Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic

Mohammed Attia, Lamia Tounsi, Josef van Genabith

Abstract

We describe the design and implementation of large-scale data processing techniques for the automatic acquisition of lexical resources for Modern Standard Arabic (MSA) from annotated and un-annotated corpora and demonstrate their usefulness for creating a wide-coverage, general-domain lexicon. Modern lexicographic principles (Atkins and Rundell, 2008) emphasise that the corpus is the only viable evidence that a lexical entry still exists in a speech community. Unlike most available Arabic lexicons which are based on previous historical (and dated) dictionaries, our lexicon starts off with corpora of contemporary texts. The lexicon is encoded in Lexical Markup Framework (LMF) which is a metamodel that provides a standardized framework for the construction of electronic lexical resources. The aim of LMF is to optimize the production and extension of electronic lexicons and to facilitate the exchange of data between all aspects of lexical resources and the interoperability with NLP applications. This lexical resource will serve as the core of our Arabic annotation tools: morphological analysis, tokenization, diacritization and base phrase chunking.

1 Introduction

A lexicon lies at the heart of most morphological analysers of Arabic (Dichy and Fargaly, 2003; Attia, 2006; Buckwalter. 2002; Beesley, 2001). The quality and coverage of the lexical database will determine the quality and coverage of the morphological analyser, and any limitations found in the database will make their way through to the morphological analyser. The literature abounds with discussions about the design of a morphological analyser, yet little effort has gone into the investigation of the nature of the database at the core of all these systems, and what design decisions have been taken in their development. Some of the valid questions that we need to ask are: what makes a word eligible to be included in the lexicon, how do we decide the possible inflections of word forms, and what sort of information that we need to accommodate? Even more important than these questions is the question of what variety of Arabic (Classical Arabic, Modern Standard Arabic, or Colloquial Arabic) do we cover, and what tests do we conduct to make sure that the word we include do really belong to our target language variety?

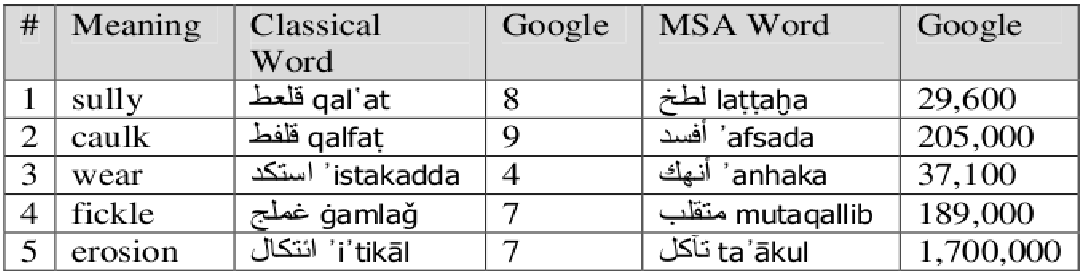
Al-Sulaiti (2006) emphasises that most existing dictionaries of Modern Standard Arabic are not corpus based. Even earlier, Ghazali and Braham (2001) lamented the fact that traditional Arabic dictionaries are based on historical perspectives and they tend to include fossilized words and meanings that are of little or no use to the language learners. They stressed the need for new dictionaries on an empirical approach that makes use of contextual analyses of language corpora.

In one exception to the inefficient traditional approach of Arabic dictionary making, Van Mol (2000) developed in Arabic-Dutch learner's dictionary in what he considered as the first attempt to build a COBUILD-style dictionary of 17,000 entries for Arabic based solely and entirely on corpus data which were used to derive information on contemporary usage, meanings and collocations. Van Mol, however, relied on intensive laborious manual work over many years to tag and translate a three million words corpus word by word in context and looking through concordances. Repeating the process would require repeating the same manual labour. In contrast our approach is automated which means that the process of acquiring lexical resources from new corpora would entail less cost. One more difference is that Van Mol's target users are language learners while our target user is NLP applications. The difference in target user entails considerable disparity in the type of information included and the way such information is presented.

Van Mol (2000) criticized the much-celebrated Arabic-English dictionary of Hans Wehr and estimates that about 5% of frequent new words and meanings were not found in the dictionary and that the great majority of the words in the dictionary are not used so frequently anymore in Modern Standard Arabic. Van Mol maintains that Hans Wehr dictionary contains about 45,000 entries, but his new Arabic-Dutch dictionary covers almost the whole range of the actual vocabulary as evidenced by his corpus with only 17,000 entries. Van Mol also argues that the fine grained word senses in Hans Wehr are mostly not appropriate for modern usage. For example the Arabic verb عمل ‘amala ‘to do’has 36 sense in Hans Wehr, while the corpus gives evidence only to 8 senses in context.

The Buckwalter Arabic Morphological Analyser (BAMA) is widely used in the Arabic NLP research community and has even been considered as a model and described as the “most respected lexical resource of its kind” (Hajic et al., 2005). It is used in the LDC Arabic POS-tagger, Penn Arabic Treebank, and the Prague Arabic Dependency Treebank. It is designed as a main database of word forms interacting with other concatenation databases. Every word form is entered separately. It takes the stem as the base form, and information on the root is also provided. Buckwalter’s morphology reconstructs vowel marks and provides English glossary. Yet there are many drawbacks in Buckwalter's morphological database that discredits it as a truthful representation of Modern Standard Arabic. These drawback are listed below.

1. Buckwalter includes some obsolete lexical items (the amount of which is yet to be determined) which gives us a hint that he relied on some older Arabic dictionaries in the compilation of his own database. This is clear from the examples below which show the classical words included in Buckwalter morphological analyser and the Google score, along with the equivalent MSA word again with the Google score.



#	Meaning	Classical Word	Google	MSA Word	Google
1	sully	قلعط qalʿat	8	لطخ laṭṭaẖa	29,600
2	caulk	قلفط qalfaṭ	9	أفسد ʾafsada	205,000
3	wear	استكد ʾistakadda	4	أنهك ʾanhaka	37,100
4	fickle	غملج ġamlağ	7	متقلب mutaqallib	189,000
5	erosion	انتكال ʾiʾtikāl	7	تآكل taʾākul	1,700,000

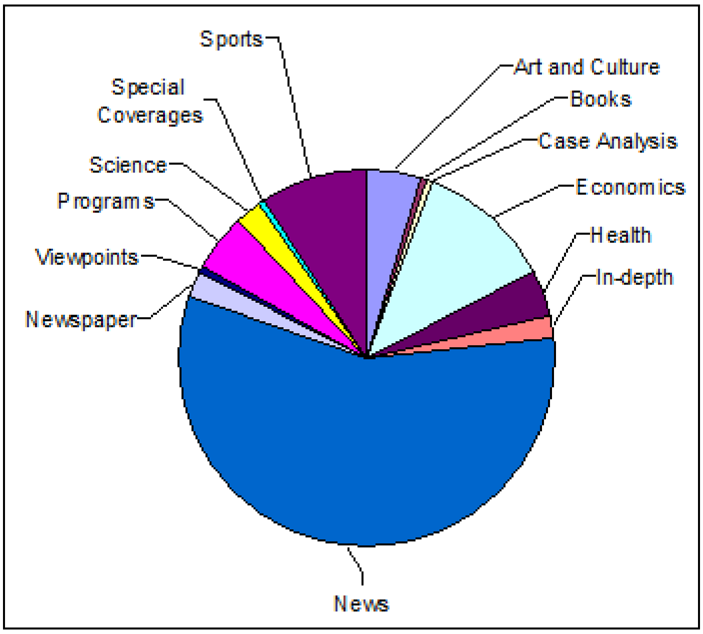
Table 2. Google score for Classical vs. MSA entries

2. To gain an estimation of the size of the problem we conducted some statistics from Al-Jazeera. The total count of lemmas in Buckwalter 2.0 is 40205. After removing diacritics (text in Al-Jazeera is not diacritised) they are reduced to 31,359.

Frequency Range	**0**	**1-100**	**101-1000**	**Over 1000**
Number of Occurrences	7312	13563	6606	3878
Per Cent	23.31%	43.25%	21.06%	12.36

The table shows that 23% of the lemmas in Buckwalter have no occurrences on Al-Jazeera web site. There is a possibility that some lemmas are used in MSA but happen to find no representation in Al-Jazeera, yet we believe that this statistics gives a very reliable indicator on the size of the problem. Al-Jazeera Channel was launched in 1996 and soon became the most dominant Arabic news channel attracting over 40 million viewers (as of 2006 statistics). It has been described as the Arab CNN (Soliman and Feuilherade, 2006). It covers news, analysis, discussion forums of different topics from politics, to economics, sports, book reviews, etc. with writers from all over the Arab region. Al-Jazeera has become the most popular and most influential media channel in the Arab world. Feuilherade (2004), the BBC reporter, states that Al-Jazeera station is “probably the only institution of its kind able to reach so many Arab hearts and minds.” Al-Jazeera employs presenters and reporters from across the spectrum of the Arabic-speaking countries.

The website hosts not only news but a variety of other topics as shown by the pie chart for the corpus we collected from Al-Jazeera between 2001 and 2004.



Not only this but we also expect the statistics to be abundant of false positives, that is word recognized by the search engine but not really the words intended in Buckwalter database. Because diacritics are ignored in writing we stripped the diacritics off before we put the lemmas to the search engine this means that the two forms kataba and kattaba will have exactly the same form. The claim that kattaba is not part of MSA is even harder to verify. We can make two small tests by checking for the diacritized form in A Word Count of Arabic Words and in the ATB. Due to the relatively small size of the two sources, we cannot make definitive conclusions but we can give considerable weight to certain assumptions.

We can also assume that Al-Jazeera covers only news jargon while the Buckwalter database covers both literary and journalize lexical items. To test this claim we collected 11768 word forms from the literary section of Corpus of Contemporary Arabic. The number of words not found in Al-Jazeera was 766 that is 6.5%. We can also note that the search is this case is for full-form words which usually yields less than the lemma. For example the form أتقولين ataqulin 'do you.fem say' in the CCA has no occurrences in Al-Jazeera but the lemma قال qala has 57600 hits.

3. Insufficient coverage of imperative and passive inflections. Regarding the imperative forms: Out of 9198 verbs, only 22 verbs (0.002%) have imperative forms. This is far less than the 32% allowed in our morphology.

   With respect to passive morphology, out of 9198 verbs, only 1404 verbs (15%) are allowed

to have a passive form. In our system, 36% of verbs can have a passive form. Buckwalter’s passive forms are also restricted by tense. Only 110 of them have a passive form in the past (perfective) tense. There are even passive forms for verbs with low probability. The first word has only one occurrence in Al-Jazeera and the second 5.
ʼyumāt ‘be made to dieيمات
يعاش yu ʿāš ‘be lived’

While قوبل qubila “is met”, which has 910 occurrences in Al-Jazeera, is not allowed in the passive in Buckwalter.

4. Some proper names are associated with senses that are no longer used in the language.
   حسام Husam / sword
   حنيفة Hanifah / orthodox

5. Buckwalter’s system does not handle multiword expressions (MWEs).

6. Buckwalter’s system does not give syntactic information on subcategorization frames.

In recent years there has been a growing tendency to standardise lexical resources by specifying the architecture of the lexical resource and the component parts of this database. It also needs to specify how these components are interconnected and how the lexical resource as a whole exchanges information with other NLP applications. LMF has emerged as an ISO standard that lays the specifications of the lexical database not for a particular language, but presumably for all the languages of world.

In our work we will adopt the LMP framework.

The paper henceforth will proceed as follows. In the following subsection we will explain what is meant by MSA and how it is different from CA. Then we will briefly summarize some of the basic ideas in modern lexicography which we will use in the construction of our lexical resource. Then we will review Arabic dictionaries and dictionary making strategies across different historical periods.

1.1 Modern Standard Arabic vs. Classical Arabic

MSA, the subject of our research, is the language of modern writing, prepared speeches and the language of the news. It is the language universally understood by Arabic speakers. MSA stands in contrast both to Classical Arabic and vernacular Arabic dialects. CA is the language which appeared in the Arabian Peninsula centuries before the emergence of Islam and continued to be the standard language until the medieval times. CA continues to the present day as the language of religious teaching, poetry and scholarly literature.

MSA is different from Classical Arabic on the lexical, morphological and syntactic levels. On the lexical level there is a significant expansion of the lexicon to cater for the needs of modernity. New words are constantly coined or borrowed from foreign languages. The coinage of new words does not necessarily abide by the traditional rules of derivation, which frequently leads to contention between writers and conformist philologists. On the morphological and syntactic levels there is a less visible degree of variation. In general MSA conforms to the rules of CA, but in MSA there is a greater tendency for simplification and modern writers use only a subset of the full range of structures, inflections and derivations available in CA. There is now no strict abidance by case ending rules which led some structures to die out, while some syntactic structures which were marginal in CA started to have more salience in MSA. For example the classical word order of object-verb-subject (OVS) is hardly found in MSA. Further, to avoid ambiguity and improve readability, there is also a tendency to avoid passive verb forms where the active readings are also possible, as in the words قدم quddima 'offered', نظم nuzzima 'organized', وثق wuthiqa 'documented'. Instead of the passive form, the alternative syntactic construction of تم tamma (performed/done) + verbal noun is used تم تقديمه tamma taqdimuhu 'it was offered', تم تنظيمه tamma tanzimuhu 'it was organized', and تم توثيقه tamma tawthiquhu 'it was documented'. The relatively marginal word order of subject-verb-object in Classical Arabic is gaining more weight in MSA. This is confirmed by Van Mol (2003) who quotes Stetkevych (1972) as pointing out the fact the MSA word order has shifted balance, as the subject precedes the verb more frequently, breaking from the classical default word order of verb-subject-object.

However, apart from Van Mol's (2003) study of the variations in complementary particles, comparisons between MSA and CA have been usually based on personal observation and subjective judgements. No profound quantifiable studies have been conducted to check how big or small the difference between MSA and CA is either on the morphological, lexical or syntactic levels.

1.2 Modern Principles of Lexicography

Before we start building a lexical database, or dictionary, we need to find answers to questions related to the nature of a dictionary, what constitutes an evidence for a lexical entry, what are the best practices and methods used in dictionary compiling, what is the role a corpus plays in a dictionary and what are the characteristics of such corpus. This is a brief description of the principles involved dictionary making. The main source of information here is *The Oxford Guide to Practical Lexicography* by Atkins and Rundell (2008).

Definition of a dictionary

A dictionary is defined as a description of the vocabulary used by members of a speech community. The job of a general-domain dictionary is to describe linguistic conventions, that is the way people normally use and understand words, rather than trying to account for idiosyncrasies, rarities and violations of the norms of the language.

Lexical evidence

The starting point for the process of dictionary building is gathering the evidence of what the members of the speech community do when they communicate. Subjective evidence, either through introspection or informant-testing (asking the opinion of some speakers), cannot form the basis of a reliable dictionary as it records only linguistic knowledge of a limited number of individuals which is ultimately partial and incomplete. The only objective evidence that we can rely on is a corpus, as it allows us to observe what people actually do when they communicate with one another. A corpus allows us to provide “typifications” of the language, that is deciding whether a given utterance is typical and therefore worth including in the dictionary, or idiosyncratic and therefore outside our scope. A typical lexical entry means that it is both “frequent” (occurs frequently in a corpus) and “well-dispersed” (found in a variety of text-types), and hence can confidently be regarded as belonging to the stable “core” of the language.

Corpora and Lexicography

The Brown Corpus of current American English, developed in the early 1960s, was the first electronic corpus of English. Its goal was to collect one million words of text. This corpus was used as a citation base for the *American Heritage Dictionary*, first appeared in 1969.

The Birmingham Collection of English Text (BCET) in the early 1980s had 20 million words was used in the compilation of the Cobuild English Dictionary. The British National Corpus (BNC) in the 1990s collected 100 million words. The BNC was a well-balanced, carefully-encoded corpus which helped set the standards of corpus collection for subsequent projects. The Oxford English Corpus (OEC) is used in the making of Oxford English Dictionary. In the 2000s, this corpus reached over one billion words.

Characteristics of a reliable corpus

A corpus provides very large volumes of data that allow us to calculate frequency statistics and observe the normal language events that are “recurrent”. In this way we can confidently distinguish between what is conventional and what is idiosyncratic, what is probable and what is possible. Building corpus for lexicography is not an exact science, yet there are some general principles (or characteristics) that, when followed carefully, can improve the corpus value and usability. They are summarised below, some of these principles may overlap or express facts from different perspectives.

a. The corpus does not favour high class language

Lexicographers working in the prescriptive tradition typically aim at preserving the “purity” of the language, and so they favour works by writers of the first-class reputation. By contrast a mainstream descriptive lexicographic corpus must be provide a true and genuine snapshot of the language as it is actually used by whole spectrum of language users.

b. The Corpus should be large and diverse

A corpus designed for use in dictionary-making should cover large and wide-ranging text-types. Corpora usually vary in size, sometime they are less than a million words or running above one billion, and there is no approved limit or minimum size for a corpus, but the frequency characteristics observed by the Zipf’s Law indicate that a few words occur very frequently while many words occur only rarely. This means that a language consists of a small number of very common words, and a large number of very infrequent or rare words. Therefore if we want to be able to adequately investigate rarer and less frequent words, we need larger and larger amounts of text.

c. The corpus should be either synchronic or diachronic

Before starting corpus collection a decision must be made whether to include texts from different historical ages (diachronic) or from a specific contemporary period (synchronic). Obviously a historical dictionary requires a diachronic corpus, while dictionaries designed for learners or ordinary users need a synchronic corpus that tells how the language is used at the present time.

d. The corpus should be well-balanced

It is not possible in corpus to follow the standard scientific way of collecting a “random sample” because the subject of our sampling is language which is a living and dynamic object. The dynamicity of the language prevents us from fully understanding it nature or determining its limits. Therefore, what corpus compilers aspire at achieving is creating a “balanced” corpus. A well-known strategy that allows us to create a balanced corpus is using “stratified sampling”, which means to break down the text into a number of text-types or subject fields. Then it will be easier to collect random samples from each of these subject fields. Moreover, the balance should not only be type, but in proportion as well, that is we need to decide on the amount we can take from each text type.

e. The corpus should avoid skewing

Skewing means that there is bias in the corpus data towards either over- or under-emphasizing a particular feature in the text to the degree that it is no longer possible to make credible generalizations. A typical example is when a corpus consists of a single type of text (such as news only, or literary works only). Such a corpus will reflect only the linguistic features of that particular genre, and will be considered skewed as it fails to satisfactorily represent the diversity of the language as a whole.

Lexical Profiling

In order to be able to gain adequate and sufficient understanding of a word, a lexicographer need to have access to the following essential information.

1. Word POS

A word class is the most central information that is directly related to its meaning. Words are usually classified into nine categories: nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions, articles, and interjections).

2. Valency Information

Valency in this context means the way a word combines systematically with other words or phrases. This will not only include the word's argument structures (or subcategorization frames, such as Subject, Object, Oblique, etc.) but also includes other grammatical constructions in which it participates in an obligatory and optional fashion (such as complements, modifiers, adjuncts, etc.), and the types of phrases that fill certain syntactic positions. For example the verb *watch* takes an object, but this object can be an NP, *watch the children*, or an NP followed by a verb in the infinitive, *watch the children play*, or an NP followed by a verb in the present participle, *watch the children playing*.

3. Collocations

This term refers to the observable tendency of certain words to occur with certain other words more often than by chance. This can be seen in nouns that tend to co-occur with certain verbs “to commit a crime”, or to be modified by certain adjectives “vast knowledge”, etc. Collocations are an important factor in determining the word's meaning.

4. Colligational preferences

This is the observable tendency of some words to have particular morphological forms or occur in a particular syntactic position. If we find a particular verb that is almost always passivised or a particular noun comes usually in the plural form, then we have an obvious case of colligational behaviour.

In order to have access to this information lexicographers usually refer to concordencers. A concordance is usually helpful in viewing lexical information, but it becomes neither practical nor efficient when the frequency hits grows larger.

A lexical profiling software (such as the Word Sketch (Kilgarriff,;), allows to avoid the disadvantages of using a concordance by using statistical methods to reveals the salient facts about the way a word most typically combines with other words, such as collocations, grammatical functions, phrase types, etc. Lexical-profiling software only works well for lemmas with at least 500 corpus hits (preferably far more). For lexical-profiling software the first requirement is a POS-tagged corpus. Word Sketches (a well-established type of lexical profile) produce a statistical summary of a lexicographically relevant information such as the word’s grammatical and collocational behaviour.

1.3 History of Arabic Lexicography

Kitab al-'Ain by al-Khalil bin Ahmed al-Farahidi (died 789) is the first complete Arabic monolingual dictionary. It was a comprehensive descriptive record of the lexicon of the contemporary Arabic language at the time. It did not register only for the high level formal language of the Koran, Prophet's sayings, poetry, and memorable pieces of literature and proverbs, but it also included truthful account of common words and phrases used by Bedouins and common people.

The other dictionaries that compiled in the centuries following *al-'Ain* typically included either refinement, expansion, correction, or orgainizational improvements of the previous dictionaries[1]. These dictionaries include *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980), *al-Muheet* by al-Sahib bin 'Abbad (died 995), *Lisan al-'Arab* by ibn Manzour (died 1311), al-Qamous al-Muheet by al-Fairouzabadi (died 1414) and Taj al-Arous by Muhammad Murtada al-Zabidi (died 1791) (Owens, 1997).

Even modern dictionaries such as *Muheet al-Muheet* (1869) by Butrus al-Bustani and *al-Mu'jam al-Waseet* (1960) by the Academy of the Arabic Language in Cairo did not start from scratch nor did they try to overhaul the process of dictionary compilation or make any significant change. Their aim was merely to preserve the language, refine older dictionaries and accommodate accepted modern terminology. Some researchers criticize Arabic dictionaries for representing a fossilized version of the language with each new one reflecting the content of the preceding dictionaries (Ghazali and Braham, 2001).

Serious work in bilingual Arabic lexicography was done by Arabists, most notable among them were Edward William Lane in the nineteenth century and Hans Vehr in the twentieth century. Edward William Lane's *Arabic-English Lexicon* (compiled between 1842 and 1876) was hugely indebted, as admitted by Lane himself (Lane, 1863), to previous Arabic monolingual dictionaries, chiefly the *Taj al-ʿArus* by Muhammad Murtada al-Zabidi (1732-1791). Lane spent 7 years in Egypt acquiring materials for his dictionary and ultimately helped preserve the decaying and mutilated manuscripts he relied on (Arberry, 1960).

The most renowned and well-celebrated Arabic-English dictionary in the modern time is Hans Wehr's Dictionary of Modern Written Arabic (first published 1961). The work started as an Arabic-German dictionary *Arabisches Wörterbuch für die Schriftsprache der Gegenwart*, published 1952, and later translated to English and revised and extended.

The dictionary compilers stated (Wehr, 1976) as its primary goal to follow descriptive and scientific principles by including only words and expressions that are attested in context of the corpus they collected.

> “From its inception, this dictionary has been compiled on scientific descriptive principles. It contains only words and expressions which were found in context during the course of wide reading in literature of every kind or which, on the basis of other evidence, can be shown to be unquestionably a part of the present-day vocabulary.”

This was an ambitious goal indeed, but was the application up to the stated standard? We find mainly three defects in the practice that defeated the declared purpose of the dictionary. These defects are in data collection, use of secondary sources and the approach to idiosyncratic classicisms. The material for the dictionary was collected between 1940 1948 and included 45,000 slips containing citations from Arabic sources. The primary source materials consisted of selected

[1] http://lexicons.sakhr.com
http://www.almeshkat.net/books/list.php?cat=16
http://www.angelfire.com/tx4/lisan/lex_zam/dilalahessays/lexicons.htm

works by poets, literary critics and writers immersed in classical literature and renowned for their high flying language such as Taha Husain, Muhammad Husain Haikal, Taufiq al-Hakim, Mahmoud Taimur, al-Manfalauti, Jubran Khalil Jubran and Amin ar-Raihani (as well as some newspapers, periodicals and specialized handbooks). These writers appeared at a time known in the history of Arabic literature as the period of *Nahda*, which means revival or Renaissance. A distinctive feature of many writers in this period was that they tried to emulate the famous literary works in the pre-Islamic era and the flourishing literature in the early centuries after Islam. This makes the data obviously skewed by favouring literary, imaginative language.

The dictionary compilers used as “secondary sources”, that is some of the then available Arabic-French and Arabic English dictionaries. Items in the secondary sources for which there were no attestations in the primary sources left to the judgment of an Arabic native speaker collaborator in such a way that word known to him, or already included in older dictionaries, were incorporated. The use of secondary sources in this way was a serious fault and was enough to damage the reliability of the Hans Wehr's dictionary as true representation of the contemporary language.

The third setback was the dictionary compilers' approach to what they defined as the problem of classicisms, or rare literary words. Despite their full understanding of the nature of these archaic forms, the decision was to include them in this dictionary, even though it was sometimes evident that they “no longer form a part of the living lexicon and are used only by a small group of well-read literary connoisseurs”. The inclusion of these rarities inevitably affected the representativeness of the dictionary and marked a significant bias towards literary forms.

Not too far away from the domain of lexicography, two Arabic word count studies appeared in 1940 and 1959 but did not receive the attention they deserve by Arabic lexicographers, perhaps because the two works were intended for pedagogical purposes to aid in the vocabulary selection for primers and graded readers. The first was Moshe Brill's work (Brill, 1940) which was a pioneering systematic study in Arabic word count. Brill conducted word count on 136,000 running words from the Arabic daily press, and the results were published as *The Basic Word List of the Arabic Daily Newspaper* (1940). This word count was used as a basis for a useful Arabic-Hebrew dictionary compiled by Brill's two assistants.

Landau (1959) tried to make up for what he perceived as a technical shortcoming in Brill's work: the count covered only the language of the daily press. So he complemented Brill's work by conducting a word count on an equal portion of 136,000 running words from Arabic prose based on 60 twentieth-century Egyptian books on a various selection of topics and domains including fiction, literary criticism, history, biography, political science, religion, social studies and economics with some material on the borderline between fiction and social sciences, e.g. travels and historical novels. It seems that Landau went into great length to collect this well-balanced corpus, which predates the emergence of the discipline of corpus linguistics and the first electronic corpus, the Brown Cropus in 1960s. Landau combined Brill's work in his book and compared it to his work, thus we have the results of two counts: Brill's count of the press usage, and Landau's count of literary usage. The former showed close to 6,000 separate words; the latter over 11,000, and the combined list gave 12,400 specific words (Perlmann, 1960).

Through this frequency study, Landau was able to deduce insightful results from frequency statistics which basically complied with Zipf's law. He noted that the first 25 words with the highest frequency represented 25% of the total number of running words, the first 100, more than 38%, the first 500, 58.5%, and the first 1000, 70%. He also found that 1134 words occurred each only once in the press, and 3905 words which occurred only once in literature, which reflects the abundance of rare words in literary works.

The only obvious weakness of this study was that the number of running words counted (only 272,000 words) was inadequately small, as admitted by the author himself, in comparison with the contemporary word count for other languages such that of Thorndike and Lorge in English (25,000,000), of Kaeding in German (11,000,000).

2 Our Project: Aim and Methodology

In the previous sections we noted how Modern Standard Arabic is different from Classical Arabic, summerized the lexicographical principles involved in dictionary making and reviewed the current state of Arabic dictionaries, what methodology they followed and what goal they tried to achieve. Now we turn to our project and explain what it tries to achieve and what methods it is going to follow to achieve these goals.

2.1 Aim of the Project

Our aim is the acquisition of Arabic lexical resources and the production of new lexical sets from these resources. In order to make sure that the lexical items we acquire reflect the modern usage and to avoid classical forms we rely on a selection of corpora that that represent both modern language in varied domains. The lexical data accumulated will be stored in a MySQL database as a convenient pivot to facilitate any further exploitation and manipulation on them such as manual validation, exporting into LMF format, and exchange with other NLP applications or Machine Readable Dictionaries (MRDs). We address two main challenges in this paper: acquiring lexical resources from corpora and inducing the lexical profile for each lemma or entry that will make the overall structure of our lexical database compatible with LMF specifications.

If we take the Buckwalter database (used in Buckwalter morphological analyser) as a baseline and compare our work to it, the following points will summarize what advantages our lexical database will have.

- We include only lexical entries that have been attested in a corpus. We don't include classical or archaic words, thus eliminating the noise and significantly reducing spurious ambiguity.
- We include subcategorization frames for verbs and verbal nouns, and each equivalent in English will be linked to the right subcat frame.
- We include +/-human semantic information for nouns.
- We include information about the root.
- We include more detailed information about derived nouns/adjectives, stating if the form is an active or passive participle or a verbal noun, *masdar*.
- We include multi-word expressions, which is entirely lacking in Buckwalter.
- We include better classification of proper nouns: person, place, organization, etc.

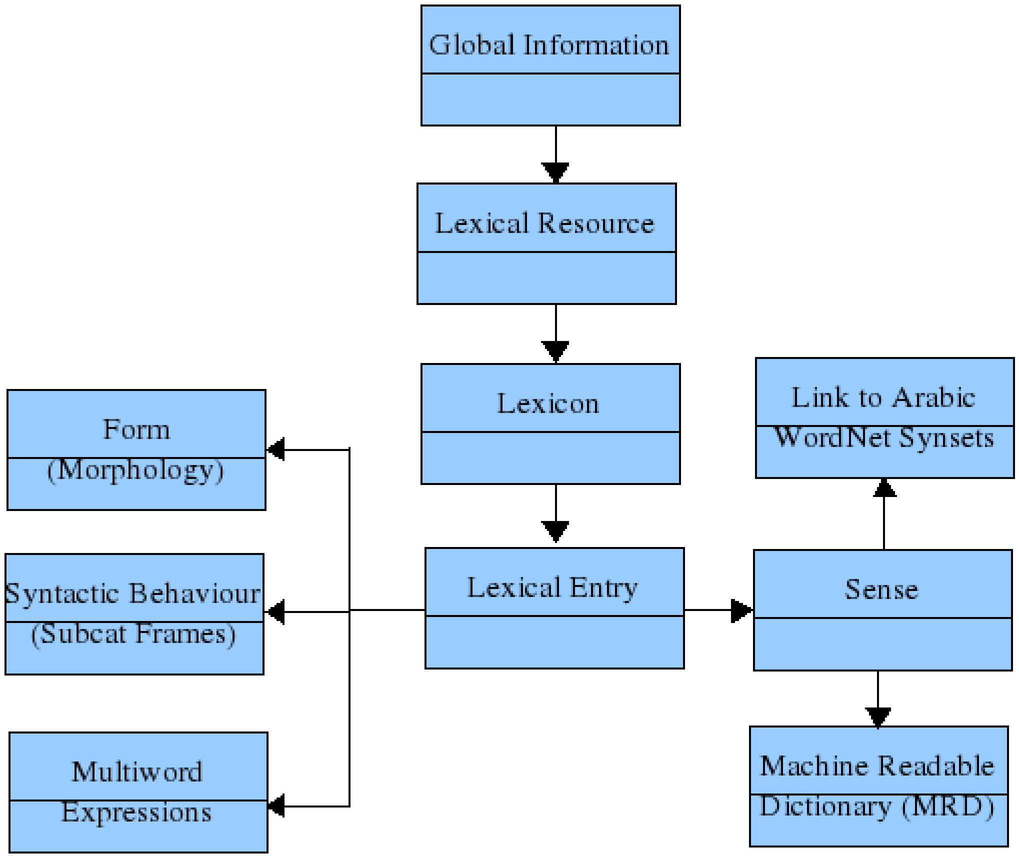
Compatibility with LMF

LMF is an ISO standard that facilitates the exchange of lexical information between different lexical resources on the one hand between lexical resources and NLP applications on the other (ISO 24613: 2007; Francopoulo et al., 2008; Khemakhem et al. 2009; Loukil et al., 2008; Salmon-Alt et al. 2005; Maks 2008). LMF specified XML as the encoding formatting of the electronic lexicons. It also specifies naming convention and a hierarchical structure of the components of the lexical resources. It also takes into account the particular needs of languages with rich and complex morphology, such as Arabic. LMF covers five main topics (ISO 24613: 2007):

1.Morphology extension
2.Machine Readable Dictionary extension

3.NLP syntax extension
4.NLP semantics extension
5.NLP multiword expression patterns extension

These can be represented graphically as in the figure below.


Global Information
Lexical Resource
Lexicon
Link to Arabic WordNet Synsets
Form (Morphology)
Syntactic Behaviour (Subcat Frames)
Lexical Entry
Sense
Multiword Expressions
Machine Readable Dictionary (MRD)

A sample encoding for the Arabic verb 'kata' (or write) in XML according to LMF will look like this:

```
<LexicalResource dtdVersion="14">
 <GlobalInformation
  <feat att="languageCoding" val="ISO 1256"/>
 </GlobalInformation>
 <Lexicon>
  <feat att="language" val="arb"/>
  <LexicalEntry>
    <feat att="partOfSpeech" val="verb"/>
    <Lemma>
      <feat att="writtenForm" val="katab"/>
    </Lemma>
    <WordForm>
      <feat att="writtenForm" val="kataba"/>
      <feat att="grammaticalNumber="singular"/>
    </WordForm>
   <Sense>
       <Sense SenseNumber="#">
```

```
        <Equivalent Language=“eng”>
            <WordForm>
              <feat att="writtenForm" val="write"/>
            </WordForm>
        </Equivalent>
    </Sense>
    <SyntacticBehaviour subcategorizationFrames="regularSVO"/>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

The information structure and presentation format is compatible with the LMF specifications.

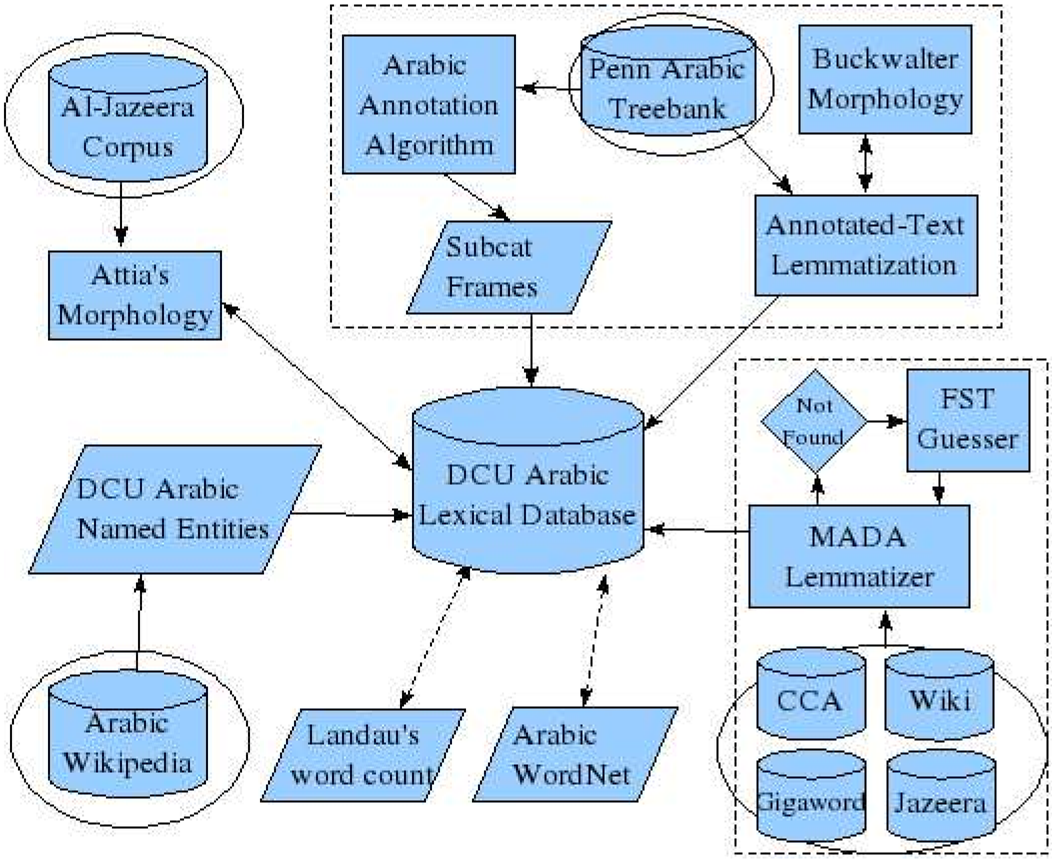
- Morphological information: word root, lemma, form, diacritics, frequency, citations. This information will be extracted from the Arabic Treebank (ATB) and Buckwalter Arabic Morphological Analyser (BAMA), and Attia's Finite State Morphological Analyser.
- Syntactic information: Subcategorization frames. This information will be automatically extracted from the DCU Arabic dependency annotated treebank, and the Arabic XLE grammar.
- Semantic information: linking to Arabic WordNet.
- Dictionary information: translation in English
- Multi-word Expression (MWE) and named entity. This will be taken from Attia's Morphology and the Arabic Named Entity Lexicon (ANEL) project, still in progress.

2.2 Lexical Acquisition Architecture

It is more complex to explore Arabic corpora due to its derivational and inflectional nature, lack of diacritics (vowel marks), and the employment of cliticization, or affixation of function words to content words. Here lemmatization proves to be an essential prerequisite in the acquisition of lexical resources for Arabic.

We build a core base of lexical items from hand-annotated corpora and then move on to extend this base by processing larger and more domain-varied un-annotated corpora.

The Penn Arabic Treebank (ATB) is a morphologically/syntactically annotated corpus of modern texts taken from the newswire. Due to the fact that it is tokenized and diacritized, and the POS tags were manually reviewed by human annotators, it constitutes a valuable resource for lexicographic purposes.



We have developed an annotation algorithm that automatically builds dependency treebank from the ATB. We have also automatically collected subcat frames in order to handle long-distance dependencies. These subcat frames will be automatically added to the relevant lexical entry in our dictionary.

Attia developed a morphological analayser with detailed information on word classes and morpho-syntactic behaviour. For instance he gives information on whether a verb is transitive or intransitive, whether the noun denote human or non-human entity, whether a form is active or passive participle or a verbal noun. He has also built a list or MWEs and well-classified list of proper nouns, stating whether the proper noun is the name of a place, organization or person, and whether the person is feminine or masculine. Attia built his morphological analyser from scratch from a corpus of news items taken from Al-Jazeera website. Therefore it meets our criteria for inclusion in the dictionary.

3. Preliminary Results

In a way of comparison we need to state that Buckwalter Arabic Morphological Analyser (BAMA) contains 40,222 lemmas (including 2034 proper nouns). Derived word are usually associated with roots, and the lexicon contains 7,614 such roots. Attia's morphology, on the other hand, contains 10799 lemmas (1532 verbs, 8923 nouns and adjectives (including 3098 proper nouns), and 344 function words) and 2818 multiword expressions. The Arabic WordNet (AWN) consists of 11,269 synsets2 containing a total of 23,481 Arabic expressions. This number includes 1,142 NEs which were extracted automatically and checked by the lexicographers.

Here is also some statistics from the Penn Arabic Treebank (ATB). The ATB consists of 23,611 sentences, 553,363 words, and 428,761 content words (nouns, verbs, adjectives and adverbs). The number of NEs in the ATB reaches 54,398.

We collected 73,115 types (unique combinations of POS and word forms) from the ATB for the open class categories, common nouns, adjectives and verbs. The vocalization and POS from Buckwalter were matched against these collected from the treebank. We found 58,810 matches for these words in BAMA with full information on word form, lemma, vocalization and translation. These were further reduced to 12,039 unique lemmas. We also collected 10,500 unique proper nouns. Yet proper names in the ATB are not classified according to type, so that it is not possible to say whether the proper noun is a person name, a country, object or an organization. We also still need to investigate what happened to the 14,305 word forms that were not matched.

In summary, in this initial stage we instantly created a full lexicon of 12,039 lemmas. Yet this can be increased by investigating the unmatched words, which constitute 20% of the data, and by applying the technique to the smaller ATBs that we have. The table below gives a more fine-grained account of the results.

	Unique POS-Word Combinations	Matched Words	Unmatched Words	Unique Lemmas
Nouns	41,183	37,328	3,855	7,184
Adjectives	14,044	9,950	4,094	2,540
Verbs	17,888	11,532	6,356	2,315

	Unique POS-Word Combinations	Unique Lemmas
Nouns	41,183	7,184
Adjectives	14,044	2,540
Verbs	17,888	2,315
Total	73,115	12,039

Then we devised a multi-layered matching mechanism by relaxing the matching conditions from exact matching to different stages of flexibility starting with small differences and ending with more severe ones.

Phase 2: Addition from Attia's Morphology

Attia's morphology adds 9.14% to the core (common nouns, adjectives and verbs) of MSALex. It also adds a ready-made list of MWEs and a list of better-classified proper nouns.

Matching with Attia's morphology
Number of Attia Nominals = 5806
Number of MSALex nominals (nouns and adjectives) = 9755
Nominals found in Attia's Morphology (Intersection MSALex and Attia) = 5508
Nominals not found in Attia's Morphology (MSALex Only) = 4247
Nominals not found in MSALex Morphology (Attia Only) = 298 – through calculation
Nominals not found in MSALex Morphology (Attia Only) = 933 – actual

Number of Attia Verbals = 1532
Number of MSALex Verbals = 2585
Verbals not found in Attia's Morphology (MSALex Only) = 1195
Verbals found in Attia's Morphology (Intersection MSALex and Attia) = 1390
Verbals not found in MSALex Morphology (Attia Only) = 142 – through calculation
Verbals not found in MSALex Morphology (Attia Only) = 195 – actual

Phase 3 Lexicon from Free Text

In this analysis we use MADA (Habash et al., 2005; Roth et al., 2008) for pre-processing.
The size of CCA is 516,798 words
“So far the CCA consists of over 843,000 words in 416 files covering a wide range of categories. ” Al-Sulaiti (2006) but this depends on how the words are counted
MADA:
NO-ANALYSIS = 69425
WORDS = 546564
Coverage 87%

6106 items from the CCA had no analysis by Mada (out of 69425) when removing punctuation and numbers.
These were reduced to 4902 items when sorted unique word and analysis combination (no repetition)
After comparing to an FST Guesser these were further reduced to 3379 after removing common spelling errors:

```
! GuessLemma.matches(".*ة.+") taa marbouta in the middle
! GuessLemma.matches(".*ى.+") Alif maqsoura in the middle in the middle
! GuessLemma.matches(".*اا.*") two alifs anywhere
```

MadaInput: *1.000000 wAlmtswqyn=[wAlmtswqyn_0 POS:AJ Al+ +ACC w+ +DEF MOOD:NA +PL]=NO-ANALYSIS
Mada Features after Conversion: +adj+pl+acc+defArt+conj

والمتسوقين	+adjوالمتسوق+Guess+dual+acc+gen@	2
والمتسوقين	+adjوالمتسوق+Guess+masc+pl+acc+gen@	3
والمتسوقين	+adjوالمتسوقين+Guess+sg@	1
والمتسوقين	+nounوالمتسوق+Guess+dual+acc+gen@	0
والمتسوقين	+nounوالمتسوق+Guess+masc+pl+acc+gen@	0
والمتسوقين	+nounوالمتسوقين+Guess+sg@	0
والمتسوقين	و+conj@ال+defArt@+adjمتسوق+Guess+dual+acc+gen@	6
والمتسوقين	و+conj@ال+defArt@+adjمتسوق+Guess+masc+pl+acc+gen@	7
والمتسوقين	و+conj@ال+defArt@+adjمتسوقين+Guess+sg@	5
والمتسوقين	و+conj@ال+defArt@+nounمتسوق+Guess+dual+acc+gen@	0
والمتسوقين	و+conj@ال+defArt@+nounمتسوق+Guess+masc+pl+acc+gen@	0
والمتسوقين	و+conj@ال+defArt@+nounمتسوقين+Guess+sg@	0
والمتسوقين	و+conj@+adjالمتسوق+Guess+dual+acc+gen@	4
والمتسوقين	و+conj@+adjالمتسوق+Guess+masc+pl+acc+gen@	5
والمتسوقين	و+conj@+adjالمتسوقين+Guess+sg@	3
والمتسوقين	و+conj@+nounالمتسوق+Guess+dual+acc+gen@	0
والمتسوقين	و+conj@+nounالمتسوق+Guess+masc+pl+acc+gen@	0
والمتسوقين	و+conj@+nounالمتسوقين+Guess+sg@	0

GuessLemma: 7@والمتسوقين@متسوق@AJ
GuessLemma: 6@والمتسوقين@متسوق@AJ
GuessLemma: 5@والمتسوقين@متسوقين@AJ
GuessLemma: 5@والمتسوقين@المتسوق@AJ
GuessLemma: 4@والمتسوقين@المتسوق@AJ
GuessLemma: 3@والمتسوقين@والمتسوق@AJ
GuessLemma: 3@والمتسوقين@المتسوقين@AJ
GuessLemma: 2@والمتسوقين@والمتسوق@AJ
GuessLemma: 1@والمتسوقين@والمتسوقين@AJ

Formula for giving weight to the guessing output:
wordWeight = ((curFormsRep * 2) + (curLemaRep * 1)) / 2;
Word Weight = ((# of different forms having the same lemma * 2)
(+ # of same forms having the same lemma * 1))
/ 2
57% from the top are valid for inclusion in a dictionary as is
6% from the bottom are valid for inclusion in a dictionary as is

Number of unique nominals in CCA: 12502
Number of unique verbals in CCA: 4245
Nominals in CCA: 240236
Verbals in CCA: 67812

Arabic Wikipedia contains 40 million (42,459,952) words (excluding tags, links and references)
In a subsection of 2,000,000 words
75,000 were not found in MADA (BAMA) - The coverage is 96%.
36,000 unique words not found
of them 26,000 words have frequency of one.
22164 verbs, nouns and adjectives were collected from the first portion
10712 were not found in the ATB
7763 Nominals not found in the ATB
2949 verbs not found in the ATB

6312 items (nominals and verbs) from the CCA were not found in the ATB
out of a total of 16747 words (nominals and verbs)

In CCA No analysis in MADA is mostly caused by punctuation marks. When these are removed we are left with 6118 words which have no coverage in MADA (and BAMA as well). These are reduced to 4149 unique words (after removing repetitions)

Testing CCA on Aljazeera
We looked for the 6312 lexical items that were extracted from the CCA and had no match in the ATB on Aljazeera web site
When searching by lemmas: 543 were not found (9%)
When searching by full forms: 941 were not found (15%)
When combining them together: 240 lexical items were not found when searching either by lemma or full form (4%).

12340 from ATB
1128 from Attia
6312 from CCA
Total: 19780

Aljazeera corpus developed in-house contains 88 million words.

Arabic Gigawords contains about 200 million words.

Testing and Evaluation

Reason for choosing Aljazeera web site for testing.
Search Engine on the web is misleading. The web is polluted with dirty data.
These are Google and CNN statistics for 27/1/2010

Misspellings	Google Score	CNN Score	Right Form	Google Score	CNN Score
arround	1,200,000	3	around	780,000,000	44,555
vedio	4,450,000	0	video	2,590,000,000	131,845
resaercher	6,200	0	researcher	26,500,000	19,729
possebility	31,100	0	possibility	95,100,000	38,163
bilieve	29,200	0	believe	349,000,000	44,330
perfromance	195,000	0	performance	459,000,000	17,085
mesjudge	80	0	misjudge	278,000	196
gtfrde	1,750	0			
ghgh	233,000	0			

We collected 12340 lemmas from the ATB. After removing tashkil these were reduced to 10071. We ran the list of lemmas on Al-Jazeera search engine. We found that 208 lemmas (2%) were not found. By analysing the errors we found that there are one of three possibilities: either the lemma has the wrong form (error in Buckwalter morphology) or the treebank has the wrong annotation, or the lemma is legitimate but has no occurrence in Al-Jazeera.

```
0	أبلق	>abolaq
frst: mtch_1	ADJ+CASE_DEF_GEN	baloqA'+i	>abolaq_1
(NP (NOUN+NSUFF_FEM_SG+CASE_DEF_GEN saEAd+at+i-)(POSS_PRON_3MP -him)))(PP
(NP (ADJ+CASE_DEF_GEN baloqA'+i)(NP (NOUN+CASE_DEF_GEN zumalA'+i)(NP
(DET+NOUN+CASE_DEF_GEN Al+>amos+i))
```

This is mistagging in the annotation it should be “bi-liqA'” as a preposition and noun not as an adjective

```
0	أبرشي >abora$iy~
frst: mtch_1	NOUN+NSUFF_FEM_SG+CASE_DEF_GEN	>abora$iy~+ap+i
	>abora$iy~_1
```

Obviously, Buckwalter gives the wrong lemma it should be >abora$iy~+ap not >abora$iy~

```
0	أدكن	>adokan
frst: mtch_1	NOUN+CASE_INDEF_GEN	dakonA'+a	>adokan_1
```

neither >adokan nor dakonA' were found in Al-Jazeera

Further we tested Attia's morphology on Al-Jazeera. Out of 7338 lemmas 31 (0.42%) were not found. Some are misspellings while the others are frozen inflected words that do not usually appear in the lemma form.

4. Future Work

We need to progress from the ATB corpus to the Corpus of Contemporary Arabic (CCA). This will allow us to extend the coverage of the lexicon and improve its representativeness by relying on data from the CCA.

5. Conclusion

We developed a model for the automatic acquisition of lexical information from texts and apply this model to construct a large lexical resource for Modern Standard Arabic from corpora. This is a multi-faceted lexical resource for Arabic that has high potentials distribution as a machine readable dictionary or as a core for bootstrapping projects in lexicography, or NLP annotation task.

References

Al-Sulaiti, Latifa; Atwell, Eric. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, vol. 11, pp. 135-171. 2006.

Arberry, Arthur John. (1960). *Oriental essays: portraits of seven scholars* . London: George Allen and Unwin.

Atkins, B. T. S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford University Press.

Attia, Mohammed A. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *Challenge of Arabic for NLP/MT Conference*, The British Computer Society, London, UK, pp. 48-67.

Beesley, Kenneth R. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.

Brill, Moshe. 1940. *The Basic Word List of the Arabic Daily Newspaper*. The Hebrew University Press Association: Jerusalem

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. In *Linguistic Data Consortium. Catalog number LDC2002L49, and ISBN 1-58563-257-0.*

Dichy, Joseph, and Fargaly, Ali. 2003. Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built? In *The MT-Summit IX workshop on Machine Translation for Semitic Languages,* New Orleans, USA.

Feuilherade, Peter. 2004. Al-Jazeera debates its future. http://news.bbc.co.uk/1/hi/world/middle_east/3889551.stm

Francopoulo, Gil, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet and Claudia Soria. 2008. Multilingual resources for NLP in the lexical markup framework (LMF). Language Resources and Evaluation (revue) ISSN 1574-020X (print) + 1572-0218 (online) Springer Netherlands

Ghazali, Salem and Braham, Abdelfattah. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. Arabic NLP Workshop at ACL/EACL. Toulouse, France.

Habash, Nizar and Rambow, Owen. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. Proceedings of the 43rd Annual Meeting of the Association for

Computational Linguistics (ACL'05), pp. 573--580.

ISO 24613: 2007 Language Resource Management Lexical Markup Framework (draft version), ISO Switzerland.

Khemakhem, Aida, Imen Elleuch, Bilel Gargouri and Abdelmajid Ben Hamadou. 2009. Towards an Automatic Conversion Approach of Editorial Arabic Dictionaries into LMF-ISO 24613 Standardized Model. Proceedings of the Second International Conference on Arabic Language Resources and Tools. Cairo, Egypt

Landau, Jacob M. 1959. *A Word Count of Modern Arabic Prose.* American Council of Learned Societies, New York.

Lane, Edward William (1863). "Preface", in Arabic-English Lexicon. London: Williams and Norgate.

Loukil, Noureddine, Haddar, Kais and Ben Hammadou, Abdelmajid (2008). Towards a syntactic lexicon of Arabic verbs. In HLT and NPL within the Arabic World: Arabic Language and Local Languages Processing – Status Updates and Prospects (LREC 2008). Marrakech, Morocco.

Maks, Isa, Tiberius Carole, Van Veenendaal Remco. 2008 Standardising bilingual lexical resources according to the Lexicon Markup Framework. LREC. Marrakech

Owens, Jonathan. 1997. "The Arabic Grammatical Tradition," *The Semitic Languages*. London: Routledge. Pg 56.

Perlmann, Moshe. 1960. “Review of *A Word Count of Modern Arabic Prose*”. Middle East Journal, Vol. 14, No. 1 (Winter, 1960), pp. 106-107. Published by: Middle East Institute

Roth, Ryan and Rambow, Owen and Habash, Nizar and Diab, Mona and Rudin, Cynthia. 2008. Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. Proceedings of ACL-08: HLT, Short Papers, pp. 117--120.

Salmon-Alt, Susanne; Akrout, Amine; Romary, Laurent. 2005. Proposals for a normalized representation of Standard Arabic full form lexica. Proceedings of the International Conference on Machine Intelligence (ICMI'05). Tozeur: Tunisia

Soliman, Amani and Peter Feuilherade. 2006. Al-Jazeera's popularity and impact. http://news.bbc.co.uk/2/hi/middle_east/6106424.stm

Van Mol, Mark (2003) *Variation in Modern Standard Arabic in Radio News Broadcasts, A Synchronic Descriptive Investigation in the use of complementary Particles*, Leuven, OLA 117, 324 p.

Van Mol, Mark. (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). Stuttgart, 8–12 August. (http://www.ilt.kuleuven.ac.be/ilt/arabic/_pdf/stuttgart.pdf)

Wehr, Hans (1976). "Introduction", in Hans Wehr & J M. Cowan *Dictionary of Modern Written Arabic*, pp. VII-XV. Ithaca, N.Y.: Spoken Language Services.