# Lexical Profiling for Arabic

**Mohammed Attia, Pavel Pecina, Lamia Tounsi, Antonio Toral and Josef van Genabith**
School of Computing
Dublin City University, Dublin, Ireland
E-mail: {mattia, ppecina, atoral, ltounsi, josef}@computing.dcu.ie

**Abstract**

We provide lexical profiling for Arabic by covering two important linguistic aspects of Arabic lexical information, namely morphological inflectional paradigms and syntactic subcategorization frames, making our database a rich repository of Arabic lexicographic details. First, we provide a complete description of the inflectional behaviour of Arabic lemmas based on statistical distribution. We use a corpus of 1,089,111,204 words, a pre-annotation tool, knowledge-based rules, and machine learning techniques to automatically acquire lexical knowledge about words' morpho-syntactic attributes and inflection possibilities. Second, we automatically extract the Arabic subcategorization frames (or predicate-argument structures) from the Penn Arabic Treebank (ATB) for a large number of Arabic lemmas, including verbs, nouns and adjectives. We compare the results against a manually constructed collection of subcategorization frames designed for an Arabic LFG parser. The comparison results show that we achieve high precision scores for the three word classes. Both morphological and syntactic specifications are combined and connected in a scalable and interoperable lexical database suitable for constructing a morphological analyser, aiding a syntactic parser, or even building an Arabic dictionary. We build a web application, AraComLex (Arabic Computer Lexicon), available at: http://www.cngl.ie/aracomlex, for managing and maintaining the standardized and scalable lexical database.

**Keywords**: Arabic; subcategorization frames; morphological analysis; morphological paradigms

## 1. Introduction

In a typical dictionary entry of a word, it is expected to find basic information pertaining to the word's morphology (possible inflections) and syntax (part of speech, whether it is transitive or intransitive, in the case of verbs, and what prepositions it can co-occur with). Yet, existing Arabic dictionaries have several limitations. Most of them do not rely on a corpus for attesting the validity of their entries (as in a COBUILD approach (Sinclair, 1987)), but they typically include either refinements, expansions, corrections, or organisational improvements over the previous dictionaries. Therefore, they tend to include obsolete words not in contemporary use. Furthermore, they often do not explicitly state all the possible inflection paradigms, and they do not provide sufficient syntactic information on word's obligatory combinations (or argument list).

The aim here is to attempt to resolve these shortcomings by automatically providing a complete description of the inflectional and syntactic behaviour of Arabic lexical entries based on statistical distribution in treebanks and un-annotated corpora. The work described in this paper is divided into two major parts. The first is focused on examining the statistical distribution of inflection paradigms for lexical entries in a large corpus pre-annotated with MADA (Roth et al., 2008), a tool which performs morphological analysis and disambiguation using the Buckwalter morphological analyser (Buckwalter, 2004) and machine learning. The second is related to the automatic extraction of syntactic information, or subcategorization frames, from the Arabic Treebank (ATB) (Maamouri and Bies, 2004).

To the best of our knowledge, this is the first attempt at extracting subcategorization frames from the ATB. The subcategorization requirements of lexical entries are important type lexical information, as they indicate the argument(s) a predicate needs in order to form a well-formed syntactic structure. Yet producing such resources by hand is costly and time consuming. Moreover, as Manning (1993) indicates, dictionaries produced by hand will tend to lag behind real language use because of their static nature. Therefore a complete, or at least complementary, automatic process is highly desirable.

This paper is structured as follows. In the introduction we describe the motivation behind our work. We differentiate between Modern Standard Arabic (MSA), the focus of this research, and Classical Arabic (CA) which is a historical version of the language. We briefly explain the current state 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 present the results obtained to date in building and extending the lexical database using a data-driven filtering method and machine learning techniques. We also explain how we use knowledge-based pattern matching in detecting and extracting broken plural forms. In Section 3, we explain the method we followed in extracting and evaluating the subcategorization frames for Arabic verbs, nouns and adjectives. In Section 4, we describe AraComLex, a web application we built for curating and combining our lexical resources. 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 originated 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 (Bin-Muqbil, 2006).

MSA is different from CA at the lexical, morphological, and syntactic levels (Watson, 2002; Elgibali and Badawi, 1996; Fischer, 1997). 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 while many words from CA have become obsolete. 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 (2003) who pointed out 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.

## 1.2 The 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 (2006) emphasises that existing dictionaries are not corpus-based. Ghazali and Braham (2001) stress the need for new dictionaries based on an empirical approach that makes use of contextual analysis of modern language corpora. They 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. The inclusion of these rarities inevitably affects the representativeness of dictionaries and marks a significant bias towards historical or literary forms. In recent years, some advances have been made (Van Mol, 2000; Boudelaa and Marslen-Wilson, 2010), but they are not enough in terms of size or the breadth of linguistic description.

The Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) is widely used by 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" (Hajič et al., 2005). 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 (Maamouri et al., 2010).

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 2.2.1. SAMA suffers from a legacy of heavy reliance on older Arabic dictionaries, particularly Wehr's Dictionary (Wehr Cowan, 1976), in the compilation of its lexical database.

Therefore, there is a strong need to compile a lexicon for MSA that follows modern lexicographic conventions (Atkins and Rundell, 2008) in order to make the lexicon a reliable representation of the language and to make it a useful resource for NLP applications dealing with MSA. Our work represents a further step to address this critical gap in Arabic lexicography. We use a large corpus of more than one billion words to automatically create a lexical database for MSA. We enrich the lexicon with syntactic information by extracting subcategorization frames and significant preposition collocates from the ATB.

## 1.3 Arabic Morphological System

Arabic morphology is well-known for being rich and complex. Arabic morphology 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 (Beesley and Karttunen, 2003). 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 (Kiraz, 2001). Arabic is considered as a typical example of a language that employs both concatenative and non-concatenative morphotactics. For example, the verb استعملوها {istaEomaluwha[1] 'they-used-it' and the noun والاستعمالات wAl{istiEomAlAt 'and-the-uses' both originate from the root عمل Eml.

Figure 1 shows the layers and tiers embedded in the representation of 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

[1] All examples are written in Buckwalter Transliteration.

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.

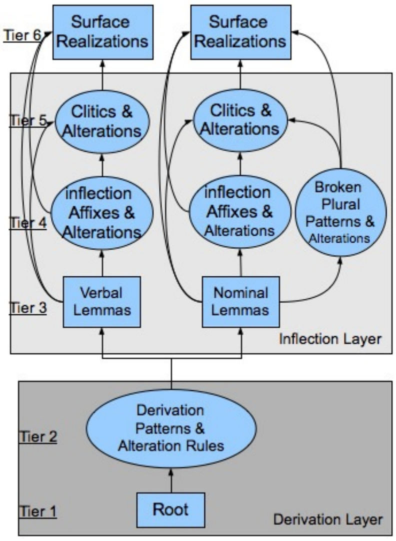


Figure 1: Arabic Morphology's Multi-tier Structure

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 (Beesley and Karttunen, 2003). An example is shown in Table 1.

| Root | درس<br>drs | | | |
|---|---|---|---|---|
| POS | V | V | N | N |
| Pattern | R$_1$aR$_2$aR$_3$a | R$_1$aR$_2$R$_2$aR$_3$a | R$_1$AR$_2$iR$_3$ | muR$_1$aR$_2$~iR$_3$ |
| Stem | darasa<br>'study' | darrasa<br>'teach' | dAris<br>'student' | mudar~is<br>'teacher' |

Table 1. Root and Pattern Interdigitation

## 2. Extending the Existing Lexicon

In this section, we describe the small-scale, manually-constructed lexical resources that we had, and how we managed to significantly extend these resources. We explain how we filter out obsolete words, how we use machine learning to acquire knowledge on morphological paradigms (or continuation classes) for new entries, and how we extract broken plural forms from our corpus. The corpus we use contains 1,089,111,204 words, consisting of 925,461,707 words from the Arabic Gigaword (Parker et al., 2009), in addition to 163,649,497 words from news articles we collected from the Al-Jazeera web site.[2]

### 2.1 Existing Lexical Resources

There are three key components in the Arabic morphological system: root, pattern and lemma. For accommodating these components, we acquire three lexical databases: one for lemmas, one for word patterns, and one for lemma-root lookup. The lemma database is collected from Attia (2006) which was developed manually. It includes 5,925 nominal lemmas (nouns and adjectives) and 1,529 verb lemmas. The advantage of the lemma entries in this resource is that they are fully specified with necessary morpho-syntactic information. In addition to the usual specification of gender, number and person, it provides information on continuation classes for nominals (as shown in Table 2), whether the noun indicates a human or non-human entity. For verbs it gives details on the transitivity, whether the passive voice is allowed or not, and whether the imperative mood is allowed or not.

We automatically create the lemma-root lookup database relying on the SAMA database. We manually developed a database for Arabic patterns that includes 490 patterns (456 for nominals and 34 for verbs). These patterns can be used as indicators of the morphological inflectional and derivational behaviour of Arabic words. Patterns are also powerful in the abstraction and coarse-grained categorisation of word forms.

### 2.2 Extending the Lexical Database

In extending our lexicon, we rely on Attia's manually-constructed lexicon (Attia, 2006) and the lexical database in SAMA 3.1 (Maamouri et al., 2010). 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, BAMA, were developed over more than a decade, and at least seven people were involved in updating and maintaining the morphology.

Our objective here is to automatically extend Attia's lexicon (Attia, 2006) using SAMA's database. In order to do this, 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 the lexicon only with lexical items that are still in current use. Second, Attia's lexicon 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.

#### 2.2.1 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 frequency statistics[3] on lemmas from three web search engines (Al-Jazeera,[4] Arabic Wikipedia,[5] and the Arabic BBC website[6]), we find that 7,095 lemmas in SAMA have zero hits.

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

[3] Statistics were collected in January 2011.

[4] http://aljazeera.net/portal

[5] http://ar.wikipedia.org

[6] http://www.bbc.co.uk/arabic/

| | Masculine Singular | Feminine Singular | Masculine Dual | Feminine Dual | Masculine Plural | Feminine Plural | Continuation Class |
|---|---|---|---|---|---|---|---|
| 1 | معلم muEal~im, ‘teacher’ | معلمة muEal~imap | معلمان muEal~imAn | معلمتان muEal~imatAn | معلمون muEal~imuwn | معلمات muEal~imAt | F-Mdu-Fdu-Mpl-Fpl |
| 2 | طالب TAlib, ‘student’ | طالبة TAlibap | طالبان TAlibAn | طالبتان TAlibatAn | - | طالبات TAlibAt | F-Mdu-Fdu-Fpl |
| 3 | تحضيري, taHoDiyriy~, ‘preparatory’ | تحضيرية taHoDiyriy~ap | تحضيريان taHoDiyriy~An | تحضيريتان taHoDiyriy~atAn | - | - | F-Mdu-Fdu |
| 4 | - | بقرة baqarap ‘cow’ | - | بقرتان baqaratAn | - | بقرات baqarAt | Fdu-Fpl |
| 5 | تنازل, tanAzul ‘concession’ | - | - | - | - | تنازلات tanAzulAt | Fpl |
| 6 | - | ضحية DaHiy~ap ‘victim’ | - | ضحيتان DaHiy~atAn | - | - | Fdu |
| 7 | محض maHoD ‘mere’ | محضة maHoDap | - | - | - | - | F |
| 8 | امتحان {imotiHAn, ‘exam’ | - | امتحانان {imotiHAnAn | - | - | امتحانات {imotiHAnAt | Mdu-Fdu |
| 9 | طيار Tay~Ar, ‘pilot’ | - | طياران Tay~ArAn | - | طيارون Tay~Aruwn | - | Mdu-Mpl |
| 10 | كتاب kitAb, ‘book’ | - | كتابان kitAbAn | - | - | - | Mdu |
| 11 | ديمقراطي diymuqoratiy~, ‘democrat’ | - | - | - | ديمقراطيون diy-muqoratiy~uwn | - | Mpl |
| 12 | خروج xuruwj, ‘exiting’ | - | - | - | - | - | NoNum |
| 13 | مباحث mabAHiv, ‘investigators’ | - | - | - | - | - | Irreg_pl |

Table 2: Arabic Continuation Classes based on the inflection grid

Frequency statistics from our corpus show that 3,604 lemmas are not used in the corpus at all, and 4,471 lemmas occur less than 10 times. Combining frequency statistics from the web and the corpus, we find that there are 29,627 lemmas that returned at least one hit in the web queries and 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 somewhat idiosyncratic. We consider the refined list as representative of the lexicon of MSA as attested by our statistics.

| No. | Classes | Features | P | R | F |
|---|---|---|---|---|---|
| | Nominals | | | | |
| 1 | Continuation Classes: 13 classes | number, gender, case, clitics | 0.62 | 0.65 | 0.63 |
| 2 | Human: yes, no, unspecified | | 0.86 | 0.87 | 0.86 |
| 3 | POS: noun, adjective | | 0.85 | 0.86 | 0.85 |
| | Verbs | | | | |
| 4 | Transitivity: transitive, intransitive | number, gender, person, aspect, mood, voice, clitics | 0.85 | 0.85 | 0.84 |
| 5 | Allow passive: yes, no | | 0.72 | 0.72 | 0.72 |
| 6 | Allow imperative: yes, no | | 0.63 | 0.65 | 0.64 |

Table 3: Results of the Classification Experiments.

### 2.2.2 Feature Enrichment

To address the second problem, we use a machine learning classification algorithm, the Multilayer Perceptron (Haykin, 1998). The main idea of machine learning is to automatically learn complex patterns from existing (training) data and make intelligent decisions on new (test) data.

In our case, we have a seed lexicon (Attia, 2006) with lemmas manually annotated with classes, and we want to build a model for predicting the same classes for each new lemma added to the lexicon. The classes (second column in Table 3) for nominals are continuation classes (or inflection paths), the semantico-grammatical feature of humanness, and POS (noun or adjective). The classes for verbs are transitivity, allowing the passive voice, and allowing the imperative mood. From our seed lexicon we extract two datasets of 4,816 nominals and 1,448 verbs. We feed these datasets with frequency statistics from our pre-annotated corpus and build the statistics into a vector grid. The features (third column in Table 3) for nominals are number, gender, case and clitics; 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 3.6.4.7[7]. We split each dataset into 66% for training and 34% for testing. We conduct six classification experiments to provide the classes that we need to include in our lexical database. 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 (above 80%) are achieved for ‘Human’, ‘POS’, and ‘Transitivity’. Typically one would assume that these features are hard to predict with any reasonable

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

accuracy without taking the context into account. It was surprising to obtain such good prediction results based on statistics on morphological features alone. We also note that the f-measure for 'Continuation Classes' is comparatively low, but considering that here we are classifying for 13 classes, the results are in fact quite acceptable. Using the machine learning model, we annotate 12,974 new nominals and 5,034 verbs.

### 2.3 Handling Broken Plurals

Broken plurals are an interesting phenomenon in Arabic where the plural is formed not through regular suffixation, but by changing the word pattern. In our seed lexicon (Attia, 2006), we have 950 broken plurals which were collected manually and clearly tagged. In SAMA, however, broken plurals are rather poorly handled. SAMA does not mark broken plurals as "plurals" either in the source file or in the morphology output. There is no straightforward way to automatically collect the list of all broken plural forms from SAMA. For example, the singular form جانب jAanib "side" and the broken plural jawAnib "sides" are analysed as in (1) and (2) respectively.

```
(1) <lemmaID>jAnib_1</lemmaID>
<voc>jAnib</voc> <pos>jAnib/NOUN</pos>
<gloss>side/aspect</gloss>
(2) <lemmaID>jAnib_1</lemmaID>
<voc>jawAnib</voc> <pos>jawAnib/NOUN</pos>
<gloss>sides/aspects</gloss>
```

The only tags that distinguish the singular from the broken plural form are the gloss (or translation) and voc (or vocalisation). We also note that MADA passes this problem on unsolved, and broken plurals are all marked num=s, meaning that the number is singular. We believe that this shortcoming can have a detrimental effect on the performance of any syntactic parser based on such data.

To extract broken plurals from our large MSA corpus (which is annotated with SAMA tags), we rely on the gloss of entries with the same LemmaID. We use Levenshtein Distance which measures the similarity between two strings. For example, using Levenshtein Distance to measure the difference between "sides/aspects" and "side/aspect" will give a distance of 2. When this number is divided by the length of the first string, we obtain 0.15, which is within a threshold (here set to 0.4). Thus the two entries pass the test as possible broken plural candidates. Using this method, we collect 2,266 candidates. We believe, however, that many broken plural forms went undetected because the translation did not follow the assumed format. For example, the word حرب harb has the translation "war/warfare" while the plural form huruwb has the translation "wars".

To validate the list of candidates, we use Arabic word pattern matching. For instance, in the jAnib example above, the singular form (vocalisation) follows the pattern fAEil (or the regular expression .A.il) and the plural form follows the pattern fawAEil (or .awA.i.). In our manually developed pattern database we have fawAEil as a possible plural pattern for fAEil. Therefore, the matching succeeds, and the candidate is considered as a valid broken plural entry. We compiled a list of 135 singular patterns that choose from a set of 82 broken plural patterns. The choice, however, is not free, but each singular form has a limited predefined set of broken plural patterns to select from. From the list of 2,266 candidates produced by Levenshtein Distance, 1,965 were validated using the pattern matching, that is 87% of the instances. When we remove the entries that are intersected with our 950 manually collected broken plurals, 1,780 forms are left. This means that in our lexicon now we have a total of 2,730 broken plural forms.

There are some insights that can be gained from the statistics on Arabic plurals in our corpus. The corpus contains 5,570 lemmas which have a feminine plural suffix, 1,942 lemmas with a masculine plural suffix (out of these 1,273 forms intersect with the feminine plural suffix), and about 1,965 lemmas with a broken plural form. This means that the broken plural formation in Arabic is as productive as the regular plural suffixation. Currently, we cannot explain why the feminine plural suffix enjoys this high preference, but we can point to the fact that masculine plural suffixes are used almost exclusively with the natural gender, while the feminine plural suffix, as well as broken plurals, are used liberally with the grammatical gender in addition to the natural gender.

## 3. Automatic Extraction of Subcategorization Frames

The encoding of syntactic subcategorization frames is an essential requirement in the construction of computational and paper lexicons alike. In English, the construction and extraction of subcategorization frames received a lot of attention, one example is the specialized lexicon COMLEX (Grishman et al., 1994) which is an extensive computational lexicon containing syntactic information for approximately 38,000 English headwords, with detailed information on subcategorization, containing 138 distinct verb frames for 5,662 active verbs lemmas.

For Arabic, the attention has been directed, almost exclusively, to the construction and automatic extraction of semantic roles (Palmer et al., 2008; Attia et al. 2008). Semantic roles are related to syntactic functions and surface phrase structures, but the three are at totally different and distinct layers of analysis. Grammatical functions are in the intermediary position between phrase structures and semantic roles. It is a major concept in semantic role labelling to make greater level of generalization. There is an emphasis on that the semantic labels do not vary in different syntactic constructions (Palmer et al., 2008). For example, the Arabic verb لاحظ lAHaZ "noticed" has two subcategorization frames: <subj,obj> for لاحظ الفرق lAHaZa Al-faroq "He noticed the difference" and <subj,comp> for لاحظ أن المحصول ينقص lAHaZa >an~a Al-maHoSuwla yanoquS "He noticed that the crop is decreasing" Yet, in the Arabic Propbank annotation[8] both frames have the same roleset:

Arg0: observer
Arg1: thing noticed or observed

To our knowledge, the only resource that currently exists for Arabic subcategorization frames is the lexicon manually developed for the Arabic LFG Parser (Attia, 2008). It is published as an open-source resource under the GPLv3 license[9]. It contains 64 frame types, 2,709 lemmas types, and 2,901 lemma-frame types, averaging 1.07 frames per lemma. The resource incorporates control information and details of specific prepositions with obliques. We use this resource in the evaluation of our automatically induced lexicon of Arabic subcategorization frames.

## 3.1 LFG Subcategorization Frames

The LFG syntactic theory (Dalrymple, 2001) distinguishes between governable (subcategorizable) and non- governable (non-subcategorizable) grammatical functions. The governable grammatical functions are the arguments required by some predicates in order to produce a well-formed syntactic structure, and they are SUBJ(ect), OBJ(ect), $\mathrm{OBJ}_{\Theta}$, OBL(ique) $_{\Theta}$, COMP(lement) and XCOMP. The non-governable grammatical functions are not required in the sentence to form a well-formed structure, and they are ADJ(junct) and XADJ. The subcategorization requirements in LFG are expressed in the following format (O'Donovan et al., 2005):

$$\pi\langle gf_1, gf_2, \ldots gf_n\rangle$$

where π is the lemma (predicate or semantic form) and *gf* is a governable grammatical function. The value of the argument list of the semantic form ensures the well-formedness of the sentence. For example, in the sentence اعتمد الطفل على والدته {iEotamada Al-Tifolu EalaY wAlidati-hi "The child relied on his mother", the verb {iEotamada "to rely" has the following argument structure {iEotamada<(↑SUBJ)( ↑$\mathrm{OBL}_{\text{>alaY}}$)>. By including a subject and an oblique with the preposition >alaY, we ensure that the verb's subcategorization requirements are met and that the sentence is well-formed, or syntactically valid.

## 3.2 Extracting Subcategorization frames from the Arabic Treebank

We follow here the successful model by the previous language resource extraction efforts for other languages including English (O'Donovan et al., 2005) and German (Rehbein and van Genabith, 2009) taking into consideration the specifics of the Arabic language and the resources available for evaluation. We automatically extract the Arabic syntactic-function based subcategorization frames by utilizing an automatic Lexical-Functional Grammar (LFG) f-structure annotation algorithm for Arabic developed in (Tounsi et al., 2009). The syntactic annotations in the ATB provides explicit information on deep representation in the phrase structure such as dealing with traces in the case of pro-dropped arguments which helped the automatic extraction of subcategorization frames to be complete. After we extract the surface forms we lemmatize all forms by re-analysing all the words using the Buckwalter morphology and then choosing the analysis where the word diacrization and the tag set in the ATB match those in the Buckwalter analysis.

We provide information on the prepositions for obliques, distinguish between active and passive frames, and provide information on the probability score for each frame and the frequency count for each lemma. We extract 240 frame types for 3,295 lemmas types, with 7,746 lemma-frame types (for verbs, nouns and adjectives), averaging 2.35 frames per lemma. We make this resource available under the open-source license GPLv3 [10]. Table 5 shows the list of grammatical functions included in our frames with examples. We compare and evaluate the complete set of subcategorization frames extracted against the manually developed subcategorization frames in the Arabic LFG Parser.

Our extraction algorithm deals with the passive voice and its effect on subcategorization behaviour. We find that in Arabic the passive forms stand at 12% of the active forms compared to 31% in English (O'Donovan et al., 2005), as shown in Table 4. Our explanation of the low frequency of the use of passive in Arabic is that there is a tendency to avoid passive verb forms when the active readings are also possible in order to avoid ambiguity and improve readability. For example, the verb form نظم nZm "organize" can have two readings, one for active and one for passive depending on diacritization, or how the word is pronounced. Therefore, instead of the ambiguous passive form, the alternative syntactic construction تم tam~a "performed/done" + verbal noun is used, giving تم تنظيمه tam~a tanZiymuhu "lit. organizing it has been done / it was organized". One evidence for the validity of our explanation is that the verb tam~a is the seventh most frequent verb in the ATB following كان kAn "be", قال qAla "say", أعلن >aEolana "declare", أكد >ak~ada "confirm", أضاف >aDAfa "add" and اعتبر {iEotabar "consider".

[8] http://verbs.colorado.edu/propbank/framesets-arabic

[9] http://arasubcats-lfg.sourceforge.net

[10] http://arabicsubcats.sourceforge.net

| | Active | Passive | Passive % |
|---|---|---|---|
| Arabic verb frames | 5,915 | 681 | 12 |
| English verb frames | 16,000 | 5,005 | 31 |

Table 4: Comparing active and passive subcategorization frames in Arabic and English

| | | Treebank Tag | Source | Meaning | Example |
|---|---|---|---|---|---|
| 1 | subj | -SBJ | L-T | subject | جاء الوقت jaA'a Al-waqt<br>lit. came the time, "The time came" |
| 2 | obj | -OBJ | L-T | object | عرفت الطريق Earaftu Al-Tariyq<br>"I knew the way" |
| 3 | obj2 | -DTV/-BNF | L-T | secondary object | أعطاه طعاما >aEoTA-hu TaEAmA<br>"gave him food" |
| 4 | obl | -CLR | L-T | oblique | اعتمد على والدته {iEotamad EalaY wAlidi-hi<br>"relied on his father" |
| 5 | obl2 | | L | secondary oblique | تنافس معه في السباق<br>tanAfasa maEa-hu fi Al-sibAq<br>"competed with him in the race" |
| 6 | obl-betweenAnd | | L | oblique for between … and | تنقل بين العراق والكويت<br>tanaq~ala bayona Al-EirAq wa-Al-kwiyt<br>"moved between Iraq and Kuwait" |
| 7 | obl-fromTo | | L | oblique for from … to | سافر من العراق إلى الكويت<br>sAfara min Al-EirAq <ilA Al-kwiyt<br>"travelled from Iraq to Kuwait" |
| 8 | obl-dir | -DIR | T | oblique for direction | شحنها إلى جدة $aHana-hA <ila jad~ap<br>"shipped it to Jeddah" |
| 9 | compL | | L | light complementizer >an | أمكنه أن يراها >amkana-hu >an yarAhaA<br>"became possible for him to see it" |
| 10 | compH | | L | heavy complementizer >an~a | أذاع أنهم هربوا >a*aEa >an~a-hum harabuwA<br>"announced that they escaped" |
| 11 | vcomp | | L | verb complement | بدأ يسقط bada>a yasoquT<br>lit. started fall, "started to fall" |
| 12 | xcomp | | L | obligatory control | أراد أن يسافر >arAda >an yusAfir<br>"wanted to travel" |
| 13 | xcomp-pred | -PRD | T | copular complement | كان مريضا kAna mariDA<br>"was sick" |
| 14 | xcomp-verb | (VP) | T | verb complement | same as 11 |
| 15 | comp-sbar | (SBAR) | T | complement with complementizer | same as 9 and 10 |
| 16 | comp-nom | (S-NOM) | T | gerund (masdar) complement | نفى علمه بالواقعة<br>nafaY Eilma-hu bi-Al-wAqiEap<br>"denied knowing the incident" |
| 17 | comp-s | (S) | T | sentential complement | قال لابد من التشاور<br>qAla lAbud~a min Al-ta$Awur<br>"he said there must be negotiations" |

L: LFG Parser, T: Treebank
Table 5: List of Arabic subcategorization frames suffixed with phrase structure information

### 3.3 Estimating the Subcategorization Probability

In order to estimate the likelihood of the occurrence of a certain argument list with a predicate (or lemma), we compute the conditional probability of subcategorization frames based on the number of token occurrences in the ATB, according to the following formula (O'Donovan et al., 2005);

$$P(ArgList \mid \Pi) = \frac{count(\Pi\langle ArgList\rangle)}{\sum_{i=1}^{n} count(\Pi\langle ArgList_i\rangle)}$$

where ArgList$_1$ … ArgList$_n$ are all the possible argument lists that co-occur with Π. Because of the variations in verbal subcategorization, probabilities are useful for discriminating prominent frames from accidental ones. An example is shown in table 6 for the verb شاهد $Ahada "watch" which has a frequency of 40 occurrences in the ATB.

| Lemma with argument list | Conditional Probability |
|---|---|
| $Ahad_1([subj,obj,comp-s]) | 0.0250 |
| $Ahad_1([subj,obj,comp-sbar]) | 0.0500 |
| $Ahad_1([subj,passive]) | 0.1000 |
| $Ahad_1([subj,obj]) | 0.8000 |
| $Ahad_1([subj]) | 0.0250 |

Table 6: Subcategorization frames with probabilities.

### 3.4 Evaluating the Subcategorization Frames

We compare our resource on subcategorization frames against a manually created subcategorization frames lexicon used in a rule-based LFG Parser. The Arabic LFG Parser has detailed subcategorisation information for lexical entries that includes the preposition of obliques, control relationships (or XCOMPs), and the type of complementizer in verbs that have complements. The number of subcategorization frames collected in the ATB induced resource is comparable to the manually constructed frames in the Arabic LFG parser for nouns and adjectives, but it is almost four times larger for verbs, as shown in Table 7. Figure 2 compares the size of the two resources in proportional intersecting circles. The circles on the left represent the treebank-induced resource, and the circles on the right represent the manually constructed resource.

| | Verbs | Nouns | Adjectives |
|---|---|---|---|
| lemma-subcat pairs in ATB | 6596 | 855 | 295 |
| lemma-subcat pairs in the LFG Parser | 1621 | 991 | 289 |
| Common lemmas | 1447 | 268 | 70 |

Table 7: Number of subcat frames in the ATB and the Arabic LFG Parser

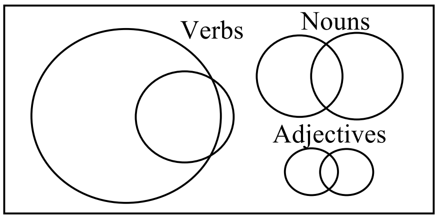


Figure 2: Intersecting circles of ATB subcategorization frames (left) and the LFG Parser (right)

We compare the subcategorization frames in terms of precision, defined here as the number of exact matches of the argument list divided by the number of common lemmas. Table 8 shows results of matching on all grammatical functions and on selected grammatical relations. We conduct the evaluation experiment at four levels: (1) we match the full argument list between the two data sets, (2) we remove the value of the preposition in obliques, (3) we also remove COMPs and XCOMPs, and (4) we only leave SUBJs, OBJs and OBJ2s. Number (4) denotes transitivity, or the most important type of argument. The smaller the number, the less important the argument type is considered in our perspective.

| | | Precision | | |
|---|---|---|---|---|
| | | Verbs | Nouns | Adjectives |
| 1 | Full argument list | 0.78 | 0.50 | 0.53 |
| 2 | Without preps | 0.82 | 0.52 | 0.66 |
| 3 | Without preps, comps and xcomps | 0.84 | 0.54 | 0.67 |
| 4 | Without obls, comps and xcomps | 0.97 | 0.73 | 0.86 |

Table 8: Evaluating the Tree-induced subcategorization frames against the resource in the Arabic LFG Parser.

Table 8 shows that, at level 4, there is a high level of agreement between the two resources. At level 1, although the precision is comparatively low for nouns and adjectives, we notice that the precision is high for verbs which constitute the largest portion of the data and the most important type of predicates when dealing with subcategorization frames.

## 4. AraComLex Lexical Management Application

In order to manage our lexical database, we have developed the AraComLex (Arabic Computer Lexicon) authoring system which provides a Graphical User Interface for human lexicographers to review, modify and update the automatically derived lexical and morphological information. We use AraComLex for storing the lexical resources mentioned in this paper as well as generating data for other purposes, such as frequency counts and data for extending our morphological transducer.

The data used in the AraComLex is stored in a relational database, with all various tables connected together as shown in Figure 3 which presents a diagram of the entity relationship (Chen, 1976) of the database. 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.

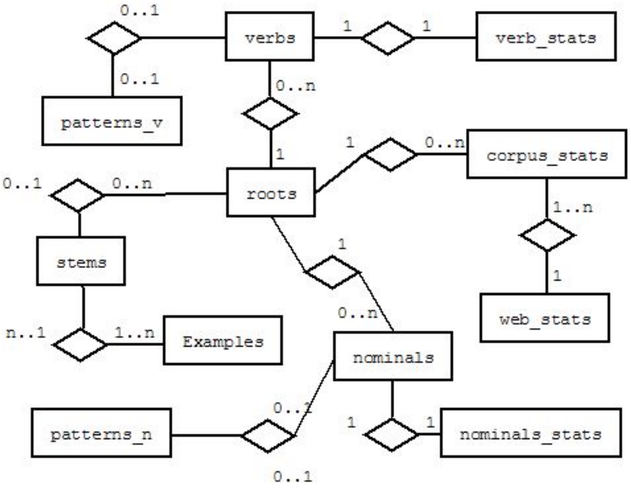
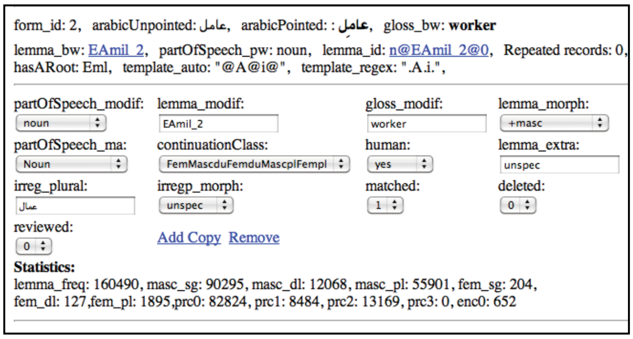


Figure 3: Entity Relationship diagram of AraComLex

AraComLex lists all 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 as shown in Table 2 above. Statistics show the total frequency of the lemma in the corpus and the weights of each morpho-syntactic feature.


form_id: 2, arabicUnpointed: عامل, arabicPointed: : عامِل, gloss_bw: **worker**

lemma_bw: EAmil_2, partOfSpeech_pw: noun, lemma_id: n@EAmil_2@0, Repeated records: 0,
hasARoot: Eml, template_auto: "@A@i@", template_regex: ".A.i.",

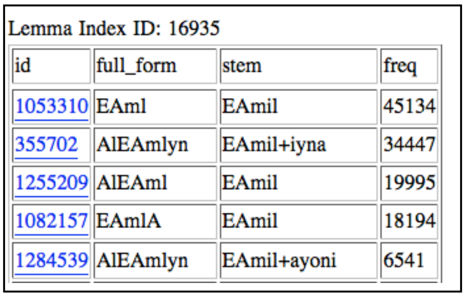
| | | | |
|---|---|---|---|
| partOfSpeech_modif:<br>noun | lemma_modif:<br>EAmil_2 | gloss_modif:<br>worker | lemma_morph:<br>+masc |
| partOfSpeech_ma:<br>Noun | continuationClass:<br>FemMascduFemduMascplFempl | human:<br>yes | lemma_extra:<br>unspec |
| irreg_plural:<br>عمال | irregp_morph:<br>unspec | matched:<br>1 | deleted:<br>0 |
| reviewed:<br>0 | Add Copy Remove | | |

**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

Figure 4: A nominal entry in AraComLex

Figure 4 shows the features specified for nominal lemmas in AraComLex. The feature "lemma_morph" feature can be either 'masc' or 'fem' for nouns and can also be 'unspec' (unspecified) for adjectives. Following SAMA, "partOfSpeech" can be 'noun', 'noun_prop', 'noun_quant', 'noun_num', 'adj', 'adj_comp', or 'adj_num'.

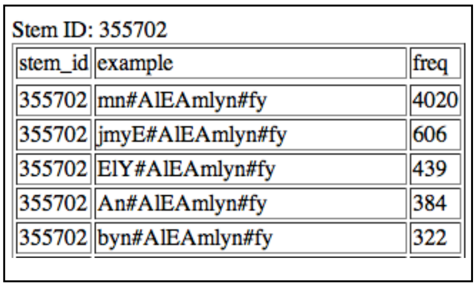
For lexicographic purposes, a lexicographer can review the lemma in detail by looking into the stems and full forms, as shown in Figure 5.


Lemma Index ID: 16935

| id | full_form | stem | freq |
|---|---|---|---|
| 1053310 | EAml | EAmil | 45134 |
| 355702 | AlEAmlyn | EAmil+iyna | 34447 |
| 1255209 | AlEAml | EAmil | 19995 |
| 1082157 | EAmlA | EAmil | 18194 |
| 1284539 | AlEAmlyn | EAmil+ayoni | 6541 |

Figure 5: Lemma Stems

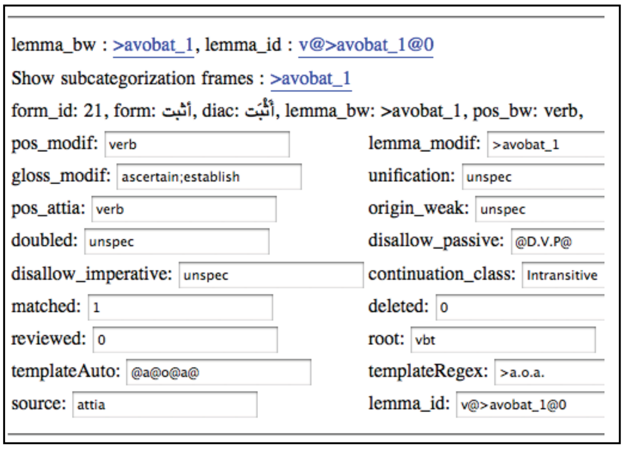
The lexicographer can go even further by reviewing the examples in which the words occurred, sorted according to frequency, as shown in Figure 6. For practical reasons and to keep the size of the database within reasonable bound, we only keep records of the word's bigrams, which in most cases are enough to provide a glimpse of the context and possible collocates.


Stem ID: 355702

| stem_id | example | freq |
|---|---|---|
| 355702 | mn#AlEAmlyn#fy | 4020 |
| 355702 | jmyE#AlEAmlyn#fy | 606 |
| 355702 | ElY#AlEAmlyn#fy | 439 |
| 355702 | An#AlEAmlyn#fy | 384 |
| 355702 | byn#AlEAmlyn#fy | 322 |

Figure 6: Word Examples

For verb lemmas, as shown in Figure 7, we provide information on whether the verb is transitive or intransitive and whether it allows passive and imperative inflection, as well as the usual information on the template and the root. One of the features that can be highly valuable for a lexicographer is the link to subcategorization frames.


lemma_bw : >avobat_1, lemma_id : v@>avobat_1@0

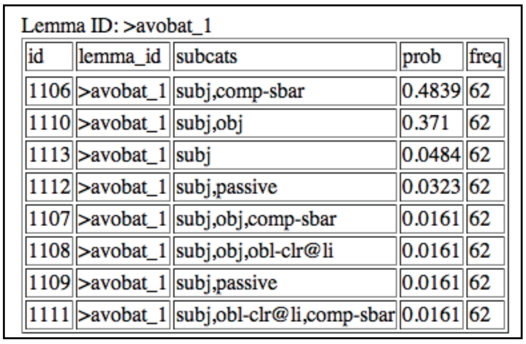
Show subcategorization frames : >avobat_1

form_id: 21, form: أثبت, diac: أَثْبَت, lemma_bw: >avobat_1, pos_bw: verb,

| | |
|---|---|
| pos_modif: verb | lemma_modif: >avobat_1 |
| gloss_modif: ascertain;establish | unification: unspec |
| pos_attia: verb | origin_weak: unspec |
| doubled: unspec | disallow_passive: @D.V.P@ |
| disallow_imperative: unspec | continuation_class: Intransitive |
| matched: 1 | deleted: 0 |
| reviewed: 0 | root: vbt |
| templateAuto: @a@o@a@ | templateRegex: >a.o.a. |
| source: attia | lemma_id: v@>avobat_1@0 |

Figure 7: A verb entry in AraComLex

The subcategorization frames, as shown in Figure 8, are sorted by probability, ensuring that more frequent subcategorization frames appear on the top. As the figure shows, information on passive occurrences and prepositions for obliques are also included.



Lemma ID: >avobat_1

| id | lemma_id | subcats | prob | freq |
|---|---|---|---|---|
| 1106 | >avobat_1 | subj,comp-sbar | 0.4839 | 62 |
| 1110 | >avobat_1 | subj,obj | 0.371 | 62 |
| 1113 | >avobat_1 | subj | 0.0484 | 62 |
| 1112 | >avobat_1 | subj,passive | 0.0323 | 62 |
| 1107 | >avobat_1 | subj,obj,comp-sbar | 0.0161 | 62 |
| 1108 | >avobat_1 | subj,obj,obl-clr@li | 0.0161 | 62 |
| 1109 | >avobat_1 | subj,passive | 0.0161 | 62 |
| 1111 | >avobat_1 | subj,obl-clr@li,comp-sbar | 0.0161 | 62 |

Figure 8: Verb Subcategorization Frames

## 5. Conclusion

We build a lexicon for MSA focusing on the problem that existing lexical resources tend to include a large subset of obsolete lexical entries, no longer attested in contemporary data, and they do not contain sufficient syntactic information. We start 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 predicting inflectional paradigms, successfully extending the lexicon to 30,587 lemmas. We also provide essential lexicographic information by automatically building a lexicon of subcategorization frames from the ATB. We develop a lexicon authoring system, AraComLex,[11] to aid the manual revision of the lexical database by lexicographers. As an output of this project, we create and distribute an open-source finite-state morphological transducer.[12] We also distribute a number of open-source resources that are of essential importance for lexicographic work, including a list of Arabic morphological patterns,[13] subcategorization frames,[14] and Arabic lemma frequency counts.[15]

## 6. Acknowledgements

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).

## 7. References

Al-Sulaiti, L., Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11(1), pp. 135-171.

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

Attia, M., Rashwan, M., Ragheb, A., Al-Badrashiny, M., Al-Basoumy, H. & Abdou, S. (2008). A Compact Arabic Lexical Semantics Language Resource Based on the Theory of Semantic Fields. In B. Nordström, A. Ranta (eds.) *GoTAL '08 Proceedings of the 6th international conference on Advances in Natural Language Processing.* Götenburg, Sweden: Springer-Verlag Berlin, Heidelberg, pp. 65-76.

Attia, M. (2006). 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, UK.

Attia, M. (2008). *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. Ph.D. Thesis. The University of Manchester, Manchester, UK.

Beesley, K.R., Karttunen, L. (2003). *Finite State Morphology*: CSLI studies in computational linguistics. Stanford, California.: CSLI.

Bin-Muqbil, M. (2006). *Phonetic and Phonological Aspects of Arabic Emphatics and Gutturals*. Ph.D. thesis in the University of Wisconsin, Madison.

Boudelaa, S., Marslen-Wilson, W.D. (2010). Aralex: A lexical database for Modern Standard Arabic. *Behavior Research Methods*, 42(2), pp. 481-487.

Buckwalter, T. (2004). Buckwalter Arabic Morphological Analyzer (BAMA) Version 2.0. Linguistic Data Consortium (LDC) catalogue number: LDC2004L02, ISBN1-58563- 324-0.

Chen, P.P. (1976). The Entity-Relationship Model: Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1, pp. 9-36.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Volume 34 of Syntax and Semantics. New York: Academic Press.

Elgibali, A., Badawi, E.M. (1996). *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said M. Badawi*. Egypt: American University in Cairo Press.

Fischer, W. (1997). Classical Arabic. In R. Hetzron (ed.) *The Semitic Languages*. London: Routledge, pp. 397-405.

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

Grishman, R., MacLeod, C. & Meyers, A. (1994). COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, pp. 268–272.

Hajič, J., Smrž, O., Buckwalter, T. & Jin, H. (2005). Feature-Based Tagger of Approximations of

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

[12] http://aracomlex.sourceforge.net

[13] http://arabicpatterns.sourceforge.net

[14] http://arabicsubcats.sourceforge.net

[15] http://arabicwordcount.sourceforge.net

Functional Arabic Morphology. In *The 4th Workshop on Treebanks and Linguistic Theories* (TLT 2005), Barcelona, Spain.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation* (2 ed.). Prentice Hall.

Kiraz, G.A. (2001). Computational Nonlinear Morphology: With Emphasis on Semitic Languages. Cambridge: Cambridge University Press.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S. & Kulick, S. (2010). LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.0. LDC Catalog No. LDC2010L01. ISBN: 1-58563-555-3.

Maamouri, M., Bies, A. (2004). Developing an Arabic Treebank: Methods, guidelines, procedures, and tools. In *Workshop on Computational Approaches to Arabic Script-based Languages, COLING*.

Manning, C. (1993). Automatic acquisition of a large subcategorisation dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 235–242.

O'Donovan, R., Burke, M., Cahill, A., van Genabith, J. & Way, A. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics*, 31(3), pp. 329-366.

Owens, J. (1997). The Arabic Grammatical Tradition. In R. Hetzron, ed.) *The Semitic Languages*. London: Routledge, pp. 46-48.

Palmer, M., Bies, A., Babko-Malaya, O., Diab, M., Maamouri M., Mansouri, A. & Zaghouni, W. (2008). A pilot Arabic Propbank. In *Proceedings of LREC*, Marrakech, Morocco.

Parker, R., Graff, D., Chen, K., Kong, J. & Maeda, K. (2009). Arabic Gigaword Fourth Edition. LDC Catalog No. LDC2009T30. ISBN: 1-58563-532-4.

Rehbein, I., van Genabith, J. (2009). Automatic Acquisition of LFG Resources For German - As Good As It Gets. In *Proceedings of the LFG09 Conference*. Cambridge, UK.

Roth, R., Rambow, O., Habash, N., Diab, M. & Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In *Proceedings of Association for Computational Linguistics (ACL)*, Columbus, Ohio.

Sinclair, J. M. (ed.) (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

Tounsi, L., Attia, M. & van Genabith, J. (2009). 'Automatic Treebank-Based Acquisition of Arabic LFG Dependency Structures.' EACL Workshop on Computational Approaches to Semitic Languages, Athens, Greece.

Van Mol, M. (2000). The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the ninth EURALEX International Congress*, Stuttgart, pp. 831-836.

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

Watson, J. (2002). *The Phonology and Morphology of Arabic*. New York: Oxford University Press.

Wehr, H., Cowan, J.M. (1976). *Dictionary of Modern Written Arabic*, pp. VII-XV. Ithaca, New York: Spoken Language Services.