Arab J Sci Eng (2018) 43:7373–7383
https://doi.org/10.1007/s13369-017-2938-8

RESEARCH ARTICLE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Arabic Solid-Stems for an Efficient Morphological Analysis

Moulay Ibrahim El-Khalil Ghembaza[1,3] · **Abdel-Halim Smai**[1] · **Khalid Saleh Aloufi**[2,3]

Received: 6 November 2016 / Accepted: 24 October 2017 / Published online: 7 February 2018

Abstract Arabic natural language processing "NLP" researchers have not yet reached a consensus on a unified definition of stop-words, which is a challenging issue and a limitation in the domain of Arabic NLP in general and Arabic morphological analysis in particular. In this research work, we start by giving a detailed definition and classification of solid-stem words which renames stop-words; we then propose a linguistic-based morphological analysis approach to process this class of words in the Arabic language. A solid-stem word has a unique morphological form and is characterized by the fact that the inflectional ending of the word does not change no matter where the word is in the sentence. A solid-stem word can be a constructed noun (pronoun, indeclinable noun, and verbal noun), an invariable verb, or a particle. A new classification of solid-stems is given, based on their type of affixations. The proposed approach distinguishes between variable and invariable solid-stem types, identifies all possible affixes, and generates all possible morphological variants of each word in a systematic way. For this purpose, we propose a formula for solid-stem affixation and describe in detail affixes schemas using finite-state machine. Building the work on strong linguistic bases lays the foundation for building efficient Arabic search engines and special-purpose information retrieval systems.

Keywords Arabic natural language processing · Computational linguistic · Morphological analysis · Solid-stem words · Affixes schema

1 Introduction

Arabic morphological analysis techniques [1–7] and their use in application such as information retrieval [8–11] have been one of the most popular research areas in Arabic natural language processing "NLP" systems in the last two decades.

The design of Arabic morphological analyzers محللات صرفية عربية is a challenging research area because of the structure and morphological complexity of the Arabic language. In particular, morphological analysis is a crucial part when developing NLP applications such as information retrieval. In such applications, detection and removal of the so-called stop-words is central to the accuracy and performance of the application. We argue that stop-words removal in the Arabic language should be dealt with differently from the English language, for the following two main reasons:

(1) Unlike in the English language, some stop-words in Arabic are in the form of affixes, like the letter "ف" in the word "فأنا";
(2) Some stop-words in Arabic are derivable; for example, the word "أنانية" can be derived from the word "أنا",

✉ Moulay Ibrahim El-Khalil Ghembaza
mghembaza@taibahu.edu.sa

Abdel-Halim Smai
asmai@taibahu.edu.sa

Khalid Saleh Aloufi
koufi@taibahu.edu.sa

[1] Department of Computer Science, College of Computer Science and Engineering, Taibah University, P.O. Box 344, Medina 41411, Kingdom of Saudi Arabia

[2] Department of Computer Engineering, College of Computer Science and Engineering, Taibah University, P.O. Box 344, Medina 41411, Kingdom of Saudi Arabia

[3] IT Research Center for the Holy Quran and Its Sciences (NOOR), College of Computer Science and Engineering, Taibah University, P.O. Box 344, Medina 41411, Kingdom of Saudi Arabia

unlike in the English language where there are no such word generations from stop-words.

We advocate that stop-words in the Arabic language need stemming approaches instead of stop-list generation based on various approaches [8,9,12–14]. Stemming can be applied to obtain a complete and accurate detection of stop-words in information retrieval applications. This stemming will first extract a stop-word and then generates all its possible variants which in turn can be passed on to a removal process if necessary.

From Sect. 4 and on, we are going to use the term solid-stem [15] instead of the term stop-word. Solid-stem words are also called inanimate words or inert words. An extended definition of the concept of solid-stem is given, and based on this definition it is shown how to detect all the solid-stem words and to generate all their morphological variants in a systematic way. The rest of this paper is organized as follows: Related features of the Arabic language are presented in Sect. 2. The concept of stop-words is discussed in Sect. 3. An extended definition of solid-stem is given in Sect. 4. Morphological variants of solid-stems are discussed in Sect. 5. Section 6 describes the design of the Arabic affixes schema for solid-stems. A conclusion is presented in Sect. 7.

2 Related Features of the Arabic Language

Arabic words can be divided into three grammatical types: nouns أسماء, verbs أفعال, and particles حروف معاني [16]. The structure of a word consists of its pattern and its lexical and grammatical forms. A change in the structure of Arabic words can be done for different reasons: either to obtain a word with a new meaning or a word with a new lexical category.

Obtaining a word with a new meaning is like changing from the singular form to dual or plural, and changing to a diminutive or relative adjectives and derived words from verbs or gerund. Obtaining a word with a new lexical category can be done, for instance, by removing or adding one or more letters from a word, or alternating letters.

The nature and structure of the Arabic words are both derivational and inflectional [15]. Derivational means that we can form a new word on the basis of an existing one. Inflectional means that a word can be altered by the addition of an affix or by changing the form of its base, i.e., stem جذع. There are three types of affixes in the Arabic language: inflectional affixes لواصق تصريفية, derivational affixes لواصق اشتقاقية and non-functional affixes لواصق زائدة also called augmentative or inoperative affixes, i.e., without grammatical and morphological functions.

3 Related Work on Stop-Words

A list of stop-words or a stop-list is a list of words that do not reflect the meaning of the content of a given document and may include prepositions, pronouns, and conjunctions. Such a list is also referred to as a functional or structural word list [11]. In some literature, stop-words are called common words because of their repeated occurrence in texts [8,17,18]. In other literature, stop-words are also called non-context bearing or syncategorematic words, i.e., words that need other terms in order to make a meaningful constituent of language [19]. Stop-words are also called excluded words because they are excluded in the mechanism of the language processing.

Removing stop-words can cause some problems in some applications such as in information retrieval when the search applies on key sentences or expressions which include these stop-words. For example, considering names of numbers in Arabic (واحد, إثنان ...) as stop-words can affect the search result in a text on economics. Another example, the name of months becomes key words in text on history [17,18].

There are two different approaches to process stop-words. The first approach is to extract these words manually and then generate all possible forms from each word and include these forms in a list or a lexicon; in this case, the size of the lexicon will continuously increase [8,9,20]. The second approach is to use finite-state machines "FSM" proposed in Al-Shalabi et al. work [12] to identify these words without any predefined list or lexicon and load them directly into memory for processing. It should be noted that FSM have already been used by Beesley in the context of Arabic morphological analysis and generation [2]. However, in both approaches, researchers have not reached a consensus on a unified and limited set of stop-words [8,9]. In fact, the lexicon-based approach suffers from performance degradation and is sometimes not used in Arabic language analyzers. The FSM-based approach in [12] uses a data set of stop-list of more than 1000 words collected from several sources and completed by translation of English stop-list into Arabic. Moreover, this data set is based on one type of words only, which are function words. This gathering of data set is not systematic and can lead to incorrect and incomplete stop-list words. For instance, in [12], the word "قولكما" is identified as a stop-word, which is not correct because this word is derived from the verb "قال". Moreover, generating Arabic stop-word by a translation from English [9] is not always appropriate.

In the next section, we are going to redefine the term solid-stem and give the motivation behind this redefinition, and contrast this new definition with the definition given by Ryding [15]. The various solid-stem variants are described in Sect. 5.

4 An Extended Definition of Solid-Stems

Table 1 Grouping of Solid-Stems

Group1	**Pronouns**	**الضمائر**
	a) Independent personal pronouns	الضمائر المنفصلة
	b) Enclitic personal pronouns	الضمائر المتصلة
Group2	**Indeclinable Nouns**	**الأسماء غير المتمكنة**
	a) Demonstrative Nouns	أسماء الإشارة
	b) Relative Nouns	الأسماء الموصولة
	c) Conditional Nouns	أسماء الشرط
	d) Interrogative Nouns	أسماء الاستفهام
	e) Metonymies	الكنايات
	f) Adverbs	الظروف
Group3	**Verbal Nouns and Onomatopoeia**	**أسماء الأفعال وأسماء الأصوات**
Group4	**Invariable Verbs (inert verbs)**	**الأفعال غير المتصرفة (الأفعال الجامدة)**
Group5	**Particles**	**حروف المعاني**

Ryding has proposed a definition of solid-stems as "words which cannot be reduced or analyzed into the root-pattern paradigm," and that "they consist of primarily three sets in Arabic: pronouns, function words, and loanwords" [15].

We propose an extended definition of solid-stems, based on a systematic linguistic-based classification. Our proposed definition of solid-stems is as follows: Solid-stem words are words which cannot be reduced or analyzed into the root-pattern paradigm (so far as the definition of Ryding [15]), and which have a unique morphological form, and are characterized by the fact that the inflectional ending of the word does not change no matter where the word is in the sentence. They consist of primarily five groups of Arabic words as discussed in the rest of this section.

The extended definition is based on the fact that in the Arabic linguistic analysis, words can be divided into several classes; and one of them (the one that interests us in this paper) is the class of words which obey a unique form الكلمات التي تلازم حالة واحدة i.e., words with a single morphological shape الكلمات غير المتصرفة. By our definition, these words are called solid-stems. The majority of these words do not accept any inflection and derivation (that is a word from which no other words can be derived) but can have affixes (prefixes سوابق and or suffixes لواحق).

In addition, the words in this class are all non-desinential or constructed مبنية in desinence إعراب, which means that their inflectional ending does not change no matter the grammatical function and its position in the sentence. Therefore, this class consists of constructed words الكلمات المبنية which are composed of constructed nouns (groups 1, 2 and 3 in Table 1: pronouns, indeclinable nouns, verbal nouns and onomatopoeia(, invariable verbs which are also called inert verbs (group 4 in Table 1) and particles (group 5 in Table 1). These five groups are shown in Table 1. This class also includes all the above words with all their possible affixes as shown in Table 2.

The motivation behind this definition is to be able to make a classification according to the three grammatical types in Arabic: nouns, verbs, and particles, which usually accept affixes, and to have a finite set of words for NLP applications.

Let us contrast the types of solid-stem words in Ryding's definition and our definition. In Ryding's definition, solid-stems "consist of primarily three sets in Arabic: pronouns, function words, and loanwords" [15]. The first set which is the set of pronouns is the same in both definitions. The second set relates to the set of function words that consist of solid-stems such as prepositions and conjunctions according to Ryding's definition [15]. In our proposed definition, we extend this set of solid-stem words to also include all particles (interrogations, conditionals, prepositions and conjunctions), as well as demonstrative nouns, relative nouns, conditional nouns and interrogative nouns (part of indeclinable nouns). Our definition takes into consideration all of these words, even if it does not consider function words as a group on its own in Table 1.

Our definition also includes all other indeclinable nouns, verbal nouns and onomatopoeia, and invariable verbs as solid-stems since they cannot be reduced or analyzed into the root-pattern paradigm.

As for the third set which consists of loanwords كلمات دخيلة such as استديو, like in Ryding's definition, our definition of solid-stem words also includes loanwords (even if they are not discussed in depth in this paper). Nevertheless, a different approach from what is described in this work should be applied to process this type of solid-stem words. It is worth mentioning that it is practically difficult to list all loanwords, as there will always be new ones. Loanwords are not discussed any further in the rest of this paper.

In addition, our definition also includes all the variants of these solid-stems words as described in Sects. 5 and 6, which is not included in Ryding's definition.

As for comparing our work with the related work in Sect. 3, the lexicon-based approach does not define and include the stop-words in a systematic way and it is not based on Arabic linguistic rules as in our work. Nevertheless, although the approaches are rather different, some stop-words are similar or the same. For example, pronouns are included in both the lexicon-based approach and our approach. Similarly, the FSM approach in the related work is neither systematic nor based on Arabic linguistic rules; besides, it considers only functions words. Overall, our approach is linguistically and computationally more formal and more complete than the related work discussed in Sect. 3.

In the rest of this paper, solid-stems are defined according to our new definition given and discussed in this section.

5 Morphological Variants of Solid-Stems

In this section, we briefly discuss the solid-stems and the generation of their variants.

The Arabic language is characterized by three levels to create new words from existing ones, as follows:

(1) The first level is adding non-functional affixes;
(2) The second level is word inflection which in turn is divided into two sublevels:
  (a) Declension for nouns تمكين الأسماء;
  (b) Conjugation for verbs تصريف الأفعال;
(3) The third level is word derivation.

Although solid-stems cannot be reduced or analyzed into the root-pattern paradigm, they can accept the three types of affixes (not to confuse with affixes categories such as prefix, infix and suffix):

(1) Non-functional (or non-category changing) affixes;
(2) Inflectional affixes and;
(3) One kind of derivational affixes (which is the suffix "ية" for the artificial gerund المصدر الصناعي).

Furthermore, a solid-stem word can accept more than one affix.

On the other hand, derivational words accept also all the three types of affixes (non-functional, inflectional affixes, and derivational affixes).

Infixes are specific for derivational words only, although solid-stem words also accept infixes but in the form of germination (consonant doubling, حركة شدّة) which has no calligraphic letter in Arabic language and often does not appear in the modern Arabic text.

Note that in general, inflection and derivation are performed through affixation. Inflection and derivation can be used to generate morphological variants of solid-stem words in particular. Inflection is performed through affixation, whereas derivation is performed through suffixation.

5.1 Levels of Word Creation

In this section, we discuss how words in general and solid-stem variants in particular can be created through affixation.

5.1.1 First Level: Non-functional Affixes

Non-functional affixes include the interrogation particle حرف الاستفهام "أ", coordinating conjunctions particles حرفا العطف "ف"، "و", preposition particles حروف الجر "ب"، "ك"، "ل", letter of protection نون الوقاية "ن", enclitic (attached) personal pronouns الضمائر المتصلة. In addition, the following non-functional affixes: imperfective pronominal affixes أحرف المضارعة "أ"، "ت"، "ن"، "ي", and the future particle حرف الاستقبال "س" are specific to conjugation verbs.

Non-functional affixes for solid-stems can be the interrogation particle, coordinating conjunctions particles, preposition particles, letter of protection "ن", enclitic pronouns. These affixes are governed by the following identified linguistic rules:

(1) The preposition cannot be a prefix for particles and verbs;
(2) Preposition cannot be placed before coordinating conjunctions particles;
(3) The interrogation particle must be before any coordinating conjunctions particle;
(4) The interrogation particle is always the first prefix (i.e., antefix);
(5) Enclitic pronouns are always the last suffix (i.e., postfix).

5.1.2 Second Level: Inflection Affixes

There are two types of Arabic words in inflection [16]:

(a) Declension, which is the inflection of nouns (declinable and indeclinable), it includes gender (masculine/feminine), number (singular/dual/plural), cases (nominative/genitive/accusative), definiteness, relative adjectives, and diminutives.
(b) Conjugation, which is the inflection of verbs (verbs are either conjugable or invariable), it includes tense (past/present/imperative), voice (passive/active), mode (indicative, subjunctive, jussive), gender (feminine or masculine), number (singular, dual, plural), and person (first, second, third).

In inflection, affixes have grammatical functions (in the form of grammatical categories). Inflectional grammatical categories are one type of morphological variants of solid-stem words. These solid-stem words are not totally inflectional.

There are a few solid-stem words (pronouns and some indeclinable nouns: interrogative nouns, metonymies, and adverbs) that can be inflected according to the following inflectional grammatical categories:

(1) Masculinity or feminization;
(2) Singularity, duality, or plurality;
(3) Nominative, genitive, or accusative;
(4) Definiteness;
(5) Relative adjectives.

It is important to note that the above-mentioned pronouns and indeclinable nouns accept the above inflectional grammatical categories after creating the artificial gerund.

However, we do not consider any inflection category for verbs because in the class of solid-stem words we consider only invariable verbs which do not accept conjugation in all tenses and all pronouns.

5.1.3 Third Level: Derivation Affixes

In general, word derivation in Arabic can be from verbs, nouns, and particles. (The amount of derivation is most for verbs and nouns and least for particles).

There are a few solid-stem words that can be derived, and for such words, there is only one type of derivation which is achieved through adding the suffix "ـية", to create the artificial gerund.

5.2 Types of Solid-Stems

In this section, a new classification of solid-stems is given, based on their type of affixations.

5.2.1 Invariable Solid-Stems

We call invariable solid-stem words solid-stem words which accept non-functional affixes only. New words created from invariable solid-stem are of the same grammatical word type.

5.2.2 Variable Solid-Stems

For solid-stem words, the first level, in Sect. 5.1, can be applied to all the words in this class of words. In contrast, levels 2 and 3 are applied to a very few solid-stem words which we call variable solid-stems. Variable solid-stems accept all the three types of affixes (non-functional, inflectional, and derivational). Adding such affixes to a given variable solid-stem results in creating new words which can be of the same or a different grammatical word type of the given variable solid-stem.

5.2.3 Derivable Solid-Stems

From the following groups of solid-stems: adverbs (subset of the indeclinable nouns), verbal nouns and onomatopoeia, and particles, a few words can be transformed into a derivable form. For example, for the adverb "صباح" we create the gerund "تصبُّحٌ" and the verb "صبَّح" from which many words can be derived such as "مصباح". For the verbal noun "أف", we create the gerund "تأفُّفٌ" and the verb "أفَّف"و"تأفَّف" from which many words can be derived such as "متأفّف" و "أفّاف" For the particle "سوف", we create the gerund "تسويفٌ" and the verb "سوَّف" from which many words can be derived such as "مسوِّف". These words accept all types of affixes, in addition to all types of derivations.

These kinds of words become part of the class of derivable words in Arabic and therefore are not treated as solid-stems.

To summarize, we have identified two types of solid-stems in Arabic: variable and invariable solid-stems. Together, these types of solid-stem words satisfy the following affixes formula:

```
[Antefix | Prefix1 | Prefix2 | Prefix3] +
              Solid-Stem +
 [Suffix1 | Suffix2 | Suffix3| Postfix]
```

Formula 1: Affixation for solid-stem words.

Table 2 summarizes all the types of Arabic affixes and their categories for solid-stems and all the different possibilities to generate morphological variants of these solid-stem words.

6 Arabic Affixes Schemas for Solid-Stems

Here, we describe the affixes schemas for the variable and invariable solid-stems. The schemas identify all the possible associated affixes and generate all possible morphological variants of each solid-stem word.

These schemas are used for both detection and generation of variable and invariable solid-stem words.

6.1 Affixes Schemas for Invariable Solid-Stems

In this section, we describe the possible affixes for the invariable solid-stems in each group in Table 1, based on the affixation levels defined in Sect. 5.1, and affixes defined in Table 2.

An example of generation without changing the grammatical word type is as follows: From the pronoun "أنت", we can add the interrogation particle "أأنت", we can add a coordinating conjunctions particle like "وأنت", we can add the letter "لـ" to obtain "لأنت", and we can combine all the above to obtain "ولأنت"، "أوأنت" ، "أولأنت". All these obtained words are pronouns.

Table 2 Arabic affixes for solid-stem words

Types of Arabic Affixes for Solid-Stems								
Prefix Category				Suffix Category				
Antefix	Prefix1	Prefix2	Prefix3	Suffix1	Suffix2		Sufix3	Postfix
non-functional	non-functional	non-functional	Inflectional	non-functional	Inflectional	Derivational	Inflectional	non-functional
Particle of interrogation "أ"	Particle"و" with all its types OR Particle"ف" with all its types	Particle "ل" with all its types OR Particle "ك" with all its types OR Particle "ب" with all its types	Definite article "ال"	Protection "ن"	Relative Adjective "ي"	Artificial Gerund "يّة/ يّت"	Duality and plurality particles "ات"، "ون"، "و"، "ين"، "ان"، "ا"، "ة/ت"	Enclitic pronouns "هن"، "هم"، "كن"، "كم"، "نا"، "هما"، "كما"، "ها"، "ـه"، "ك"، "ي"

Below, the affixes schemas for the invariable solid-stems in the five groups in Table 1 are described.

Group 1 This group consists of independent personal pronouns and enclitic personal pronouns. The following automata define all possible affixes for independent personal pronouns. However, enclitic personal pronouns are always and only suffixes; for this reason, no specific automata are defined for such pronouns. (In Arabic language, enclitic personal pronouns cannot be independent words).

Independent personal pronouns can be subject pronouns or object pronouns, as shown in Fig. 1a, b, respectively.

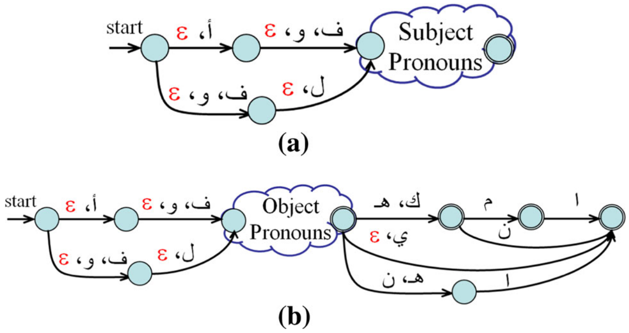

Fig. 1 Finite automaton for independent personal pronouns. **a** Subject independent pronouns. **b** Object independent pronouns

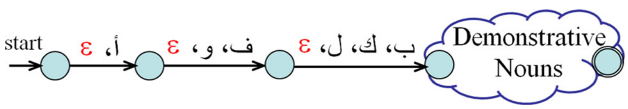

Fig. 2 Finite automaton for demonstrative nouns

Subject independent personal pronouns (for example, أنا and أنت) accept as prefixes the interrogation particle "أ", the coordinating conjunctions particles "و" ،"ف", and the particle "ل" with all its possible types (such as jurative particle لام القسم), but it does not accept any suffixes.

Object independent personal pronouns accept as prefixes the interrogation particle, the coordinating conjunctions particles, and the particle "ل" with all its possible types (such as jurative particle). All object pronouns have in common the root "إيّا" (for example, "إيّاي" and "إيّاك") Object pronouns are obtained by appending the enclitic pronouns (see Table 2) as a postfix to the root "إيّا". No other suffixes are accepted.

Group 2 This group is composed of six subgroups: demonstrative nouns, relative nouns, conditional nouns, interrogative nouns, metonymies, and adverbs.

Demonstrative nouns (for example, "هذا" and "هذه") accept as prefixes the interrogation particle, the coordinating conjunctions particles and the preposition particles "ل" ،"ك" ،"ب" and do not accept any suffixes, as shown in Fig. 2.

Relative nouns (for example, "الذي" and "التي") accept as prefixes the coordinating conjunctions particles and the preposition particles, and do not accept any suffixes, as shown in Fig. 3.

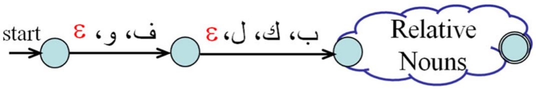

Fig. 3 Finite automaton for relative nouns

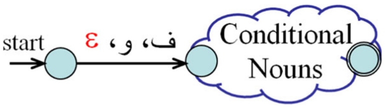

Fig. 4 Finite automaton for conditional nouns

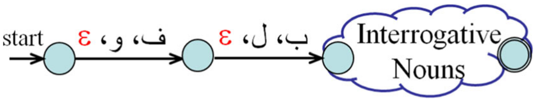

Fig. 5 Finite automaton for interrogative nouns

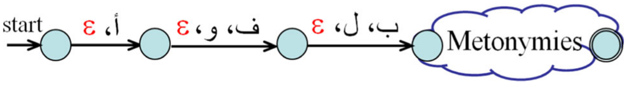

Fig. 6 Finite automaton for metonymies

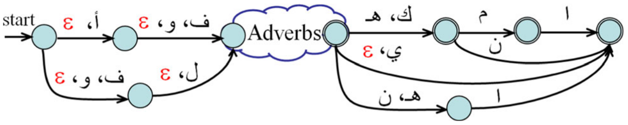

Fig. 7 Finite automaton for adverbs

Conditional nouns (for example, "مَن" and "ما") accept as prefixes only the coordinating conjunctions particles and do not accept any suffixes, as shown in Fig. 4.

Interrogative nouns (for example, "ماذا" and "متى") accept as prefixes the coordinating conjunctions particles, and the preposition particles "ل" ،"ب" and do not accept any suffixes, as shown in Fig. 5.

Metonymies (for example, "كذا" and "كيت") accept as prefixes the interrogation particle, the coordinating conjunctions particles, and the preposition letters "ل" ،"ب"، and do not accept any suffixes as shown in Fig. 6.

Adverbs (for example "قبل" and "بعد") accept as prefixes the interrogation particle, the coordinating conjunctions particles, and the particle "ل" with all its possible types and accept as postfix the enclitic pronouns as shown in Fig. 7.

Group 3 This group consists of verbal nouns and onomatopoeia. Some verbal nouns (for example, "هيهات" and "شتّان") and onomatopoeia (for example, "كخ" and "بخ") cannot be incorporated with the enclitic pronouns and accept as prefixes only the coordinating conjunctions particles as shown in Fig. 8a, b.

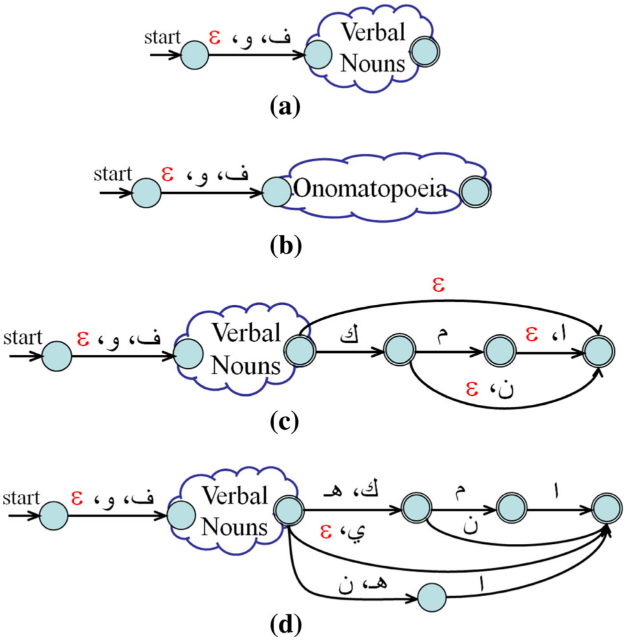

Fig. 8 Finite automaton for Verbal nouns and onomatopoeia

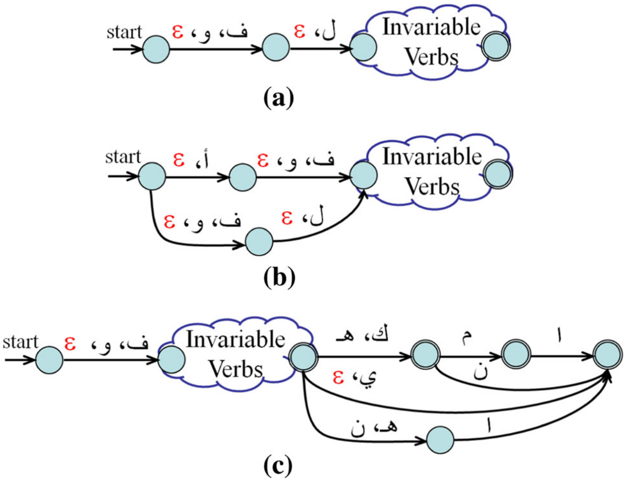

Fig. 9 Finite automaton for invariable verbs. **a** Invariable verbs of praise and blame. **b** Invariable verbs طالما و قلّما. **c** Invariable exception verbs

Another subgroup of verbal nouns (for example, "دونك" and "هاك") can be incorporated with the postfix enclitic pronouns "كن" ،"كم" ،"كما" ،"ك" only and accept as prefixes only the coordinating conjunctions particles as shown in Fig. 8c.

The last subgroup of verbal nouns (for example, "ويكأنّ") can be incorporated with the postfix enclitic pronouns and accept as prefixes only the coordinating conjunctions particles as shown in Fig. 8d.

Group 4 This group is composed of invariable verbs; the first subgroup is composed of verbs of praise and blame أفعال المدح والذم (for example, "حبذا" and "ساء"), accepts as prefixes the coordinating conjunctions particles and the jurative particle, and does not accept any suffixes, as shown in Fig. 9a.

The second subgroup is composed of the invariable verbs "طالما" and "قلّما", accepts as prefixes the interrogation particle, the coordinating conjunctions particles, and the jurative particle, and does not accept any suffixes, as shown in Fig. 9b.

The third subgroup is composed of the exception verbs أفعال الاستثناء (for example, "حاشـا" and "خـلا") and accepts as prefixes only the coordinating conjunctions particles and as postfix the enclitic pronouns, as shown in Fig. 9c.

Group 5 This group consists of particles such as prepositions, coordinating conjunctions, conditionals, and interrogations.

The first subgroup is composed of words such as interrogation particles and exception particles حروف الاستثناء like "إلاّ" and accepts as prefixes only the coordinating conjunctions particles as shown in Fig. 10a.

The second subgroup is composed of words such as causality particles حروف التعليل like "كـي", particle of future حـرف الاستقبال "سـوف", and particle of certainty حرف التحقيق "قد" and accepts as prefixes only the coordinating conjunctions particles and the particle "ل" as shown in Fig. 10b.

The third subgroup is composed of words such as particles of jussive حروف الجزم like "لـم", and particles of negation حروف النفي like "لـن" and accept as prefixes only the interrogation particle and the coordinating conjunctions particles as shown in Fig. 10c.

The fourth subgroup is composed of words such as confirmative particles حروف التوكيد like "أنّ", accepts as prefixes the interrogation particle, the coordinating conjunctions particles, and the preposition particles "ل" ،"ب", and can be incorporated with the postfix enclitic pronouns "كما" ،"هما" only, as shown in Fig. 10d.

The fifth subgroup is composed of words such as preposition particles حـروف الجـر, and accepts as prefixes the interrogation particle, the coordinating conjunctions particles and the particle "ل“, and can be incorporated with the postfix enclitic pronouns, as shown in Fig. 10e.

The sixth subgroup is composed of words such as sisters of “Indeed” إنّ وأخواتها, and accepts as prefixes the interrogation particle, the coordinating conjunctions particles, the particle “ل” with all its possible types and the preposition particle "ب", and can be incorporated with the postfix enclitic pronouns, as shown in Fig. 10f.

6.2 Affixes Schemas for Variable Solid-Stems

In this section, we describe the possible affixes for the variable solid-stems in groups 1, 2, and 4 in Table 1, based on the affixation levels defined in Sect. 5.1, and affixes defined in Table 2. Recall that groups 3 and 5 do not contain any variable solid-stems.

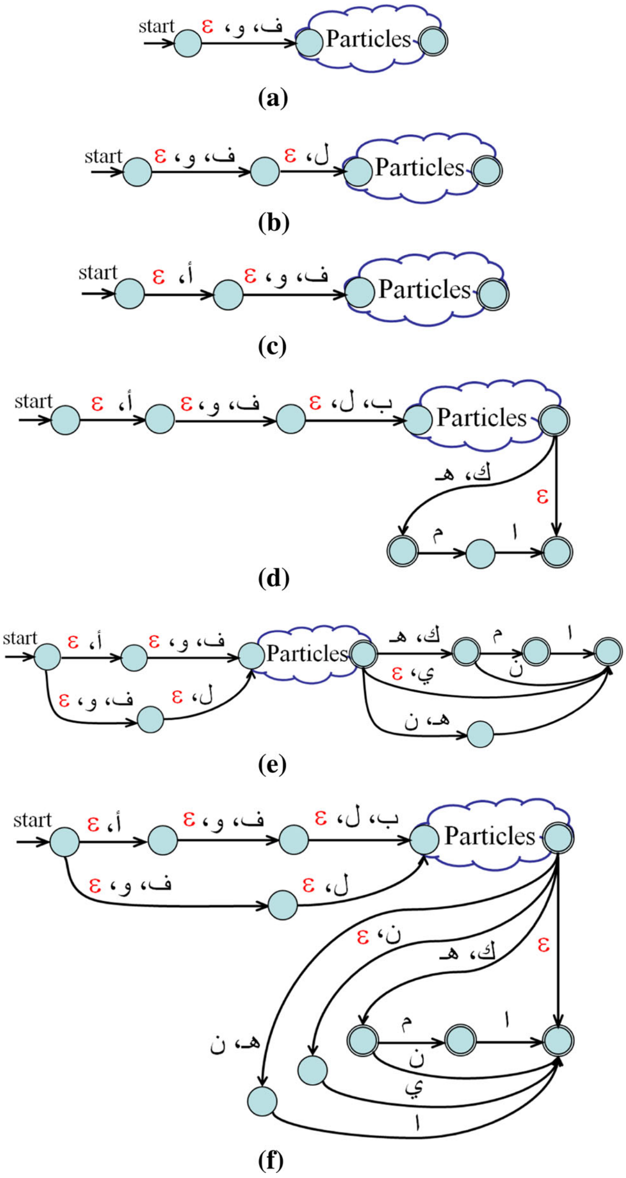
(a)
(b)
(c)
(d)
(e)
(f)

Fig. 10 Finite automaton for particles

An example of generation with changing grammatical word type is as follows: From the pronoun "أنا", we can derive the artificial gerund "أنانيـة", we can inflect it to relative adjective "أنـاني", as well as define it to become "الأنـاني", we can put it in its dual form "أنانيان" and transform it into its all its plural form "أنانيون" and "أنانيات", and we can add to it the enclitic pronouns like "أنانيتي".

In addition from the same pronoun "أنـا" we can generate new words of the same grammatical word type as follows: We can add the interrogation particle "أأنا", we can add the coordinating conjunctions particle like "وأنا", we can add the

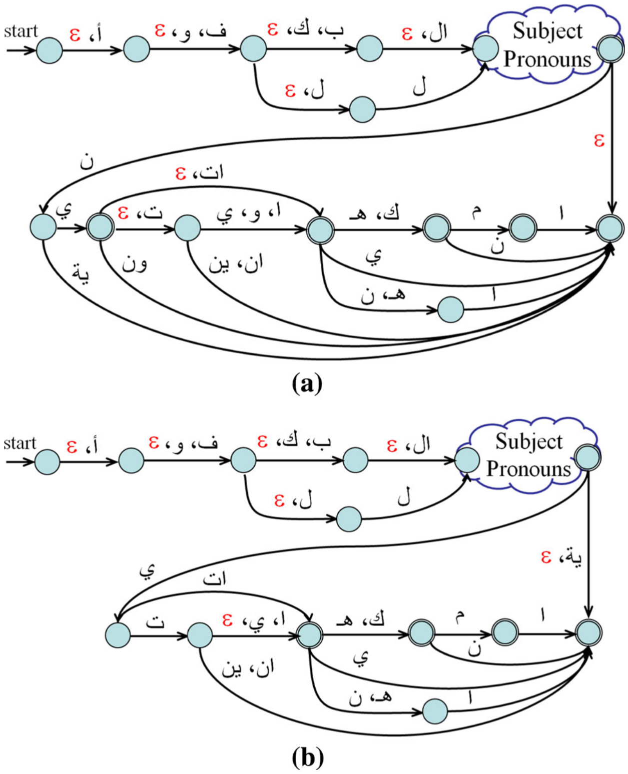

Fig. 11 Finite automaton for subject pronouns

letter "لـ" to obtain "لأنا", and we can combine all the above to obtain "ولأنا"، "أوأنا"، "أولأنا".

Below, the affixes schemas for the variable solid-stems in groups 1, 2, and 4 are described.

Group 1 As mentioned in Sect. 5.1, in this group we consider only subject independent personal pronouns. Object independent personal pronouns accept only non-functional affixes (inflectional and derivational affixes are not accepted) as described in Sect. 5.1. Enclitic personal pronouns are always and only suffixes, and therefore, no specific automata are defined for such pronouns.

Moreover, only three of the subject independent personal pronouns ("أنا"، "هو"، "هي") accept inflectional and derivational affixes, in addition to the non-functional affixes. These three pronouns accept all types of prefixes: the interrogation particle, coordinating conjunctions particles, prepositions, and the definite article "ال" and accept all types of suffixes: protection particle نون الوقاية only for the pronoun "أنا" (as shown in Fig. 11a), the relative adjective article "ي", the artificial gerund "يـة", duality particles, and all plurality particles except the particles "و"، "ون"، "ين" for the pronoun "هو" (see Table 2), and as postfix enclitic pronouns.

The automata below define all possible affixes for subject independent personal pronouns as shown in Fig. 11a, b.

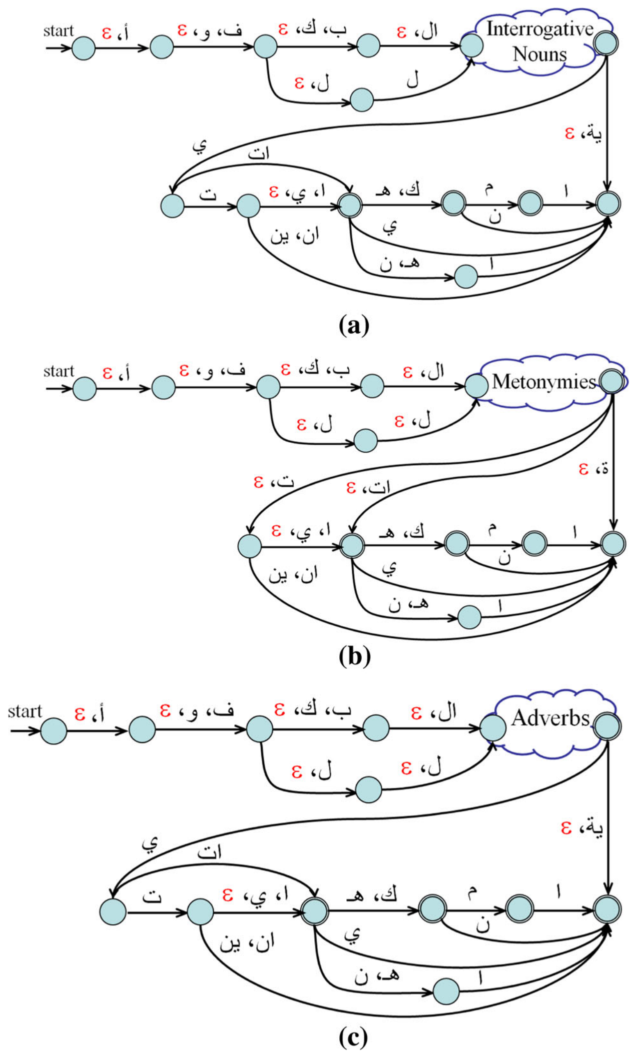

Fig. 12 Finite automaton for indeclinable nouns. **a** Interrogative nouns. **b** Metonymies. **c** Adverbs

It should be noted that the pronoun "هـي" is incorporated with the prefix "مـا" and the suffix "ة" to form the word "ماهـية" and this word accepts the same affixes as the pronoun "هو" as shown in Fig. 11b.

Group 2 In this group, only two of the interrogative nouns "كم"، "كيف", two of metonymies "فلان"، "بضع", and some adverbs like "حيث" accept inflectional and derivational affixes, in addition to the non-functional affixes. These three types of nouns (i.e., interrogative, metonymy, and adverb) accept all types of prefixes: the interrogation particle, coordinating conjunctions particles, prepositions, and the definite article "ال" and accept as suffixes: the relative adjective particle "ي" except for metonymies, the artificial gerund "ية" except for metonymies, duality particles and plurality particles except the particles "و"، "ون"، "ين" (see Table 2), and as postfix enclitic pronouns.

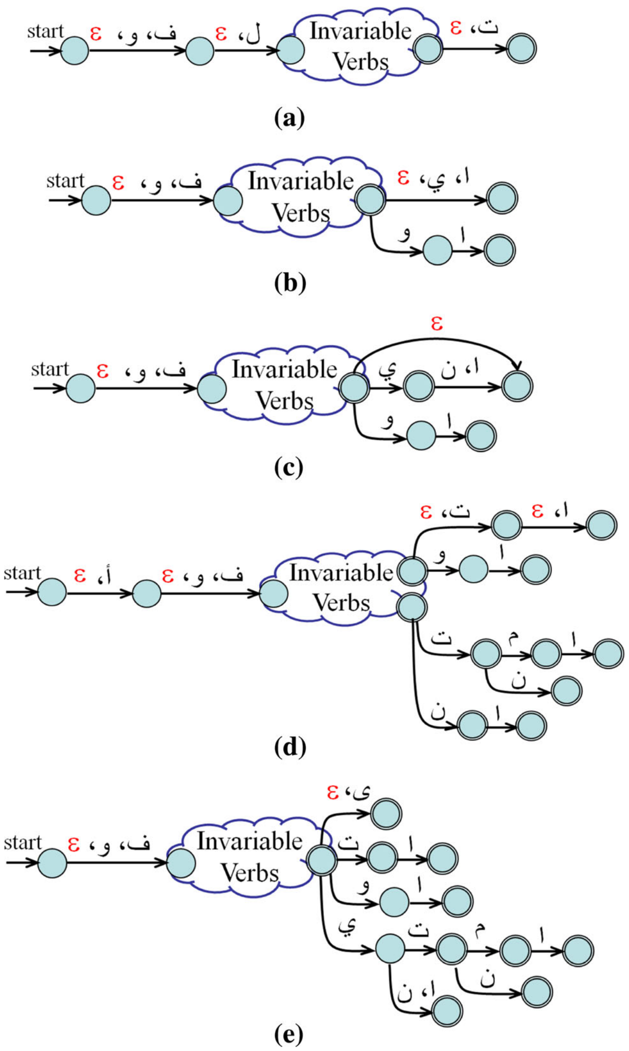

Fig. 13 Finite automaton for invariable verbs. **a** Invariable verbs نِعم، بِئس. **b** Invariable verb هلّم. **c** Invariable verb هات. **d** Invariable verb ليس. **e** Invariable verb عسى

The automata below define all possible affixes for these three types of nouns as shown in Fig. 12a–c.

Group 4 Invariable verbs are verbs which are not conjugable in all tenses and all pronouns. They may accept conjugation in the past tense only or in the imperative tense only or with singular third person pronouns ضمائر الغائب المفرد "هو" و "هي" only [21]. Therefore, variants can be generated from some invariable verbs.

For example, the invariable verbs "نِعم"، "بِئس" are conjugated only for "هـو" و "هـي", as shown in Fig. 13a. The invariable verbs "هلّـم"، "هـات" are conjugated only in the imperative tense, as shown in Fig. 13b, c. The invariable verbs "عسى"، "ليس" are conjugated only in the past tense, as shown in Fig. 13d, e.

7 Conclusion

We gave a linguistic-based and formal definition of the concept of solid-stem words (called stop-words in previous related work). We demonstrated how to distinguish between variable and invariable solid-stem words and how to generate all possible solid-stem variants. We proposed a formula for solid-stem affixation and provided a detailed description of the affixation schemas that detect any solid-stem as per our extended definition and that generate all the possible morphological variants of any detected solid-stem. This work represents a good effort on the direction of a deep understanding and a unified definition of the concept of solid-stem words in the Arabic Language. Building the work on strong linguistic bases lays the foundation for building efficient Arabic search engines and special-purpose information retrieval systems. We also intend to investigate the particular case of solid-stem words that can be transformed into a derivable form and therefore become part of the class of derivable words in Arabic.

Acknowledgements The authors would like to thank the Deanship of Scientific Research at Taibah University for their support to conduct this research work under research grant no. 432/669.

References

1. Al-Sughaiyer, I.A.; Al-Kharashi, I.A.: Arabic morphological analysis techniques: a comprehensive survey. J. Am. Soc. Inf. Sci. Technol. **55**(3), 189–213 (2004)
2. Beesley, K.R.: Arabic finite-state morphological analysis and generation. In: Proceedings of 16th International Conference on Computational Linguistics, vol. 1, pp. 89–94 (1996)
3. Belkridem, F.Z.; El Sebai, A.: An ontology based formalism for the Arabic language using verbs and their derivatives. Commun. IBIMA **11**(5), 44–52 (2009)
4. Boudlal, A.; et al.: A Markovian approach for arabic root extraction. Int. Arab J. Inf. Technol. **8**(1), 91–98 (2011)
5. Buckwalter, T.: Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data Consortium (2004)
6. Darwish, K.: Building a shallow morphological analyzer in one day. In: Proceedings of Association for Computational Linguistics, Workshop on Computational Approaches to Semitic languages, pp. 1–8 (2002)
7. Rafea, A.; Shaalan, K.: Lexical analysis of inflected arabic words using exhaustive search of an augmented transition network. Softw. Pract. Exp. **23**(6), 567–588 (1993)
8. Abu El-Khair, I.: Effect of stop words elimination for Arabic information retrieval: a comparative study. Int. J. Comput. Inf. Sci. **4**(3), 119–133 (2006)
9. Chen, A.; Gey, F.: Building an Arabic Stemmer for information retrieval. In: Proceedings of Eleventh Text Retrieval Conference (2002)
10. Larkey, L.S.; Ballesteros, L.; Connell, M.E.: Improving Stemming for Arabic information retrieval: light Stemming and Co-occurrence analysis. In: Proceedings of 25th Annual International Conference on Research and Development in Information Retrieval, pp. 275–282 (2002)

11. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, London (1989)
12. Al-Shalabi, R., et al.: Stop-word removal algorithm for Arabic Language. In: Proceedings of International Conference on Information and Communication Technologies: From Theory to Applications (2004)
13. Alajmi, A.; Saad, E.M.; Darwish, R.R.: Toward an Arabic stopwords list generation. Int. J. Comput. Appl. **46**(8), 8–13 (2012)
14. Yuang, C.T.; Banchs, R.E.; Siong, C.E.: An empirical evaluation of stop word removal in statistical machine translation. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (2012)
15. Ryding, K.C.: A Reference Grammar of Modern Standard Arabic. Cambridge University Press, Cambridge (2005)
16. El-Dahdah, A.: A dictionary of Arabic grammar in charts and tables. Lebanon Library (1996)
17. ArabEyes. http://wiki.arabeyes.org. Last Visited (2014)
18. Arabic Stop-Words Project. http://arabicstopwords.sourceforge.net/. Last Visited (2014)
19. Van Rijsbergen, C.J.: Information retrieval. J. Am. Soc. Inf. Sci. **30**(6), 374–375 (1979)
20. Khoja, S.: Stemming Arabic Text. http://zeus.cs.pacificu.edu/shereen/research.htm (1999)
21. Al-Afghani, S.: Summary of the Arabic language grammar. Dar El-Fikr Beirut, Beirut (1968)