authors	Electronic websites	25,000 tagged words	Available for free	Manually Tagged

3. Review of Arabic Language Corpus

For Arabic language, there is a problem in that no tagged corpora are available for free. Most of the Arabic corpora were created to serve the researchers in their personal research and not for public use. Some of the well-known Arabic language corpora are: LDC Arabic newswire corpus, CCA: Corpus of Contemporary Arabic, Penn Arabic Treebank Corpus, Nijmegen Corpus, Arabic Corpus, An-Nahar Newspaper Text Corpus and Al-Hayat newspaper corpus[15]. In this paper, a review of the available corpora was conducted in order to determine the limitations of the current resources of the Arabic corpora. The summary of the available corpora is presented in Table 1.

Corpus of Contemporary Arabic was developed by[16,21]. It was collected and published in 2004 after an in depth survey of the literature to review the need for a free available corpus for the public. The main purpose of the corpus was to provide material for Teaching Arabic As A Foreign Language (TAFL) and to provide an authentic data source that can help in developing tools that serve Arabic language use. The corpus was collected using wide range of written and spoken text resources from several websites. The number of words in this corpus was around one million, and is available online for free[16].

In[17] developed an Arabic corpus of 131,836 words by building an application to download RSS feeds automatically. Her work was motivated by the lack of Arabic corpus that she faced as a researcher in computational linguistics field. The focus in her corpus was in colloquial form of Arabic according to the lack of the linguistic tools for this form. Thirty seven blogs about the death of Saudi Arabian journalist were chosen and collected[17].

Essex Arabic Summaries Corpus was developed by[18] in University of Essex for research purposes. This corpus was free and available to the public in 2010. The corpus was collected from 153 Arabic articles and 765 summaries of these articles that were developed using Mechanical Turk[18].

Arabic Newswire Corpus was developed by[19] to be applied in information retrieval and language modelling. This corpus was collected from written articles from Agence France Presse (AFP) from 1994 to 2000. The corpus contained 666,094 Arabic words that is available for free only to Linguistic Data Consortium (LDC) members[20].

In[20] Arabic Corpus was developed by[20] from 1986 to 2003. The corpus contains 2.5 to 3 billion words which were collected from written resources in the website. This corpus is not free for public use[20].

Mark van at Leuven at Catholic University developed Leuven Corpus from1990 to 2004. It was collected from written and spoken Arabic words from different web resources, radio and school books to form a three million Arabic word corpus. The main purpose of this corpus was to use it in Arabic to Dutch and Dutch to Arabic dictionary. This corpus is not free to the public[21].

CALLFRIEND Corpus was developed in University of Pennsylvania LDC by. It was collected from 60 telephone conversations of the speech produced by telephone lines from Egyptian native speakers to serve as a base for speech recognition study. The corpus is not freely available to the public.

Nijmegen Corpus (1996) was developed at Nijmegen University by[22] with about two million words from magazines and fiction books that served as a base for Arabic to Dutch and Dutch to Arabic dictionary. The source materials focused on written texts, because at that time web sources were not rich with Arabic texts and were mainly presented as images. The corpus is not freely available to the public[22].

CLARA corpus (1997) was compiled to serve as a base for Arabic-Czech dictionary development. This project started at Charles University, and was financed by Grant Agency of the Czech Republic and the Grant Agency of Charles University. The corpus contained 37 million words of Modern Standard Arabic that were collected from published material beginning in 1975. Texts in this corpus covered several wide subjects of agriculture, art, finance and so on. The corpus is not available for free[16].

Egypt (1999) is a parallel corpus of the Qur'an in English and Arabic that was developed at John Hopkins University[16]. Researchers reported that this corpus has some problems related to the length of the Arabic sentences, which are considered very long in the Arabic version compared to the short sentences in the English version. This corpus is not free to the public[16].

Broadcast News Speech corpus (2000) was developed at University of Pennsylvania LDC from spoken materials of more than 110 broadcasts to be used for speech recognition research. The source of this corpus was the radio show Voice of America and it is not freely available to the public[16].

DINAR Corpus (2000) was developed at Nijmegen University and consists of about 10 million words. The main purpose of this corpus is for natural language processing researches and general research. One of the main achievements of this project was the production of Arabic Raw Corpora for lexical purposes and a tagged reference corpus of 200,000 words. This corpus is not available to the public[16].

An-Nahar Newspaper Text Corpus of 140 million words was collected from different Arabic Lebanon Articles from 1995 to 2000. This work was done by European Language Resources Association. The corpus is not available for free[23].

Open Source Arabic Corpus (OSAC) was developed by[24] in 2010 from written Colloquial Arabic (CA) and Modern Standard Arabic (MSA). This corpus contains 18 million words that were categorized according to 10 categorizations. It is available to download but the problem of this corpus is that it is untagged[24].

"Arabic newspapers corpus" (2012) was developed by[25] from written modern standard Arabic and it is freely available to download. The corpus was collected from newspapers and it includes several subjects from cultural, religion, sport and politics[25].

In[26] developed an Arabic corpus for event mining (EventCorp) in 2013. The main target of the developed corpus is to collect the events which were published in the media in one corpus[26].

KACST was developed by[27] in 2015. Around 700 million words were collected from both classical Arabic and modern standard Arabic. The main objective of this corpus is to develop a large corpus that is freely available for different purposes[27].

As we can see from the literature, there is a huge effort in developing Arabic corpus for different purposes. The main contribution of our research is to investigate the literature and as we can see in Table 1, there are a few corpora that are available for free. This induced us to provide a free Arabic tagged corpus that is specifically designed to serve as training and testing data for Arabic part of speech tagger. This corpus can also be used in other natural language processing applications.

4. Designed Corpus

The huge spread of search engines in Arabic language and the use of digital Arabic texts over the Internet in the last decade simplify the process of collecting digital Arabic texts from different sources. To build an Arabic corpus of 25,000 words different Arabic web-sites that are written in Modern Standard Arabic (MSA) were used to extract the words. The extracted words are not limited to a particular subject, so they could cover a wide range of subjects and provide a more accurate training and testing process. Some of the topics that are covered in the collected texts are: health care, science, sports, cooking and religion. Some of the websites that were used to extract the words are:

1. *http://aafeaa.blogspot.com*: This website presents topics about the health care.
2. *http://www.addustour.com*: This is Al-Dustor newspaper website, which is a Jordanian daily newspaper that presents several aspects and topics.
3. *http://www.madawi.com/vb/t9267.html*: This website presents different sport topics in Arabic language.
4. *http://www.islamonline.net/Arabic/Science/200*: This website presents different science topics in Arabic language.

After extracting the word and using Arabic grammar books these words were given a specific tag from the tag set. The Arabic tag set that has been used in this study is based on the traditional Arabic categorizations that state that all Arabic words are derived from nouns, verbs or particles. Table 2 shows the main POS-tagging categorizations. Noun is categorized into five main tags: original noun, agent noun, patient noun, adjective noun and superlative noun. Table 3 shows these tags. A Verb is a word that describes an action and it is similar to that in English language. Verbs could be categorized into three main tags: perfect verb, imperfect verb and imperative verb. Table 4 shows these tags.

Table 2. Main part of speech tags.

Main Part of Speech Tags	Tag (Main)	Example
Noun	N	بِبِتاكَ
Verb	V	لصو
Particle	P	نْ أَ

For the purpose of this study, we generated our own tag symbols. This is shown in the last column of Table 5. Each tag consists of three tuples [T, S, G] where:

1. *T:* This tuple represents the main tag of the word.
2. *S:* This tuple represents the sub-class of the main tag.
3. *G:* This tuple represents the gender (masculine or feminine).

Table 3. Description of the main tags of the Noun.

Main POS Noun	Definition	Example	Tag
Original Noun	The word that describes an event that is not connected with any time period.	ةعيابم	[N1O2]
Agent Noun	The noun that is used to describe the person how has done an action or the subject of the verb.	بردملا	[N1A2]
Patient Noun	The noun that is used to describe the object of the verb.	تيبملا	[N1P2]
Adjective Noun	The adjective noun in Arabic language is actually similar to it in English language; it is the noun that is used to describe something.	ٌميسَو	[N1D2]
Superlative Noun	The superlative is the noun that used to compare two nouns in a specific adjective.	ُرَغصَأ	[N1S2]

Table 4. Description of the main tags of the Verb.

Main POS Verb	**Definition**	**Example**	**Tag**
Perfect/ Past Verb	Describes an action that happened in the past.	ٌبَتَك	[V1P2]
Imperfect/ Progress Verb	Describes an action that is happening in the present time.	ُبُتْكَي	[V1I2]
Imperative Verb	Describes an order to do some action.	ُبُتْكا	[V1M2]

Table 5. Examples of the Annotated Corpus.

Word	**Tag 1 (Main)**	**Tag2**	**Tag3**
ةعيابم	N	ON	F
تملع	V	PV	F
رون	N	ON	M

For example, the word: [ٌبِتاَك] has the tag [N1P2M3] ,which means that this word is a patient noun and the] gender of this noun is masculine. The tuple [N] refers to the noun, the tuple [P] for patient, and the tuple [M] for masculine.

The developed tagged corpus (annotated corpus) contains four fields: the first field for the Arabic words itself, while, the rest of the fields are for the tags of the word. These words are collected and tagged using the Arabic language grammar books and with a help of an Arabic linguistic specialist. Examples of Arabic words and their associated tags are presented in Table 4.

5. Results and Discussion

Despite the growing awareness of the importance of Arabic corpora, this research area still has some limitations. Few corpora are available for free and no tagged Arabic corpus is available for free. Tagged corpus is very important for the development of part of speech taggers. POS-tagging is usually the first step in linguistic analysis. In addition, building many natural language processing applications is a very important intermediate step. It could be used in spell checking and correcting systems, speech recognition systems, information retrieval systems and text-to-speech synthesis systems.

In order to conduct this research, an in-depth of the current Arabic corpora was explored. This research is not the only work of this area but the shortage of the availability of other corpora induced us to conduct this research that may help other educators and learners. Free access to tagged Arabic corpus is very important for researchers in the area of Part-of speech tagger development.

Developing the Arabic corpus of 25,000 tagged words is the first step of this research. This corpus has the advantage of being manually tagged which guarantees that it is accurate and it can be used for training and testing Arabic part-of-speech taggers. The developed corpus was used in the training and testing of Arabic part-of- speech tagger system which was developed by the authors. The developed tagger correctly tagged 98.94% of the words in the training dataset. This result indicates that the developed tagger is trained well. Improving the annotated Arabic corpus can be done by adding additional Arabic words and their tags. Using the developed corpus in the training and testing process of other natural language processing applications is our future work.

6. Conclusion

Developing the corpus is not just a process of collecting written or spoken texts from electronic and other resources; it is about processing these texts and preparing them to be used in different natural language processing applications. Arabic language gained the interest of researchers in the last decade. Researchers tried to build systems that can manipulate the Arabic language. Several universities and organizations developed Arabic corpora and used it as a data source for their research in Arabic language.

Although the huge effort that has been done in the area of linguistics is very extensive, there are some limitations in the available corpus for Arabic language. Most of these corpora are developed for specific target and they are not available for free. The work that is provided in this research tried to fill this gap by building a tagged Arabic corpus that can be directly used in training and testing Arabic part of speech taggers. The developed corpus still needs to be extended and examined through multiple part-of-speech taggers.

7. References

1. Hays DG. Gerald Penn, Philosophy of Linguistics. 2012 Jan; 14:143.
2. Leech G. 100 million words of English, English Today. 1993: 9(01):9–15.
3. McEnery AM, McEnery T. Computational Linguistics: A Handbook and Toolbox for Natural Language Processing, Sigma Press, 1992.
4. Altunyurt L, Orhan Z. Part-of-Speech Tagger for Turkish, Technical Report, Department of Computer Engineering, Jun 2006.
5. Jurafsky D, Speech MJ. Language Processing. International Edition, 2008, p. 66–7.
6. Ku H, Francis WN. Computational Analysis of Present-Day {A}merican {E}nglish. Brown University Press, 1967, p. 424.
7. The Arabic Language: The Glue that Binds the Arab World. Date Accessed: 2016. Available at: http://www.sharjah-bookfair.com/en/newsdetails/625.
8. Khoja S. APT: An Automatic Arabic Part-of-Speech Tagger (Doctoral Dissertation, Lancaster University), 1963, p. 1–6.
9. Jiyad M. A Hundred and One Rules! A Short Reference for Arabic Syntactic, Morphological and Phonological Rules for Novice and Intermediate Levels of Proficiency, 2006, p. 2–4.
10. Alqrainy S, Ayesh A. Developing a Tag Set for Automated POS Tagging in Arabic, WSEAS Transactions on Computers. 2006 Nov; 5(11):2787–92.
11. Aosh M. Learning Language at a Distance: An Arabic Initiative, Foreign Language Annals. 2001 Jul 1; 34(4):347–54.
12. Allen R, Allouche A. Let's Learn Arabic: A Proficiency-Based Syllabus for Modern Standard Arabic, University of Pennsylvania; 1988.
13. Habash N. Arabic Morphological Representations for Machine Translation, In: Arabic Computational Morphology Springer Netherlands, 2007, p. 263–85.
14. Attia M. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modeling Finite State Networks, In: Challenges of Arabic for NLP/MT Conference, The British Computer Society, London: UK, 2006 Oct 23; 10(1):72.
15. Alqrainy S. A Morphological-Syntactical Analysis Approach for Arabic Textual Tagging, Doctor of Philosophy in Computer Science, 2008, p. 228.
16. Al-Sulaiti L, Atwell E. Designing and Developing a Corpus of Contemporary Arabic, In: Proceedings of the Sixth TALC Conference, 2004 Mar, p. 1–92.
17. Khoja S. An RSS Feed Analysis Application and Corpus Builder, Interface: The Journal of Education, Community and Values. 2009; 9(3):115–18.
18. El-Haj M, Kruschwitz U, Fox C. Using Mechanical Turk to Create a Corpus of Arabic summaries, Research Repository, 2010, p. 1–4.
19. Graff D, Walker K. Arabic Newswire Part 1. Linguistic Data Consortium, Philadelphia. LDC Catalog Number LDC2001T55 and ISBN, 2001.
20. Buckwalter T. Issues in Arabic Orthography and Morphology Analysis. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages, Association for Computational Linguistics, 2004 Aug 28, p. 31–34.
21. Canavan A, Zipperlen G. CALLFRIEND Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia, 1996.
22. Hoogland J. Lexical Gaps in Arabic Lexicography with Evidence from Arabic Dictionaries, Approaches to Arabic Linguistics: Presented to Kees Versteegh on the Occasion of his Sixtieth Birthday. 2007 Oct; 15:455–73.
23. Maamouri M, Bies A, Buckwalter T, Mekki W. The Pennarabic Tree Bank: Building a Large-Scale Annotated Arabic Corpus. In: NEMLAR Conference on Arabic Language Resources and Tools, 2004 Sep 22-27, 466–67.
24. Saad MK, Ashour W. Osac: Open Source Arabic Corpora. In: 6th Arch Eng Int. Symposiums, EEECS, 2010 Nov 25; 10:112–17.
25. Khorsheed MS, Al-Thubaity AO. Comparative Evaluation of Text Classification Techniques using a Large Diverse Arabic Dataset, Language Resources and Evaluation. 2013 Jun 1; 47(2):513–38.

26. Alasfour AA, Trausan-Matu S. Developing an Arabic Corpus for Event Mining. In: System Theory, Control and Computing (ICSTCC), 17th International Conference, 2013 Oct 11, p. 21–28.

27. Al-Thubaity AO. A 700M+ Arabic Corpus: KACST Arabic Corpus Design and Construction, Language Resources and Evaluation. 2015 Sep 1; 49(3):721–51.