Arabic Morphological Analysis Techniques
A Survey and Classification

Ameerah Alothman1, AbdulMalik Alsalman2
Department of Computer Science, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

Abstract—Recently, activity surrounding Arabic natural language processing has increased significantly. Morphological analysis is the basis of most tasks related to Arabic natural language processing. There are many scientific studies on Arabic morphological analysis, yet most of them lack an accurate classification of Arabic morphology and fail to cover both recent and traditional techniques. This paper aims to survey Arabic morphological analysis techniques from 2005 to 2019 and to organize them into a reasonable and expandable classification system. To facilitate and support new research, this paper compares the currently available Arabic morphological analyzers, reaches certain conclusions, and proposes some promising directions for future research in Arabic morphological analysis.

Keywords—Arabic analyzer; Arabic lexicon; classification morphology; morphological analysis; natural language processing

I. INTRODUCTION

Since the advent of the computing era, researchers have been trying to develop systems which can interact with humans; these systems play an essential role in facilitating human life by saving time and improving the quality of work. Morphological analyzers are one such system and constitute an important component of many applications dealing with natural language processing (NLP), machine translation, information search and retrieval, and more.

Morphology is a challenge in Arabic natural language processing (ANLP), and a somewhat complex task. This is because the most important characteristic of Semitic languages is their nonconcatenative nature. Arabic words are composed of roots, derived from certain patterns extracted from stems and their affixes. One root and a small number of patterns with several affixes can form many stems (word formations).

Accordingly, it is necessary to study and classify the techniques of Arabic morphological analyses, because doing so may contribute to greater understanding and improved construction of morphological methodologies, and will pave the way for future researchers in the field of ANLP.

The main purpose of this article is to survey Arabic morphological analysis techniques and bridge the gap in scientific survey studies from 2005 to 2019. This paper is organized as follows: In the second section, we provide basic definitions for this article’s most frequently used terms. In the third section, we propose a classification of Arabic morphological analysis techniques and describe some of the shortcomings of earlier classifications. In the fourth section, we present a survey of Arabic morphological analysis techniques. The fifth section presents a discussion of the comparative study undertaken. Finally, we conclude and summarize some important future directions for Arabic morphological analysis techniques. We adopt Buckwalter [1] for the transliteration of Arabic characters, providing transliterations in brackets where relevant.

II. BASIC DEFINITIONS

There are many terms related to Arabic morphological analyses, and many papers have made great efforts towards the Arabization and standardization of these terms. The book Introduction to Arabic Language Processing, [2] as well as its translation into the Arabic language [3], is one of the most important references in this field of study. Table I presents the meanings and translations of the most frequently used terms in this research.

Fig. 1 Root-Pattern Morphology Process.
TABLE I. BASIC TERMS USED FREQUENTLY IN ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES (ARRANGED ALPHABETICALLY)

<table>
<thead>
<tr>
<th>Term</th>
<th>Translation</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affix</td>
<td></td>
<td>Three types that attach to the root: prefixes, suffixes, and infixes.</td>
<td></td>
</tr>
<tr>
<td>Basic Arabic letter used in</td>
<td></td>
<td>The three basic letters used to construct a pattern in Arabic, which are:</td>
<td></td>
</tr>
<tr>
<td>patterns</td>
<td></td>
<td>[I،E،ل] آ - إ - ل.</td>
<td></td>
</tr>
<tr>
<td>Inflected stem</td>
<td></td>
<td>A stem that may have a prefix and/or suffix to provide meaningful</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>context, also known as a surface word.</td>
<td></td>
</tr>
<tr>
<td>Lexeme</td>
<td></td>
<td>The smallest part of the lexicon that has meaning.</td>
<td></td>
</tr>
<tr>
<td>Long vowel</td>
<td></td>
<td>Also called the “weak letters set” (حرف العلة،) consists of three</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>letters of the alphabet. (“أ”، “إ”، “ي”)</td>
<td></td>
</tr>
<tr>
<td>Morpheme</td>
<td></td>
<td>The smallest unit of the language that has meaning.</td>
<td></td>
</tr>
<tr>
<td>Morphological analysis</td>
<td></td>
<td>The process used to determine all possible morphological analyses of a</td>
<td></td>
</tr>
<tr>
<td>techniques</td>
<td></td>
<td>word.</td>
<td></td>
</tr>
<tr>
<td>Pattern</td>
<td></td>
<td>Abstract CV-template (C: Consonant, V: Vowel) representation of how</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>to order the root and short vowels (and some affixes) to generate the</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>stem. It conveys a grammatical meaning, such as part of speech (POS)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>and tense.</td>
<td></td>
</tr>
<tr>
<td>Root</td>
<td></td>
<td>A sequence of three (most commonly), four (less commonly), or five</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(rarely) consonants. It can be derived based on various patterns. It</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>identifies the general meaning of a word.</td>
<td></td>
</tr>
<tr>
<td>Short vowel or Diacritic</td>
<td></td>
<td>Includes diacritics, which are marks usually written above or below a</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>letter. Diacritics include: 1) three short vowels (“أ”، “إ”، “ي”)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>and “غ” (غرس), and the absence of any vowel (“ف”)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) three mutations (ة، تمين) occurring in the final positions of a</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>word in nominals only; and 3) Shadda (“ـ”)</td>
<td></td>
</tr>
<tr>
<td>Stem</td>
<td></td>
<td>The core of concatenative morphology, it is a surface word generated</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>by inserting the radicals of roots and short vowels into the pattern</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>template slots (e.g., the interdigitating of roots with the patterns).</td>
<td></td>
</tr>
</tbody>
</table>

III. CLASSIFICATION OF ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES

Many scientific papers have tackled the classification of Arabic morphology, and several reviews exist of the most cited Arabic morphological analyzers [4-9]. These studies have many shortcomings, including the following: 1) They are very general in their classification process, and most existing analyzers are classified under one category, “linguistic”. 2) They are somewhat outdated (especially in terms of classification methods) and do not take new techniques into consideration. 3) The authors of these review papers do not provide a standard or basis for the construction of their morphological analyzers or define the approaches that were used to analyze words.

Our aim is to bridge the gaps in the previous studies. Therefore, we have classified Arabic morphology in a more detailed and precise manner than previous studies, in terms of the units used in the analysis. This is based on the approach adopted (linguistic or data-driven lexicon) to adequately clarify the variation in work (see Fig. 2). We also limit ourselves to morphology work carried out after 2004, so that our research will complement the comprehensive survey conducted in this field by Al-Sughaier and Al-Kharashi [4] in 2004.

According to [4], the classification of Arabic morphological analysis techniques falls into four main approaches, namely, pattern-based, combinational, template and morphological. This classification neglects the core unit of how to build lookup tables or linguistic rules (i.e. What should they be based on – root, stem, or lexeme?). As we know, Semitic languages are rich in morphology, and therefore the unit of Arabic used in the analysis must first be specified. Moreover, this classification ignores the machine learning approaches that have received more attention in the latest research. In addition, it does not differentiate between different levels of linguistics and does not take Arabic syntax into consideration. Lastly, it includes a pattern-based approach, which can be more accurately described as part of an approach rather than a separate approach in itself. In the next section, we present in greater detail the proposed classification, which is legitimate and covers all recent and traditional techniques.

III. CLASSIFICATION OF ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES

Many scientific papers have tackled the classification of Arabic morphology, and several reviews exist of the most cited Arabic morphological analyzers [4-9]. These studies have many shortcomings, including the following: 1) They are very general in their classification process, and most existing analyzers are classified under one category, “linguistic”. 2) They are somewhat outdated (especially in terms of classification methods) and do not take new techniques into consideration. 3) The authors of these review papers do not provide a standard or basis for the construction of their morphological analyzers or define the approaches that were used to analyze words.

Our aim is to bridge the gaps in the previous studies. Therefore, we have classified Arabic morphology in a more detailed and precise manner than previous studies, in terms of the units used in the analysis. This is based on the approach adopted (linguistic or data-driven lexicon) to adequately clarify the variation in work (see Fig. 2). We also limit ourselves to morphology work carried out after 2004, so that our research will complement the comprehensive survey conducted in this field by Al-Sughaier and Al-Kharashi [4] in 2004.

According to [4], the classification of Arabic morphological analysis techniques falls into four main approaches, namely, pattern-based, combinational, template and morphological. This classification neglects the core unit of how to build lookup tables or linguistic rules (i.e. What should they be based on – root, stem, or lexeme?). As we know, Semitic languages are rich in morphology, and therefore the unit of Arabic used in the analysis must first be specified. Moreover, this classification ignores the machine learning approaches that have received more attention in the latest research. In addition, it does not differentiate between different levels of linguistics and does not take Arabic syntax into consideration. Lastly, it includes a pattern-based approach, which can be more accurately described as part of an approach rather than a separate approach in itself. In the next section, we present in greater detail the proposed classification, which is legitimate and covers all recent and traditional techniques.

III. CLASSIFICATION OF ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES

Many scientific papers have tackled the classification of Arabic morphology, and several reviews exist of the most cited Arabic morphological analyzers [4-9]. These studies have many shortcomings, including the following: 1) They are very general in their classification process, and most existing analyzers are classified under one category, “linguistic”. 2) They are somewhat outdated (especially in terms of classification methods) and do not take new techniques into consideration. 3) The authors of these review papers do not provide a standard or basis for the construction of their morphological analyzers or define the approaches that were used to analyze words.

Our aim is to bridge the gaps in the previous studies. Therefore, we have classified Arabic morphology in a more detailed and precise manner than previous studies, in terms of the units used in the analysis. This is based on the approach adopted (linguistic or data-driven lexicon) to adequately clarify the variation in work (see Fig. 2). We also limit ourselves to morphology work carried out after 2004, so that our research will complement the comprehensive survey conducted in this field by Al-Sughaier and Al-Kharashi [4] in 2004.

According to [4], the classification of Arabic morphological analysis techniques falls into four main approaches, namely, pattern-based, combinational, template and morphological. This classification neglects the core unit of how to build lookup tables or linguistic rules (i.e. What should they be based on – root, stem, or lexeme?). As we know, Semitic languages are rich in morphology, and therefore the unit of Arabic used in the analysis must first be specified. Moreover, this classification ignores the machine learning approaches that have received more attention in the latest research. In addition, it does not differentiate between different levels of linguistics and does not take Arabic syntax into consideration. Lastly, it includes a pattern-based approach, which can be more accurately described as part of an approach rather than a separate approach in itself. In the next section, we present in greater detail the proposed classification, which is legitimate and covers all recent and traditional techniques.

III. CLASSIFICATION OF ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES

Many scientific papers have tackled the classification of Arabic morphology, and several reviews exist of the most cited Arabic morphological analyzers [4-9]. These studies have many shortcomings, including the following: 1) They are very general in their classification process, and most existing analyzers are classified under one category, “linguistic”. 2) They are somewhat outdated (especially in terms of classification methods) and do not take new techniques into consideration. 3) The authors of these review papers do not provide a standard or basis for the construction of their morphological analyzers or define the approaches that were used to analyze words.

Our aim is to bridge the gaps in the previous studies. Therefore, we have classified Arabic morphology in a more detailed and precise manner than previous studies, in terms of the units used in the analysis. This is based on the approach adopted (linguistic or data-driven lexicon) to adequately clarify the variation in work (see Fig. 2). We also limit ourselves to morphology work carried out after 2004, so that our research will complement the comprehensive survey conducted in this field by Al-Sughaier and Al-Kharashi [4] in 2004.

According to [4], the classification of Arabic morphological analysis techniques falls into four main approaches, namely, pattern-based, combinational, template and morphological. This classification neglects the core unit of how to build lookup tables or linguistic rules (i.e. What should they be based on – root, stem, or lexeme?). As we know, Semitic languages are rich in morphology, and therefore the unit of Arabic used in the analysis must first be specified. Moreover, this classification ignores the machine learning approaches that have received more attention in the latest research. In addition, it does not differentiate between different levels of linguistics and does not take Arabic syntax into consideration. Lastly, it includes a pattern-based approach, which can be more accurately described as part of an approach rather than a separate approach in itself. In the next section, we present in greater detail the proposed classification, which is legitimate and covers all recent and traditional techniques.
IV. SURVEY OF ARABIC MORPHOLOGICAL TECHNIQUES

This section reviews the main approaches to building Arabic morphological analyzers found in the existing literature. Additionally, it lists the morphological systems that have adopted these approaches. Table II provides a summary of the approaches surveyed.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Author and reference</th>
<th>Date</th>
<th>Known as</th>
<th>Test Data</th>
<th>Result (%)</th>
<th>Language coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Elghamry [10]</td>
<td>2005</td>
<td>A constraint-based algorithm</td>
<td>2,700 unique words</td>
<td>Percentage of correct root = 92%</td>
<td>Information not available (N/A)</td>
</tr>
<tr>
<td>Supervised</td>
<td>Daya et al. [11]</td>
<td>2008</td>
<td>Identifying Semitic roots</td>
<td>N/A</td>
<td>Precision: 87.92%; Recall: 92.19%</td>
<td>MSA</td>
</tr>
<tr>
<td>Supervised</td>
<td>Rodrigues and Cavar [13]</td>
<td>2007</td>
<td>Learning Arabic morphology using statistical constraints</td>
<td>10,000 words from BAMA1 dataset</td>
<td>Root predicted with 75% precision</td>
<td>Non-vowelized</td>
</tr>
<tr>
<td>Supervised</td>
<td>Snyder and Barzilay [14]</td>
<td>2008</td>
<td>Unsupervised multilingual learning</td>
<td>Snyder &amp; Barzilay (S&amp;B) dataset</td>
<td>Performance of automatic segmentation: Precision = 67.75%; Recall = 77.29%</td>
<td>N/A</td>
</tr>
<tr>
<td>Supervised</td>
<td>Poon et al. [15]</td>
<td>2009</td>
<td>Unsupervised with log-linear models</td>
<td>– S&amp;B dataset – Arabic Treebank (ATB)</td>
<td>– S&amp;B: F1 = 90 – ATB: F1 = 80.2</td>
<td>N/A</td>
</tr>
<tr>
<td>Data-driven</td>
<td>Botha and Blunsom [16]</td>
<td>2013</td>
<td>Adaptor grammar for learning</td>
<td>– BW corpus (without diacritics) – BW’ with diacritics – Quranic Arabic (QA)</td>
<td>Trilateral root identification accuracy: BW = 67.1%; BW’ = 0.7% – Segmentation: BW = 73.66%; BW’ = 74.54% – QA has a low performance (excluded from comparison)</td>
<td>Vowelized and non-vowelized</td>
</tr>
<tr>
<td>Data-driven</td>
<td>Fullwood and O’Donnell [17]</td>
<td>2013</td>
<td>Learning nonconcatenative morphology</td>
<td>N/A</td>
<td>Accuracy = 92.3%</td>
<td>Vowelized</td>
</tr>
<tr>
<td>Data-driven</td>
<td>Khalil and Carroll [18, 19]</td>
<td>2013</td>
<td>Unsupervised induction of Arabic root and pattern lexicons</td>
<td>Quranic Arabic Corpus (QAC)</td>
<td>Root extraction accuracy = 87.2%</td>
<td>Non-vowelized</td>
</tr>
<tr>
<td>Root-pattern</td>
<td>Gridach and Chenfour [20]</td>
<td>2014</td>
<td>Developing a new system for Arabic morphological analysis and generation</td>
<td>ALECSO Corpus</td>
<td>Accuracy = 95.08%</td>
<td>MSA</td>
</tr>
<tr>
<td>Root-pattern</td>
<td>Habash and Rambow [21]</td>
<td>2006</td>
<td>MAGEAD</td>
<td>Penn Arabic Treebank (PATB), Levantine Arabic Treebank (LATB)</td>
<td>MSA: Context type recall (CTyR) = 52.9%; Context token recall (CToR) = 60.4%; LEV: CTyR = 95.4%; CToR = 94.2%</td>
<td>MSA &amp; Levantine</td>
</tr>
<tr>
<td>Lexeme</td>
<td>Smrz [22]</td>
<td>2007</td>
<td>ElixirFM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Lexeme</td>
<td>Habash [23]</td>
<td>2007</td>
<td>ALMOR</td>
<td>Im Arabic words from the United Nations Arabic-English corpus</td>
<td>Precision = 99.61%; Recall = 87.78%</td>
<td>MSA</td>
</tr>
<tr>
<td>Lexeme</td>
<td>Attia et al. [24]</td>
<td>2011</td>
<td>AraComLex</td>
<td>– 400,000 words from general news – 400,000 semi-literary words</td>
<td>– 87.13% coverage rate on words from the general news – 85.73% coverage rate on semi-literary words</td>
<td>MSA</td>
</tr>
</tbody>
</table>
A. Linguistic Lexicon-Based Approach

In the linguistic lexicon-based approach, solid linguistic rules represented in the heavy lexicon are the core data upon which analysis depends. The lexicon contains two main sections: the first comprises word roots and/or patterns and/or stems, grouped in morphological ways, and the second contains any information related to these contents that the system shows in the results. This approach follows the steps in Fig. 3, with some variations depending on the lexicon and its analyses. The following shows the four basic linguistic lexicon-based approaches:

1) Root-pattern morphology: In brief, morphology is the study of the relationship between meaning and form. It is one of the most challenging tasks in Semitic languages like Arabic, Maltese, and Hebrew. For the most part, Arabic morphology is not concatenative (also called discontiguous or nonlinear). Arabic words are generated from their base roots [5]. In linguistics, there are several nonconcatenative methodologies for deriving the stems of words, because they provide linguistic information [6]. Root-pattern is one of these methodologies.

It is useful to briefly review one of the most important theories of nonconcatenative morphology. In 1979, McCarthy [31, 32] proposed a theorem accepted by linguists (especially computational) to form a stem through a derivational integration of roots and patterns. This mechanism is important for representing the structure of a word in Semitic language morphology.

McCarthy’s [32] work depends on autosegmentalizing the vowels and placing them in a separate tier from the pattern. It has three tiers, as seen in Fig. 4, where C stands for Consonant, V for Vowel:

1) Root tier: refers to consonantal segments, including the meaning of a lexeme, such as (كِتَابُ = كِتَةَبَةَ = k t b), which means “write”.

2) Pattern tier: refers to a prosodic template associated with a particular meaning or grammatical function such as (كَتَبَ = CVCVC =CaCaC), which means, “he wrote”.

3) Vocalization tier: represents pronounced letters and involves grammatical information such as tense, number, and derivational functions.

![Fig. 3. The basic Steps of the Linguistic Lexicon-based Approach.](image)

![Fig. 4. An Example of McCarthy’s [32] Work.](image)
To form an abstract stem, association rules are matched between consonants from the root tier and the pattern tier, and between vowels from the vocalization tier and from the pattern tier. There have been many systems attempting to model Arabic morphology based on McCarthy’s theorem. Most of these systems adopted finite-state language modelling tools [33].

Root-pattern morphology depends on the root and pattern of the word entered for analysis (see Table III). The method involves building lexicons of roots and patterns (or lists of Arabic roots and affixes to cover all prefixes, suffixes, and infixes). Continuous research is being done to extract words that belong to one of the entries in these lists. This process is meant to output analysis of stem forms. Fig. 1 illustrates the main steps followed in this morphology.

One of the earliest published works to adopt this morphology was a system proposed by Hlal [34] and Hegazi and El-Sharkawi [35, 36]. It was also adopted by the Xerox lexicon [37], whose entries depend on root and pattern morphemes. Gridach and Chenfour [20] adopted this morphology with some variations in building their lexicon, depending on XML-based morphological definition language (XMODEL) for its construction.

2) Stem-based morphology: Dichy and Farghaly [6] and Farghaly and Senellart [38] support the claim that building a stem-based lexicon is more intuitive, efficient, and easy to develop and extend compared to a lexicon based on roots.

On the other hand, earlier Arabic morphologies were only responsible for the analysis and/or generation of the correct formations of Arabic words. Many Arabic NLP systems, such as machine translations and automatic summarizations, need linguistic information related to each lexical entry to ascribe elaborate knowledge to each word, in order to become more efficient. This information involves the tense of the verb, number, gender, and part of speech (POS), as well as syntactic features such as the type of subject or object, the count of nouns, and so on. In this context, one adds semantic information, such as the categorization of the noun as human, time, place, and so on. This linguistic information is associated with the stems, which are neither roots nor patterns nor a combination of them [6].

According to the above, Arabic stem-based morphology can achieve a more effective morphological strategy by reducing the complexity of word formations and granting linguistic and semantic information to each entry, thus eliminating the greater lexical gaps.

Two approaches have been built based on this morphology: 1) stems based on root-pattern morphology; and 2) stem-based morphology, including root patterns and syntactic features.

a) Stems based on root-pattern morphology: Briefly, this morphology can be described as follows: each existing lexical entry is checked against candidate entries integrating root and pattern (to generate a stem), in addition to prefix or suffix combinations. Therefore, if the lexicon in this morphology contains, for example, X root and Y pattern, then the XY root-pattern virtual links represent all possible stems, which must be severely restricted to give a reasonable number of meaningful words [33].

The major difference between root-pattern morphology and stems based on root-pattern morphology lies in their analysis mechanisms (see Fig. 1 and 5). The former uses the root and pattern morphemes themselves, while the latter uses stems based on root and pattern morphemes [39].

In this regard, the most famous Arabic analyzer to adopt this morphology is the Buckwalter Arabic Morphological Analyzer (BAMA) [1, 28]. BAMA is based on Buckwalter’s lexicon, which is integrated with the Xerox lexicon [38].

Currently, there are three main versions of BAMA. BAMA 1.0 is available for public use, while BAMA 2.0 and Standard Arabic Morphological Analyzer (SAMA) 3.0 [29] are available through the Linguistic Data Consortium (LDC).

b) Stem-based morphology, including root patterns and syntactic features: Dichy and Farghaly [33] present the significance of syntactic features in Arabic computational morphology in detail. Systems based on this method produce a higher level of morphological analyzers, called morpho-syntactic analyzers. As we know, there are six linguistic levels: phonetics, phonology, morphology, syntax, semantics, and pragmatics (see Fig. 6). This approach takes advantage of the features of the syntax and morpheme levels.

This morphology differs from previous approaches because it applies the additional grammatical features step to results such as prepositions "<b>" and "<k>, which only appear in the genitive case with nouns. These features play an important role in ensuring proper insertion of lexical entries, especially the main ones, such as nouns and verbs.

<table>
<thead>
<tr>
<th>Root</th>
<th>Pattern</th>
<th>In Arabic</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>دَرِسُ (d r s)</td>
<td>السا/ن (C a C a)</td>
<td>تَدَرِسْ (darasa)</td>
<td>درس نار (da-ras)</td>
</tr>
<tr>
<td>(C a C i C)</td>
<td>فَعَّلَ (C a C i C)</td>
<td>تَدَرِسْ (dar-as)</td>
<td>دراس (d ar-as)</td>
</tr>
<tr>
<td>(C a C i C)</td>
<td>قَالَ (C a C i C)</td>
<td>تَدَرِسْ (d a r-s w n)</td>
<td>دارسون (d ar-s w n)</td>
</tr>
</tbody>
</table>

Fig. 5. Stems based on the Root-Pattern Morphology Process.
Standard Arabic Language Morphological Analysis (SALMA) tools [7, 41] fall under this morphological approach. They include SALMA–Tagger, SALMA–ABCLexicon, and SALMA–Tag Set. AlKhalil morphological analyzer [30, 42] also depends on this morphology.

3) Lexeme-based morphology: Typically, lexemes differ only in inflection and cliticization. To put it simply, more than one word can be formed from one lexeme. For example, the lexeme (bayt) includes (bayt), (lilbayt), and (buyuwt). Therefore, the lexeme is not equivalent to a word in any language. It is considered an important abstraction used in linguistic morphology, and is the smallest part of the lexicon that has meaning (or semantic content). Additionally, a lexeme has a morphological form and syntactic category [2].

The claim that the stem is a morphological part with greater relevance to the lexeme is the premise underpinning lexeme-based morphology. This methodology depends on the crucial information of the stem, which must be extracted from the word in the right way. Soudi et al. [43] develop a lexeme-based morphology and present an Arabic version of a morphology rule compiled in the MORPHE tool (MORPHE is a general computational engine that works based on transformational rules and a discrimination hierarchy which must be constructed for each language).

In the lexeme-based methodology, the primary representation is made for the stem (including all operations on the stem, such as transformational rules applied to a stem to handle stem variation issues in several contexts of prefixes and/or suffixes). In other words, this methodology adopts a computational implementation of a non-sub-fragmented lexicon. Thus, this methodology differs from the root-pattern methodology, which gives equal consideration and separate lexicons to each constituent of a word (i.e., sub-lexicons for the root, for the pattern, and for vocalization) [5].

Many works on Arabic morphological analyzers adopt this methodology. Among these works are the following: a) a prototype lacking broad coverage, such as the MORPHE tool [43, 44]; and b) large-scale systems such as:

- ElixirFM [22], which reused the Buckwalter lexicon [1, 28].
- MAGEAD [21, 45] CALIMA, both of which handle Arabic dialects (MAGEAD entirely manually designed while CALIMA manually verified the annotated data lexicon using several computational techniques). There are three versions of CALIMA: CALIMAEGY [25], CALIMAGLF [26], and CALIMAstar [27]. Respectively, these cover Egyptian Arabic, Gulf Arabic, and all variants of MSA and Arabic dialects.

- AL-MORGEANA (abbreviated to ALMOR) [23], which extends the BAMA morphological databases with the lexeme and feature keys that are used in the analysis. For example, ALMOR uses the BAMA lexicon but changes the mode of analysis to produce a lexeme-and-feature format as output, rather than the stem-and-affix format, which is the Buckwalter output. It is important to mention here that ALMOR is the analyzer used in the MADA [46] tool. In addition, the new version of MADA is called MADAMIRA [47]. It is a Java NLP tool combining MADA with a shallow syntactic parser called AMIRA [48].

- AraComLex [24], which is based on the MSA lexical database¹, was specifically constructed for this purpose using a corpus of more than one million words.

4) Syllable-based morphology: Most syllable-based morphology work has been performed on European languages such as German, English, and Italian. Cahill [49] asserts the possibility of analyzing the Semitic languages using syllable-based morphology in a way that is not significantly different from that applied to European languages.

However, to our knowledge, there have been no attempts to build an Arabic morphological analyzer adopting this morphology to substantiate or reject this claim.

B. Data-Driven Lexicon-Based Approach

Machine learning techniques underpin these morphologies. These techniques are fast and do not require extensive linguistic knowledge because they depend on the annotated or unannotated corpus used in the training stage. Dinh et al. [50] claim that doubts could be raised around purely data-driven systems (which do not possess any linguistic base), but they are based on a hybrid. The new techniques prove this claim to be untrue. Recently, many supervised and unsupervised learning techniques have proved valuable in this area, as we will demonstrate in the two following subsections. Thus, we predict a promising future for these morphologies.

1) Supervised learning morphology: This approach attempts to infer parameter values from labeled resources without linguistic expertise about data. Supervised learning resources involve lexica of affixes and pairs of inflected words with their roots [51].

Supervised approaches are not famous in the domain of nonconcatenative morphology acquisition. These approaches require a massive lexicon in the training stage to achieve high precision. Some researchers take pride in their ability to avoid these massive lexica, but the disadvantages can be seen in their results, which have many limitations and are therefore not

---

¹ http://arabiconly.com/aracomlex/form_nominals.php
highly precise in general. However, this is the reality of any new technique. This method will become more promising as more annotated data becomes available.

The existing literature on Arabic morphology that uses this approach to identify Arabic roots is limited. There are two types which adopt some supervised learning: a) learning that is based on pre-existing dictionaries using Hidden Markov Models (HMM) [12] or neural network (NN) models [52], and b) learning that only uses rule constraints [10] or multi-class classifier models [11].

2) Unsupervised learning morphology: Unsupervised learning morphology, in essence, is the process of acquiring intra-word structures and the rules by which they merge to generate word forms [16]. In other words, morphology is induced without prior knowledge, based on training that uses large volumes of unannotated data, without supplying an example of the expected output. This research field began in the mid-1990s and continues today. Researchers consider unsupervised approaches attractive because of the large quantities of unlabelled data available on the Internet [15]. In recent years, unsupervised learning of concatenative language morphology (e.g., stem-affix morphology) has received more attention than nonconcatenative language morphology (e.g., root and pattern morphology) [53].

There are few studies in this field, but they vary according to the objectives of their algorithms. Some aim to learn segmentation [13-15], which means transforming a given word into its stem and affix(es), whereas others aim to learn lexica and patterns [16, 17], which means providing a list of the patterns and assigning each pattern the lexicon information related to all stems belonging to it.

In a significant contribution to this field of research, Khaliq and Carroll [18, 19] have built a morphological analyzer based on roots and patterns induced from the lexicon, based on learning from an unannotated corpus rather than linguistic rules, as noted in the section of this paper dealing with root-pattern morphology. This analyzer achieved good accuracy with root extraction, achieving 94% after many iterative reinforcement stages.

V. DISCUSSION

As shown in the previous survey section, there are multiple morphological analyzers, with varying accuracy and features. No analyzer provides perfect performance, and none has been adopted as standard. Therefore, choosing one of these existing analyzers is difficult and represents a challenge in NLP tasks.

In this section, we compare the analyzers available for public use. Most relevant morphological analyzers achieved acceptable results (according to their developers) but were not available for reuse or evaluation.

To the best of our knowledge, the most recent and efficient morphological analyzers to achieve good accuracies for Arabic morphology are ALKhalil, AraComLex, and ALMOR. ALMOR is no longer available for download. It was distributed as part of MADA Distribution from Columbia University. A new version of MADA, called MADAMIRA, is now available. MADAMIRA is a morphological analyzer and a POS tagger (i.e., MADAMIRA operates within a word context while AlKhalil and AraComLex operate outside of a word context). Table IV compares these analyzers according to various attributes.

<table>
<thead>
<tr>
<th>Attribute name</th>
<th>ALMOR</th>
<th>AraComLex</th>
<th>AlKhalil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different configurations (pre-/post-processing)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Performance metric</td>
<td>Precision &amp; recall</td>
<td>Coverage rate</td>
<td>Coverage rate</td>
</tr>
<tr>
<td>Running through</td>
<td>Application Programming Interface (API)</td>
<td>API</td>
<td>Graphical User Interface (GUI)</td>
</tr>
<tr>
<td>Directionality</td>
<td>Analysis and generation</td>
<td>Analysis only</td>
<td>Analysis only</td>
</tr>
<tr>
<td>Expected input</td>
<td>Text only (works on diacritized text, but no consideration of these diacritics)</td>
<td>Non-vowelized word (does not work on a diacritized word)</td>
<td>Fully or partially diacritized text</td>
</tr>
<tr>
<td>Accuracy (sample of 50 words) [8]</td>
<td>88%</td>
<td>56%</td>
<td>90%</td>
</tr>
<tr>
<td>Engine</td>
<td>Code-based (Java and Perl languages)</td>
<td>Finite-state machinery</td>
<td>Code-based (Java and Perl languages)</td>
</tr>
<tr>
<td>Input format</td>
<td>Word or text</td>
<td>Just one word per query</td>
<td>Word or text</td>
</tr>
<tr>
<td>Output format</td>
<td>Text of (feature: value) pairs</td>
<td>In one line, separate between features by (+)</td>
<td>Table (like CSV file) of features</td>
</tr>
<tr>
<td>Tag set</td>
<td>About 36 basic tag sets</td>
<td>About 14 tag sets</td>
<td>About 118 tag sets</td>
</tr>
<tr>
<td>Transliteration schemes for results</td>
<td>Buckwalter</td>
<td>UTF-8</td>
<td>UTF-8</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

Many scientific studies discuss Arabic morphological analysis techniques, reviews, and analyzer tools, but they lack a specific and accurate classification of traditional and recent methods. In fact, the linguistic lexicon-based and data-driven lexicon-based approaches are the two main approaches for morphological analysis techniques. All techniques found in the existing literature align with these approaches. This classification can guide us towards standard Arabic morphological analysis techniques.

A linguistic lexicon-based approach depends on solid linguistic rules derived from the lexicon. It covers four types of morphology based on analysis process terms: root-pattern, stem, lexeme, and syllable. The data-driven lexicon-based approach depends on an annotated or unannotated corpus to
undergo a training process on data, in order to collect rules which are then used to output word forms.

Most of the systems mentioned in this survey are not available for public use. We highlighted the most recent available systems, and compared them on various aspects.

It is important that future research in Arabic morphological analysis investigate the following issues:

- Developing a gold standard Arabic corpus that can be used to compare morphological analysis systems.
- Developing a large annotated Arabic corpus to be used in the promising data-driven approach morphologies.
- Developing a hybrid approach using linguistic and data-driven morphologies to merge the advantages and strengths of these two approaches.
- Using a unified standard of performance metrics in evaluation systems to compare approaches.
- Building a multicomponent toolkit for Arabic morphological analyzers to integrate these analyzers’ results and choose the one with the best performance.
- Building a multicomponent toolkit for Arabic morphological analyzers in order to facilitate a selection process for the one that best fits the researcher’s/user’s needs.

REFERENCES


