Constructing and Using Broad-Coverage Lexical Resource for Enhancing Morphological Analysis of Arabic

Majdi Sawalha, Eric Atwell
School of Computing,
University of Leeds, Leeds, LS2 9JT, UK
E-mail: sawalha@comp.leeds.ac.uk, eric@comp.leeds.ac.uk

Abstract

Broad-coverage language resources which provide prior linguistic knowledge must improve the accuracy and the performance of NLP applications. We are constructing a broad-coverage lexical resource to improve the accuracy of morphological analyzers and part-of-speech taggers of Arabic text. Over the past 1200 years, many different kinds of Arabic language lexicons were constructed; these lexicons are different in ordering, size and aim or goal of construction. We collected 23 machine-readable lexicons, which are freely available on the web. We combined lexical resources into one large broad-coverage lexical resource by extracting information from disparate formats and merging traditional Arabic lexicons.

To evaluate the broad-coverage lexical resource we computed coverage over the Qur'an, the Corpus of Contemporary Arabic, and a sample from the Arabic Web Corpus, using two methods. Counting exact word matches between test corpora and lexicon scored about 65-68%; Arabic has a rich morphology with many combinations of roots, affixes and clitics, so about a third of words in the corpora did not have an exact match in the lexicon. The second approach is to compute coverage in terms of use in a lemmatizer program, which strips clitics to look for a match for the underlying lexeme; this scored about 82-85%.

1. Introduction

Lexicography is the applied part of lexicology. It is concerned with 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). On the other hand, lexicology is 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). Long-term efforts lexicographic projects have been greatly accelerated since the advent and use of computers which is known as computational lexicography. However, constructing a large-scale broad-coverage lexicon involves long-time development of specifications, design, collection of lexical data, information structuring, and user-oriented presentation formatting (Eynde & Gibbon, 2000).

Modern English dictionaries are stored using computerized lexicographic databases. The most-widely and accepted lexicographic database representation is lexical text markup using SGML (Standard Generalised Markup Language) such as; XML. Other Database Management System (DBMS) can be used such as; relational databases, object-oriented DBMS with inheritance mechanisms, and hybrid object-oriented/relational databases.

Traditional Arabic lexicons are not available in computerized lexicographic databases. Moreover, traditional Arabic lexicons have different arrangement methodologies than modern English dictionaries. Common English dictionaries list lexical entries, which are words, arranged alphabetically; followed by the meaning of that word, while Arabic lexicons are mainly arranged by selecting the root as main lexical entries. The roots are followed by a definition part which may span several pages. The definition part is written as a unit or an article which defines all the derived words of a certain root. These lexical entries are not arranged or distinguished with special formatting.

A study of a traditional Arabic lexicon called *al-qāmūs al-muḥīṭ* القاموس المحيط "The comprehensive lexicon" showed three major drawbacks of traditional Arabic lexicons. First, it does not represent language development periods in different times. Second, the ambiguity of defining and explaining lexical meaning of the words. Third, the unorganized way of ordering the derivations of lexical entries and the absence of the origin of the derivations. The researcher highlighted the importance of ordering the derivations of each lexical entry to directly access the meaning of the derivations, and to show the origin of the Arabic word and its specifications (Khalil, 1998).

2. Traditional Arabic lexicography

Arabic lexicography is one of the original and deep-rooted arts of Arabic literature. The first lexicon constructed was *kitāb al-'ayn* كتاب العين 'al-'ayn lexicon' by *al-farāhīdī* (died in 791). Over the past 1200 years, many different kinds of Arabic language lexicons were constructed; these lexicons are different in ordering, size and aim or goal of construction. Many Arabic language linguists and lexicographers studied the construction, development and the different methodologies used to construct these lexicons.

Traditional Arabic lexicons distinguish between four classes of ordering lexical entries in the lexicon. First, *al-ẖalīl* methodology is developed by الخليل بن أحمد الفراهيدي *al-ẖalīl bin aḥmad al-farāhīdī* (died in 791). His lexicon is called *kitāb al-'ayn* كتاب العين. 'The *al-'ayn'* lexicon lists the lexical entries phonologically according to exits of

letters sounds from the mouth and throat, from the farthest letter exit to the nearest. Second, *abī 'ubayd* methodology is developed by *abī 'ubayd al-qāsim bin sallām* أبي عُبيدٍ القاسم بن سلاّم (died in 838). He wrote many small books, each of which describes one subject or meaning, such as books describing horses, milk, honey, flies, insects, palms, and human creation. Then he collated all these small books into one large lexicon called *al-ğarīb al-muṣnnaf fī al-luğah* الغريب المُصنّف في اللغة 'The Irregular Classified Language'. Third, *al-ğawharī* methodology is developed by *'ismā'īl bin ḥammād al-ğawharī* (died in 1002) and his lexicon is called *aṣ-ṣiḥāḥ fī al-luğah* الصحاح في اللغة 'The Correct Language'; this uses alphabetical order for ordering the lexical entries. However, he arranged the lexical entries of his lexicon depending on the last letter of the word, and then the first letter. Finally, the *al-barmakī* methodology is developed by *abu al-ma'ālī moḥammad bin tamīm al-barmakī* أبو المعالي محمد بن تميم البرمكي, who lived in the same time period as *al-ğawharī*. *al-barmakī* did not construct a new lexicon; but he alphabetically re-arranged a lexicon called *aṣ-ṣiḥāḥ fī al-luğah* الصحاح في اللغة 'The Correct Language' by *al-ğawharī*. He added little information to that lexicon.

Figure 1a and 2 show a sample of text taken from traditional Arabic lexicons; the target lexical entries are underlined and highlighted in blue. Figure 1b shows the human translation of the sample of figure 1a, the target lexical entries are highlighted by square brackets. Figure 3 is a sample of the Arabic-English lexicon by Edward Lane (Lane, 1968) volume 7, pages 117-119, the target lexical entries are underlined.

كتب: الكِتابُ: معروف، والجمع كُتُبٌ وكُتْبٌ. كَتَبَ الشيءَ يَكْتُبه كَتْباً وكِتاباً وكِتابةً، وكَتَّبَه: خَطَّه؛ قال أَبو النجم: أَقْبَلْتُ من عِنْدِ زِيادٍ كالخَرِفْ، تَخُطُّ رِجْلايَ بَخَطٍّ مُخْتَلِفْ، تُكَتِّبانِ في الطَّرِيقِ لامَ أَلِفْ قال: ورأَيت في بعض النسخ يَكْتِبان، بكسر التاء، وهي لغة بَهْرَاء، يَكْسِرون التاء، فيقولون: تِعْلَمُونَ، ثم أَتْبَعَ الكافَ كسرةَ التاء. والكِتابُ أَيضاً: الاسمُ، عن اللحياني. الأَزهري: الكِتابُ اسم لما كُتب مَجْمُوعاً؛ والكِتابُ مصدر؛ والكِتابةُ لِـمَنْ تكونُ له صِناعةً، مثل الصِّياغةِ والخِـياطةِ. والكِتْبةُ: اكْتِتابُك كِتاباً تنسخه. ويقال: اكْتَتَبَ فلانٌ فلاناً أَي سأَله أَن يَكْتُبَ له كتاباً في حاجة. واسْتَكْتَبه الشيءَ أَي سأَله أَن يَكْتُبَه له. ابن سيده: اكْتَتَبه كَكَتَبه. وقيل: كَتَبه خَطَّه؛ واكْتَتَبه: اسْتَمْلاه، وكذلك اسْتَكْتَبه. واكْتَبه: كَتَبه، واكْتَتَبْته: كَتَبْته. وفي التنزيل العزيز: اكْتَتَبَها فهي تُمْلى عليه بُكْرَةً وأَصِـيلاً؛ أَي اسْتَكْتَبَها. ويقال: اكْتَتَبَ الرجلُ إذا كَتَبَ نفسَه في دِيوانِ السُّلْطان ...

Figure 1a: A sample of text from the traditional Arabic lexicon "*lisān al-'rab*", the target lexical entries are underlined and highlighted in blue.

3. Processing steps for Arabic Lexicons

Twenty three lexicons have been collected from different resources from the web where all of them are freely available. *maktaba*^{t} *al-miškā*^{t} *al-'islāmyya*^{h} [1] مكتبة المشكاة الاسلامية provides most of these lexicons which are written in MS-Word files. Each lexicon is written in a different format and has its own arrangement methodology of its lexical entries. After manually converting each lexicon text into a unified format by choosing the most common format for all the root entries in the lexicon, information such as roots, words and meaning are automatically extracted using specialized programs. The results are stored in separate dictionaries which include roots, words, and meanings. A combination algorithm combines the disparate lexicon information into one large broad-coverage lexical resource.

Common processing steps were applied to all lexicons. First, all lexicons' files were converted from MS-Word or HTML web pages into standard text files in Unicode 'utf-8' encoding. Second, a statistical analysis computed the word's frequency and the vocabulary size for both vowelized and non-vowelized text of each lexicon. The lexicons' texts contain 14,369,570 words, 2,184,315 vowelized word types and 569,412 non-vowelized word types. Table 1 shows the summary of the statistical analyses of the lexicons' texts used to construct the broad-coverage lexical resource.

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

Table 1: statistical analysis of the lexicons' text used to construct the broad-coverage lexical resource

k t b: [*Alkitab*] the book; is well known. The plural forms are [*kutubun*] and [*kutbun*]. [*kataba Alshay'*] He wrote something. [*yaktubuhu*] the action of writing something. [*katban*], [*kitaban*] and [*kitabatan*] means the art of writing. And [*kattabahu*] writing it means draw it up. Abu Al-Najim said: I returned back from Ziyad's house [after meeting him] and behaved demented, my legs drawn up differently (means walking in a different way). They wrote [*tukattibani*] on the road the letters of *Lam Alif* (describing how he was walking crazily and in a different way). He said: I saw in a different version, the word "they wrote" [*tikittibani*] using the short vowel *kasrah* on the first letter [taa], as it is used by Bahraa' (Arab tribe) dialect. They say: (ti'lamuwn) (you know). Then the short vowel kasrah is propagated to the following letter (kaf). Moreover, [*Alkitab*] the book is a noun. Al-lihyani Al-Azhari definition is: [*Alkitab*] The book is the name of a collection of what has been written (a collection of written materials or texts). And the book has gerund [*Alkitabatu*] writing (art of writing) for whoever has a profession, similar to drafting and sewing. And [*Alkitabatu*]: is copying a book [copying a book in several copies]. It is said: [*iktataba*] someone subscribed another means; he asked to write him a letter in something. [*istaktabahu*] He dictated someone something means to write him something. Ibn Sayyedah: [*Iktatabahu*] is similar to [*katabahu*]. It is said: [*katabahu*] write something down means draw up. And [*Iktatabahu*] writing something down means dictate someone something, which is the same meaning of [*Istaktabahu*]. [*Iktatabahu*] registering (masculine), and [*Iktatabathu*] registing (feminine). In the Qur'an: [*Iktatabaha*] He registered it, he has dictated it every sunrise and sunset, which means dictating it. It is said: [*Iktataba Al-rajul*] The man registered, if he registered himself in the Sultan's office …

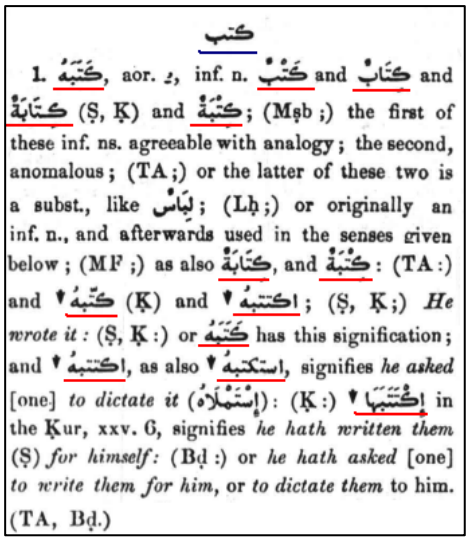
Figure 1b: A Human translation of the sample of text from the traditional Arabic lexicons "*lisān al-'rab*", the target lexical entries are highlighted using square brackets,

[1] http://www.almeshkat.net

(ك ت ب):

(كَتَبَهُ) كِتْبَةً وَكِتَابًا وَكِتَابَةً وَقَوْلُهُ وَإِذَا كَانَتْ السَّرِقَةُ صُحُفًا لَيْسَ فِيهَا كِتَابٌ أَيْ مَكْتُوبٌ (وَفِي حَدِيثِ أُنَيْسٍ) وَاحْكُمْ بِكِتَابِ اللَّهِ أَيْ بِمَا فَرَضَ اللَّهُ مِنْ كَتَبَ عَلَيْهِ كَذَا إذَا أَوْجَبَهُ وَفَرَضَهُ (وَمِنْهُ) الصَّلَاةُ الْمَكْتُوبَةُ وَأَمَّا قَوْلُهُ – صَلَّى اللَّهُ عَلَيْهِ وَآلِهِ وَسَلَّمَ - [مَا بَالُ أَقْوَامٍ يَشْتَرِطُونَ شُرُوطًا لَيْسَتْ فِي كِتَابِ اللَّهِ تَعَالَى] فَقِيلَ الْمُرَادُ قَوْله تَعَالَى {أُدْعُوهُمْ لِآبَائِهِمْ} إلَى أَنْ قَالَ وَمَوَالِيكُمْ فِيهِ أَنَّهُ نَسَبَهُمْ إلَى مَوَالِيهِمْ كَمَا نَسَبَهُمْ إلَى آبَائِهِمْ فَلَمَّا لَمْ يَجُزْ التَّحَوُّلُ عَنْ الْآبَاءِ لَمْ يَجُزْ عَنْ الْأَوْلِيَاءِ وَيَجُوزُ أَنْ يُرَادَ بِكِتَابِ اللَّهِ قَضَاؤُهُ وَحُكْمُهُ عَلَى لِسَانِ رَسُولِ اللَّهِ – صَلَّى اللَّهُ عَلَيْهِ وَآلِهِ وَسَلَّمَ – إنَّ الْوَلَاءَ لِمَنْ أَعْتَقَ (وَأَكْتُبَ الْغُلَامَ وَكَتَبَهُ) عَلَّمَهُ الْكِتَابَ (وَمِنْهُ) سَلَّمَ غُلَامَهُ إلَى مُكَتِّبٍ أَيْ إلَى مُعَلِّمِ الْخَطِّ رُوِيَ بِالتَّخْفِيفِ وَالتَّشْدِيدِ (وَأَمَّا الْمَكْتَبُ) وَالْكُتَّابُ فَمَكَانُ التَّعْلِيمِ وَقِيلَ الْكُتَّابُ الصِّبْيَانُ (وَكَاتَبَ) عَبْدَهُ مُكَاتَبَةً وَكِتَابًا قَالَ لَهُ حَرَّرْتُك يَدًا فِي الْحَالِ وَرَقَبَةً عِنْدَ أَدَاءِ الْمَالِ (وَمِنْهُ) قَوْله تَعَالَى {وَالَّذِينَ يَبْتَغُونَ الْكِتَابَ} وَقَدْ يُسَمَّى بَدَلُ الْكِتَابَةِ مُكَاتَبَةً وَأَمَّا الْكِتَابَةُ فِي مَعْنَاهَا فَلَمْ أَجِدْهَا إلَّا فِي الْأَسَاسِ وَكَذَا تَكَاتَبَ الْعَبْدُ إذَا صَارَ مُكَاتَبًا وَمَدَارُ التَّرْكِيبِ عَلَى الْجَمْعِ (وَمِنْهُ كَتَبَ النَّعْلَ وَالْقِرْبَةَ) خَرَزَهَا (وَالْكُتْبُ الْخُرَزُ) الْوَاحِدَةُ كُتْبَةٌ (وَمِنْهُ كَتَبَ الْبَغْلَةَ) إذَا جَمَعَ بَيْنَ شَفْرَتَيْهَا بِحَلْقَةٍ (وَالْكَتِيبَةُ) الطَّائِفَةُ مِنْ الْجَيْشِ مُجْتَمِعَةً (وَبِهَا سُمِّيَ) أَحَدُ حُصُونِ خَيْبَرَ (وَقَوْلُهُمْ) سُمِّيَ هَذَا الْعَقْدُ مُكَاتَبَةً لِأَنَّهُ ضَمُّ حُرِّيَّةِ الْيَدِ إلَى حُرِّيَّةِ الرَّقَبَةِ أَوْ لِأَنَّهُ جَمَعَ بَيْنَ نَجْمَيْنِ فَصَاعِدًا ضَعِيفٌ جِدًّا وَإِنَّمَا الصَّوَابُ أَنَّ كُلًّا مِنْهُمَا كَتَبَ عَلَى نَفْسِهِ أَمْرًا هَذَا الْوَفَاءَ وَهَذَا الْأَدَاءَ.

Figure 2: A sample of text from the traditional Arabic lexicon "*al-muğrab fī tartīb al-mu'rab*", the target lexical entries are underlined and highlighted in blue.



كتب

1. كَتَبَهُ, aor. ـُ, inf. n. كَتْبٌ and كِتَابٌ and كِتَابَةٌ (S, K) and كِتْبَةٌ; (Msb;) the first of these inf. ns. agreeable with analogy; the second, anomalous; (TA;) or the latter of these two is a subst., like لِبَاسٌ; (Lḥ;) or originally an inf. n., and afterwards used in the senses given below; (MF;) as also كِتَابَةٌ, and كِتْبَةٌ: (TA:) and ↑ كَتَّبَهُ (K) and ↑ اكتتبهُ; (S, K;) *He wrote it*: (S, K:) or كَتَّبَهُ has this signification; and ↑ اكتتبهُ, as also ↑ استكتبهُ, signifies *he asked* [one] *to dictate it* (إِسْتَمْلَاهُ): (K:) ↑ إِكْتَتَبَهَا in the Ḳur, xxv. 6, signifies *he hath written them* (S) *for himself*: (Bḍ:) or *he hath asked* [one] *to write them for him*, or *to dictate them* to him. (TA, Bḍ.)

Figure 3: A Sample of the definition of the root كتب *k-t-b* 'wrote' from an Arabic-English Lexicon by Edward Lane, http://www.tyndalearchive.com/TABS/Lane/

4. Analyzing lexicons' text separately

Each lexicon was constructed in different way of arranging its roots and lexical entries. Moreover, Lexicons are typed into machine-readable files in different formats but without using any computerized lexicographic representations. These factors add more processing challenges. Therefore, each lexicon is processed separately using specialized programs. An important preprocessing step converts each lexicon text into a unified format by choosing the most common format for all the root entries in the lexicon. This step is done manually which needs to go through all the text in the lexicon files and re-format the root entries that do not follow the selected format. The common structure of all lexicons is root-definition structure, where each root entry in the lexicon is followed by the definition part that groups all the derived words and their meaning. After that, a program is written to extract the roots and words derived from that root. The tokenizing module in the program must specify the root entries and their definition parts. Then, a bag of words is extracted from the definition text. The bag of words stores pairs of word-root where each word appearing in the definition part is associated with the root of that part.

The definition parts of the roots are articles that define each root and defines the lexical entries derived from a certain root. The writing style of the definition part connects the lexical entries and their meanings together without following any ordering methodology. The writing style of the definition parts show the lexical entries conjoined with all kinds of clitics and affixes. Clitics, such as conjunctions and pronouns, are used to connect the definitions of the lexical entries together as one unit.

The use of clitics and affixes adds more challenge to the construction of the broad-coverage lexical resource. We used modules of the morphological analyzer for Arabic text (Sawalha & Atwell, 2009a) (Sawalha & Atwell, 2009b), to separate the lexical entries from the clitics and affixes attached to that word. The morphological analyzer generates all possible combinations of clitics, affixes and stem for the analyzed word. Only the analyses that match the clitics and affixes with the clitics and affixes lists used by the morphological analyzer are selected as candidate analyses.

Many words appearing on the definition part are not relevant to the root associated with that definition. Such words are found in the bag of words that root. A normalization analysis that verifies the word-root pairs is done by applying linguistic knowledge that governs the derivation process of words from their roots. These conditions are simply described as the following:

Condition 1 (check consonants): If all consonant letters constructing the root appear in the analyzed word, then check condition 2.
Condition 2 (consonants order): If all root letters appear in the same order as the word's letters, then word-root combination might be correct.

In the first condition (check consonants), we classified Arabic letters into four groups, letters that appear in clitics or affixes, vowels, *hamza* and letters that might be changed in derivation due to substitution إقلاب *'iqlāb* to simplify the pronunciation of the word. Then, a procedure is applied to verify each letter of the word. Another procedure is applied to match the order of the letters of both the analyzed word and its root. The analyses that meet the two conditions are candidate analyses and are stored in the lexicon database. The information of clitics, affixes and stem are also stored with the word-root combination.

5. Combining the processed lexicons in one broad-coverage lexical resource

After analyzing each lexicon, a combination algorithm is applied to construct the broad-coverage lexicon. The algorithm starts by selecting a large lexicon called لسان العرب *lisān al-'rab* 'Arab tongue' as a seed to the broad-coverage lexicon. Then, the lexicons are combined one by one to the broad-coverage lexicon. Figure 4 shows the first 60 lexical entries of the root كتب *k-t-b* 'wrote' stored in the broad-coverage lexicon. After, combining each lexicon the percentage of records added to the broad-coverage lexicon is computed. The percentage starts by 100% for the seed lexicon and decreases during the combination process. The percentage will tell us when the combination process should stop, and which lexicons are better to construct a broad-coverage lexical resource. Table 2 shows the number of records extracted from 7 analyzed lexicons so far, and the number and the percentage of records combined to the broad-coverage lexicon.

#	Lexicon	Word types[B]	Records inserted [A]	Percentage	
				(A/B)%	(A/C)%
1	*lisān al-'rab*	207,992	207,992	100.00%	47.80%
2	*mu'ğam al-muḥīṭ fī al- luğat*	74,507	61,113	82.02%	14.04%
3	*taǧ al-'arūs min ğawāhir al-qāmūs*	128,119	95,415	74.47%	21.93%
4	*muḫtār aṣ-ṣiḥāḥ*	19,540	16,573	84.82%	3.81%
5	*al-muğrab fī tartīb al-mu'rab*	12,396	9,805	79.10%	2.25%
6	*kitābu al-'ayn*	30,292	18,878	62.32%	4.34%
7	*al-mu'ğam al-wasīṭ*	36,660	25,364	69.19%	5.83%
	Totals	***509,506***	***435,140* [C]**	**85.40%**	***100.00%***

Table 2: the number of records extracted from 7 analyzed lexicons, and the number and the percentage of records combined to the broad-coverage lexicon.

أكتبه	*'aktabahu*	الكتاب	*al-kitāb*	الكُتْبَةُ	*al-kutba^{tu}*
أَكْتَبَ	*'aktaba*	الكتابة	*al-kitāba^{t}*	الكُتْبَةُ	*al-kutba^{tu}*
أَكْتَبْتُ	*'aktabtu*	الكتابَة	*al-kitāba^{ta}*	الكِتاب	*al-kitāb*
أَكْتِبْنِي	*'aktibnī*	الكتابَة	*al-kitāba^{t}*	الكِتابةُ	*al-kitāba^{tu}*
إكْتاباً	*'iktāb^{an}*	الكتاتيب	*al-katātīb*	الكِتابَ	*al-kitāba*
استكتبه	*'istaktabahu*	الكتبة	*al-kitba^{t}*	الكِتابَةُ	*al-kitāba^{tu}*
اسْتَكْتَبَه	*'istaktabahu*	الكتيبة	*al-katība^{t}*	الكِتابُ	*al-kitābu*
اسْتَكْتَبَها	*'istaktabahā*	وكَتيبة	*wa katība^{t}*	الكِتاب	*al-kitābi*
اكتتب	*'iktataba*	الكَتائِبَ	*al-katā'iba*	المكاتب	*al-mukātib*
اكْتَتَبَ	*'iktataba*	الكَتائِبُ	*al-katā'ibu*	المكاتبة	*al-mukātiba^{t}*
اكْتَتَبَه	*'iktatabahu*	الكَتيبةُ	*al-katība^{ta}*	المكتب	*al-maktab*
اكْتَتَبَها	*'iktatabahā*	الكَتائبَ	*al-katā'iba*	المكتبة	*al-maktaba^{t}*
اكْتُبْ	*'uktub*	الكَتبة	*al-kataba^{t}*	المكتوبة	*al-maktūba^{t}*
اكْتُبِتْ	*'uktutibtu*	الكَتْبُ	*al-katbu*	الْكُتَّابُ	*al-kuttābu*
اكْتِبائك	*'iktitābuk*	الكَتْبِ	*al-katbi*	الْكِتَابَ	*al-kitāba*
اكْتِبائكَ	*'iktitābuka*	الكُتُبُ	*al-kutabu*	الْكِتَابَةُ	*al-kitāba^{tu}*
الاكْتِتابُ	*al-'iktitābu*	الكُتَيْبَةُ	*al-kutaybatu*	الْكِتَابَةِ	*al-kitāba^{ti}*
التكاتب	*at-takātubu*	الكُتَّابَ	*al-kuttāba*	الْمَكْتَبُ	*al-maktabu*
الكاتب	*al-kātib*	الكُتَّاب	*al-kuttābi*	الْمَكْتُوبَةُ	*al-maktūba^{tu}*
الكاتِبُ	*al-kātibu*	الكُتْبة	*al-kutba^{t}*	اِسْتَكْتَبَ	*'istaktaba*

Figure 4: The first 60 lexical entries of the root كتب k-t-b 'wrote' stored in the broad-coverage lexical resource.

6. Evaluation

The evaluation process shows the coverage of the broad-coverage lexical resource on different types of text corpora. The Qur'an, the Arabic Web Corpus[2] and the Corpus of Contemporary Arabic are used to compute the coverage of the broad-coverage lexical resource in two ways. First; exact match where each non-vowelized word in the test corpora is searched in the lexicon. Table 3 shows the coverage percentage using exact match method scores about 65-57%.

Corpus	Tokens	Words	Words covered by lexicon	Coverage
Qur'an	77,800	77,799	52,536	67.53%
CCA	684,726	594,664	389,133	65.44%
Web	1,128,114	833,916	546,880	65.58%

Table 3: The coverage of the lexicon using exact match method.

Arabic words in any text come up with many different forms of clitics attached to it, which makes the matching process of the word and the lexical entries of the lexicon not an easy task and decreases the coverage percentage. The second method is to compute the coverage of the broad-coverage lexical resource through an application that depends on it. We have developed a lemmatizer for Arabic text to be used to process large and real data; the Arabic Web Corpus which consists of 100 million words of Arabic web pages. The lemmatizer depends on the broad-lexical resource to extract the lemma and the root of the word. Each word is tokenized into different forms consisting of proclitics, stem and enclitics, and then each stem is searched in the lexicon. If the stem is found in the lexicon then the root and the vowelized stems stored in the broad-coverage lexicon are retrieved. When a correct analysis is retrieved from the lexicon then we count it as a valid lexicon reference. The coverage of the lexicon is computed by the percentage of valid lexicon references to the number of words in the test sample. The lemmatizer uses other three linguistic lists; list of function words (stop words) which have fixed syntactic analysis in any context (Diwan, 2004), named entities list (Benajiba et al, 2008) and list of broken plurals[3]. We computed the coverage of the broad-coverage lexical resource one time with the inclusion of these functional words, and another time without including the functional words in the test. Table 4 and 5 show the coverage percentage of the lexicon computed using the lemmatizer program. The coverage percentage scored about 85% of the words, including functional words, and about 82% of the words excluding functional words, referenced the lexicon and retrieved valid analysis.

We studied the common words which are not covered by the broad-coverage lexical resource. We found that common not covered words belongs to; functional words (stop words) which are easily included to the lexicon along with their syntactical and morphological analysis by collecting them from traditional Arabic grammar books such as (Diwan, 2004). The other category of common not covered words are the new Arabic terms, and borrowed words (Arabaized words) which are foreign words transliterated into Arabic by writing the word in Arabic letters. This is a common problem found in news paper and web pages text. The lack of updating Arabic lexicons and the lack of the correct translation of the borrowed words will increase the frequency of this type of word in contemporary Arabic text. Figure 5 shows a sample of common words not covered by the broad-coverage lexical resource.

Corpus	Tokens	Words	Words covered by lexicon	Coverage
Qur'an	77,804	77,803	64,065	82.34%
CCA	685,161	595,099	507,943	85.35%
Web	1,128,624	834,426	708,101	84,86%

Table 4: Coverage including function words.

Corpus	Tokens	Words	Words covered by lexicon	Coverage
Qur'an	77,804	54,004	42,532	78.76%
CCA	685,161	411,482	338,790	82.33%
Web	1,128,624	576,407	476,190	82.61%

Table 5: Coverage excluding function words.

ذَلِكَ	ḏālika	التي	allatī
السَّمَاوَاتِ	assamāwāti	الإنسان	al'insān
إِنَّهُمْ	'innahum	الإيميل	al'imayl
بِاللَّهِ	billāhi	التليفون	attilifūn
عَنْهُمْ	ʿanhum	الفلسطيني	al-falasṭīnī
بِالْحَقِّ	bilḥaqqi	دردشة	dardašaᵗ
فَأُوْلَئِكَ	fa'ulā'ika	انقر	'unqor
فَبِأَيِّ	fabi'ayyi	الأمريكية	al-'amrīkyyaᵗ
وَإِلَى	wa-'ilā	الداخلية	ad-dāḫilyyaᵗ
فَسَوْفَ	fasawfa	الانتخابات	Al-'intiḫābāt
المتحدة	al-muttaḥidaᵗ	الولايات	al-wilāyāt
الدكتور	Ad-duktūr	الاجتماعية	al-iğtimāʿiyyaᵗ
السياحية	as-siyāḥyyaᵗ	الإنترنت	al-'intarnit
الغربية	al-ḡarbyyaᵗ	التنمية	at-tanmiyaᵗ
الاقتصادية	al-'iqtiṣādyyaᵗ	الثقافية	aṯ-ṯaqāfiyyaᵗ

Figure 5: a sample of common words which are not covered by the lexicon.

7. The corpus of lexicons

Al-Sulaiti and Atwell (2006) developed the Corpus of Contemporary Arabic. This corpus contains 1 million words taken from different genres collected from newspapers and magazines. It contains the following domains; Autobiography, Short Stories, Children's Stories, Economics, Education, Health and Medicine, Interviews, Politics, Recipes, Religion, Sociology, Science, Sports, Tourist and Travel and Science. Similar to most Arabic corpora, the text of the Corpus Contemporary Arabic is

[2] http://corpus.leeds.ac.uk/internet.html

[3] http://sites.google.com/site/elghamryk/arabiclanguageresources

taken from newspapers and magazines text. Our lexicons' text can be used as an Arabic corpus of dictionaries, which has different domain than the existing corpora. The Arabic corpus of dictionaries covers a period of more than 1200 years and consists of large number of words and word types. It also has both vowelized and non-vowelized text. Figure 6 shows the number of words and word types and the 25 words of highest frequency.

Partially-vowelized		Non-vowelized	
Word	**Frequency**	**Word**	**Frequency**
في	292,396	من	322,239
من	269,200	في	301,895
قال	172,631	قال	190,918
و	120,060	أي	132,635
على	108,252	و	130,809
ما	89,195	على	119,639
وقال	88,233	إذا	115,842
عن	82,027	وقال	99,601
إذا	81,479	ابن	94,980
أي	78,622	ما	94,530
وهو	75,149	بن	92,213
لا	69,737	عن	87,064
ابن	58,334	وهو	80,375
به	53,343	لا	73,066
وفي	53,197	أبو	72,231
وقد	50,648	أن	65,419
أبو	47,915	أو	62,298
بن	46,880	الله	59,511
أي	46,788	به	58,941
هو	45,916	يقال	58,062
يقال	45,794	وفي	55,077
عليه	44,786	وقد	53,992
ولا	42,190	عليه	50,906
الله	39,961	هو	49,785
أو	39,210	إلى	48,363

Figure 6: The number of words and word types and the part of the frequency list of the corpus of lexicons text.

8. Conclusion

In this paper we showed the process of constructing a broad-coverage lexicon for Arabic to be used in NLP applications such as lemmatizers, morphological analyzers and part-of-speech taggers. We described the traditional Arabic lexicons, arranging methodologies and the challenges and drawbacks of these lexicons.

We described the development of constructing a broad-coverage lexical resource by combining extracted information from disparate lexical resources formats and merging Arabic lexicons. Processing steps of constructing the broad-coverage lexical resource involve; first, analyzing lexicons' text separately by manually converting each lexicon text into a unified format by choosing the most common format for all root entries. Then, for each lexicon a specialized program extracts 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 resource.

The evaluation of the broad-coverage lexical resource is done by computing the coverage of it. The coverage is computed using two methods; first methodology computes the coverage by matching the words of the test corpora to the words in the lexicon which scored about 67%. The second methodology uses a lemmatizer program to compute the coverage, and scored about 82%.

This is the first version of the broad-coverage lexical resource. We will extend the lexicon by including the full morphological analyses of the lexical entries and other useful information that will enhance the accuracy of NLP applications. Online access method to the contents of the broad-coverage lexical resource and downloadable version will be released.

9. References

Al-Suliti, L., Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics* 11, pp.135--171.

al-ğawharī "ابو النصر اسماعيل بن حماد الجوهري الفارابي" (died in 1009), *aṣ-ṣiḥāḥ fī al-luğa*^{h} *الصحاح في اللغة 'The Correct Language'*, al-miškā^{t} Islamic Library (online-library) http://www.almeshkat.net/books/archive/books/alsehah%20g.zip

Benajiba, Y., Diab, M., Rosso, P. (2008) Arabic named entity recognition using optimized feature sets. *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii: Association for Computational Linguistics, pp.284--293

Diwan, A. (2004) *al-mu'ğam an-naḥwī li-mufradāt al-luğa*^{h} *al- 'arabiyya*^{h} *المعجم النحوي لمفردات اللغة العربية*, Aleppo, Syria: fusselat lil-dirasāt wa at-tarğamah wa an-našir.

Eynde, V.E, Gibbon, D (2000) *Lexicon development for speech and language processing*, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Khalil, H. (1998), *dirāsāt fi al-luğa*^{h} *wa al-ma'āğim* " *دراسات في اللغة والمعاجم* " *Studies of language and lexicons*. First Edition, Beirut, Lebanon: Dar al-nahḍa^{h} al-'arabia^{h}.

Lane, E. W. (1968). *An Arabic-English Lexicon*. Beirut, Librarie Du Liban.

Sawalha, M., Atwell, E. (2009a). Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. *Proceedings of the 5th International Corpus Linguuistics Conference CL2009* Liverpool, UK.

Sawalha, M. and Atwell, E. (2009b). توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية (Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). *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.