# A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer

Mohammed Attia, Pavel Pecina, Antonio Toral, Lamia Tounsi, and Josef van Genabith

School of Computing, Dublin City University, Dublin, Ireland
{mattia,ppecina,atoral,ltounsi,josef}@computing.dcu.ie
http://computing.dcu.ie

**Abstract.** Current Arabic lexicons, whether computational or otherwise, make no distinction between entries from Modern Standard Arabic (MSA) and Classical Arabic (CA), and tend to include obsolete words that are not attested in current usage. We address this problem by building a large-scale, corpus-based lexical database that is representative of MSA. We use an MSA corpus of 1,089,111,204 words, a pre-annotation tool, machine learning techniques, and knowledge-based templatic matching to automatically acquire and filter lexical knowledge about morpho-syntactic attributes and inflection paradigms. Our lexical database is scalable, interoperable and suitable for constructing a morphological analyser, regardless of the design approach and programming language used. The database is formatted according to the international ISO standard in lexical resource representation, the Lexical Markup Framework (LMF). This lexical database is used in developing an open-source finite-state morphological processing toolkit.[1] We build a web application, AraComLex (Arabic Computer Lexicon),[2] for managing and curating the lexical database.

**Keywords:** Arabic Lexical Database, Modern Standard Arabic, Arabic morphology, Arabic Morphological Transducer.

## 1 Introduction

Lexical resources are essential in most Natural Language Processing (NLP) applications such as text summarisation, classification, indexing, information extraction, information retrieval, machine-aided translation and machine translation. A lexicon is a core component of any morphological analyser [1,2,3,4]. The quality and coverage of the lexical database determines the quality and coverage of the morphological analyser, and limitations in the lexicon will cascade through to higher levels of processing. A lexical database intended for NLP purposes differs from traditional dictionaries in that information on inflection and derivation in the former needs to be represented in a formal and fully explicit way.

Existing Arabic dictionaries are not corpus-based (as in a COBUILD approach [5]), but rather reflect historical and prescriptive perspectives, making no distinction between

[1] http://sourceforge.net/projects/aracomlex/

[2] http://www.cngl.ie/aracomlex

C. Mahlow and M. Piotrowski (Eds.): SFCM 2011, CCIS 100, pp. 98–118, 2011.

entries from Modern Standard Arabic (MSA) and Classical Arabic (CA). Therefore, they tend to include obsolete words that have no place in current usage.

Current computational resources, such as the Buckwalter Arabic Morphological Analyzer (BAMA) [3] and its successor, the Standard Arabic Morphological Analyzer (SAMA) [6], have inherited this drawback from older dictionaries. For example, SAMA contains several thousands of entries that are hardly ever encountered by modern Arabic speakers, such as قَلعَطِ *qalʿaṭ* ‘to stain’, قَلفَط *qalfaṭ* ‘to spoil’, إِستَكَدَّ *istakadda* ‘to exhaust’, and غملَج *ġamlaǧ* ‘unstable’. As a consequence morphological analyses for MSA texts contain many “spurious” interpretations that increase the ambiguity level and complicate text processing.

We address this problem at the lexicographic and computational levels by deriving a specialised MSA lexical resource and generating a Finite State Technology (FST) morphological transducer based on that resource. We start with a manually crafted small MSA seed lexical resource [2] which we take as a model. We extended our lexical resource using SAMA’s database. We use web search queries and statistics from a large automatically annotated MSA corpus (containing 1,089,111,204 words) as two separate filters to determine which lexical information in SAMA is truly representative of MSA and which is CA. Words attested in the MSA data are included in the lexicon, while the others are filtered out.

The filtering stage results in a raw MSA lexical resource that does not contain all the information we need in order to build a complete computational lexicon. For example, our small, hand-crafted seed lexicon [2] includes information about the continuation classes (or inflection paradigms) and humanness for nominal entries, and transitivity and passive/imperative transformations for verbs. This information, however, is missing in the SAMA entries. To solve this problem we use machine learning techniques to help add the new features.

We develop a web application, AraComLex, for curating MSA lexical information. AraComLex provides an interface between the human lexicographer and the lexical database, and provides facilities for editing, maintaining and extending the list of entries. AraComLex complies with the LMF standard in naming convention and hierarchical structure. AraComLex is also used to automatically extend the Arabic FST morphological transducer. We test our morphological transducer for coverage and number of analyses per word, and compare the results to the older version of the transducer as well as to SAMA.

This paper is structured as follows. In the introduction we describe the motivation behind our work. We differentiate between MSA, the focus of this research, and CA which is a historical version of the language. We give a brief history of Arabic lexicography and describe how outdated words are still abundant in current dictionaries. Then we outline the Arabic morphological system to show what layers and tiers are involved in word derivation and inflection. In Section 2, we explain the corpus used in the construction of our lexical database and the standards and technologies employed, mainly Lexical Markup Framework (LMF) and FST. In Section 3, we describe AraComLex, a web application we built for curating our lexical resource. We present the results obtained so far in building and extending the lexical database using a data-driven filtering method and machine learning techniques. We outline how the lexical database is used in creating an open-source FST morphological analyser and evaluate the results. In Section 4, we point to future work, and finally, Section 5 gives the conclusion.

## 1.1 Modern Standard Arabic vs. Classical Arabic

Modern Standard 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 around the world. MSA stands in contrast to both Classical Arabic (CA) 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 a direct descendent of CA and is used today throughout the Arab World in writing and in formal speaking [7].

MSA is different from Classical Arabic at the lexical, morphological, and syntactic levels [8], [9], [10]. At 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 classical morphological rules of derivation, which frequently leads to contention between modern writers and more traditional philologists. Although MSA conforms to the general rules of CA, MSA shows a tendency for simplification, and modern writers use only a subset of the full range of structures, inflections, and derivations available in CA. For example, Arabic speakers no longer strictly abide by case ending rules, which led some structures to become obsolete, while some syntactic structures which were marginal in CA started to have more salience in MSA. For example, the word order of object-verb-subject, one of the classical structures, is rarely found in MSA, while the relatively marginal subject-verb-object word order in CA is gaining more weight in MSA. This is confirmed by Van Mol [11] who quotes Stetkevych [12] as pointing out the fact that MSA word order has shifted balance, as the subject now precedes the verb more frequently, breaking from the classical default word order of verb-subject-object. Moreover, to avoid ambiguity and improve readability, there is a tendency to avoid passive verb forms when the active readings are also possible, as in the words نُظِّمَ *nuẓẓima* ‘to be organised’ and وُثِّقَ *wuṯṯiqa* ‘to be documented’. Instead of the passive form, the alternative syntactic construction تَمَّ *tamma* ‘performed/done’ + verbal noun is used, تَمَّ تَنظِيمُهُ *tamma tanẓiymuhu* ‘lit. organising it has been done / it was organised’, and تَمَّ تَوثِيقُهُ *tamma tawṯiyquhu* ‘lit. documenting it has been done / it was documented’.

To our knowledge, apart from Van Mol’s [11] study of the variations in complementary particles, no extensive empirical studies have been conducted to check how significant the difference between MSA and CA is either at the morphological, lexical, or syntactic levels.

## 1.2 A Brief 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 just record the high level formal language of the Koran, Prophet's sayings, poetry, and memorable pieces of literature and proverbs, but it also included a truthful account of common words and phrases as used by Bedouins and common people.

The other dictionaries that were compiled in the centuries following *al-'Ain* typically included either refinement, expansion, correction, or organisational improvements of the previous dictionaries. 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) [13].

Even relatively 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 mostly to preserve the language, refine older dictionaries, and accommodate accepted modern terminology. Some researchers criticise Arabic dictionaries for representing a fossilised version of the language with each new one reflecting the content of the preceding dictionaries [14]. Until today, to our knowledge, these remarks are still true.

Noteworthy work in bilingual Arabic lexicography was done by Arabists, most notable among them were Edward William Lane in the nineteenth century and Hans Wehr in the twentieth century. Edward William Lane's *Arabic–English Lexicon* (compiled between 1842 and 1876) was strongly indebted, as admitted by Lane himself [15], to previous Arabic monolingual dictionaries, chiefly the *Taj al-Arous* by Muhammad Murtada al-Zabidi (1732-1791). Lane spent seven years in Egypt acquiring materials for his dictionary and ultimately helped preserve the decaying and mutilated manuscripts he relied on [16].

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

The dictionary compilers, Wehr and Cowan, stated that their primary goal was to follow the descriptive and scientific principles by including only words and expressions that were attested in the corpus they collected [17]:

> "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 three main defects that, in practice, defeated the declared purpose of the dictionary. The first is in data collection, the second is in the use of secondary sources, and the third is in their approach to idiosyncratic classicisms. Data collection was conducted between 1940 and 1948, and the data included 45,000 slips containing citations from

Arabic sources. These sources consisted of selected works by poets, literary critics, and writers immersed in classical literature and renowned for their grandiloquent 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 specialised 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 second defect is that the dictionary compilers used some of the then available Arabic–French and Arabic–English dictionaries as "secondary sources". Items in the secondary sources for which there were no attestations in the primary sources, i.e. corpus data, were left to the judgement of an Arabic native speaker collaborator in such a way that words known to him, or already included in older dictionaries, were incorporated. The use of secondary sources in this way was a serious fault because of the subjectivity of decisions, and this was enough to damage the reliability of Wehr's dictionary as a true representation of the contemporary language.

The third drawback 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 the 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" [17]. 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 deserved from 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 [18] which was a pioneering systematic study in Arabic word count. Brill conducted a 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 two assistants of Brill.

In 1959 Jacob Landau 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. 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 selection of various 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 detail in collecting this well-balanced corpus which pre-dates the discipline of corpus linguistics and the first electronic corpus, the Brown Corpus [19]. Landau combined his work with Brill's work in a book called *A Word Count of Modern Arabic Prose* [20]. The outcome was the result of two word counts: Brill's count of the press usage, and Landau's count of literary usage. The former included close to 6,000 separate words; the latter over 11,000, and the combined list 12,400 words.

Through this frequency study, Landau was able to bring useful insights from the 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 1,000, 70%. He also found that 1,134 words occurred only once each in the press, and 3,905 words occurred only once in literature, which reflects the abundance of rare words in literary works.

An obvious weakness of this study, as admitted by the author himself, was that the number of running words counted (only 272,000 words) was inadequately small in comparison to the word count for other languages at the time such that for English (25,000,000), and German (11,000,000).

### 1.3 Current State of Arabic Lexicography

Until now, there is no large-scale lexicon (computational or otherwise) for MSA that is truly representative of the language. Al-Sulaiti [21] emphasises that most existing dictionaries are not corpus-based. Ghazali and Braham [14] point out the fact that traditional Arabic dictionaries are based on historical perspectives and that they tend to include obsolete words that are no longer in current use. They stress the need for new dictionaries based on an empirical approach that makes use of contextual analysis of modern language corpora.

The Buckwalter Arabic Morphological Analyzer (BAMA) [3] is widely used in the Arabic NLP research community. It is a *de facto* standard tool, and has been described as the "most respected lexical resource of its kind" [22]. It is designed as a main database of 40,648 lemmas supplemented by three morphological compatibility tables used for controlling affix-stem combinations. Other advantages of BAMA are that it provides information on the root, reconstructs vowel marks and provides an English glossary. The latest version of BAMA is renamed SAMA (Standard Arabic Morphological Analyzer) version 3.1 [6].

Unfortunately, there are some drawbacks in the SAMA lexical database that raise questions for it to be a truthful representation of MSA. We estimate that about 25% of the lexical items included in SAMA are outdated based on our data-driven filtering method explained in Section 3.2. SAMA suffers from a legacy of heavy reliance on older Arabic dictionaries, particularly Wehr's Dictionary [17], in the compilation of its lexical database. Therefore, there is a strong need to compile a lexicon for MSA that follows modern lexicographic conventions [23] in order to make the lexicon a reliable representation of the language.

There are only a few recorded attempts to break the stagnant water in Arabic lexicography. Van Mol in 2000 [24] developed an Arabic–Dutch learner's dictionary of 17,000 entries based on corpus data (3,000,000 words), which were used to derive information on contemporary usage, meanings and collocations. He considered his work as the first attempt to build a COBUILD-style dictionary [5]. More recently, Boudelaa and Marslen-Wilson in 2010 [25] built a lexical database for MSA (based on a corpus of 40 million words) which provided information on token and type frequencies for psycholinguistic purposes.

Our work represents a further step to address this critical gap in Arabic lexicography. We use a large corpus of one billion words to automatically create a lexical database for MSA. We follow the LMF naming conventions and hierarchical structures, and we provide complete information on inflection paradigms, root, patterns, humanness (for nouns), and transitivity (for verbs). This lexical database is interoperable with a finite-state morphological transducer.

### 1.4 Arabic Morphotactics

Arabic morphology is well-known for being rich and complex. The reason behind this complexity is the fact that it has a multi-tiered structure where words are originally derived from roots and pass through a series of affixations and clitic attachments until they finally appear as surface forms. Morphotactics refers to the way morphemes combine together to form words [26], [27]. Generally speaking, morphotactics can be concatenative, with morphemes either prefixed or suffixed to stems, or non-concatenative, with stems undergoing internal alterations to convey morpho-syntactic information [28]. Arabic is considered as a typical example of a language that employs both concatenative and non-concatenative morphotactics. For example, the verb إِستَعمَلُوهَا *istaꞛmaluw-hā* 'they-used-it' and the noun وَالاستَعمَالَات *wa-'l-istaꞛmālāt* 'and-the-uses' both come from the root عمل *ꞛml*.

Figure 1 shows the layers and tiers embedded in the Arabic morphological system. The derivation layer is non-concatenative and opaque in the sense that it is a sort of abstraction that affects the choice of a part of speech (POS), and it does not have a direct explicit surface manifestation. By contrast, the inflection layer is more transparent. It applies concatenative morphotactics by using affixes to express morpho-syntactic features. We note that verbs at this level show what is called 'separated dependencies' which means that some prefixes determine the selection of suffixes. From the analysis point of view, we note that stemming is conducted in the inflection layer and it can be done at two levels: either by stripping off tier 5 alone, producing إِستَعمَلُو *istaꞛmaluw* and إِستِعمَالَات *istiꞛmālāt* in our examples, or tier 5 along with tier 4, producing إِستَعمَل *istaꞛmal* and إِستِعمَال *istiꞛmāl*. It must be noted that automatic stemming that removes clitics and/or affixes will tend to produce forms that do not resemble actual words unless there is a way to ameliorate the effect of alterations.

According to our analysis of a small subset of data (from the Arabic Gigaword Corpus) containing 1,664,181 word tokens, we found that there are 125,282 unique types, or full form words from among the open classes: nouns, verbs, and adjectives. Stemming at the clitics level (tier 5) produces 42,145 unique stems, that is a reduction of 66% of the types. Lemmatisation produces 19,499 unique lemmas, that is a reduction of 54% of the stems, and a reduction of 84% of the types. This shows that lemmatisation is very effective in reducing data sparseness for Arabic.

In the derivational layer Arabic words are formed through the amalgamation of two tiers, namely root and pattern. A root is a sequence of three consonants and the pattern is a template of vowels with slots into which the consonants of the root are inserted. This process of insertion is called interdigitation [4]. An example is shown in table 1.

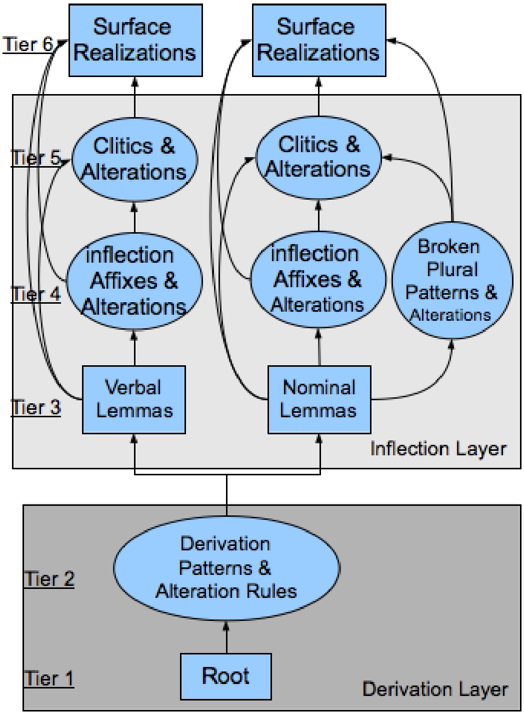


**Fig. 1.** Arabic morphology's multi-tier structure

## 2 Methodology

In this section, we explain the techniques and standards we followed in the construction of our lexical resource.

### 2.1 Using Heuristics and Statistics from a Large Corpus

For the construction of a lexicon for MSA we take advantage of large and rich resources that have not been exploited in similar tasks before. We use a corpus of 1,089,111,204 words, consisting of 925,461,707 words from the Arabic Gigaword corpus fourth edition [29] in addition to 163,649,497 words from news articles we collected from the Al-Jazeera web site.[3] One concern about this corpus is that it might not be as well-balanced as could be desired due to the fact that it is taken from only one domain, namely the news domain. However, to the best of our knowledge, this is the only large-scale corpus available for Arabic to date. Moreover, newspapers and websites tend to cover a variety of topics in addition to news. For example, the Al-Jazeera website covers, beside news, topics such as science, sports, art and culture, book reviews, economics, and health.

[3] `http://aljazeera.net/portal`. Collected in January 2010.

**Table 1.** Root and Pattern Interdigitation

| **Root** | درس<br>**drs** | | | |
|---|---|---|---|---|
| **Patterns** | $R_1aR_2aR_3a$ | $R_1aR_2R_2aR_3a$ | $R_1\bar{a}R_2iR_3$ | $muR_1aR_2R_2iR_3$ |
| **POS** | V | V | N | N |
| **Stem** | darasa | darrasa | dāris | mudarris |
| | ‘study’ | ‘teach’ | ‘student’ | ‘teacher’ |

We pre-annotate the corpus using MADA [30,31,32], a state-of-the-art tool for morphological processing. MADA combines SAMA and SVM classifiers to choose the best morphological analysis for a word in context, doing tokenisation, lemmatisation, diacritisation, POS tagging, and disambiguation. MADA is reported to achieve high accuracy (above 90%) for tokenisation and POS tagging tested on the Arabic Penn Treebank, but no evaluation of lemmatisation is reported. We use MADA and a data-driven filtering approach, described in Section 3.2, to identify core MSA lexical entries.

For the annotated data we collect statistics on lemma features and use machine learning techniques, also described in Section 3.2, in order to extend a manually constructed seed lexicon. We use machine learning to specifically predict new features that are not provided either by SAMA or MADA such as continuation classes, humanness, and transitivity.

### 2.2 Using State-of-the-Art Standards for Lexical Resource Representation

Over the past decade, there has been a growing tendency to standardise lexical resources by specifying the architecture and the component parts of the lexical model. There is also a need to specify how these components are interconnected and how the lexical resource as a whole exchanges information with other NLP applications. Lexical Markup Framework (LMF) [33,34] has emerged as an ISO standard that provides the specifications of the lexical database not for a particular language, but presumably any language.

LMF provides encoding formats, naming conventions, and a hierarchical structure of the components of lexical resources to ensure consistency. LMF was published officially as an international standard in 2008 and is now considered the state of the art in NLP lexical resource management. The purpose of LMF is to facilitate the exchange of lexical information between different lexical resources on the one hand and between lexical resources and NLP applications on the other. LMF takes into account the particular needs of languages with rich and complex morphology, such as Arabic. Figure 2 shows the Arabic root management in LMF and how verbs and nouns are linked through the common root.

### 2.3 Using Finite State Technology

One of our objectives for constructing the lexical resource is to build a morphological analyser and generator using bidirectional finite state technology (FST). FST has been used successfully in developing morphologies for many languages, including Semitic

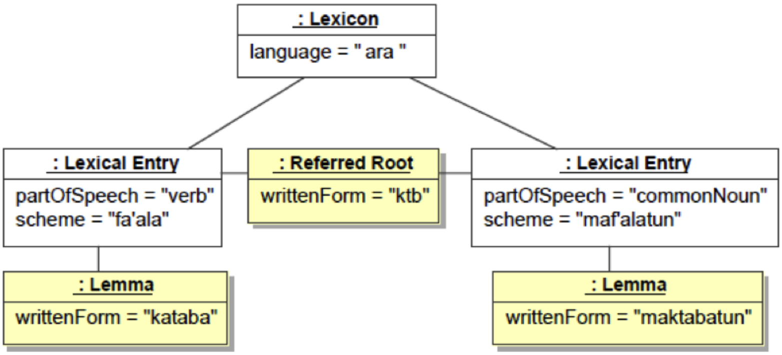


**Fig. 2.** LMF Arabic root management (adapted from [34])

languages [27]. There are a number of advantages of this technology that makes it especially attractive in dealing with human language morphologies; among these are the ability to handle concatenative and non-concatenative morphotactics, and the high speed and efficiency in handling large automata of lexicons with their derivations and inflections that can run into millions of paths.

## 3 Results to Date

In this section we present the results we obtained so far in building and extending the lexical database. We describe a web application, AraComLex, we built for maintaining and curating our lexical resource. We also outline our test case, namely an open-source FST morphological analyser which is based on our lexical database.

### 3.1 Building Lexical Resources

There are three key components in the Arabic morphological system: root, pattern, and lemma. In order to accommodate these components, we create four lexical databases: one for nominal lemmas (including nouns and adjectives), one for verb lemmas, one for word patterns, and one for root-lemma lookup. From a manually created MSA lexicon [2] we construct a seed database of 5,925 nominal lemmas and 1,529 verb lemmas. At the moment, we focus on open word classes and exclude proper nouns, function words, and multiword expressions which are relatively stable and fixed from an inflectional point of view.

We build a database of 380 Arabic patterns (346 for nominals and 34 for verbs) which can be used as indicators of the morphological inflectional and derivational behaviour of Arabic words. Patterns are also powerful in the abstraction and course-grained categorisation of word forms. In our lexicon, we account for 93.2% of all nominals using a set of 94 pre-defined patterns implemented as regular expressions.

We create a lemma-root look up database containing all lemmas and their roots. On the surface, nominals and verbs have different, unrelated inflection paradigms. But in

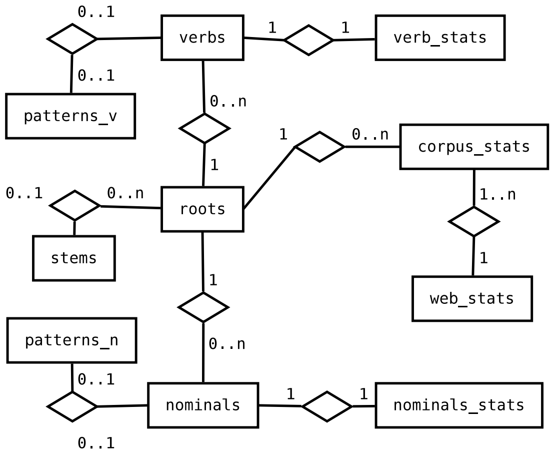


**Fig. 3.** Entity-relationship diagram of AraComLex

fact, they are closely interconnected through the common root. For example, if a root is capable of producing a transitive verb, it can also produce a passive participle.

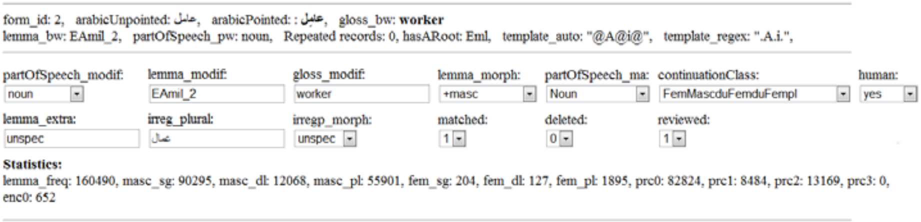
### 3.2 AraComLex Lexical Management Application

In order to manage our lexical database, we have developed the AraComLex lexicon authoring system[4] which provides a graphical user interface for human lexicographers to curate the automatically derived lexical and morphological information. We use AraComLex for storing the lexical resources mentioned in Section 3.1 as well as generating data for machine learning, storing extensions to the lexicon, and generating data for the morphological transducer, as explained in the following subsections. Figure 3 shows the entity-relationship diagram [35] of the database used in the AraComLex application. In this diagram, entities are drawn as rectangles and relationships as diamonds. Relationships connect pairs of entities with given cardinality constraints (represented as numbers surrounding the relationship). Three types of cardinality constraints are used in the diagram: 0 (entries in the entity are not required to take part in the relationship), 1 (each entry takes part in exactly one relationship) and *n* (entries can take part in an arbitrary number of relationships). Entities correspond to tables in the database, while relationships model the relations between the tables.

In AraComLex, we provide the key lexical information for bootstrapping and extending a morphological processing engine.

AraComLex covers the LMF's morphology extension by listing the relevant morphological and morpho-syntactic features for each lemma. We use finite sets of values implemented as drop-down menus to allow lexicographers to edit entries while ensuring consistency, as shown in figure 4. Two of the innovative features added are the "±human" feature and the 13 continuation classes which stand for the inflection grid, or all possible inflection paths, for nominals.

[4] `http://www.cngl.ie/aracomlex`



form_id: 2, arabicUnpointed: عامل, arabicPointed: : عامِل, gloss_bw: **worker**
lemma_bw: EAmil_2, partOfSpeech_pw: noun, Repeated records: 0, hasARoot: Eml, template_auto: "@A@i@", template_regex: ".A.i.",

| | | | | | | |
|---|---|---|---|---|---|---|
| partOfSpeech_modif: | lemma_modif: | gloss_modif: | lemma_morph: | partOfSpeech_ma: | continuationClass: | human: |
| noun | EAmil_2 | worker | +masc | Noun | FemMascduFemduFempl | yes |
| lemma_extra: | irreg_plural: | irregp_morph: | matched: | deleted: | reviewed: | |
| unspec | عمال | unspec | 1 | 0 | 1 | |

**Statistics:**
lemma_freq: 160490, masc_sg: 90295, masc_dl: 12068, masc_pl: 55901, fem_sg: 204, fem_dl: 127, fem_pl: 1895, prc0: 82824, prc1: 8484, prc2: 13169, prc3: 0, enc0: 652

**Fig. 4.** AraComLex Lexicon Authoring System for nominals with support statistics

Figure 4 shows the features specified for nominal lemmas in AraComLex. The feature "partOfSpeech" can be either 'noun', 'noun_prop', 'noun_quant', 'noun_num', 'adj', 'adj_comp', and 'adj_num'. The "lemma_morph" feature can be either 'masc' or 'fem' for nouns and can also be 'unspec' (unspecified) for adjectives. The "human" feature can be either 'yes', 'no', or 'unspec'. There are 13 continuation classes for Arabic nominals as shown in table 3 which represent the inflection grid (or all possible inflection paths) for nominals.

For verb lemmas we provide information on whether the verb is transitive or intransitive and whether it allows passive and imperative inflection.

For patterns we specify whether a pattern is nominal or verbal, whether it indicates a broken plural or singular form. Overall, we have 12 types of patterns: verbs, singular nouns, broken plural nouns, *masdar* (verbal noun), names of instruments, active participles, passive participles, *marrah* (instance), *mubalaghah* (exaggeration), comparative adjectives, *mushabbahah* (semi-adjectives), and names of places.

### 3.3 Extending the Lexical Database

In extending our lexicon, we rely on Attia's manually-constructed finite state morphology [2] and the lexical database in SAMA 3.1 [6]. Creating a lexicon is usually a labour-intensive task. For instance, Attia took three years in the development of his morphology, while SAMA and its predecessor, Buckwalter's morphology, were developed over more than a decade, and at least seven people were involved in updating and maintaining the morphology.

In this project we want to automatically extend Attia's finite state morphology using SAMA's database, but we need to solve two problems. First, SAMA suffers from a legacy of obsolete entries and we need to filter out these outdated words, as we want to enrich our lexicon only with lexical items that are still in current use. Second, our lexical database requires features (such as humanness for nouns and transitivity for verbs) that are not provided by SAMA, and we want to automatically induce these features.

**3.3.1 Lexical Enrichment.** To address the first problem we use a data-driven filtering method that combines open web search engines and our pre-annotated corpus. Using statistics[5] from three web search engines (Al-Jazeera,[6] Arabic Wikipedia,[7] and

[5] Statistics were collected in January 2011.

[6] `http://aljazeera.net/portal`

[7] `http://ar.wikipedia.org`

**Table 2.** Arabic Inflection Grid and Continuation Classes

| | **Masculine Singular** | **Feminine Singular** | **Masculine Dual** | **Feminine Dual** | **Masculine Plural** | **Feminine Plural** | **Continuation Class** |
|---|---|---|---|---|---|---|---|
| 1 | مُعَلِّم *muʿallim*<br>‘teacher’ | مُعَلِّمَة<br>*muʿallimat* | مُعَلِّمَان<br>*muʿallimān* | مُعَلِّمَتَان<br>*muʿallima-tān* | مُعَلِّمُون<br>*muʿallimuwn* | مُعَلِّمَات<br>*muʿallimāt* | Fem-Mascdu-Femdu-Mascpl-Fempl |
| 2 | طَالِب *ṭālib*<br>‘student’ | طَالِبَة *ṭālibat* | طَالِبَان *ṭālibān* | طَالِبَتَان<br>*ṭālibatān* | — | طَالِبَات *ṭālibāt* | Fem-Mascdu-Femdu-Fempl |
| 3 | تَعلِيمِيّ<br>*taʿliymiyy*<br>‘educational’ | تَعلِيمِيَّة<br>*taʿliymiyyat* | تَعلِيمِيَّان<br>*taʿliymiyyān* | تَعلِيمِيَّتَان<br>*taʿliymiyya-tān* | — | — | Fem-Mascdu-Femdu |
| 4 | — | بَقَرَة *baqarat*<br>‘cow’ | — | بَقَرَتَان<br>*baqaratān* | — | بَقَرَات *baqarāt* | Femdu-Fempl |
| 5 | تَنَازُل *tanāzul*<br>‘concession’ | — | — | — | — | تَنَازُلَات<br>*tanāzulāt* | Fempl |
| 6 | — | ضَحِيَّة *ḍaḥiyyat*<br>‘victim’ | — | ضَحِيَّتَان<br>*ḍaḥiyyatān* | — | — | Femdu |
| 7 | مَحض *maḥḍ*<br>‘mere’ | مَحضَة *maḥḍat* | — | — | — | — | Fem |
| 8 | اِمتِحَان *imtiḥān*<br>‘exam’ | — | اِمتِحَانَان<br>*imtiḥānān* | — | — | اِمتِحَانَات<br>*imtiḥānāt* | Mascdu-Femdu |
| 9 | طَيَّار *ṭayyār*<br>‘pilot’ | — | — | — | طَيَّارُون<br>*ṭayyāruwn* | — | Mascdu-Mascpl |
| 10 | كِتَاب *kitāb*<br>‘book’ | — | كِتَابَان *kitābān* | — | — | — | Mascdu |
| 11 | دِيمُقرَاطِيّ<br>*diymuqrāṭiyy*<br>‘democrat’ | — | — | — | دِيمُقرَاطِيُّون<br>*diymuqrā-ṭiyyuwn* | — | Mascpl |
| 12 | خُرُوج *ḫuruwǧ*<br>‘exiting’ | — | — | — | — | — | NoNum |
| 13 | مَبَاحِث *mabāḥiṯ*<br>‘investigators’ | — | — | — | — | — | Irreg_pl |

the Arabic BBC website[8]), we find that 7,095 lemmas in SAMA have zero hits, leaving only 33,553 as valid forms in MSA. Corpus statistics from our corpus, described in Section 2.1, show that 3,604 lemmas are not used in the corpus at all, leaving 37,044 lemmas that have at least one instance in the corpus. Of those, 4,471 lemmas occur less than 10 times, leaving 32,573 as more stable lemmas. Combining web statistics and corpus statistics, we find that there are 30,739 lemmas that returned at least one hit in the web queries and occurred at least once in the corpus. Only 29,627 lemmas are left if we consider lemmas that have at least one hit on the web data and that occurred at least 10 times in the corpus. Using a threshold of 10 occurrences here is discretionary, but the aim is to separate the stable core of the language from instances where the use of a word is perhaps accidental or idiosyncratic. We consider the refined list as representative of the lexicon of MSA as attested by our statistics.

[8] `http://www.bbc.co.uk/arabic/`

**Table 3.** Results of the classification experiments

| **Classes** | **Features** | **P** | **R** | **F** |
|---|---|---|---|---|
| **Nominals** | | | | |
| Continuation Classes (13 classes) | number, gender, case, clitics | 0.62 | 0.65 | 0.63 |
| Human (yes, no, unspec) | | 0.86 | 0.87 | 0.86 |
| POS (Noun, Adjective) | | 0.85 | 0.86 | 0.85 |
| **Verbs** | | | | |
| Transitivity (Transitive, Intransitive) | number, gender, person, aspect, mood, voice, clitics | 0.85 | 0.85 | 0.84 |
| Allow Passive (yes, no) | | 0.72 | 0.72 | 0.72 |
| Allow Imperative (yes, no) | | 0.63 | 0.65 | 0.64 |

We note that web statistics and corpus statistics are substantially different. Web searching does not allow diacritisation and does not have form disambiguation, while the corpus is automatically diacritised and disambiguated using MADA. Therefore, we believe that both statistics are complementary: web queries will show how likely the word form is in current use, and the corpus statistics will indicate how likely it is that a certain interpretation is valid.

It is problematic, if at all necessary, to manually create a gold standard that indicates whether a word belongs to MSA or CA. A human decision in this regard can be highly biased, subjective and dependent on each annotator's perception and background education. Therefore, we assume that the safest criteria for selecting MSA entries is to investigate whether or not a certain entry is attested in a large and representative modern corpus.

**3.3.2 Feature Enrichment.** To address the second problem, we use a machine learning classification algorithm, the Multilayer Perceptron [36,37], to build a model for predicting the required features for each new lemma. We use two manually annotated datasets of 4,816 nominals and 1,448 verbs. We feed these datasets with statistics from our pre-annotated corpus and build these statistics into a vector grid. The features that we use for nominals are number, gender, case and clitics; and for verbs: number, gender, person, aspect, mood, voice and clitics. For the implementation of the machine learning algorithm we use the open-source application Weka version 3.6.4.[9] We split each dataset into 66% for training and 34% for testing. We conduct six experiments to classify for six features that we need to include in our lexical database. For nominals we predict which continuation class (or inflection path) each nominal is likely to take. We predict the grammatical feature of humanness. We also classify nominals into nouns

[9] `http://www.cs.waikato.ac.nz/ml/weka/`

and adjectives. As for verbs, we classify them according to transitivity, and whether or not to allow the inflection for the passive voice and the imperative mood. Table 3 gives the results of the experiments in terms of precision, recall and f-measure.

The results show that the highest f-measure scores were achieved for 'Human', 'POS' and 'Transitivity'. Typically one would assume that these features are hard to predict with any reasonable accuracy without taking the context into account. So, it was surprising to get such good prediction results based only on statistics of morphological features. We could assume that the 'clitics' feature provides some clues about the context. However, removing the 'clitics' feature results in a drop from 0.86 to 0.83 in f-Measure for 'Human', which is not a big drop. This means that Arabic morphological features are powerful predictors of what the entry is likely to be with regards to unknown features. We also note that the f-measure for 'Continuation Classes' is comparatively low, but considering that here we are classifying for 13 features, we assume that the results are acceptable.

### 3.4 An Open-Source FST Arabic Morphological Analyser

The Xerox XFST System [27] is a well-known finite state compiler, but the disadvantage of this tool is that it requires a license to access full functionality, which limits its use in the larger research community.

Fortunately, there is an attractive alternative to the Xerox compiler, namely Foma [38], an open-source finite-state toolkit that implements the Xerox lexc and xfst utilities. Foma is largely compatible with the Xerox/PARC finite-state tools. It also embraces Unicode fully and supports a number of operating systems. We have developed an open-source morphological analyser for Arabic[10] using the Foma compiler allowing us to easily share and distribute our morphology to third parties. The database, which is being edited and validated using the AraComLex tool, is used to automatically extend and update the morphological analyser, allowing for greater coverage and better capabilities. In this section we explain our system design and report on evaluation results.

**3.4.1 System Design and Description.** There are three main strategies for the development of Arabic morphological analysers depending on the initial level of analysis: root, stem or lemma. In a root-based morphology, such as the Xerox Arabic Morphological Analyser [4], analysing Arabic words is based on a list of roots and a list of patterns interacting together in a process called interdigitation, as explained earlier. In a stem-based morphology, such as SAMA [3,6], the stem is considered as a base form of the word. A stem is a form between the lemma and the surface form. One lemma can have several variations when interacting with prefixes and suffixes. Such a system does not use alteration rules and relies instead on listing all stems (or form variations) in the database. For example, in SAMA's database, the verb شَكَرَ *šakara* 'to thank' has two entries: شَكَرَ *šakara* for perfective and شكُر *škur* for the imperfective. In a lemma-based morphology words are analysed at the lemma level. A lemma is the least marked form of a word, that is the uninflected word without suffixes, prefixes, proclitics or enclitics. In Arabic, this is usually the perfective, $3^{rd}$ person, singular verb, and in the case of

[10] http://sourceforge.net/projects/aracomlex/

nouns and adjectives, the singular indefinite form. In a lemma-based morphology there is only one entry for the verb شَكَرَ *šakara*, for example, that is the perfective form. The imperfective along with other inflected forms are generated from the lemma through alteration rules.

In our implementation we use the lemma as the base form. We believe that a lemma-based morphology is more economical than the stem-based morphology as it does not list all form variations and relies on generalised rules. It is also less complex than the root-based approach and less likely to overgenerate [1,2]. This leads to better maintainability and scalability of our morphology.

In a standard finite state system, lexical entries along with all possible affixes and clitics are encoded in the lexc language which is a right recursive phrase structure grammar [4,27]. A lexc file contains a number of lexicons connected through what is known as "continuation classes" which determine the path of concatenation. In example (1), the lexicon 'Proclitic' has a form 'wa' which has a continuation class 'Prefix'. This means that the forms in 'Prefix' will be appended to the right of 'wa'. The lexicon 'Proclitic' also has an empty string, which means that 'Proclitic' is optional and that the path can proceed without it. The bulk of lexical entries are listed under 'Root' in the example.

```
(1) LEXICON Proclitic
wa              Prefix;
                Prefix;
LEXICON Prefix
ya              Root;
LEXICON Root
shakara         Suffix;
kataba          Suffix;
LEXICON Suffix
una             Enclitic;
LEXICON Enclitic
ha              #;
```

With inflections and concatenations, words usually become subject to changes or alterations in their forms. Alterations are the discrepancies between underlying strings and their surface realisations [26], and alteration rules are the rules that relate the surface forms to the underlying forms. In Arabic, long vowels, glides and the glottal stop are the subject of a great deal of phonological (and consequently orthographical) alterations like assimilation and deletion. Many of the challenges an Arabic morphological analyser faces are related to handling these issues. In our system there are about 130 replace rules to handle alterations that affect verbs, nouns, adjectives and function words when they undergo inflections or are attached to affixes and clitics. Alteration rules are expressed in finite state systems using XFST replace rules of the general form shown in (2).

```
(2) a -> b || L _ R
```

The rule states that the string `a` is replaced with the string `b` when `a` occurs between the left context `L` and the right context `R`.

In our system, nouns are added by choosing from a template of continuation classes which determine what path of inflection each noun is going to select, as shown in example (3) (gloss is included in square brackets for illustration only).

```
(3) LEXICON Nouns
+masc+human^ss^ مُعَلِّم['teacher']^se^       FemMascduFemduFemplMascpl;
+masc+human^ss^طَالِب['student']^se^       FemMascduFemduFempl;
+masc+nonhuman^ss^كِتَاب['book']^se^    Mascdu;
+fem+nonhuman^ss^كُرَّاسَة['notebook']^se^ DualFempl;
```

These continuation class templates are based on the facts in table 2 above, which shows the inflection choices available for Arabic nouns according to gender (masculine or feminine) and number (singular, dual or plural).

As for verbs in our lexc file, the start and end of stems are marked to provide information needed in conducting alteration operations, as shown in example (4). The tags are meant to provide the following information:

- The multi-character symbol `^ss^` stands for stem start and `^se^` for stem end.
- The flag diacritic `@D.V.P@` means "disallow the passive voice", `@D.M.I@` means "disallow the imperative mood".
- `Transitive` and `Intransitive` are the continuation classes for verbs.

```
(4) LEXICON Verbs
^ss^شَكَرَ['thank']^se^             Transitive;
^ss^فَرِحَ['be-happy']^se^@D.V.P@ Intransitive;
^ss^أَمَرَ['order']^se^@D.M.I@      Transitive;
^ss^قَالَ['say']^se^                Intransitive;
```

**3.4.2 Morphology Evaluation.** In this section we test the coverage and rate of analyses per word in our morphological analyser compared to an earlier version (the baseline) and SAMA. We build a test corpus of 800,000 words, divided into 400,000 of what we term as Semi-Literary text and 400,000 for General News texts. The Semi-Literary texts consist of articles collected from columns, commentaries, opinions and analytical essays written by professional writers who tend to use figurative and metaphorical language not commonly used in ordinary news. This type of text exhibits the characteristics of literary text, especially the high ratio of word tokens to word types: out of the 400,000 tokens there are 60,564 types. The General News text contrasts with the literary text in that the former has a lower ratio of word tokens to word types: out of the 400,000 tokens there are 42,887 types. This observation is similar to the finding of Jacob Landau in his book *A Word Count of Modern Arabic Prose* [20] where he conducted a word count on 136,000 running words from Arabic prose and an equal portion from the daily press. The former resulted in over 11,000 unique words and the latter close to 6,000.

**Table 4.** Coverage and Rate per word test results

| Morphology | No. of | General News | | Semi-Literary | |
|---|---|---|---|---|---|
| | Lemmas | Coverage | Rate per word | Coverage | Rate per word |
| Baseline | 10,799 | 79.68% | 1.67 | 69.37% | 1.62 |
| AraComLex | 28,807 | 86.89% | 2.10 | 85.14% | 2.09 |
| SAMA | 40,648 | 88.13% | 5.32 | 86.95% | 5.3 |

Table 4 compares the coverage and rate per word results for AraComLex against the baseline, that is the morphology developed in [2], and LDC's SAMA, version 3.0.

The results show that for Semi-Literary texts we achieve a considerable improvement in coverage in AraComLex over the baseline; rising from 69.37% to 85.14%, that is 15.77% absolute improvement. Yet, for the General News texts, we achieve less improvement: from 80% to 87% coverage, that is 7% absolute improvement.

Compared to SAMA, AraComLex has 1.24% (absolute) less coverage on General News, and 1.81% (absolute) less coverage on the Semi-Literary texts. At the same time we notice that the average rate of analyses per word (ambiguity rate) is significantly lower in AraComLex (2.1) than in SAMA (5.3).

Testing shows that we achieve coverage results comparable to SAMA's morphology, and our ambiguity level (rate of analyses per word) is about 60% lower than SAMA. We assume that the lower rate of ambiguity in AraComLex is mainly due to the fact that we excluded obsolete words and morphological analyses from our lexical database.

## 4 Future Work

In extending our lexical database, we have been dependent mainly on SAMA. We filtered the SAMA lexicon through open web search engines and corpus data pre-annotated with MADA. In our corpus there are more than 700,000 types (or unique words) that are not recognised by SAMA. Now we need to devise a methodology to validate and include stable lemmas not included in SAMA's database. This will entail using a morphological guesser and applying heuristics. We will also include a large list of named entities [39] and multiword expressions [40] in a separate database that can be automatically embedded in our morphological analyser.

## 5 Conclusion

We build a lexicon for MSA that provides the information necessary for the construction of a morphological analyser, independent of design, approach, implementation strategy and programming language. We focus on the problem that existing lexical resources tend to include a subset of obsolete analyses and lexical entries, no longer attested in MSA. We start off with a manually constructed lexicon of 10,799 MSA lemmas and automatically extend it using lexical entries from SAMA's lexical database, carefully excluding obsolete entries and analyses. We use machine learning on statistics derived from a large pre-annotated corpus for automatically extending and complementing the SAMA-based lexical information, resulting in a lexicon of 28,807 lemmas for MSA. We follow the LMF standard for lexical representation, which aims at facilitating the interoperability and exchange of data between the lexicon and NLP applications. We develop a lexicon authoring system, AraComLex, to aid the manual revision of the lexical database by lexicographers. We use the database to automatically update and extend an open-source finite state morphological transducer. Evaluation results show that our transducer has coverage similar to SAMA, but at a significantly reduced average rate of analysis per word, due to avoiding outdated entries and analyses.

**Acknowledgments.** This research is funded by Enterprise Ireland (PC/09/037), the Irish Research Council for Science Engineering and Technology (IRCSET), and the EU projects PANACEA (7FP-ITC-248064) and META-NET (FP7-ICT-249119).

## References

1. Dichy, J., Ali, F.: 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 (2003)
2. Attia, M.: An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In: Challenges of Arabic for NLP/MT Conference. The British Computer Society, London (2006)
3. Buckwalter, T.: Buckwalter Arabic Morphological Analyzer (BAMA) Version 2.0. Linguistic Data Consortium (LDC) catalogue numberLDC2004L02,ISBN1-58563-324-0 (2004)
4. Beesley, K.R.: 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 (2001)
5. Sinclair, J.M. (ed.): Looking Up: An Account of the COBUILD Project in Lexical Computing. Collins, London (1987)
6. Maamouri, M., Graff, D., Bouziri, B., Krouna, S., Kulick, S.: LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.0. LDC Catalog No. LDC2010L01 (2010) ISBN: 1-58563-555-3
7. Bin-Muqbil, M.: Phonetic and Phonological Aspects of Arabic Emphatics and Gutturals. Ph.D. thesis in the University of Wisconsin, Madison (2006)
8. Watson, J.: The Phonology and Morphology of Arabic. Oxford University Press, New York (2002)
9. Elgibali, A., Badawi, E.M.: Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said M. Badawi. American University in Cairo Press, Egypt (1996)
10. Fischer, W.: Classical Arabic. In: The Semitic Languages. Routledge, London (1997)
11. Van Mol, M.: Variation in Modern Standard Arabic in Radio News Broadcasts, A Synchronic Descriptive Investigation in the use of complementary Particles. Leuven, OLA 117 (2003)
12. Stetkevych, J.: The modern Arabic literary language: lexical and stylistic developments. Publications of the Center for Middle Eastern Studies, vol. (6). University of Chicago Press, Chicago (1970)
13. Owens, J.: The Arabic Grammatical Tradition. In: The Semitic Languages. Routledge, London (1997)

14. Ghazali, S., Braham, A.: Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. In: Arabic NLP Workshop at ACL/EACL, Toulouse, France (2001)
15. Lane, E.W.: Preface. In: Arabic–English Lexicon. Williams and Norgate, London (1863)
16. Arberry, A.J.: Oriental essays: portraits of seven scholars. George Allen and Unwin, London (1960)
17. Wehr, H., Cowan, J.M.: Dictionary of Modern Written Arabic, pp. VII-XV. Spoken Language Services, Ithaca (1976)
18. Brill, M.: The Basic Word List of the Arabic Daily Newspaper. The Hebrew University Press Association, Jerusalem (1940)
19. Kuŏcera, H., Francis, W.N.: Computational Analysis of Present-Day American English. Brown University Press, Providence (1967)
20. Landau, J.M.: A Word Count of Modern Arabic Prose. American Council of Learned Societies, New York (1959)
21. Al-Sulaiti, L., Atwell, E.: The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics 11 (2006)
22. Hajič, J., Smrž, O., Buckwalter, T., Jin, H.: Feature-Based Tagger of Approximations of Functional Arabic Morphology. In: The 4th Workshop on Treebanks and Linguistic Theories (TLT 2005), Barcelona, Spain (2005)
23. Atkins, B.T.S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford (2008)
24. Van Mol, M.: The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In: Heid, U., Evert, S., Lehmann, E., Rohrer, C. (eds.) Proceedings of the Ninth EURALEX International Congress, Stuttgart, pp. 831–836 (2000)
25. Boudelaa, S., Marslen-Wilson, W.D.: Aralex: A lexical database for Modern Standard Arabic. Behavior Research Methods 42(2) (2010)
26. Beesley, K.R.: Arabic Morphological Analysis on the Internet. In: The 6th International Conference and Exhibition on Multilingual Computing, Cambridge, UK (1998)
27. Beesley, K.R., Karttunen, L.: Finite State Morphology: CSLI studies in computational linguistics. CSLI, Stanford (2003)
28. Kiraz, G.A.: Computational Nonlinear Morphology: With Emphasis on Semitic Languages. Cambridge University Press, Cambridge (2001)
29. Parker, R., Graff, D., Chen, K., Kong, J., Maeda, K.: Arabic Gigaword Fourth Edition. LDC Catalog No. LDC2009T30 (2009) ISBN: 1-58563-532-4
30. Habash, N., Rambow, O., Roth, R.: MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: The 2nd International Conference on Arabic Language Resources and Tools (MEDAR 2009), Cairo, Egypt, pp. 102–109 (2009)
31. Habash, N., Rambow, O.: Arabic Tokenization, Morphological Analysis, and Part- of-Speech Tagging in One Fell Swoop. In: Proceedings of the Conference of American Association for Computational Linguistics (ACL 2005). The University of Michigan, Ann Arbor (2005)
32. Roth, R., Rambow, O., Habash, N., Diab, M., Rudin, C.: Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In: Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio (2008)
33. Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., Soria, C.: Multilingual resources for NLP in the lexical markup framework (LMF). Language Resources and Evaluation (2008) ISSN 1574-020X
34. ISO 24613: Language Resource Management Lexical Markup Framework (draft version), ISO Switzerland (2007)
35. Chen, P.P.: The Entity-Relationship Model: Toward a Unified View of Data. ACM Transactions on Database Systems 1, 9–36 (1976)

36. Rosenblatt, F.: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC (1961)
37. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall, Englewood Cliffs (1998)
38. Hulden, M.: Foma: a finite-state compiler and library. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), Association for Computational Linguistics, Stroudsburg (2009)
39. Attia, M., Toral, A., Tounsi, L., Monachini, M., van Genabith, J.: An automatically built Named Entity lexicon for Arabic. In: LREC 2010, Valletta, Malta (2010)
40. Attia, M., Toral, A., Tounsi, L., Monachini, M.: van Genabith. Automatic Extraction of Arabic Multiword Expressions. In: COLING 2010 Workshop on Multiword Expressions: from Theory to Applications, Beijing, China (2010)