An Open-Source Finite State Morphological Transducer for Modern Standard Arabic

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

Abstract

We develop an open-source large-scale finite-state morphological processing toolkit (AraComLex) for Modern Standard Arabic (MSA) distributed under the GPLv3 license.[1] The morphological transducer is based on a lexical database specifically constructed for this purpose. In contrast to previous resources, the database is tuned to MSA, eliminating lexical entries no longer attested in contemporary use. The database is built using a corpus of 1,089,111,204 words, a pre-annotation tool, machine learning techniques, and knowledge-based pattern matching to automatically acquire lexical knowledge. Our morphological transducer is evaluated and compared to LDC's SAMA (Standard Arabic Morphological Analyser).

1 Introduction

Due to its complexity, Arabic morphology has always been a challenge for computational processing and a hard testing ground for morphological analysis technologies. A lexicon is a core component of any morphological analyser (Dichy and Farghaly, 2003; Attia, 2006; Buckwalter, 2004; Beesley, 2001). 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.

In this paper, we present an approach to automatically construct a corpus-based lexical database for Modern Standard Arabic (MSA), focusing on the problem that existing lexical resources tend to include obsolete lexical entries no longer attested in contemporary use. The database is used as the lexical component of a large-scale finite state morphological analyser. We specify the morpho-syntactic features and inflection paradigms that need to be explicitly stated for the morphological analyser and show how this information can be learned through machine learning techniques. We explain how broken plural forms are extracted from the corpus using Levenshtein Distance and pattern matching. We report the results of our experiments and evaluate and compare our system against LDC's SAMA (Standard Arabic Morphological Analyser) (Maamouri et al., 2010) showing a substantial reduction of the number of analyses per input word due to avoiding obsolete interpretations no longer present in MSA.

This paper is structured as follows. In the introduction, we differentiate between MSA, the focus of this research, and Classical Arabic (CA) which is a historical version of the language. We also give a brief account of the current state of Arabic morphological analysis and outline the structure of the Arabic morphological system, showing what layers and tiers are involved in word derivation and inflection. Section 2 explains the methodology in constructing our morphological analyser and the lexical database. Section 3 presents the results obtained so far in building and extending the lexical database by using our MSA data-driven filtering method and machine learning techniques. We outline how broken plurals are extracted and handled in our morphology. In Section 4, we evaluate the morphology, and finally, Section 5 gives the conclusion.

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

Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing, pages 125–133,
Blois (France), July 12-15, 2011. ©2011 Association for Computational Linguistics

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 (Bin-Muqbil, 2006).

MSA is different from Classical Arabic 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 quotes Stetkevych (1970) 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 نُظَّمَ 'to be organised'. Instead of the passive form, the alternative syntactic construction تَمَّ 'performed/done' + verbal noun is used, تَمَّ تَنظِيمُهُ 'lit. organising it has been done / it was organised'.

To our knowledge, apart from Van Mol's (2003) 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 The Current State of Arabic Morphological Analysis

Existing Arabic lexicons are not corpus-based (as in a COBUILD approach (Sinclair, 1987)), but rather reflect historical and prescriptive perspectives, making no distinction between entries for MSA and CA (Classical Arabic). Therefore, they tend to include obsolete words not in contemporary use.

The Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) is a *de facto* standard tool which is widely used in the Arabic NLP research community. The latest version of BAMA is renamed SAMA version 3.1 (Maamouri et al., 2010), and it contains 40,648 lemmas. However, SAMA suffers from a legacy of heavy reliance on older Arabic dictionaries, particularly Wehr's Dictionary (Wehr and Cowan, 1976). We estimate that about 25% of the lexical items included in SAMA are outdated based on our data-driven filtering method presented in Section 3.2.

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 and morphological analysis. We use a large corpus of more than one billion words to automatically create a lexical database for MSA.

1.3 Arabic Morphotactics

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 mor-

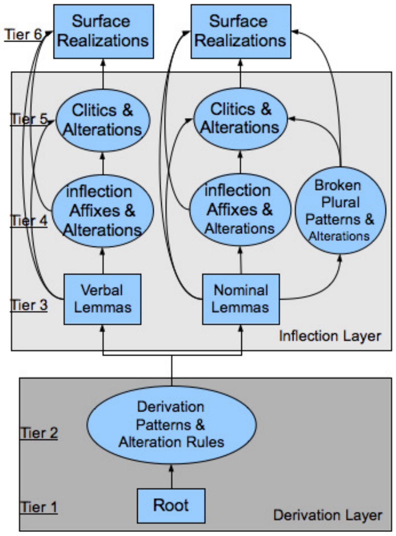

Figure 1: The Multi-tier Structure of the Arabic Morphology.

phemes combine together to form words (Beesley, 1998; 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 إِستَعمَلُوهَا ‘they-used-it’ and the noun وَالاستَعمَالَات ‘and-the-uses’ both originate from the root عمل.

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 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.

2 Methodology

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

2.1 Using Finite State Technology for Arabic

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 languages (Beesley and Karttunen, 2003). 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.

The Xerox XFST System (Beesley and Karttunen, 2003) is a well-known finite state compiler, but the disadvantage of this tool is that it is a proprietary software, which limits its use in the larger research community. Fortunately, there is an alternative, namely Foma, (Hulden, 2009), which is an open-source finite-state toolkit that implements the Xerox lexc and xfst utilities. We have developed an open-source morphological analyser for Arabic using the Foma compiler allowing us to share our morphology with third parties. The lexical database, which is being edited and validated, is used to automatically extend and update the morphological analyser, allowing for greater coverage and better capabilities.

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 (or vowels with consonants) with slots into which the consonants of the root are inserted. This process of insertion is called interdigitation (Beesley, 2001). An example is shown in Table 1.

Root	درس **drs**		
Pattern	$R_1aR_2aR_3a$	$R_1aR_2R_2aR_3a$	$R_1\bar{a}R_2iR_3$
POS	V	V	N
Stem	d a r a s a ‘study’	d a r r a s a ‘teach’	d ā r i s ‘student’

Table 1: Root and Pattern Interdigitation.

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 (Beesley, 2001), 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 (Buckwalter, 2004; Maamouri et al., 2010), 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 nouns and adjectives, the singular indefinite form. Other inflected forms are generated from the lemma through alteration rules.

In our implementation of the Arabic finite state transducer, 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 (Dichy and Farghaly, 2003; Attia, 2006). This leads to better maintainability and scalability of our morphology.

In an XFST 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 (Beesley, 2001; Beesley and Karttunen, 2003). A lexc file contains a number of lexicons connected through what is known as "continuation classes" which determine the path of concatenation. Example (1) gives a snapshot of some verbs in our lexc file. 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", and `@D.M.I@` means "disallow the imperative mood".
- `Transitive` and `Intransitive` are used as the continuation classes for verbs.

```
(1) 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;
```

Similarly, nouns are added by choosing from a set of continuation classes which determine what path of inflection each noun is going to select, as shown in example (2) (gloss is included in square brackets for illustration only). These continuation classes (13 in total) are based on the facts in Table 2, which shows the inflection choices available for Arabic nouns according to gender (masculine or feminine) and number (singular, dual or plural).

```
(2) LEXICON Nouns
+m+human^ss^ مُعَلِّم['teacher']^se^       FMduFduFplMpl;
+m+human^ss^طَالِب['student']^se^         FMduFduFpl;
+m+nonhuman^ss^كِتَاب['book']^se^     Mdu;
+f+nonhuman^ss^بَقَرَة['cow']^se^       DuFpl;
```

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 (Beesley, 1998), and alteration rules are the rules that relate the surface forms to the underlying forms. Alteration rules are expressed in finite state systems using XFST replace rules of the general form shown in (3).

```
(3) 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 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

	Masculine Singular	Feminine Singular	Masculine Dual	Feminine Dual	Masculine Plural	Feminine Plural	Continuation Class
1	مُعَلِّم *muʿallim* ‘teacher’	مُعَلِّمَة *muʿallimat*	مُعَلِّمَان *muʿallimān*	مُعَلِّمَتَان *muʿallimatān*	مُعَلِّمُون *muʿallimuwn*	مُعَلِّمَات *muʿallimāt*	F-Mdu-Fdu-Mpl-Fpl
2	طَالِب *ṭālib* ‘student’	طَالِبَة *ṭālibat*	طَالِبَان *ṭālibān*	طَالِبَتَان *ṭālibatān*	-	طَالِبَات *ṭālibāt*	F-Mdu-Fdu-Fpl
3	تَحضِيرِيّ *taḥḍiyriyy* ‘preparatory’	تَحضِيرِيَّة *taḥḍiyriyyat*	تَحضِيرِيَّان *taḥḍiyriyyān*	تَحضِيرِيَّتَان *taḥḍiyriyyatān*	-	-	F-Mdu-Fdu
4	-	بَقَرَة *baqarat* ‘cow’	-	بَقَرَتَان *baqaratān*	-	بَقَرَات *baqarāt*	Fdu-Fpl
5	تَنَازُل *tanāzul* ‘concession’	-	-	-	-	تَنَازُلَات *tanāzulāt*	Fpl
6	-	ضَحِيَّة *ḍaḥiyyat* ‘victim’	-	ضَحِيَّتَان *ḍaḥiyyatān*	-	-	Fdu
7	مَحض *maḥḍ* ‘mere’	مَحضَة *maḥḍat*	-	-	-	-	F
8	اِمتِحَان *imtiḥān* ‘exam’	-	اِمتِحَانَان *imtiḥānān*	-	-	اِمتِحَانَات *imtiḥānāt*	Mdu-Fdu
9	طَيَّار *ṭayyār* ‘pilot’	-	-	-	.tayyAruwn	-	Mdu-Mpl
10	كِتَاب *kitāb* ‘book’	-	كِتَابَان *kitābān*	-	-	-	Mdu
11	دِيمُقرَاطِيّ *diymuqrāṭiyy* ‘democrat’	-	-	-	دِيمُقرَاطِيُّون *diymuqrāṭiyyuwn*	-	Mpl
12	خُرُوج *ḫuruwǧ* ‘exiting’	-	-	-	-	-	NoNum
13	مَبَاحِث *mabāḥiṯ* ‘investigators’	-	-	-	-	-	Irreg_pl

Table 2: The Arabic Inflection Grid and Continuation Classes.

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 when they are attached to affixes and clitics.

2.2 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 (Parker et al., 2009), in addition to 163,649,497 words from news articles we collected from the Al-Jazeera web site.[2]

We pre-annotate the corpus using MADA (Roth et al., 2008), 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 the annotated data to collect statistics on lemma features and use machine learning techniques, described in Section 3.2.2, in order to extend a manually constructed seed lexicon (Attia, 2006). We also use the annotated data to extract a list of broken plurals, as described in Section 3.3.

3 Results to Date

In this section, we present the results obtained so far in building and extending the lexical database.

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

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 (Attia, 2006) 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 490 Arabic patterns (456 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 coarse-grained categorisation of word forms.

3.2 Extending the Lexical Database

In extending our lexicon, we rely on Attia's manually-constructed finite state morphology (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, Buckwalter's morphology, 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 finite state morphology (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 our lexicon only with lexical items that are still in current use. Second, our lexical database and the FST morphology require features (such as humanness for nouns and transitivity for verbs) that are not provided by SAMA, and we want to automatically induce these features.

3.2.1 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] 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. 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.

3.2.2 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

[3] Statistics were collected in January 2011.

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

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

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

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.

version 3.6.4.[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 are 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. It was surprising to obtain such good prediction results based only 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.

3.3 Handling Broken Plurals

In our seed morphology (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 جَانِب *ǧānib* "side" and the broken plural جَوَانِب *ǧawānib* "sides" are analysed as in (4) and (5) respectively.

```
(4) <lemmaID>jAnib_1</lemmaID>
    <voc>jAnib</voc> <pos>jAnib/NOUN</pos>
    <gloss>side/aspect</gloss>

(5) <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 is the gloss (or translation) and voc (or vocalisation). We also note that MADA passes this problem on unsolved, and broken plurals are all marked with `num=s`, which means 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 حَرب *ḥarb* has the translation "war/warfare" while the plural form حُرُوب *ḥuruwb* has the translation "wars".

To validate the list of candidates, we use Arabic word pattern matching. For instance, in the above example, 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

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

Morphology	No. of Lemmas	General News		Semi-Literary	
		Coverage	Rate per word	Coverage	Rate per word
AraComLex 1.0	10,799	79.68%	1.67	69.37%	1.62
AraComLex 2.0	28,807	86.89%	2.10	85.14%	2.09
AraComLex 2.1	30,587	87.13%	2.09	85.73%	2.08
SAMA	40,648	88.13%	5.32	86.95%	5.30

Table 4: Coverage and Rate Per Word Test Results.

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 list 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 (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.

4 Morphology Evaluation

In this section, we test the coverage and rate per word (or the average number 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.

Table 4 compares the results of coverage and rate per word for AraComLex 2.1 against the baseline (AraComLex 1.0), that is the morphology originally developed in (Attia, 2006); AraComLex 2.0, which does not contain the broken plural extension; and LDC's SAMA version 3.1.

The results show that for the Semi-Literary text, we achieve a considerable improvement in coverage for AraComLex 2.1 over the baseline, increasing from 69.37% to 85.73%, that is 16.36% absolute improvement. However, for the General News text, we achieve less improvement: from 79.68% to 79.68% coverage, that is 7.45% absolute improvement.

Compared to SAMA, AraComLex 2.1 has 1.00% (absolute) less coverage on General News, and 1.22% (absolute) less coverage on the Semi-Literary text. However, the rate per word is significantly lower in AraComLex (2.08) than in SAMA (5.30). 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.

5 Conclusion

We build a large-scale open-source finite state transducer for MSA (AraComLex) distributed under the GPLv3 license. We start off with a manually constructed lexicon of 10,799 MSA lemmas and automatically extend it to 30,587 lemmas, carefully excluding obsolete entries and analyses that are not attested in contemporary data, that is a large MSA corpus containing more than one billion words. We successfully use machine learning to predict morphosyntactic features for newly acquired words. We also use Levenshtein Distance and Arabic word pattern matching to extract broken plurals. 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

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

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.

Beesley, K. R. 1998. Arabic Morphological Analysis on the Internet. In: The 6th International Conference and Exhibition on Multilingual Computing, Cambridge, UK.

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

Beesley, K. R., and Karttunen, L. 2003. Finite State Morphology: CSLI studies in computational linguistics. Stanford, Calif.: Csli.

Bin-Muqbil, M. 2006. Phonetic and Phonological Aspects of Arabic Emphatics and Gutturals. Ph.D. thesis in the University of WisconsinMadison.

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

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

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

Fischer, W. 1997. *Classical Arabic*. In: *The Semitic Languages*. London: Routledge.

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

Hulden, M. 2009. Foma: a finite-state compiler and library. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09). Stroudsburg, PA, USA.

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

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., and Kulick, S. 2010. LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.1. LDC Catalog No. LDC2010L01. ISBN: 1-58563-555-3.

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

Roth, R., Rambow, O., Habash, N., Diab, M., and 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.

Stetkevych, J. 1970. *The modern Arabic literary language: lexical and stylistic developments*. Publications of the Center for Middle Eastern Studies, No. 6. Chicago and London: University of Chicago Press.

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. and Cowan, J. M. 1976. Dictionary of Modern Written Arabic, pp. VII-XV. Ithaca, N.Y.: Spoken Language Services.