Abstract

Arabic lexicography is a well-established and deep-rooted art of Arabic literature. Computational lexicography, invests computational and storage powers of modern computers, to accelerate long-term efforts in lexicographic projects. A collection of 23 machine-readable dictionaries, which are freely available on the web, were used to build the Corpus of Traditional Arabic lexicons (the TAL-Corpus). The purpose for constructing the TAL-Corpus is to collect and organize well-established and long traditions of traditional Arabic lexicons which can also be used to create new corpus-based Arabic dictionaries.

The compilation of the TAL-Corpus followed standard design and development criteria that informed major decisions for corpus creation. The corpus building process involved extracting information from disparate formats and merging traditional Arabic lexicons. As a result, the TAL-Corpus contains more than 14 million words and over 2 million word types (different words).

The TAL-Corpus was applied to create useful morphological database. This database was automatically constructed using a new algorithm which is informed by Arabic linguistics theory. The newly developed algorithm processed the text of the TAL-Corpus and as result it extracted 2,781,796 entries. These entries were stored in the morphological database where each represents a word-root pair (i.e., an Arabic word and its root).

A comparative evaluation of the TAL-Corpus and other three Arabic corpora showed that the lexical diversity of its vocabulary scored higher. Moreover, its coverage was computed by comparing words and lemmas against their equivalents of other corpora where it scored about 67% when comparing words and 82% when comparing lemmas.

Keywords: lexicography, traditional Arabic lexicon, corpora, dictionary building, the TAL-Corpus

1. Introduction

Lexicography is the applied part of lexicology. It is concerned with the design and construction of lexicons. Lexicography defines the process of collating, ordering of entries, derivations and their meaning, depending on the aim of the lexicon to be constructed and its size. Lexicography is defined as “…the branch of applied linguistics concerned with the design and construction of lexica for practical use.” (Eynde & Gibbon 2000). Moreover, lexicology is also defined as “…the branch of descriptive linguistics concerned with the linguistic theory and methodology for describing lexical information, often focusing specifically on issues of meaning.” (Eynde & Gibbon 2000).

Corpora have been used to construct dictionaries since the release of the Collins-Birmingham University International Database COBUILD. Computer technology was used in the four stages of constructing COBUILD: data-collection, entry-selection, entry construction and entry-arrangement (Ooi 1998). Similarly, the Oxford English Corpus was created to construct the Oxford English Dictionaries. It consists of 2.5 billion words of 21st century English which provides evidences of language use and development. It also draws an accurate picture of the language today. It contains text from literary, novels, specialist journals, everyday newspapers, magazines, blogs, emails, and Internet message boards. These texts were collected from all parts of the world, namely: the UK,
the United States, Ireland, Australia, New Zealand, the Caribbean, Canada, India, Singapore, and South Africa (The Oxford Dictionaries, 2018).

Arabic corpora started to appear in the late 1980s. They differ in size, type, purpose of development, and the materials used to develop them (Al-Sulaiti & Atwell 2006). Freely available Arabic corpora were surveyed by Zaghounani (2014). The survey categorized 66 freely available Arabic corpora into: (i) Raw Text Corpora such as monolingual corpora, multilingual corpora, dialectal corpora, and web-based corpora. (ii) Annotated Corpora such as named entities, error annotation, part-of-speech, syntax, semantic, and anaphora. (iii) Lexicons such as lexical databases and words lists. (iv) Speech Corpora such as audio recordings and transcribed data. (v) Handwriting Recognition Corpora such as scanned and annotated documents. (vi) Miscellaneous Corpora such as questions/answers, comparable corpora, plagiarism detection and summaries.

The first freely available Arabic corpus was the Corpus of Contemporary Arabic (Al-Sulaiti & Atwell 2006). It contains 1 million words collected from newspapers and magazines text. Most monolingual raw text of Arabic corpora cover the news domain. Examples of such corpora are: OSAC: Open Source Arabic Corpora (Saad & Ashour 2010); Khaleej-2004 corpus (Abbas & Smaili 2005); Watan-2004 corpus (Abbas et al, 2011); KACST Arabic Newspaper Corpus (Al-Thubaity et al. 2013); Arabic Words Corpus which is a frequency list of 1.5 million words collected by Al-Saadi. The International Corpus of Arabic (ICA) project by Bibliotheca Alexandrina (BA) was planned to construct a corpus that contains 100 million from Press, Net articles, books and academic text sources. The ICA was planned to cover the Arabic language as being used all over the Arab world (Alansary & Nagi, 2014). Al Rabiah et al, (2013) developed the King Saud University Corpus of Classical Arabic (KSUCCA) which consists of around 50 million words. It was collected from authenticated Classical Arabic texts. It was constructed to study the language of the Qur’an using distributional lexical semantics.

Ismail, et al, (2014) developed a set of computational tools and corpus resources that would facilitate research in historical semantics and etymological lexicography. They constructed Historical Arabic Corpus (HAC) with around 45 million tokens in the first phase of development. They analyzed the collected text automatically and they annotated the text with linguistic information such as; part-of-speech, root, and morphological pattern. This corpus was collected from about 500 sources that represent 1 600 years of continuous language use which represent the entire recorded history of the Arabic language. Additionally, they developed a corpus builder that integrates a stemmer with a tagger to process and annotate documents, and then compile them into an XML-formatted corpus. They also created an indexer, a search engine, a concordancer and a dictionary editor that together facilitate searching and extraction of linguistic knowledge from HAC. Also, these tools facilitate the compilation of dictionary entries in a hypothetical dictionary on historical principles.

Since the release of COBUILD, corpora proved to be excellent resources for developing new dictionaries. In addition to lexical information, corpora can provide more useful information that would enrich dictionaries such as idioms, phrases, collocates, word sketches and thesauri of words. Arabic corpora were not yet used to construct most existing Arabic dictionaries (Al-Sulaiti & Atwell, 2006). Ghazali & Barham (2001) criticized existing Arabic dictionaries for literal translations of their lexical items and the lack of idioms; phrasal verbs; collocations; and new words entering the language. Based on a corpus of 1.5 million words, they investigated the different meaning of the verb ذُكِّرَ ‘ahada ‘he took’ on both the corpus and Al-Waseet Arabic dictionary. In addition to its literal meaning, they discovered two additional meanings using the corpus which were not mentioned in Al-Waseet Arabic dictionary.

Recently, corpora are used to build bilingual dictionaries for Modern Standard Arabic (van Mol, 2000; van Mol & Paulussen, 2001; Hoogland, 1996; Zemank, 2001). A corpus of 3 million words was constructed to develop a Dutch–Arabic/Arabic–Dutch dictionary (van Mol, 2000). Older version of the Dutch–Arabic/Arabic–Dutch dictionaries were constructed using a 2 million-word corpus (Hoogland, 1996). Likewise, a corpus of 50 million words was used to develop an Arabic–Czech dictionary (Zemank, 2001).

Moreover, The Oxford Arabic Corpus (OAC) consists of 880 million words. It is used to construct the Oxford Arabic Dictionary (OAD) (Arts & McNeil 2013, Arts 2014). They used the Oxford Arabic Corpus with Sketch Engine to eliminate unnatural examples, to add appropriate examples showing natural usage, to identify modern senses of old words, and to include new vocabulary in the constructed dictionary.

Furthermore, a jellyfish dictionary for Arabic was developed using a large-scale Arabic corpus. A jellyfish dictionary is defined as a self-updating and automatically monitoring language change. Three motivations for developing a jellyfish dictionary for Arabic are (i) discovering new words; (ii) flagging obsolete words; and (iii) recognizing new senses. The large-scale Arabic corpus used to develop the jellyfish dictionary is consisted of 1 billion words (Attia & van Genabith, 2013).
In addition, many Arabic lexical databases were constructed. The morphological Analyzer for Arabic (BAMA) (Buckwalter 2002; 2004) contains Arabic-English lexicon files. One of them is contains 82,185 stems which was reused in many Arabic NLP tools such as morphological analyzers and spell checkers. Similarly, AyaSpell a spell checker for Arabic depends on a lexicon which was built by analyzing 5 traditional Arabic lexicons. It contains more than 50,000 entries distributed on more than 10,000 verbs and more than 40,000 nouns, particles and residuals (Zarrouki & Kebdani 2009; Zarrouki & Balla 2009). A third example is the Arabic WordNet (AWN). It is a lexical resource for MSA which is based on the design and the contents of the Princeton WordNet (PWN) for English. The semantic background for the AWN were encoded in a large ontology that contains around 1,000 terms and 4,000 definition statements (Elkateb & Black 2001; Black & El-Kateb 2004; Elkateb, Black et al. 2004; Rodríguez, Farwell et al. 2008). Likewise, Arabic Verbinet is a large lexicon for Arabic verbs. It contains verb entries where each entry is a third person masculine singular perfect verb. It has 173 classes which contain 4,392 verbs and 498 frames (Mousser 2010). Aralex is a lexical database which was developed to study the cognitive processing of Arabic on relation of precise frequency counts. Aralex was built depending on a 40-million word MSA corpus which was collected from online newspapers. It provides information about orthographic forms, stems, roots and patterns and their frequencies (Boudelaa & Wilson 2010). Quranic Arabic WordNet (QAWN) is a word net for the Qur’an and consists of 6,918 synsets that were constructed from about 8,400 unique word senses, on average of 5 senses for each word (Al-maayah et al 2015). These lexical databases are designed and built for a specific purpose and for specific Arabic NLP application. They are small in size and they are designed for MSA only (Sawahla 2011).

This paper describes an important lexical resource that is constructed to improve Arabic lexicography and Arabic NLP tools. The Corpus of Traditional Arabic Lexicons (the TAL-Corpus) is constructed from the text of 23 traditional Arabic lexicons spanning the period of over 1,200 years. The TAL-Corpus will be used as part of a large lexicographic corpus of Arabic to build new modern Arabic dictionaries. The TAL-Corpus can also be used to study of the evolution of Arabic vocabulary system. The TAL-Corpus is accessible via an online interface which allows users to search for lexical entries.

2. Traditional Arabic Lexicons and Lexicography

Arabic lexicography is a well-established and deep-rooted art of Arabic literature. Arabic lexicography was founded by al-farāhīdī (died in 791) who constructed the first Arabic lexicon kitāb al-‘āyn ‘al-‘āyn lexicon’. Over the past 1,400 years, many Arabic lexicons were constructed. The lexical entries (i.e. roots) appear in Arabic dictionaries and followed by a definition part which may span several pages. The definition part is written as a unit or an encyclopaedic article which defines all the derived words from a certain root. These lexical entries are not arranged or distinguished with special formatting. Figure [1] shows samples of a lexical entry (the root كتب k-t-b) with the definition part from the traditional Arabic dictionary “lisān al-‘arab”. Figure [2] is the human English translation of the lexical entry sample listed in the first figure. The derived words in both figures are manually underlined and highlighted in blue.

Figure 1. A sample of text from the traditional Arabic dictionary “lisān al-‘arab” for the lexical entry (الكتاب) where the derived words of the root (k-t-b) are underlined and highlighted in blue.

Four main classes of ordering lexical entries in lexicons were developed and followed by authors of Arabic lexicons. Three arrangement methodologies depend on the roots of the words as lexical entries for Arabic lexicons. The fourth one groups lexical entries according to their conceptual themes or topical frames. These arrangement methodologies are different than those used in modern English dictionaries. Lexical entries of common English dictionaries, which are words (i.e. lexical entries in form of lemmas), are arranged alphabetically followed by the
The first arrangement methodology of lexical entries of Arabic lexicons is the al-ḥalil methodology. It was developed by al-ḥalil ibn ʿahmad al-farāḥidī (died in 791). The second arrangement methodology is the al-ḡawhari methodology which was developed by ʿismāʿil ibn ḥammād al-ḡawhari (died in 1002). The al-barmakī methodology is the third arrangement methodology. This arranging method was developed by abū maʿāli mohammād bin ʿamīm al-barmakī, who lived in the same time period as al-ḡawhari. al-barmakī did not construct a new lexicon; but he alphabetically re-arranged a lexicon called ʿāš-ṣiḥāḥ fī al-ḍuʿāʾ (الساح في اللاء). For these three ordering methods, roots are considered the lexical entries. The last methodology is the abū ʿubayd methodology which was developed by abū ʿubayd al-qāsim bin sallām (died in 838). The following sections discuss the arrangement methodologies for lexical entries of traditional Arabic dictionaries.

Figure 2. A human translation of the sample of text from the traditional Arabic lexicon “lisān al-ʿarab”, the target lexical entries are highlighted in blue and square brackets

2.1 The al-ḥalil Ordering Methodology

The first traditional Arabic lexicon is called kitāb al-ʿayn “al-ʿayn lexicon”. It was developed by al-ḥalil ibn ʿahmad al-farāḥidī (died in 791). The al-ḥalil ordering methodology, which was followed in constructing ‘The al-ʿayn’ lexicon, arranges the lexical entries phonologically according to places of articulation of phonemes from the mouth and throat, working forwards from glottal through to labial regions. The al-ʿayn lexicon was divided into books, where one book was dedicated for each letter. Each book was then divided into 4 sections according to their internal structure: (i) doubled biliteral roots; (ii) intact triliteral roots; (iii) doubly-defective roots; and (iv) quadriliteral and quinquiliteral roots. Many lexicons followed al-ḥalil’s methodology with slight modifications. Table [1] lists some of traditional Arabic Lexicons that followed al-ḥalil’s methodology.
2.2 The al-ğawharῑ Methodology

Ismā‘īl bin Ḥammad al-ğawharῑ (died in 1002) constructed a lexicon called as-siḥāḥ fī al-luḡaḥ “The Correct Language”. Roots are the lexical entries of this lexicon. They were alphabetically ordered according to their last letter, then the first letter. This methodology is called the al-ğawharῑ methodology. The lexicon was organized into chapters where each chapter corresponds to the last letter of the root. Each chapter includes sections corresponding to the first letter of the root, then the second letter of triliteral roots, then the third letter of quadriliteral roots, then the fourth letter in quinquiliteral roots. For example, the word baṣaṭ “spread” which is derived from the root (b-s-t) is found in chapter ط representing the last letter of the root, and in section ب representing the first letter of the root. Table [1] lists some of traditional Arabic Lexicons that followed this ordering methodology.

2.3 The al-barmakῑ Methodology

The third lexicon ordering methodology is “The al-barmakῑ methodology”. It was developed by abū al-ma‘ālī muḥammad bin tamīm al-barmakῑ (died in 1006). In this methodology, lexical entries (i.e. roots) are alphabetically arranged according to the first letter of the root. Al-barmakῑ lived in the same period as al-ğawharῑ. Al-barmakῑ did not construct a new Arabic lexicon. Instead, he re-arranged the lexical entries of as-siḥāḥ fī al-luḡaḥ, which was developed by al-ğawharῑ. The al-barmakῑ methodology was followed by az-zamaḥṣarī (died in 1143) in constructing his lexicon asās al-balāḡa “Fundamentals of Rhetoric”. Table [1] lists Arabic lexicon which followed the al-barmakῑ methodology for ordering lexical entries. The al-barmakῑ methodology for ordering lexical entries becomes the most widely used ordering methodology for Arabic lexicons.

2.4 The abū ‘ubayd Methodology

Abū ‘ubayd al-qāsim bin sallām (died in 838) developed the fourth ordering methodology for Arabic lexicons which is called “The abū ‘ubayd methodology”. This methodology arranges and groups together lexical entries according to their semantic fields. This arrangement methodology is similar to arranging lexical entries in modern thesauri. Many lexicons followed this ordering methodology. al-ğarib al-muṣannaf fī al-luḡaḥ “The Irregular Classified Language” by abū ‘ubayd al-qāsim bin sallām was the first lexicon that followed this methodology. This lexicon includes many small books that describe similar topics (i.e. group words of similar meanings) such as books describing horses, milk, honey, flies, insects, palms, and human creation. Then, more than thirty small books were collated into one large lexicon. Figure [3] shows a sample from Colours’ Book taken from al-ğarib al-muṣannaf fī al-luḡaḥ lexicon. Table [1] lists traditional Arabic lexicons that followed abī ‘ubayd methodology.

Figure 3. A sample of the Irregular Classified Language lexicon
## Table 1. Examples of Traditional Arabic Lexicons classified according to their Arrangement Methodology

| Arrangement Methodology | Traditional Arabic Lexicons following this Arrangement Methodology |
|-------------------------|---------------------------------------------------------------------|
| **1. The al-ḥalīl Methodology** |  
| 1. The intertextual arrangement method | *kitāb al-ʿayn al-ʿayn Lexicon* by Khalil bin Ahmad al-Ḥalīl (died in 1755/791 AD). |
| 2. The translational arrangement method | *muʿjam al-muḥīṭ fi al-luḡāt “The Comprehensive Language”* by as-Sāḥib bin ʿAbbād (died in 385/995 AD). |
| 3. The alphabetical arrangement method | *al-muḥākam wa al-muḥīṭ al-ʿaʿзам “The Greatest Verified and Comprehensive Lexicon* by Ibn Sayyidah, Abū al-Ḥasan (died in 458/1065 AD). |

| **2. The al-ḡawharī Methodology** |  
| 1. The alphabetical arrangement method | *al-ʾibāḥ az-zāḥīr fi al-luḡāt “The High Flood Water of Language”* by al-Hasan bin Muhammad as-Saqqānī (died in 650/1252 AD). |
| 2. The associative arrangement method | *al-ʿqāmīs al-muḥīṭ “The Comprehensive Dictionary”* by Muhammad bin ʿAbd al-Ḥafīz al-Fayrūz Abādī (died in 817/1414 AD). |
| 3. The thematic arrangement method | *tağ al-ʿarūs min ḡawāhir al-qaṃūs “Bridal Crown of Dictionaries”* by az-Zubaydī (died in 1205/1790 AD). |

| **3. The al-barmakī Methodology** |  
| 1. The alphabetical arrangement method | *muʿjam al-ɡīm “The Jīm Lexicon”* by Abū ʿAmr ʾAḥsān (died in 206/821 AD). |
| 2. The intertextual arrangement method | *gāmharat al-luḡāt “The Gathering of the Language”* by Ibn Zayd (died in 256/869 AD). |
| 3. The associative arrangement method | *muʿjam maṣyqīs al-luḡāt “The Lexicon of the Standard Language”* by Abu ʾAbdAllah ʾUthmān bin ʿAbd al-Husayn ʿAbd al-Rahīm bin Fāris bin Zayd (died in 395/1004 AD). |
| 4. The thematic arrangement method | *muʿjam māʾ istaʿgām “A Lexicon of Foreign Words”* by al-Bakrī al-Andalusi (died in 487/1094 AD). |

| **4. The al-ḡawharī Methodology** |  
| 1. The alphabetical arrangement method | *tahdīb al-ʿafāl “The Refined Verbs”* by Abu ʾAbdAllah ʾUthmān bin ʿAbd al-Rahīm bin Fāris bin Zayd (died in 515/1121 AD). |
| 2. The associative arrangement method | *al-muṣbāḥ al-ɡīm “The Selected of the Correct Language”* by Abū ʿAbbās ʿAbdAllah (died in 535/1143 AD). |
| 3. The thematic arrangement method | *asās  al-balāḡa “Fundamentals of Rhetoric”* by Abū ʾAbdAllah ʾUthmān bin ʿAbd al-Rahīm bin Fāris bin Zayd (died in 535/1143 AD). |
| 4. The intertextual arrangement method | *lāgūh al-qaṃūs ṭamāma al-ḥamāwī, Abū al-ʿAbbās (died in 666/1267 AD). |

| **5. The al-ḡawharī Methodology** |  
| 1. The alphabetical arrangement method | *muʿjam al-ʿafāl al-muṭaʿād “The Lexicon of Transitive Verbs”* by Musa bin Muhammad al-Małyānī al-ʿAḥmādī (published in 1979). |

**Notes:**
- **No. 1:** *Kitāb al-ʿayn al-ʿayn Lexicon* by Khalil bin Ahmad al-Ḥalīl (died in 1755/791 AD).
- **No. 2:** *Muṣḥāb al-ʿayn Lexicon* by Abu ʿAbdAllah ʾUthmān bin ʿAbd al-Rahīm bin Fāris bin Zayd (died in 535/1143 AD).
The motivation behind constructing the TAL-Corpus is to collect and organize well-established and long traditions of traditional Arabic lexicon in one freely available resource. The TAL-Corpus will help Arabic lexicographers to design and construct new modern Arabic dictionaries. These dictionaries can have new ordering methodology where derived words can be easily linked with their lexical entries whether they are roots or lemmas. The TAL-Corpus can be used to determine the origin of Arabic vocabulary and can track the development and changes of their meanings. The TAL-Corpus can also be used to extract useful information that supports Arabic NLP applications such as root extraction applications, morphological analyzers, semantic networks of Arabic vocabulary, WordNets, ontologies … etc.

The following sections show the design criteria followed in constructing the TAL-Corpus. Atkins et al., (1992) proposed general criteria of corpus design. These principal aspects and standards were recommended to be followed to inform major decisions for corpus creation. These criteria were designed to support high-quality and compatible corpora regardless of the corpus language, purpose, and location. Sections 3.1 to 3.5 discuss the design criteria followed to construct the TAL-Corpus.

3.1 Text

The text of the TAL-Corpus was collected from 23 freely available traditional Arabic lexicons. These lexicons are listed in Table 1. Al-Meshkat Islamic Network ḳaba‘a‘ mīshā ar-‘ilāmiyya provides most of these lexicons freely. These lexicons have been key-boarded (i.e typed) and put online in machine readable formats as MS-Word (.doc) or HTML text files.

The texts of the collected Arabic dictionaries were organized using different ordering methodologies as discussed in Section 2. However, most of these lexicons use roots as their main lexical entries. The definition of a root in each lexicon is written as an encyclopaedic article that contains the derived words from that root, their meanings, and examples of usages. These definitions vary in size from half a page to span several pages. Figure [4] shows a sample of text of a lexical entry taken from a traditional Arabic lexicon; the derived words are underlined and highlighted in blue. The text of the collected lexicons is fully vowelized, partially vowelized or non-vowelized. Texts (i.e. definitions) of similar roots from the different traditional Arabic dictionaries were grouped together in the TAL-Corpus. Then, several automatic processing steps and algorithms were applied to extract relevant linguistic information such as derived words and lemmas. Sections 3.4 and 3.5 discusses in detail these processing steps and algorithms.

For all collected lexicons, common processing steps were applied. These steps include: (i) converting the file
formats from MS Word or HTML web pages into standard text files in Unicode 'utf-8' encoding. (ii) A statistical analysis was applied that computed the words frequencies and the vocabulary size for both vowelized and non-vowelized text of the corpus. As a result, the complete TAL-Corpus contains 14,369,570 words, 2,184,315 vowelized word types and 569,412 non-vowelized word types (i.e. after removing short vowels (diacritics) from the text). Table [2] shows the summary of the statistical analyses of the lexicon texts used to construct the TAL-Corpus. Figure [5] shows the highest 25 frequent words in the TAL-Corpus of partially vowelized and non-vowelized forms of words.

Table 2. Statistical Analysis of the Lexicons’ Text used to construct the TAL-Corpus

|               | Number of files | 247          |
|---------------|----------------|--------------|
| **Size**      |                | 178.32 MB    |
| **Vowelized word** |               |             |
| Number of words |                | 14,369,570   |
| Number of word types |         | 2,184,315    |
| **Non-vowelized word** |           |             |
| Number of words |                | 14,369,570   |
| Number of word types |         | 569,412      |

Table 2. Statistical Analysis of the Lexicons’ Text used to construct the TAL-Corpus

| Partially-vowelized | Frequency | Non-vowelized | Frequency |
|---------------------|-----------|---------------|-----------|
| في      | fi “in”   | من      | min “from”   |
| من      | min “from”| في      | fi “in”     |
| قال     | qāl “he said” | قال     | qāl “he said” |
| و       | wa “and”  | أي      | ’ayy “which” |
| على     | ’alā “over”| و       | wa “and”    |
| ما       | mā “what” | على     | ’alā “over” |
| وقال    | wa qāl “and he said” | إذا | 'īhā “if” |
| عن      | ‘an “about”| و قال | wa qāl “and he said” |
| إذا      | ’īhā “if” | ابن     | ’ibn “son of” |
| أي      | ’ay “which”| ما      | mā “what”   |
| وهو     | wa huwa “and he” | بن | bin “son of” |
| لا       | là “no”   | عن      | 'an “about” |
| ابن      | ’ibn “son of” | وما  | wa huwa “and he” |
| به      | bihi “in it” | لا    | là “no”     |
| وفي     | wa fī “and in” | أبو   | abā “father” |
| وقد     | wa qad “and perhaps” | أَو  | ’aw “or” |
| أبو      | abū “father” | الله  | allāh “Allah” |
| بن      | bin “son of” | الله  | allāh “Allah” |
| أي      | ’ay “which”| bihi  | “in it”     |
| هو      | huwa “he”  | ـ  | yuqāl “it is said” |
| يقال    | yuqāl “it is said” | و في | wa fi “and in” |
| عليه    | ‘alayhi “about him” | و قد | wa qad “and perhaps” |
| ولا     | wa lā “and not” | عليه | ‘alayhi “about him” |
| اللهو    | allāh “Allah” | هو   | huwa “he”   |

Figure 5. The first 25 words of the frequency list generated from the TAL-Corpus Corpus
The analysis represented by Tables [3] and [4] and Figure [6] classifies the traditional Arabic lexicons which were included in the TAL-Corpus, according to the time of construction. The time period spans around 14 centuries since the first Arabic lexicon was created (i.e. from the second Hijri century to the fifteenth Hijri century). This time span was divided into 14 time frames where each corresponds to 100 years. These time frames were defined by the creation times of the traditional Arabic dictionaries which are indicated by the death date of dictionaries’ authors. The first time frame includes one lexicon kitābu al-'ayn which consists of 348 114 words and 141 098 word types which forms 2.42% of the text size and 3.72% of the vocabulary size of the TAL-Corpus. The lexicons from 12th century are the largest. They contain 5 215 917 words and 1 211 432 word types. They form 36.30% of the TAL-Corpus text and 31.90% of its vocabulary size. The lexicons included in this time frame are taq al-‘arās min ḡawāhir al-qaṃūs and mu’ğam tahdīb al-luḡa which represent the largest in terms of number of words and vocabulary size.

Table 3. Text and vocabulary size of the Traditional Arabic Dictionaries and their percentage in the TAL-Corpus

| Time Frame | Lexicon Name | Date (Died in) | # Words | # Types | % of Words | % of Types |
|------------|--------------|----------------|---------|---------|------------|------------|
| 1          | kitābu al-'ayn | 175H (791AD)  | 348,114 | 141,098 | 2.42%      | 3.72%      |
| 2          | mu’ğam al-ḡīm | 206H (821AD)  | 125,676 | 56,274  | 0.87%      | 1.48%      |
| 3          | al-ḡarib al-muṣannaf fī al-luḡa | 223H (838AD) | 16,541 | 7,775   | 0.12%      | 0.20%      |
|            | ḡamharat al-luḡa |            | 256H (869AD) | 396,144 | 123,576   | 2.76%      | 3.25%      |
|            | al-munaqqadāt fī al-luḡa | 310H (922AD) | 32,173 | 16,942  | 0.22%      | 0.45%      |
|            | mu’ğam al-muhūt fī al-luḡa | 385H (995AD) | 392,246 | 168,870 | 2.73%      | 4.45%      |
|            | mu’ğam maqāyīs al-luḡa | 395H (1004AD) | 445,126 | 129,838 | 3.10%      | 3.42%      |
| 4          | as-ṣīḥāh fī al-luḡa | 400H (1009AD) | 593,654 | 118,591 | 4.13%      | 3.12%      |
|            | al-muḥkam wa al-muhūt al-‘a’azam | 458H (1065AD) | 1,020,137 | 279,157 | 7.10%      | 7.35%      |
|            | al-muḥṣasas fī al-luḡa | 458H (1065AD) | 902,324 | 274,780 | 6.28%      | 7.24%      |
|            | mu’ğam ma ‘ista’qam | 487H (1094AD) | 278,713 | 43,289  | 1.94%      | 1.14%      |
| 5          | tahdīb al-aṭfā’al | 515H (1121AD) | 132,319 | 38,102  | 0.92%      | 1.00%      |
| 6          | asās al-balāḡa | 538H (1143AD) | 289,436 | 95,887  | 2.01%      | 2.52%      |
|            | al-muğrib fī tartīb al-mu’rib | 610H (1213AD) | 128,047 | 39,930  | 0.89%      | 1.05%      |
|            | al-‘ībāb az-zāḥīr fī al-luḡa | 650H (1252AD) | 261,658 | 100,536 | 1.82%      | 2.65%      |
|            | muḥtār as-ṣīḥāh | 666H (1267AD) | 171,487 | 40,295  | 1.19%      | 1.06%      |
| 7          | lisān al-‘rab | 711H (1311AD) | 2,146,545 | 507,860 | 14.94%     | 13.37%     |
|            | al-muṣbāb al-munīr fī ḡarīb al-ṣarḥ al-kabīr | 770H (1368AD) | 219,276 | 61,422  | 1.53%      | 1.62%      |
| 8          | al-qaṃūs al-muhūt | 817H | 563,460 | 203,600 | 3.92%      | 5.36%      |
| Frame | Time frame | # of dictionaries | # words | # types | % of words | % of types |
|-------|------------|-------------------|---------|---------|------------|------------|
| 1     | 100H-199H (718AD-814AD) | 1 | 348,114 | 141,098 | 2.42% | 3.72% |
| 2     | 200H-299H (815AD-911AD) | 3 | 538,361 | 187,625 | 3.75% | 4.94% |
| 3     | 300H-399H (912AD-1008AD) | 3 | 869,545 | 315,650 | 6.05% | 8.31% |
| 4     | 400H-499H (1009AD-1105AD) | 4 | 2,794,828 | 715,817 | 19.45% | 18.85% |
| 5     | 500H-599H (1106AD-1202AD) | 1 | 132,319 | 38,102 | 0.92% | 1.00% |
| 6     | 600H-699H (1203AD-1299AD) | 4 | 850,628 | 276,648 | 5.92% | 7.28% |
| 7     | 700H-799H (1300AD-1396AD) | 2 | 2,365,821 | 569,282 | 16.46% | 14.99% |
| 8     | 800H-899H (1397AD-1493AD) | 1 | 563,460 | 203,600 | 3.92% | 5.36% |
| 9     | 900H-999H (1494AD-1590AD) | 0 | - | - | - | - |
| 10    | 1000H-1099H (1591AD-1687AD) | 0 | - | - | - | - |
| 11    | 1100H-1199H, 1688AD-1784AD | 0 | - | - | - | - |
| 12    | 1200H-1299H, 1785AD-1881AD | 2 | 5,215,917 | 1,211,432 | 36.30% | 31.90% |
| 13    | 1300H-1399H, 1882AD-1978AD | 1 | 615,352 | 112,164 | 4.28% | 2.95% |
| 14    | 1400H-Today, 1979AD | 1 | 75,225 | 26,299 | 0.52% | 0.69% |

Figure 6. Traditional Arabic dictionaries included in the TAL-Corpus

3.2 Text Handling
After collecting the text of 23 traditional Arabic dictionaries, common pre-processing steps were applied. First, all dictionaries’ files were converted into standard text files using Unicode ‘utf-8’ encoding. Then, the SALMA-Tokenizer and the SALMA-root extractor and Lemmatizer (Sawalha, 2011) were used to tokenize and process Arabic words by stripping diacritics, and extracting the root and the lemma for each word in the TAL-Corpus. Third, frequency lists of both vowelized and non-vowelized word were generated (see Table [1] and Figure [5]).

Special algorithm was developed to extract the derived words of the lexical entries for the dictionaries included in the TAL-Corpus. The purpose of this algorithm is to group together roots and their definition parts and then to extract derived words of roots from their related definition articles. To achieve this goal, a specific treatment were applied to each dictionary text. The 23 collected dictionaries were originally constructed following an ordering methodology of their lexical entries as discussed in Section 2. Most of them use roots as their main head words of lexical entries. These dictionaries were typed into machine-readable files in different formats without using any lexicographic representations that can be recognized by Computers. Therefore, specialized programs were developed for each dictionary to reformat and extract useful information such as roots, definitions and derived words.

The root-definition structure is the common basic structure for most traditional Arabic dictionaries. Each lexical entry consists of the root as a head word and the definition part. The definition part is written as an encyclopaedic feature using free writing style. These encyclopaedic articles defines the root and its derived words and their linguistic attributes are specified. However, the derived words of a root within the definition part are neither structured nor ordered. This free writing style requires the authors of dictionaries to add affixes and clitics to the derived words within the definition parts. Clitics, such as conjunctions, prepositions and connected pronouns, are used to connect sentences and paragraphs of these definition articles.

For the above mentioned-reasons, the free writing style of the definition part adds extra challenges to extract the derived words and their definitions. Therefore, a dedicated algorithm was developed to extract the roots and their derived words from the dictionaries’ texts. The tokenizing module in the program specifies the boundaries of a lexical entry which is normally starts with a root followed by an article that defines that root. For each lexical entry, the algorithm extracts and pairs words from the definition part with the root and stores them in vectors (i.e. bag of words). Many of these word-root pairs are not correct matches (i.e. the word is not derived from the associated root). A normalization analysis verified these word-root pairs by throwing out pairs where the word is not derived from its associated root. The normalization procedure applies linguistic knowledge that governs the derivation process of words from their roots. These linguistic rules were used to match the consonant letters of words and roots and their order for each word-root pair. The first linguistic rule checks if all consonant letters forming the root appear in the paired word. The second rule examines if all root letters order appear in the derived word. Both rules must be applied to every word-root pair for verification. This process is applied to extract the derived words of a root and later to build a morphological lexicon (See Section 3.3.1). Figure [7] shows the process of selecting word-root pairs. Table [5] shows the number of words and the percentage of words extracted from the original text of the dictionaries.

| Word-root vector for the root كتب | كتب (كتاب، كتب) | كتب (الكتاب، كتب) | كتب (كتب، كتب) | كتب (كتب، كتب) |
|----------------------------------|-----------------|-----------------|-----------------|-----------------|
| (مختلف، كتب)                    | (عذاب، كتب)     | (خاطبة، كتب)    | (كتاب، كتب)     | (كتاب، كتب)     |
| (كتب، كتب)                      | (مارف، كتب)     | (سافر، كتب)     | (كتاب، كتب)     | (كتاب، كتب)     |
| (كتاب، كتب)                      | (قضى، كتب)     | (كانت، كتب)     | (كتاب، كتب)     | (كتاب، كتب)     |
| (كتاب، كتب)                      | (الكتاب، كتب)   | (كتاب، كتب)     | (كتاب، كتب)     | (كتاب، كتب)     |

Figure 7. Using linguistic knowledge to select word-root pairs from traditional Arabic lexicons. The selected word-root pairs are underlined and highlighted in blue.

| Lexicon name | Word types | Words extracted | Roots extracted |
|--------------|------------|-----------------|-----------------|
|              |            |                 |                 |

Table 5. Words and Roots Extracted from 8 Traditional Arabic lexicons
Table 6. Number of records extracted and inserted in the SALMA sexiliteral roots.

| #  | Lexicon | Word [B] | types | Records inserted [A] | Percentage (A/B)% | (A/C)% |
|----|---------|----------|-------|----------------------|------------------|--------|
| 1  | tağ al-'arūs min ḡawāhir al-qāmūs | 831,504 | 474,351 | 57.05% | 11,101 |
| 2  | lisān al-'rab | 507,860 | 274,305 | 54.01% | 9,355 |
| 3  | muʿgam al-muhūf fi al-luḡa h | 168,870 | 66,763 | 39.54% | 6,411 |
| 4  | kitāb al-ayn | 141,098 | 54,970 | 38.96% | 5,826 |
| 5  | al-muʿgam al-wasīf | 112,164 | 45,614 | 40.67% | 6,489 |
| 6  | al-muṣbāḥ al-munīr fi ḡarīb ʾaš-šarḥ al-kabīr | 61,422 | 29,742 | 48.42% | 2,947 |
| 7  | muḥārīr ʾaṣ-ṣīḥāh | 40,295 | 17,636 | 43.77% | 3,420 |
| 8  | al-muğrab fī tartīb al-muʿrab | 39,930 | 13,798 | 34.56% | 2,322 |

3.3 Advanced Text Handling

The TAL-Corpus implements advanced text handling tools which can automatically process linguistic information in a corpus and allow more sophisticated statistical analyses. Lexical database (i.e. the SALMA-ABCLexicon) was created using the extracted information from the TAL-Corpus text.

3.3.1 Link to Lexical Database

The TAL-Corpus was used to construct the SALMA-ABCLexicon. The SALMA-ABCLexicon is a lexical database that contains around three million word-root pairs. This lexical database was extracted from the text of the TAL-Corpus following the analyses steps as described in Section 3.2. These steps include (i) manually converting the traditional Arabic dictionaries’ text into a unified format; (ii) a specialized algorithm extracts a bag of words from the definition part text of Arabic dictionaries where word-root pairs are stored; (iii) two linguistic rules were applied to the word-root pairs to verify that words are derived from the associated roots.

Later, a specialized program combines the disparate lexicon information into one large broad-coverage lexical resource the SALMA-ABCLexicon. A lexical information of a large dictionary called lisān al-ʿrab ‘Arab tongue’ was feed to the program as a seed for the SALMA-ABCLexicon. All word-root pairs of the first dictionary were included in the SALMA-ABCLexicon which represent around 48% of the total records. Around 82% of the words and roots of the mahāzīt fi al-luḡa h dictionary were added which represents around 14% of total records. The percentage of added records decreases during the combination process. This decrement indicates the termination of the combination process and which traditional Arabic dictionaries are better to construct the morphological lexicon. Table [6] shows the traditional Arabic dictionaries that were used to construct the SALMA-ABCLexicon. It also shows the number of records and their percentage that contribute to the construction of the SALMA-ABCLexicon.

The SALMA-ABCLexicon contains 2 774 866 word-root pairs that represent 509 506 different words and 261 125 different non-vowelized words. It contains 12 729 roots that are distributed into 12 biliteral roots; 8 585 trilateral roots; 4 038 quadrilateral roots; 63 quinqueliteral roots; and 31 different sexiliteral roots. The 509 506 word types of the lexicon are distributed into; 117 word types derived from biliteral roots; 483 356 word types of trilateral roots; 30 873 word types of quadrilateral roots; 615 word types of quinqueliteral; and 335 word types of sexiliteral roots. Figure [8] shows the first 60 derived words of the root k-t-t-b ‘wrote’.

Table 6. Number of records extracted and inserted in the SALMA-ABCLexicon.

| #  | Lexicon | Word [B] | types | Records inserted [A] | Percentage (A/B)% | (A/C)% |
|----|---------|----------|-------|----------------------|------------------|--------|
| 1  | lisān al-ʿrab | 207,992 | 207,992 | 100.00% | 47.80% |
| 2  | muʿgam al-muhūf fi al-luḡa h | 74,507 | 61,113 | 82.02% | 14.04% |
| 3  | tağ al-ʿarūs min ḡawāhir al-qāmūs | 128,119 | 95,415 | 74.47% | 21.93% |
| 4  | muḥārīr ʾaṣ-ṣīḥāh | 19,540 | 16,573 | 84.82% | 3.81% |
| 5  | al-muğrab fī tartīb al-muʿrib | 12,396 | 9,805 | 79.10% | 2.25% |
| 6  | kitāb h al-ayn | 30,292 | 18,878 | 62.32% | 4.34% |
| 7  | al-muʾgam al-wasīf | 36,660 | 25,364 | 69.19% | 5.83% |

Totals 509,506 435,140 [C]
إذا أستَمْلاه وَأَنجَحَهَا الصَّواب وقيل واَكْتَبَ وَأَنْجَحُ تَنَجَّزُها وكذا النَّجْحِ إذا أو كَكَتَبَه واَكْتَبَه تِكِتَّابَه، ومن وَأَرِيت بالكسر وَسَارَ بالفَتْح، وفي النَّجَاح وَشِيَّك أَي وَاسْتَكْتَبَه كَتَبَه خَطَّه، الشَّدِيد عن الأَساس أصل ه وقِد ثم.

Figure 8. The first 60 lexical entries of the root كتاب k-t-b "wrote" stored in the SALMA-ABCLexicon

4. The TAL-Corpus Markup

Markups are introduced to the TAL-Corpus to indicate its features such as lexicon name, lexical entry, and definitions of lexical entries. The TAL-Corpus is formatted using XML technology where lexicons are reformatted and their lexical entries are alphabetically arranged. All traditional Arabic lexicons that form the TAL-Corpus are stored using XML files. XML is a markup language that facilitates the labelling or tagging of corpus features. The use of XML allows formatting and labelling the features of the TAL-Corpus. Figure 9 shows the XML structure and the labels used to format the corpus files.

Figure 9. XML structure of The Corpus of Traditional Arabic Lexicons
These corpus markups were effectively used when a web interface\(^2\) for searching the contents of the corpus was developed. The web interface allows users to access the contents of the corpus, to search for a root and to retrieve the definition parts from the traditional Arabic lexicons included in the TAL-Corpus. Figure 10 shows part of the web interface for part of the results after searching for the root “كتب” \(k-t-b\).

5. Evaluation

The purpose of constructing the TAL-Corpus is to introduce a new lexicographic corpus that contains the majority of standard Arabic vocabulary. This kind of corpus will not only help in the design and development of Arabic monolingual dictionaries but also it can support constructing Computational Linguistics resources such as; morphological dictionaries, frequency lists, lexical and morphological databases, etc. The SALMA-ABCLexicon is a lexical and morphological dictionary that was constructed using the TAL-Corpus text (see Section 3.3.1). It contains slightly under three million word-root pairs.

There are no mature standard criteria for evaluating newly constructed text corpora (Atkins et al, 1992). Therefore, our criteria for evaluating the TAL-Corpus should meet the goal for construction. We need our corpus to include the majority of standard Arabic vocabulary. Moreover, these vocabularies should be diverse and cover contemporary as well as classical ones. Lexical diversity is defined by McArthy and Jarvis (2010) as “the range of different words used in a text, with a greater range indicating a higher diversity”. Lexical diversity (LD) is computed as the token-type ratio. The lexical diversity of the TAL-Corpus scored 0.152. It was evaluated by comparing it against the LD of rival Arabic corpora. The Arabic Web 2012 (arTenTen) corpus belongs to the TenTen corpora family which was created by harvesting web pages using SpiderLing. It contains around 7.5 billion tokens which represents around 2 million word types (Arts et al, 2014). Its LD scored about 0.000263. Similarly, the Arabic Internet Corpus was developed by harvesting articles from webpages published in Arabic. It contains around 165 million tokens and more than 4 million different tokens. Its LD is computed and scored 0.025965. The third corpus used in this comparative evaluation is the Arabic Wikipedia corpus (wiki-ar)\(^3\). It contains around 16 million tokens and slightly less than 1 million types. The LD for this corpus scored 0.057. Table 7 summarizes the LD for the 4 corpora used in the comparative evaluation. It shows that the LD of the TAL-Corpus scored the highest. Although it is similar size compared to the Arabic Wikipedia Corpus, its LD is 2.7 times higher. In comparison with large Arabic corpora namely: the Arabic Internet Corpus and the Arabic Web 2012 Corpus, although these large corpora contains large amounts of texts harvested from webpages, their LD is less in magnitude of times than the LD of the TAL-Corpus.
Another criteria for evaluating the TAL-Corpus is based on the coverage of its vocabulary on different types of text corpora. The evaluation experiments were performed using the SALMA-ABCLexicon and three text corpora: the Qur'an, the Arabic Internet Corpus, and the Corpus of Contemporary Arabic. The SALMA-ABCLexicon was used because it was constructed using the TAL-Corpus and it contains all the vocabulary instances from the TAL-Corpus. The three corpora were selected to represent different types of Arabic text. The Qur’an represents Classical Arabic; the Corpus of Contemporary Corpus represents Modern Standard Arabic; and a snapshot of current Arabic language on the web is represented by the Arabic Internet Corpus.

Two experiments were conducted to compute the coverage of the TAL-Corpus. The first experiment is based on exact matching of the non-vowelized words of the three corpora with the non-vowelized words of the SALMA-ABCLexicon. The results of this experiment scored a coverage of 67.53% for the Qur’an; 65.58% for the Arabic Internet Corpus; and 67.5% for the Corpus of Contemporary Arabic. Table [8] and Figure [11] show the results of the first coverage experiment. Some tokens are not words (i.e. Arabic words) but numbers, dates, currency symbols, punctuations, HTML or XML tags and English words. Only Arabic words were selected to compute the coverage of the SALMA-ABCLexicon.

Table 8. The coverage of the lexicon using exact word-match method

| Corpus    | Tokens | Arabic words | Covered words | Coverage % |
|-----------|--------|--------------|---------------|------------|
| Qur’an    | 77 800 | 77 799       | 52 536        | 67.53%     |
| CCA       | 684 726| 594 664      | 389 133       | 65.44%     |
| Internet  | 1 128 114| 833 916   | 546 880       | 65.58%     |

Figure 11. The coverage percentage of the TAL-Corpus using exact match method

Arabic is a morphologically rich language. Therefore, most Arabic words in context are complex words. Clitics and affixes are attached to the words in context which remarkably increase the various forms of words. Clitics make the matching process with lexical entries of the SALMA-ABCLexicon not an easy task. Hence, the coverage percentage would decrease. As an alternative, the coverage of the TAL-Corpus was computed by matching the lemmas of the SALMA-ABCLexicon with the lemmas of the three corpora. The SALMA-Lemmatizer (Sawalha, 2011) was used to lemmatize the three corpora and the lexical entries of the SALMA-ABCLexicon. The SALMA-Lemmatizer also includes a list of function words. The other part of this experiment excludes function words from the coverage calculations. Tables [9] and [10] show the coverage percentage of the TAL-Corpus computed by matching lemmas including and excluding the function words respectively. Figure 12 shows a summary of the coverage of the TAL-Corpus based on matching lemmas.
Table 9. Coverage of lemmas including function words

| Corpus | Tokens | Words | Covered words | Coverage % |
|--------|--------|-------|---------------|------------|
| Qur’an | 77 804 | 77 803 | 64 065        | 82.34%     |
| CCA    | 685 161| 595 099| 507 943       | 85.35%     |
| Internet | 1 128 624 | 834 426 | 708 101      | 84.86%     |

Table 10. Coverage lemmas excluding function words

| Corpus | Tokens | Words | Covered words | Coverage % |
|--------|--------|-------|---------------|------------|
| Qur’an | 77 804 | 54 004 | 42 532        | 78.76%     |
| CCA    | 685 161| 411 482| 338 790       | 82.33%     |
| Internet | 1 128 624 | 576 407 | 476 190      | 82.61%     |

Figure 12. Coverage percentage of the TAL-Corpus using the lemmatizer

The average coverage percentage of the TAL-Corpus is 84.18% when matching the lemmas of the three corpora with the lemmas of the SALMA-ABCLexicon including function words. The coverage of the TAL-Corpus scored highest at 85.35% when computed using the CCA Corpus. The coverage scored 84.86% and 82.34% using the Internet corpus and the Qur’an respectively. The average coverage percentage of the TAL-Corpus is 81.23% after excluding function words. The highest coverage percentage was achieved using the Arabic Internet Corpus at 82.61%. Similar coverage percentage at 82.33% was achieved using the CCA corpus. Finally, 78.76% was the coverage percentage scored when the Qur’an lemmas were matched excluding function words.

The evaluation experiments of the TAL-Corpus by computing its coverage against three Arabic corpora showed that it does not fully cover words that belong to the categories; (i) function words; (ii) new Arabic terms; (iii) relative nouns; and (iv) borrowed words. Function words such as ذَلِكََ “that”; وَإَِلَى “and to”; إِنَّهُم َ “they are”; and التي “which” were not covered in the TAL-Corpus. These words can be easily added by including traditional Arabic grammar books in the corpus (Diwan 2004). Second, new Arabic terms such as “chat”; انقر “click” and الانتخابات “elections” are not covered because these words have appeared recently due to recent technical and social developments. Unfortunately, modern Arabic dictionaries are not available in machine readable format. Therefore, including these dictionaries in the TAL-Corpus requires retyping these dictionaries and reformating them in a machine readable format. Third, relative nouns والأسماءَالمنسوبة are nouns that indicate affiliation of something to these nouns. Relative nouns such as دردشة “tourism”; الاجتماعيَّ “social”; and الثقافية “cultural” have become widely used in the media and modern standard Arabic. Annexing this group of words to the TAL-Corpus can be achieved by including modern Arabic dictionaries. Fourth, borrowed words such as الـدكتور “doctor”; الـإمِيل “e-mail”; الـتلفون “telephone”; and الإنترنت “Internet” are foreign words transliterated into Arabic by using Arabic letters. Borrowed words are frequently found in newspaper and web pages text because of the lack of standard translations of them. However, Arabic
Language Academies (i.e. organizations which are responsible for standardizing Arabic) are producing specialized dictionaries and word lists that translate these technical terms into Arabic. These specialized dictionaries can be included in the TAL-Corpus to increase its coverage. Figure [13] shows a sample of words which are not covered in the TAL-Corpus.

| Arabic           | English         | Category          |
|------------------|-----------------|-------------------|
| دليلكة           | That            | Economical        |
| اسماءة            | Skies           | The human         |
| عينهم           | They            | E-mail            |
| پیشنهاد          | Swear to God    | Telephone         |
| دارسة            | By the right    | Palestinian       |
| فآینه            | And those       | Click             |
| فبی         | In what         | American          |
| وعند            | It will         | Elections         |
| اتلیفا           | which           | States            |
| siyāḥiya          | United          | Social            |
| jāmī'iyah        | Doctor          | Internet          |
| al'insan         | Tourism         | Developmental     |
| al-mutahāda      | Interior        | Cultural          |
| al-imayl         | Election        |                  |
| al-filaṣīnī      | Election        |                  |
| nhằm             | After them      |                  |
| ینعور           | Arabic          |                  |
| لاشه             | American        |                  |
| iṣṭiḥābāt        | Interior        |                  |
| یتیمیا           | Social          |                  |
| I'tamad           | Developmental   |                  |
| یتماریکیا         |                  |                  |
| Dāẖilayt         |                  |                  |
| Iṯtiḥābāt        |                  |                  |
| Sīyāḥiyah        |                  |                  |
| یتینика          |                  |                  |
| al-Qābīyya       |                  |                  |
| Qalbīyya         |                  |                  |
| al-Ṭamīyya       |                  |                  |
| یتینیقا          |                  |                  |

Figure 13. A sample of common words which are not covered by the TAL-Corpus

6. Potential Users and Uses

The purpose for constructing the TAL-Corpus was to provide a collection of traditional Arabic dictionaries that can be analysed, studied and used to create comprehensive language resources such as; new Arabic dictionaries; frequency lists; collocates; morphological dictionaries, etc. Obviously, the potential users for the TAL-Corpus are lexicographers, Arabic linguists, language learners and computational linguists. The following is a discussion of potential uses of each expected user of this corpus.

- **Lexicographers**: This corpus was constructed as a resource for building new Arabic dictionaries. Therefore, lexicographers could use it to find examples of usage for words from different periods, track the changes in meaning of a certain vocabulary, and mark the origin of words and when they first appeared. The TAL-Corpus represents a bank of citations which are essential for the construction of new Arabic dictionaries. Citations denote objective evidence of language in use (Atkins and Rundell 2008).

- **Arabic linguists**: the TAL-Corpus provides the Arabic linguists with a repository of 23 traditional Arabic dictionaries. Feature labels (i.e. annotations (See Section 4) which were added to the corpus) make the search for a word, root, phrase or idiomatic expression easier via the corpus than paper based versions of traditional Arabic dictionaries. Arabic linguists are interested in studying the structures as well as the semantic features of words. The TAL-Corpus is an excellent resource for providing both. Word structures can be studied because roots and their derived words are provided. Semantic features of words such as the senses of the words; the changes to the meaning of the word; or new usage can be investigated and tracked using the TAL-Corpus. In addition, linguists can compare between the traditional Arabic dictionaries in terms of vocabulary size, ordering methodology and definitions of words. They also can conduct a comparison of other criteria such as features included in the dictionaries. These features can be derived from the different senses of words, phrases, idioms and examples of usage.

- **Language learners**: Arabic language learners of both native and nonnative speakers use Arabic dictionaries mainly to search for words' meanings. Searching traditional Arabic dictionaries, where roots are the lexical entries, is not easy as it requires learners to know the root of the words. The TAL-Corpus provides a collection of 23 traditional Arabic dictionaries which were annotated to facilitate searching for definitions of either a word or a root. Learner can search for a word and retrieve the definition of it in addition to other linguistic information such root, lemma, derived words of the same root or lemma, examples of usage, phrases and idioms.

- **Computational linguists**: Corpora are essentially used by computational linguists to build language models for machine learning algorithms. The TAL-Corpus could be used to build language models for Arabic morphological analysers, stemmers and lemmatizers. As well as, language models for semantic analysis can
be built for Arabic using the TAL-Corpus. Computational linguists can build tracking programs that investigate the development of Arabic vocabulary and the changes of their meanings. The TAL-Corpus includes traditional Arabic dictionaries of a period that span more than 1,200 years which enables tracking the development and changes of meaning for Arabic vocabulary. In conclusion, the TAL-Corpus is an essential resource for extracting useful information that supports a wide variety of Arabic NLP applications such as: root extraction applications, morphological analysers, semantic networks of Arabic vocabulary, WordNets, ontologies … etc.

7. Discussion of the Results, Limitations and Improvement

The TAL-Corpus is constructed using text from traditional Arabic dictionaries. It is characterized by a wide coverage of Arabic words, word types and roots. The evaluation proved that the TAL-Corpus has a wide coverage of about 85% of the test corpora words. Despite the time span of 13 centuries of the traditional Arabic lexicons from which the TAL-Corpus has been derived, only 15% of the test corpora words were not captured. The latest Arabic dictionary included in the TAL-Corpus is المعجم الوسيط al-mu'ğam al-wasif which appeared in 1960s. Hence, new vocabulary items added to Arabic in the past 50 years are not covered in the TAL-Corpus. Moreover, due to the advances in telecommunication and information technology; globalization; and the wide and intensive use of social networks, words of foreign languages have been increasingly used in both spoken and written Arabic. These foreign words do not have a proper translation into Arabic, but are written using Arabic letters (i.e. transliterated). Advances in telecommunication and information technology imply new products with their original names have entered Arab countries. These products keep their original names which have been widely used and become part of the contemporary Arabic vocabulary. Moreover, the use of dialectical Arabic has increased in the written and spoken forms due to open systems such as chat rooms, blogs and forums, and social networks which allow people to write text without restrictions.

The TAL-Corpus was used to construct a broad-coverage morphological database the SALMA-ABCLexicon. This database did not involve any manual correction due to the limitations in funding. However, an automatic correction and verification procedure was applied to part of the database. The verification procedure was performed by counting how many times the word-root pairs appear in the analyzed traditional Arabic dictionaries. 976,427 word-root pairs representing 35.19% of the lexicon’s word-root pairs scored a count of 2 or more. This means that these word-root pairs appeared in different dictionaries. Therefore, these word-root pairs have a high potential to be valid and correct.

This is the first version of the SALMA-ABCLexicon. It can be extended to include the full morphological analyses of the lexical entries and other useful information that will enhance the performance of NLP applications. Special linguistic lists such as compounds, collocations, idiomatic phrases, phrasal verbs and named entities can be added to extend the lexicon. Moreover, morphological lists such as broken plurals, intransitive and transitive verbs, rational and irrational words and primitive nouns can be another extension to the lexicon. The SALMA-ABCLexicon can also be extended by adding modern and dialect vocabularies from newly constructed Arabic corpora and the web.

8. Conclusions

The Corpus of Traditional Arabic Lexicons (the TAL-Corpus) is a special corpus which is constructed from the text of 23 traditional Arabic dictionaries. These dictionaries are spanning over a period of 1,200 years. The corpus contains 14,369,570 words and 2,184,315 word types. The motivation for building the TAL-Corpus is to collect and organize well-established and long traditions of traditional Arabic lexicons. The TAL-Corpus can also be used to construct new corpus-based Arabic dictionaries. Corpora were not used to construct Arabic dictionaries and lexical databases yet. Therefore, building corpora for the purpose of building new Arabic dictionaries is needed.

Thousands of traditional Arabic dictionaries were constructed in the past 1,200 years. These dictionaries are different size, type and ordering of their lexical entries. The wide variety of traditional Arabic dictionaries represent rich base for building a corpus that can be further used and exploit to construct new corpus-based Arabic dictionary.

The TAL-Corpus followed standard design and development criteria that informed major decisions in corpus creation. The text of the TAL-Corpus is composed from the text of 23 freely available and machine readable traditional Arabic dictionaries. These dictionaries were processed to have a unified format. The unified format is based on arranging the contents of the corpus by roots (i.e. the head words for the majority of traditional Arabic dictionaries) and their definitions. Then, the SALMA-root extractor and lemmatizer were used to tokenize, strip diacritics, and extract roots and lemmas for each word in the corpus. Frequency lists of both vowelized and non-vowelized word were also generated.
The SALMA-ABCLexicon is constructed by analysing the TAL-Corpus text. The processing steps in constructing the SALMA-ABCLexicon involve; applying linguistic rules that were encoded in a specialized program to extract the root and the words derived from that root. Second, a combination algorithm merges the information extracted from the previous step into one large broad-coverage lexical database. The SALMA-ABCLexicon contains 2,781,796 vowelized word-root pairs which represent 509,506 different non-vowelized words.

The TAL-Corpus is stored and distributed using XML technology. The corpus XML files contain all markups which indicate the corpus features. The choice of using XML technology is to facilitate the distribution and the use of the corpus. The TAL-Corpus is an open-source resource which is licenced under a Creative Commons Attribution-NonCommercial 4.0 International Licence.

The evaluation of the TAL-Corpus was done by computing its coverage over three Arabic corpora; the Corpus of the Contemporary Arabic; the Qur’an text; and the Arabic Internet Corpus. The coverage was computed by matching the words of the test corpora to the words in the SALMA-ABCLexicon, which scored about 67%. A lemmatizer program was used to compute the coverage by matching the lemmas of the test corpora and the lemmas of the SALMA-ABCLexicon. This method scored a coverage of about 82%.

The potential users for the TAL-Corpus are lexicographers, Arabic linguists, language learners and computational linguists. The potential practices for TAL-Corpus are to provide a collection of traditional Arabic dictionaries that can be analysed, studied and used to create comprehensive language resources such as; new Arabic dictionaries; frequency lists; collocates; morphological dictionaries, etc.

References
Abbas, M., & Smaili, K. (2005) Comparison of Topic Identification Methods for Arabic Language, RANLP05 : Recent Advances in Natural Language Processing, pp. 14-17, 21-23 September 2005, Borovets, Bulgaria.
Abbas, M., Smaili, K., & Berkani, D. (2011) Evaluation of Topic Identification Methods on Arabic Corpora, Journal Of Digital Information Management, 9(5), 185-192.
Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, 11(2), Jan 2006, p. 135-171. https://doi.org/10.1075/ijcl.11.2.02als
Al-Thubaity, A., Khan, M., Al-Mazrua, M., & Al-Mousa, M. (2013). New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool. In Proceedings of the Asian Language Processing (IALP) Conference, pp.67-70. https://doi.org/10.1109/IALP.2013.21
Alansary, S., & Nagi, M. (2014) The International Corpus of Arabic: Compilation, Analysis and Evaluation. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 8-17, October 25, 2014, Doha, Qatar.
AlMaayah, M., Sawalha, M., & Abushariah, M. (2016) Towards an automatic extraction of synonyms for Quranic Arabic WordNet. Int J Speech Technol, 19, 177. https://doi.org/10.1007/s10772-015-9301-9
Alrabiaih, M., Al-Salaman, A. & Atwell, E. (2013). The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. In Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2), 22 Jul. 2013, Lancaster University, UK.
Arts, T. (2014). The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary. In Proceedings of XVI EURALEX International Congress: The User in Focus, 15-19 July 2014, Bolzano/Bozen.
Arts, T., & McNell, K. (2013) Corpus-based lexicography in a language with a long lexicographical tradition: The case of Arabic. In Proceedings of Second Workshop on Arabic Corpus Linguistics (WACL-2), Workshop in conjunction with the Corpus Linguistics 2013 conference, Monday 22nd July 2013 – Lancaster University, UK.
Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A., & Suchomel, V. (2014). arTenTen: Arabic Corpus and Word Sketches. Journal of King Saud University-Computer and Information Sciences, 26(4), 357-371. https://doi.org/10.1016/j.jsuci.2014.06.009
Atkins, S., & Rundell, M. (2008). The Oxford guide to practical lexicography Oxford; New York; Oxford University Press.
Atkins, S., Clear, J., & Ostler, N. (1992) Corpus Design Criteria. Lit Linguist Computing, 7(1), 1-16. https://doi.org/10.1093/lc/7.1.1
Attia, M., & van Genabith, J. (2013). A Jellyfish Dictionary for Arabic. Retrieved Nov., 1, 2017, from
https://www.researchgate.net/publication/259494599_A_Jellyfish_Dictionary_for_Arabic
Black, W. J., & ElKateb, S. (2004). A Prototype English-Arabic Dictionary Based on WordNet. *The Second Global Wordnet Conference 2004* Brno, Czech Republic, January 20-23, 2004, 67-74. Retrieved from http://www.fi.muni.cz/gwc2004/proc/95.pdf
Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0: Linguistic Data Consortium*, catalog number LDC2004L02 and ISBN 1-58563-324-0.
Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, 42(2), 481-487. https://doi.org/10.3758/BRM.42.2.48
Diwan, A. H. (2004). المعجم النحوي لمفردات اللغة العربية. *The Syntactic Lexicon of Arabic Words* (First edition). Aleppo, Syria: Fusselat Publishers.
ElKateb, S., Black, W. J., & Farwell, D. (2006). Arabic WordNet and the Challenges of Arabic. *In Proceedings of The Challenge of Arabic for NLP/MT International Conference at The British Computer Society (BCS)*, London.
ElKateb, S., & Black, W. J. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. *Arabic NLP Workshop at ACL/EACL*. Toulouse, France.
Ghazali, S. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. *Arabic NLP Workshop at ACL/EACL*. Toulouse, France.
Hoogland, J. (1996). The use of OCR software for Arabic in order to create a text corpus of Modern Standard Arabic for lexicographic purposes. In A. Ubaydli (Ed.), *Proceedings of the international conference and exhibition on multi-lingual computing* (pp. 2701-2716). Cambridge University.
Ismail, O., Yagi, S., & Hammo, B. (2014). Corpus Linguistic Tools for Historical Semantics in Arabic. *International Journal of Arabic-English Studies (IJAES)* 15, 135-152.
Khalil, H. (1998). *Dirasat fi al-lughah wa al-ma‘ajim* " دراسات في اللغة والمعاجم " *Studies of language and lexicons* (First Edition ed.). Beirut, Lebanon: Dar al-nahdhah al-arabiah.
Mousser, J. (2010). A Large Coverage Verb Taxonomy for Arabic. *In Proceedings of the Seventh conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta.
Ooi, Vincent B. Y. (1998). *Computer corpus lexicography* Edinburgh: Edinburgh University Press.
Oxford Dictionaries. (2017) The Oxford English Corpus, Retrieved Nov., 1, 2017, from https://en.oxforddictionaries.com/explore/oxford-english-corpus
Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., & Martí, M. (2008). Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. *In proceedings of the 6th Conference on Language Resources and Evaluation LREC2008*, Marrakech (Morocco). Retrieved from http://www.lsi.upc.edu/~nlp/papers/rodriguez08b.pdf
Saad, M., & Ashour, W. (2010), OSAC: Open Source Arabic Corpora, in *‘EEECS’10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science*, European University of Lefke, Cyprus, 118-123.
Sawalha, M. (2011). *Open-source Resources and Standards for Arabic Word Structure Analysis*. (PhD), University of Leeds, Leeds.
Sawalha, M., & Atwell, E. (2010). Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic. *In: Proceedings of the Language Resource and Evaluation Conference LREC 2010*, 17-23 May 2010, Valletta, Malta.
van Mol, M., & Paulussen, H. (2001). AraLat: a relational database for the development of bi-lingual Arabic dictionaries. In S. Lee (Ed.), *Proceedings of Asialex 2001, Asian Bilingual-ism and the Dictionary* (pp. 206-211), Seoul, August 2001.
van Mol, M. (2000). The development of a new learner’s dictionary for Modern Standard Arabic: the linguistic corpus approach. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (Eds.), *Proceedings of the ninth EURALEX International Congress* (pp. 831-836). Stutt-gart, 8-12 August.
Zaghouani, W. (2014) Critical Survey of the Freely Available Arabic Corpora, *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC’14 workshop*, Reykjavik, Iceland.

Zarrouki, T., & Kebdani, M. (2009). *Aya-Spell Project, An Open-source Arabic Spell Checker Dictionary, experience and Future Work*. *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy, Damascus - Syria*.

Zarrouki, T., & Balla, A. (2009). Implementation of infixes and circumfixes in the spellcheckers. *2nd International Conference on Arabic Language Resources and Tools*, Cairo - Egypt.

Zemanek, P. (2001). *Clara (Corpus Linguae Arabicae): An Overview*. In ELSNET (Ed.), *Proceedings of ACL/EACL workshop on Arabic language processing*, Toulouse, France.

**Notes**

Note 1. شبكة مشكاة الإسلامية Meshkat Islamic Network http://www.almeshkat.net

Note 2. A web interface for searching the traditional Arabic lexicons for a certain root http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py

Note 3. Frequency list of the Arabic Wikipedia corpus (wiki-ar) is found on http://corpus.leeds.ac.uk/frqc/wiki-ar.num

Note 4. Leeds collection of Internet corpora: Arabic Internet Corpus http://corpus.leeds.ac.uk/internet.html

Note 5. The text of the Qur’an used in this experiment was represented in MSA script.

Note 6. Jordanian Arabic Language Academy: Word lists of technical terms http://www.majma.org.jo/?cat=53

**Copyrights**

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).