Natural Language Engineering **1** (1): 1–28. Printed in the United Kingdom

Sarf: Fast and Application Customizable Arabic Morphological Analyzer

Fadi A. Zaraket

Ameen Jaber

Jad Makhlouta

Electrical and Computer Engineering,
American university of Beirut

(*Received 20 March 2014; revised 30 September 2014*)

Abstract

The rich nature of Arabic morphology makes morphological analysis key for Arabic natural language processing applications. Arabic morphological analyzers return several morphological solutions for a given Arabic word. Each solution consists of several morphological features such as part of speech and gloss description tags. Often times, applications need only few of those features. This paper presents Sarf, an application customizable morphological analyzer for Arabic. Sarf provides an interface that allows application developers to (1) control and prioritize the analysis, (2) refine solution features, and (3) define categories and associate them with existing morphemes.

Sarf uses agglutinative and fusional morphemes for affix representation, and refines the morpheme lexicons of SAMA and BAMA. This reduces redundant morphemes, and subsequently inconsistent morpheme tags in the lexicons. It also solves the segmentation correspondence problem between an input word and the several parts of the associated morphological solution. It uses diacritics to refine solutions, and solves the 'run-on words' problem. The implementation of Sarf efficiently encodes the morpheme lexicons. Sarf was used in several NLP applications for information extraction and provided more accurate solutions than existing solvers with faster running time.

1 Introduction

Natural language processing (NLP) applications such as *machine translation* (MT) and *information extraction* (IE) require *morphological analysis* to preprocess Arabic text due to the rich morphological nature of the Arabic language (Benajiba, Rosso, & Benedíruiz, 2007; Habash & Sadat, 2006). Arabic morphological analyzers return the internal structure of a given Arabic word composed of several *morphemes* including *affixes* (*prefixes* and *suffixes*), *clitics* (*proclitics* and *enclitics*), and *stems* (Al-Sughaiyer & Al-Kharashi, 2004). The morphological solution consists also of several *morphological features* (tags) associated with the word and its constituent morphemes such as *part of speech* (POS), transliteration, *gloss*, and *vocalized morpheme form* (VMF) tags (diacriticized form of the morpheme). The prefix and suffix attach before and after the stem, respectively. Clitics are special affixes that attach to the stem to form a word, and differ from regular affixes in that they play a syntactic role of another word (often omitted) (Habash, 2010).

This work presents Sarf, an *fast and application customizable morphological analyzer*

	suffix ونها	stem لعب	prefix وسي
POS	IVSUFF_SUBJ:MP_MOOD:I+IVSUFF_DO:3FS	VERB_IMPERFECT	CONJ+FUT+IV3MP
Transliteration	uwnahA	loEab	wa+sa+ya
Gloss	[MASC.PL.]+it/them/her.	play	and they will

Table 1. Example morphological solution for the word وسيلعبونها *wsylʿbwnhā*

for Arabic that was used in several applications for information extraction from Arabic text (Jaber & Zaraket, 2013; Makhlouta, Zaraket, & Harkous, 2012; Zaraket & Makhlouta, 2012a, 2012c). Sarf provides NLP application developers with an application programming interface (API) to control and refine morphological analysis on the fly. The developer implements the interfaces in the application. Sarf calls the interfaces on control points such as prefix, stem, suffix, and full solution matches. The Sarf API allows the application to (1) control and prioritize the analysis, (2) refine the solution features, and (3) define developer categories and associate them with existing morphemes.

Sarf is a significant extension of the work in (Zaraket & Makhlouta, 2012b). It represents Arabic affixes as agglutinative affix morphemes with fusional affix concatenation rules. Simpler agglutinative affix morphemes can be concatenated to form a more complex affix (Vajda, 2001). Fusional affix concatenation rules specify affix pairs and use regular expressions in as substitution rules to compose the resulting orthographic and semantic tags from the tags of the original morphemes (Spencer, 1991). The Sarf substitution rules are in sync with rules and examples on morpheme concatenative properties from Arabic morphology textbooks (AlRajehi, 2000a, 2000b). This representation resolves consistency, maintenance, and segmentation issues of the current approaches in BAMA and SAMA. Sarf also provides the option to use partial diacritics in disambiguating the morphological solutions of a partially diacritized word.

Sarf makes the following additional contributions:

- Sarf provides an application customizable morphological analyzer where the developer can control and refine the analysis.
- Sarf is a novel Arabic morphological analyzer with agglutinative affixes and fusional affix concatenation rules based on textbook Arabic morphological rules and on the concatenation rules of existing analyzers.
- Sarf solves inconsistencies in existing affix lexicons of BAMA and SAMA.
- Sarf solves the correspondence between the morphological solution and the morphological segmentation of the original text problem.
- Sarf is fully implemented and available online as an open source tool.[1]

We evaluated Sarf for segmentation correspondence, lexicon size, lexcion cosistency, accuracy, and runtime efficiency. Our results show that Sarf lexicons are smaller than lexicons of existing analyzers, and provide coverage for more morphological solutions. Sarf

[1] http://research-fadi.aub.edu.lb/carla/doku.php?id=sarf

Prefix	Vocalized	Category	Gloss	POS data
ف*f*	فَ*fa*	Pref-Wa	and/so	fa/CONJ+
ي*y*	يَ*ya*	IVPref-hw-ya	he/it	ya/IV3MS+
في*fy*	فَيَ*faya*	IVPref-hw-ya	and/so + he/it	fa/CONJ+ya/IV3MS+
سي*sy*	سَيَ*saya*	IVPref-hw-ya	will + he/it	sa/FUT+ya/IV3MS+
فسي*fsy*	فَسَيَ*fasaya*	IVPref-hw-ya	and/so + will + he/it	fa/CONJ+sa/FUT+ya/IV3MS+
ي*y*	يَ*ya*	IVPref-hmA-ya	they (both)	ya/IV3MD+
في*fy*	فَيَ*faya*	IVPref-hmA-ya	and/so + they (both)	fa/CONJ+ya/IV3MD+
سي*sy*	سَيَ*saya*	IVPref-hmA-ya	will + they (both)	sa/FUT+ya/IV3MD+
فسي*fsy*	فَسَيَ*fasaya*	IVPref-hmA-ya	and/so + will + they (both)	fa/CONJ+sa/FUT+ya/IV3MD+
و*w*	وَ*wa*	Pref-Wa	and	wa/CONJ+
وي*wy*	وَيَ*waya*	IVPref-hw-ya	and + he/it	wa/CONJ+ya/IV3MS+
وسي*wsy*	وَسَيَ*wasaya*	IVPref-hw-ya	and + will + he/it	wa/CONJ+sa/FUT+ya/IV3MS+
وي*wy*	وَيَ*waya*	IVPref-hmA-ya	and + they (both)	wa/CONJ+ya/IV3MD+
وسي*wsy*	وَسَيَ*wasaya*	IVPref-hmA-ya	and + will + they (both)	wa/CONJ+sa/FUT+ya/IV3MD+

Table 2. Partial BAMA v1.2 prefix lexicon

simplifies the lexicon maintenance task, provides better accuracy and improves run time efficiency as compared to existing analyzers.

1.1 Background and motivation

Current morphological analyzers such as BAMA (Buckwalter, 2002), SAMA (Kulick, Bies, & Maamouri, 2010a), Beesley (Beesley, 2001), MADAMIRA (Pasha et al., 2014), and ElixirFM (Smrž, 2007) take as input white space delimited tokens, consider them as words, and enumerate all possible morphological solutions for each word.

For example, given the word وسيلعبونها *wsylʿbwnhā* (and they will play it), an analyzer may return the solution presented in Table 1 with a prefix, a stem, a suffix, and their corresponding POS, transliteration, and gloss tags. The prefix وسيـ can be further segmented into (1) the proclitic و with POS tag `CONJ` and gloss tag `and`, (2) the proclitic سـ with POS tag `FUT` and gloss tag `will`, and (3) the prefix يـ with POS tag `IV3MP` and gloss tag `they(people)`. Similarly, the suffix ونها can be segmented into (1) the suffix ونـ, forming a circumfix with يـ, with POS tag `IVSUFF_SUBJ:MP_MOOD:I` and gloss tag `[MASC.PL.]`, and (2) the enclitic ـها with POS tag `IVSUFF_DO:3FS` and gloss tag `it/them/her`.

The exhaustive enumeration of all solutions may hurt performance and may not be necessary or appropriate in some applications as noted in (Maamouri, Bies, Kulick, Zaghouani,

et al., 2010). Thus the need of a customizable morphological analyzer that adapts to application specific requirements.

The accuracy of morphological analysis suffers from inherent difficulties of the Arabic language such as omitted diacritics and position dependent letter forms. Diacritics, i.e. short vowels, such as fatha (ـَ *a*), damma (ـُ *u*), kasra (ـِ *i*), tanween (i.e. doubled diacritic including ـً *an* , ـٌ *un* , ـٍ *in*), and sokun (ـْ) are almost always omitted in written Arabic text as they can be inferred by human readers. The mark shadda (ـّ) denotes the repetition of the marked character and is also often omitted. Partial diacritics can help disambiguate solutions. Consider the unvocalized word أكل *akl* with nine morphological solutions. Its partially vocalized version أَكِل *akil* has only two solutions; VMF أَكِل *aakil* with gloss `I+trust/put in charge`, and VMF أُكِّل *aukil* with gloss `I+make tired/wear out`.

Analyzers such as BAMA and SAMA ignore partial diacritics while other analyzers such as (Attia & Elaraby Ahmed, 2000; Beesley, 2001; Chaâben Kammoun, Hadrich Belguith, & Ben Hamadou, 2010) make use of the partial diacritics to reduce ambiguity. Sarf provides an option that enables the use of existing diacritics for disambiguation, and considers the diacritics at morpheme boundaries, to generate only the diacritic matching solutions, rather than generating all morphological solutions then filtering them.

Arabic letters have up to four different forms corresponding to their position in a word, i.e, beginning, middle, end, and separate word forms. This allows the phrase الىالمدرسة to be visually recognizable as two separate words الى *ilā* (to) and المدرسة *almdrsh* (the school) without the need of a delimiter space in between. The reason is the first word الى *ilā* ends with ى *ā* a non-connecting letter. These words, referred to as 'run-on' words (Buckwalter, 2004), occur regularly, and greatly increase the difficulty of tokenization.

Concatenative morphological analyzers (Buckwalter, 2002; Kulick et al., 2010a) are based on lexicons of prefixes L_p, stems L_s, and suffixes L_x. As shown in Table 2, each entry in a lexicon includes the morpheme, its vocalized form with diacritics, a *concatenation compatibility category* (CCC) rule, a POS tag, and a gloss tag. Separate CCC rules specify the compatibility of prefix-stem R_{ps}, stem-suffix R_{sx}, and prefix-suffix (circumfix) R_{px} concatenations. The affixes و, and يـ in the before mentioned example are valid standalone prefixes, and can be concatenated to the stem لعب to form ولعب and يلعب, respectively. The L_p and L_x lexicons contain also all final forms of concatenated affixes as shown in Table 2 for sample morphemes.

This is the source of several problems:

- L_p and L_x contain redundant entries which results in maintenance and consistency issues (Kulick, Bies, & Maamouri, 2010b; Maamouri, Kulick, & Bies, 2008).

- Augmenting L_p and L_x with additional morphemes, such as أَ *aa* (the question glottal hamza), may result in a quadratic increase in the size of the lexicons (*Hun-*

spell Manual Page., 2012). The additional morpheme may attach to exiting morphemes. Currently, this results in adding all the resulting morphemes to the BAMA and SAMA lexicons.

- The L_p and L_x lexicons are larger than needed especially that they have to account for several forms of a morpheme with varying diacritics.
- The concatenated forms in L_p and L_x contain concatenated POS and other tags. The alignment and correspondence between the original word and its morphemes with the tags of its morphological solution are essential to the success of NLP tasks such as MT and IE (Lee, Haghighi, & Barzila, 2011; Nasredine, Laib, & Fluhr, 2008). The analysis of the example للقضاء*llqḍā*ʾ , `li/PREP + Al/DET + qaDA'/NOUN`, is segmented into two tokens: `li/PREP` and `Al/DET + qaDA'/NOUN` (Maamouri et al., 2008). The best approximation of the unvocalized entry of each token is ل and القضاء, respectively, with an extra letter ا*ā* . This is not a faithful representation of the original text data and the segmentation does not correspond with that of the input text.

Alternatively, Sarf represents only atomic affix morphemes in the lexicons and generates compound affixes from the atomic ones using agglutinative and fusional rules. In this case, Sarf requires only five atomic affix morphemes and five prefix-prefix rules that form compound affixes from atomic ones to represent the entries of Table 2. Finally, we evaluated Sarf with several case studies from the applications that used it with its API (Jaber & Zaraket, 2013; Makhlouta et al., 2012; Zaraket & Makhlouta, 2012a, 2012c). Our results show that Sarf performs better than existing Arabic morphological analyzers in terms of running time, and the application case studies show the efficiency and the utility of the Sarf application customizable API.

The rest of this paper is structured as follows. In Section 2, we present an overview of Sarf. In Section 3, we present the interface provided by Sarf for the application developer to control and refine the morphological analysis. In Section 4, we present our method to build agglutinative affix morphemes with fusional affix concatenation rules. In Section 5, we present our method of using partial diacritics to reduce the morphological ambiguity. In Section 6, we compare Sarf to related work. Finally, we discuss the use of Sarf in multiple NLP tasks and present the results of comparing Sarf to other systems in terms of speed, accuracy, and consistency in Section 8.

2 Overview

The flow diagram in Figure 1 illustrates the components of Sarf. The stem lexicon of Sarf L_s extends the lexicon of Buckwalter (Buckwalter, 2002) with proper and location names extracted from different online sources as well as biblical sources. [2] The fusional and agglutinative affix rules encode the morpheme concatenation and affix/stem compatibility rules. The use of diacritics to disambiguate morphological solutions is optional. The *construct Sarf structures* process takes as input the lexicon, the fusional and agglutinative

[2] http://alasmaa.net/ , http://ar.wikipedia.org/, Genesis 4:17-23; 5:1-32; 9:28-10:32; 11:10-32; 25:1-4, 12-18; 36:1-37:2; Exodus 6:14-25; Ruth 4:18-22; 1 Samuel 14:49-51; 1 Chronicles 1:1-9:44; 14:3-7; 24:1; 25:1-27:22; Nehemiah 12:8-26; Matthew 1:1-16; Luke 3:23-38

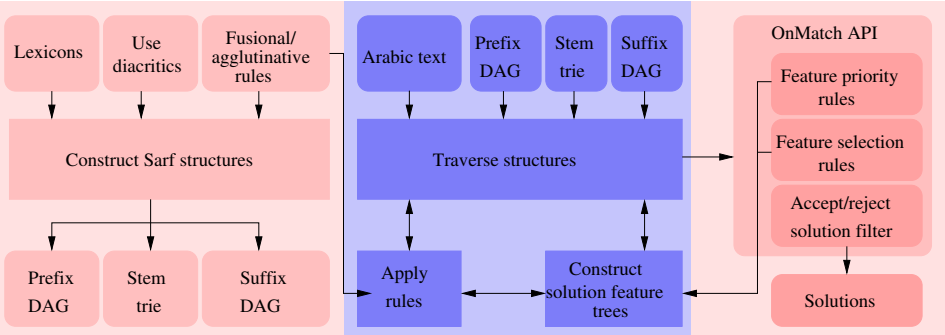

Fig. 1. Sarf flow diagrams: construction and traversal of morpheme structures, and generation of solutions.

rules, and the use diacritic option and constructs directed acyclic graph (DAG) structures that encode the affixes, and a root index trie structure that encodes the stems (Aoe, 1989).

The *traverse structures* process reads the user-provided Arabic text in question one character at a time and traverses the Sarf structures accordingly. The traversal produces a sequence of morphemes each with its morpheme features, applies the fusional and agglutinative rules on each morpheme, checks for concatenation compatibility, and constructs morphological solution feature trees. The construction of the solution trees calls the OnMatch API at every morpheme match (control point) and takes into account the feature priority and selection rules defined by the NLP application developer. A feature priority rule is an order on features that is followed when constructing the morphological solution tree. Features with higher priority appear at higher levels in the solution tree. A feature selection rule defines the morphological features to be included in the morphological solution tree; other features are ignored in the construction of the solution. The traversal reports the constructed solution trees to a filter specified by the developer that either accepts or rejects the reported solutions.

2.1 Sarf structures

Sarf represents affix lexicons and rules using directed acyclic graphs as shown in Figure 2. This provides compactness as well as linear time traversal with respect to the input text. Sarf represents stems in an efficient double array trie structure (Aoe, 1989) to benefit from the common sub-strings. In contrast, the Buckwalter analyzer considers all possible substrings and looks them up in affix hash tables, and performs several hash lookups in the stem hash tables in the order of all possible partitions of the input string. Sarf saves a binary version of the affix and stem structures which allows a fast loading time and only regenerates them if one of the lexicons, and agglutinative and fusional rules are modified.

The diagram in Figure 2 illustrates the Sarf data structures. Subfigures (a), (b), and (c) in Figure 2 represent $\mathcal{P}$ the prefix DAG, $\mathcal{S}$ the stem trie, and $\mathcal{X}$ the suffix DAG, respectively. Boxes and circles denote morpheme versus non-morpheme nodes, respectively. The edges between nodes are labeled with input letters. Morpheme nodes contain the morphological features of the matching morpheme. These features include the VMF, gloss, POS, and compatibility information for morpheme concatenation.

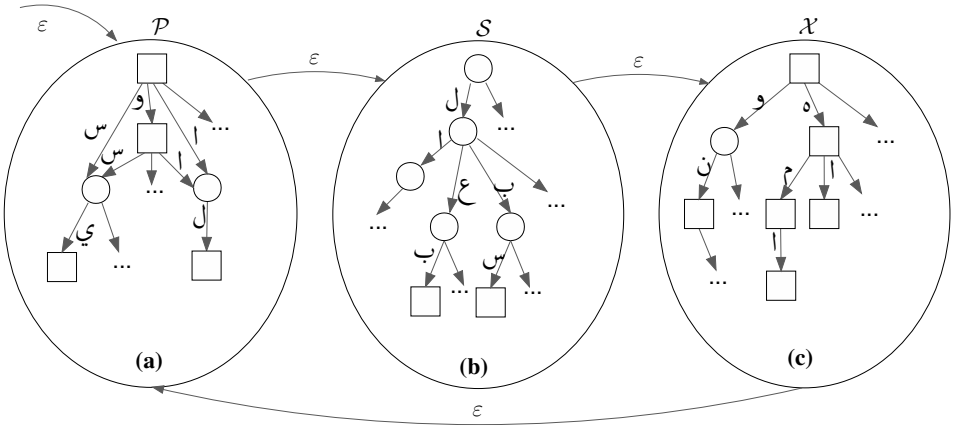

Fig. 2. Example affix DAGs and stem trie.

When we reach a morpheme node in $\mathcal{P}$, $\mathcal{S}$, or $\mathcal{X}$, we proceed with the traversal in the next data structure. We use the symbol ϵ to refer to this transition. Invalid moves from a current node given an input letter denote the absence of a valid solution through this traversal path.

2.2 Solution construction

Figure 3 illustrates the analysis of the input وأكله*w-ʾklh* . The traversal results in two sequences of morpheme nodes where each sequence refers to a valid segmentation. The first sequence is ⟨ و*w* ,أ*ʾ*,كل*kl* ,ه*h* ⟩ and the second sequence is ⟨و*w* ,أكل*ʾkl* ,ه*h* ⟩.

Sarf calls the developer-defined API at each morpheme node and uses the feature selection and priority rules to construct the solution tree of each morpheme. Each path from the root of a morpheme solution tree to one of its leaves is a morpheme solution path. Figures 3(a) to (e) show the constructed solution feature trees of the morphemes و*w* , أ*ʾ* , كل*kl* , ه*h* , and أكل*ʾkl* , respectively. For this example, the API selects the gloss, POS, and VMF features with decreasing priority. Figure 3(f) illustrates an alternative solution feature tree of the morpheme أكل*ʾkl* with POS at highest priority followed VMF and gloss. The comparison between the solution trees (e) and (f) shows how the priority rules can lead to the construction of smaller trees depending on the application at hand. For example, the solution feature tree (f) is more efficient if the developer is interested in the POS first.

The set of valid morphological solutions is composed of the solution paths that match the prefix-prefix, prefix-stem, stem-suffix, suffix-suffix, and prefix-suffix compatibility rules. For example, the first paths (paths to the leftmost leaves) of each of the solution trees in Figure 3(a), (b), (c), and (d) are compatible. The resulting morphological solution has the prefix و*w* with POS tag `CONJ+`, gloss tag `and`, and VMF tag و*wa* , the prefix أ*ʾ* with POS tag `IVIS+`, gloss tag `I`, and VMF tag أ*ʾa* , the stem كل with POS tag `VERB_IMPERFECT`, gloss tag `trust/put in charge`, and VMF tag كل*kil* , and the suffix ه with POS tag `PVSUFF:DO:3MS`, gloss tag `him/it`, and VMF ه*ho* .

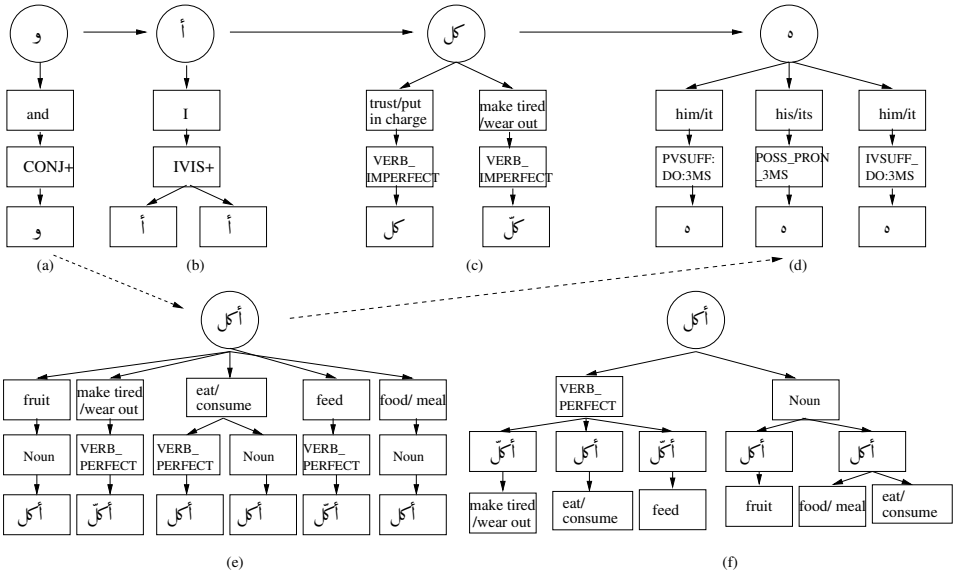

Fig. 3. Sample solution feature trees.

2.3 Running example

The diagram in Figure 2 illustrates a running example of Sarf and how it implements agglutinative and fusional affixes, and handles 'run-on' words. The traversal Φ corresponding to parsing the string اللّاعبونوسيلعبها *wsylʿb-h-ā 'l-lāʿbwn* [3] starts at the root square node in $\mathcal{P}$ which is a morpheme node. The ϵ-edge connects $\mathcal{P}$ to $\mathcal{S}$ and proceeds from any morpheme node in $\mathcal{P}$ to the root node in $\mathcal{S}$. The ϵ-edge that connects $\mathcal{S}$ to $\mathcal{X}$ follows the same behavior.

When there are two valid moves such as و*w* and ϵ in the start case, Φ spawns an exact copy of itself Ψ. Φ proceeds with و*w* in $\mathcal{P}$, and Ψ moves to $\mathcal{S}$ through ϵ. Each of Φ and Ψ represent a valid analysis path so far. A traversal path (Φ or Ψ) dies when it reaches an invalid move. In our example, if there were no stems that start with the letter و*w* , Ψ will die. In reality, Ψ will die when the input is at وسيلع*wsylʿ*. The affix DAGs allow agglutinative and fusional affixes. So for example, Φ will reach a morpheme node through و*w* and could proceed to $\mathcal{S}$. But when س*s* follows, we move to another node in $\mathcal{P}$ corresponding to the prefix وس*ws* . In fact, a valid traversal in $\mathcal{P}$ will process the sub-string وسي*wsy* as a prefix before moving to $\mathcal{S}$. The same traversal behavior applies to $\mathcal{X}$.

Consider Φ after it consumed وسي*wsy* and transitioned into the root node of $\mathcal{S}$. Now Φ will traverse with العب*lʿb* to reach a morpheme node. Before moving with the letter ب*b* to the morpheme node, Φ needs to make sure that the stem العب*lʿb* is compatible with the prefix وسي*wsy* . Sarf keeps compatibility category values as part of the accept nodes. Thus

[3] و س ي ل ع ب ه ا ا ل ل ا ع ب و ن in separate form to ease following the example

each morpheme node in Figure 2 represents more than one concrete node. If the category of العب*lEb* is compatible with the category of وسي*wsy* then Φ moves to a morpheme node. Otherwise, it moves to a regular node or dies.

Since Φ is now in a morpheme node in $\mathcal{S}$, it continues to traverse $\mathcal{S}$ and spawns Ξ which moves to the root of $\mathcal{X}$. After Φ consumes ـهـ*h* , it dies since there is no ـهـ-edge from the current node in $\mathcal{S}$. Ξ consumes ـهـ*h* and reaches a valid analysis morpheme node.

A Sarf traversal considers a full word at any morpheme node in $\mathcal{X}$ and continues the traversal using an ϵ transition to the root node of $\mathcal{P}$. This solves the 'run-on' words problem. Consider the was no space between words اللّاعبون and وسيلعبها. The traversal will transition to the root of $\mathcal{P}$ when the word وسيلعبها is fully consumed, and then the traversal of اللّاعبون will resume. As for the other transitions from $\mathcal{X}$ morpheme nodes to the root of $\mathcal{P}$ before the completion of وسيلعبها, they will result in dead traversals and they will not be reported. Sarf reports a valid traversal, including the morphological solutions of the 'run-on' words, when it reaches text delimiters, such as white space and punctuation, with valid segmentation of the input string.

3 Application specific api

Sarf provides the developer with an application programming interface (API) that allows to (1) define developer categories and associate them with specific morphemes. (2) provide rules that prioritize and filter the solution features, and (3) control and refine the morphological analysis on the fly at solution *control points*. The control points are (1) agglutinative prefix matches, (2) stem matches, (3) agglutinative suffix matches, and (4) full solution matches. The developer implements interfaces that Sarf calls at the control points. The developer processes the solution features provided at the control point and returns a value that instructs Sarf to (1) proceed with the analysis ignoring the current solution, (2) accept the solution and continue considering other solutions, and (3) accept the solution and stop the analysis. The developer can inspect at each control point the agglutinative morphemes and their compatibility category tags, the VMF, gloss, POS, lemma, and the developer-defined category tags.

Consider the task that aims to detect words with possible VERB POS tags. The API can be implemented to reject the analysis at the stem control point if the POS tag is not a VERB. This prevents the analyzer from computing insignificant full morphological solutions at early stages of the analysis. The developer can also use the feature selection filter API to disregard features such as VMF, gloss, and categories. This reduces the size of the solution trees and their corresponding traversal time. For example, given the word أكل*kl* , with gloss tags such as 'he/it eat' and 'fruit', Sarf typically returns nine possible morphological solutions with different VMF, POS tags, and gloss tags. With the feature selection filter API, Sarf considers only two solutions with two solutions with the `VERB_PERFECT` and `NOUN` POS tags.

Sarf also provides the developer with the ability to alter the structure of the morpholog-

Category 1	Category 2	Resulting Category	Substitution rules
NPref-Bi	NPref-Al	NPref-BiAl	
NPref-Li	NPref-Al	NPref-LiAl	
			*r//*ال||ال\\
Pref-Wa	none of { Pref-0, NPref-La, PVPref-La}	$2	
IVPref-li-	IVPref-*-y*	IVPref-(@1)-liy(@2)	*d//he/||him/*\\, *d//they||them*\\ . . . *d//(+2)|| to*\\

Table 3. Example rules from R_{pp}

ical solutions so that the traversal of the solutions is application specific. The developer can provide a factorization order of the solution features using the API priority rules. Sarf uses the priority rules to build the structures. This allows the developer to dismiss analysis earlier at the control points. Consider the task with an aim is to detect adjectives with positive sentiment. Since the POS feature is more limited than the gloss feature, it might give priority to the POS over the gloss. Hence, it can filter invalid solutions early in the analysis.

Moreover, Sarf enables the application developer to define categories and associate them with existing morphemes. Consider the task of detecting words that indicate family relations such as ابنibn (son), أبb (father), and أمm (mother). The developer can define a category called 'family connections' and associate it with the stems ابنibn , أبb , أمm or with their relevant glosses. The user defined categories are attached to the tags in the morpheme nodes as auxiliary tags that can be looked up in constant time.

4 Agglutinative and fusional morphemes

Sarf considers three types of affixes:

- *Atomic affix morphemes* such as يـ can be affixes on their own and can directly connect to stems using the R_{ps} and R_{sx} rules.
- *Partial affix morphemes* such as سـ can not be affixes on their own and need to connect to other affixes before they connect to a stem.
- *Compound affixes* are concatenations of atomic and partial affix morphemes as well as other smaller compound affixes. They can connect to stems according to the R_{ps} and R_{sx} rules.

Sarf forms compound affixes from atomic and partial affix morphemes using newly introduced prefix-prefix R_{pp} and suffix-suffix R_{xx} concatenation rules.

Sarf considers L_p and L_x to be lexicons of atomic and partial affix morphemes associated with their tags. Sarf forms agglutinative affixes using prefix-prefix R_{pp} and suffix-suffix R_{xx} concatenation or agglutination rules. A rule $r \in R_{pp} \cup R_{xx}$ takes the compatibility category tags of affixes a_1 and a_2 and checks whether they can be concatenated. If

so, the rule takes a_1 and a_2 and their tags and generates the affix $a = r(a_1, a_2)$ with its associated tags according to substitution rules based on regular expressions. The rules are fusional in the sense that they modify the orthography and the semantics of the resulting affixes by more than simple concatenation.

We illustrate this with the example rules in Table 3. Row 1 presents a simple rule that allows the concatenation of prefixes with category `NPref-Bi` such as بِ*bi-* and كَ*ka-* to prefixes with category `NPref-Al` such as ال, the result is the compound prefix with category `NPref-BiAl`. Since no substitution rule is specified, the tags of the resulting prefix are simple concatenations.

Row 2 presents a rule that takes prefixes with category `NPref-Li` such as لِ*li-* and prefixes with category `NPref-Al` such as ال. The substitution rule replaces the الـ with لـ resulting in للـ. The syntax of the substitution rule for the affix form is `r//(substring)||(replacement)\\`.

The rule in the Row 3 states that prefixes of category `Pref-Wa` can be concatenated with prefixes with categories that are neither of `Pref-0`, `NPref-La`, and `PVPref-La` categories. The resulting category is denoted with $2 which means the category of the second prefix.

Row 4 illustrates the use of the wild character '`*`' to capture sub-strings of length zero or more in the second category, and refers to the captured sub-strings in the resulting category using the '`@`' operator. The '`@`' operator is always followed by a number that denotes the captured '`*`' expression. Row 4 has also an example of substitution rules for the gloss (description) tag that start with the letter d. The `+2` pattern in the last substitution rule means that the `to` partial gloss description should be appended after the gloss of the second affix. Substitution rules for POS tags start with the letter p.

4.1 Building R_{pp} and R_{xx}

Our method is in line with native Arabic textbooks on morphology and syntax (AlRajehi, 2000a, 2000b; Mosaad, 2009) where only atomic and partial affixes are introduced. The textbooks also list rules to concatenate the affixes and discuss the syntax, semantic, and phonological forms of the resulting affixes. For example, the fourth rule in Table 3 is derived from a textbook rule that states `IVPref-li-` prefixes connect to all imperfect verb prefixes and transform the subject pronoun in the gloss to an object pronoun.

The method built the rules in four steps:

1. In the first step, we encoded textbook morphological rules into patterns.
2. In the second step, we inspected the BAMA and SAMA affix lexicons and extracted the atomic and partial affixes from them.
3. Then, we grouped the rest of the BAMA and SAMA affixes into the rules we collected from the textbooks.
4. We refined the rules wherever necessary, and we grouped rules that shared the same patterns.

	Affix	Vocalized	Inconsistent tag
a) missing plus in **gloss** tag of **prefix**	فل*fl*	فِل*fali*	and/so [+] for/to + the
	و بال*wbāl*	و بال*wabiāl*	and + with/by [+] the
b) missing alternative **gloss** in **prefix**	ف*f*	ف*fa*	and/so
	فب*fb*	فب*fabi*	and [/so] + with/by
	فك*fk*	فك*faka*	and [/so] + like/such as
c) gender/number qualifier omitted in **gloss** of subject **suffix**	ن*n*	ن*n*	they [fem.pl.] <verb>
	نهم*nhm*	نهم*nahom*	they [fem.pl.] <verb> them
	ات*āt*	ات*āt*	[fem.pl.]
	اتك*ātk*	اتك*ātka*	[fem.pl.] your
d) also in **gloss** of object **suffix**	انهما*ānhmā*	انهما*ānihimā*	them (both)
	تكم*tkm*	تُكُم*tkum*	it/they/she <verb> you (pl.)
e) different ways to express them (both) in **gloss** of **suffix**	تهما*thmā*	تُهُما*thumā*	it/they/she <verb> them (both)
	اهما*āhmā*	اهُما*āhumā*	we <verb> (both of) them (both)
	ناهما*nāhmā*	ناهُما*nāhumā*	we <verb> (both of) them (both)
f) '.' omitted after pl in **gloss**	م*m*	مُم*um*	you [masc.pl.] <verb>
	ونا*wnā*	ونا*uwnā*	you [masc.pl[.]] <verb> us
g) **POS** tag is not same as vocalized	ته*th*	تِه*thi*	+ti/PVSUFF_SUBJ:2FS
			+ hu/ [hi/] PVSUFF_DO:3MS

Table 4. Sample BAMA inconsistencies

We validated our work by generating all possible affixes and compared them against the BAMA and SAMA affix lexicons. The comparison resulted in discovering the BAMA and SAMA inconsistencies listed in Tables tables 4 and 5.

4.2 Redundancy

Consider the partial lexicon of prefixes in Table 2. The first five rows can be replaced with three atomic affix morphemes and one partial affix morpheme in L_p and three rules to generate compound morphemes in R_{pp}. Representing prefix يا*ya* (them/both) required four entries, three of them only differ in their dependency on the added يا*ya* . Representing prefix و*w* required the addition of five entries. With Sarf, the equivalent addition of يا*ya* (them/both) requires only two rules in R_{pp} and the addition of و*w* requires only one additional entry in L_p. The difference is much larger when we consider the full lexicon as will be shown in Section 7.

4.3 Inconsistencies

The entries in Tables tables 4 and 5 list examples of the 197 and 208 inconsistencies detected in the affix lexicons of BAMA version 1.2 and SAMA version 3.1, respectively. We

	Affix	Vocalized	Inconsistent tag
a) missing standalone alef with no hamza **prefix** forms	*ا*ā	*آ*ā	I
	*سا*sā	*سَأُ*sa'au	I
b) additional by in **gloss**	*و*w	*وَ*wa	with
	*وال*wā-	*والـ*waā-	with /by + the
c) additional space in vocalized form	*ﻑ*f-	*فـ*fa-	
d) wrong **prefix**	*وف*wf-	*وفـ*wafa-	and + so/and
e) missing definite indicator in **suffix gloss**	*آت*āt	*آتِ*āti	[fem.pl.] + [def.acc.]
	*اتك*ātk	*اتِكِ*ātiki	[fem.pl.] + [def.acc.] + your [fem.sg.]
	*آتك*ātk	*آتِكِ*ātiki	[fem.pl.] + [def. acc.] + your [fem.sg.]
f) omitted gender/num qualifier in **gloss**	*اك*āk	*اكِ*āki	we [verb] + you [fem.sg.]
	*نهم*nhm	*نَهُم*nahum	they [fem.pl.] [verb] + them
g) different ways to express them (both) in **gloss** of **suffix**	*تهما*thmā	*تُهُما*athumā	it/they/she [verb] them (both)
	*اهما*āhmā	*اهُما* āhumā	we [verb] (both of) them (both)
	*ناهما*nāhmā	*ناهما*nāhomā	we [verb] (both of) them (both)
h) leftover BAMA style tags in **gloss**	*كم*km	*كُم*kum	he/it [verb] + you [masc.pl.]
	*تكم*tkm	*تُكُم*tukum	I [verb] + you (pl.) [masc.pl.]
	*كن*kn	*كُنَّ*ukuna	I [verb] + you (women) [fem.pl.]
i) indicative **gloss** with jussive **POS**	*نا*nā	*نا*nā	IVSUFF_MOOD:J, [ind.] [jus.] + us
	*ني*ny	*نِي*niy	IVSUFF_MOOD:J, [ind.] [jus.] + me
j) omitted '.' in **gloss**		*ـُ*u	[def.nom.]
	*ه*h	*هُ*uhu	[def.nom ·]
	*تكن*tkn	*تُكُنَّ*atukuna	[fem.sg.] + [def.nom ·] + your [fem.pl ·]
k) shadda inconsistent in **POS**	*ي*y	*يَ*ya	ya/POSS_PRON_1S
	*ي*y	*يّ*iyya	. . . + ∼a/ ya/ POSS_PRON_1S

Table 5. Sample SAMA inconsistencies

found a small number of these inconsistencies manually and we computed the full list via comparing L_p and L_x with their counterparts computed using our agglutinative affixes. Most of the inconsistencies are direct results of partially redundant entries with erroneous tags. We note that SAMA corrected several BAMA inconsistencies, but also introduced several new ones when modifying existing entries to meet new standards and when introducing new entries.

The following describes the BAMA inconsistencies illustrated in Table 4:

(a) L_p omits a plus (+) symbol that indicates boundaries in compound prefixes.

(b) L_p omits the (so) alternative gloss that corresponds to *ﻑ*f- in several compound prefixes.

(c) L_x omits gender and number qualifiers that appear within within square brackets from several glosses of subject suffixes.

(d) L_x omits gender and number qualifiers from several glosses of subject suffixes that appear within parenthesis.

(e) L_x expresses the dual quantifier as 'them (both)' in the majority of the entries, and as '(both of) them' in several entries.

(f) L_x omits the dot ('.') symbol from the gloss abbreviation of plural.

(g) L_x contains POS tags that are not consistent with the semantics of the vocalized tags for compound affixes.

The following describes the SAMA inconsistencies illustrated in Table 5:

(a) L_p misses entries for Alef prefixes with omitted hamza or madda due to relaxed writing standards which are common in many documents. This is resolved in SAMA for standalone Alef prefixes via preprocessing tokens and flipping all forms of Alef into one form. We report it here since compound prefix entries with Alef are all listed, and the standalone prefixes are available but commented out.

(b) L_p contains an additional erroneous alternative gloss for*wa-* in only one compound prefix*wa-* ; while correctly not included elsewhere.

(c) L_p contains stray spaces in the vocalized tags of one of the *وَfa-* alternatives.

(d) L_p contains an entry that supports the concatenation of*wa-* and *وَfa-* conjunctions. This entry is erroneous and is illegal in Standard Arabic.

(e) L_x omits the definite indicator in the gloss of several suffixes.

(f) L_x omits gender and number qualifiers that appear within square brackets from the gloss tags of several suffixes.

(g) L_x expresses the dual quantifier as 'them (both)' in the majority of the entries, and as '(both of) them' in several other entries.

(h) L_x contains number indicators in the gloss tags still expressed in BAMA style.

(i) L_x contains entries with an *indicative* gloss mood and a *jussive* POS mood.

(j) L_x omits dot ('.') for the abbreviation of plural in several gloss tags.

(k) L_x represents a repeated consonant by a shadda in the POS tag where it should not. SAMA POS tags should spell out the repeated consonants if each belongs to an one In SAMA, the repeated consonant (of the shadda) is spelled out whenever the consonants has its separated partial POS tag.

In addition, 53 BAMA and 27 SAMA minor differences exist between L_p and L_x of BAMA and SAMA and their counterparts computed using our agglutinative affixes. For example, the BAMA gloss tags for prefixes that contain 'bi/PREP' report 'with/by' in some entries and its reverse 'by/with' in others. In addition, we detected several entries in L_p of SAMA with no category compatibility rules in R_{ps}, R_{sx}, and R_{px}.

5 Diacritics

Diacritics are short vowels that are often omitted in Arabic text and inferred by readers from context. Their omission adds to the ambiguity problem of Arabic morphological analysis. The diacritics ـَ *(fatha)*, ـُ *(damma)*, and ـْrepresent and appears above the letter. 'a' vowel, 'o' vowel, and consonant, respectively, and appear above the letter. The diacritic ـِ *i* represents a 'y' vowel and appears below the letter. The diacritics ـً *an* , ـٌ *un* , and ـٍ *in* represent the 'a', 'o', and 'y' vowels followed by a phonetically stressed نْ *n* consonant.

The shadda ـّ mark is not a diacritic but is treated typographically as one, and is also often omitted in Arabic text. It denotes a repeated letter, first as consonant, and second as vocalized. Arabic forbids two consonant diacriticized letters to follow each other.

Analyzers such as BAMA and SAMA ignore input partial diacritics because they consider them to be (1) rare in common corpora, and (2) unreliable because of dialect diversity and human errors (Attia, 2006; Elkateb et al., 2006). However, the work in (Attia & Elaraby Ahmed, 2000; Beesley, 2001; Chaâben Kammoun et al., 2010) considers partial diacritics to decrease morphological ambiguity. We inspected the ATB v3.2 corpus (Maamouri & Bies, 2004) for diacritics and we found that 1.364 percent of the words were partially diacriticized and those diacritics eventually reduced morphological ambiguity. Hence, we decided to provide an option with Sarf that enables the use of existing partial diacritics in text to eliminate morphological solutions that are not in agreement with the partial diacritization.

Key to partial diacritic analysis is a diacritic-aware consistency check that replaces standard string matching checks. The `Diacritic-aware consistency check` algorithm takes as input two words w_1 and w_2. It checks that the sequence of non-diacritic letters, ignoring the diacritics between them, are equal. It also checks that all sequences of diacritics occurring between non-diacritic letters are consistent. Two sequences of diacritics are consistent iff:

1. Both are equal, or
2. One of the sequences is empty, or
3. If one has a shadda, then the other has no sukoun, or
4. If one has a shadda and the other has no shadda, then then the rest of the diacritics are compared recursively.

Table 6 illustrates the `diacritic-aware consistency check` as compared to the standard string comparison with an example. Part (a) shows two diacritic-consistent words أَكّل *aakl* and أَكَل *kal* as a fatha ـَ *a* is compatible with an empty diacritic and a shadda ـّ is compatible with a fatha ـَ *a* . Part (b) illustrates two inconsistent diacritizations since ـِ *i* is incompatible with ـَ *a* next to the letter ك *k* .

	word	string comparison					diacritic-aware consistency check				
(a)	أَكّلaakl	أ	ـَa	ك k	ـّ	ل l	أ	ـَa	ك k	ـّ	ل l
	↕	↕	↕	↕	↕	?	↕	↗	↗	↗	
	أَكَلakal	أ	ك k	ـَa	ل		أ	ك k	ـَa	ل	
(b)	أُكِلaukil	أ	ـُu	ك k	ـِi	ل l	أ	ـُu	ك k	ـِi	ل l
	↕	↕	↕	↕	↕	?	↕	↗	↗	↗	
	أَكَلakal	أ	ك k	ـَa	ل		أ	ك k	ـَa	ل	

Table 6. Arabic string comparison with consideration of partial diacritics

	Sarf	SAMA	ElixirFM	MADA+ TOKAN	MADAMIRA	Beesley	Fassieh
Application customizable	✓	-	-	-	-	-	-
Feature selection	✓	-	-	-	✓	-	-
Run-on words	✓	-	-	-	-	-	-
Partial diacritics	✓	-	-	-	-	✓	-
Affix segmentation	✓	-	functional	tokenization schemes	statistical	-	-
Root-Pattern	-	-	✓	-	-	✓	✓
Automated disambiguation	-	-	-	SVM	SVM	-	maximum aposteriori

Table 7. Comparison of Sarf with SAMA, ElixirFM, MADA+TOKAN, MADAMIRA, Beesley, and Fassieh

6 Related work

In this section, we review work related to Arabic morphological analyzers, segmentation correspondence, partial diacritics, and application specific analyzers.

Table 7 summarizes the comparison between Sarf and related Arabic morphological analyzers. Only ElixirFM (Smrž, 2007), Beesley (Beesley & Karttunen, 2003), and Fassieh (Attia, Rashwan, & Al-Badrashiny, 2009) provide root-pattern analysis of the stem. ElixirFM, MADAMIRA (Pasha et al., 2014), and MADA+TOKAN (Habash, Rambow, & Roth, 2009) are based on BAMA and SAMA and use functional and statistical techniques to address the segmentation problem by reverse engineering the multiple tags of the affixes. Sarf differs in that the segmentation is an output of the morphological analysis and not a reverse engineering of the multi-tag affixes. Sarf is the only analyzer that addresses the 'run-on words' problem and solves it while performing the analysis. MADA+TOKAN, MADAMIRA, and Fassieh apply morphological disambiguation using support vector ma-

chines (SVM), and *maximum a posteriori* (MAP) estimation, respectively. Beesley and Sarf consider partial diacritics to eliminate morphological solutions that are not in agreement with the partial diacritization. Sarf provides an application customizable analyzer that enables the developer to control and refine the analysis on the fly and filter the solution features. MADA+TOKAN and MADAMIRA provides partial control over the output, and not the analysis, where MADA+TOKAN allows the user to selected from several segmentation schemes and MADAMIRA enables the user to select solution features.

Sarf builds upon the lexicon of Buckwalter(Buckwalter, 2002). SAMA is an updated version of BAMA with increased lexicon coverage and additional POS tags (Maamouri, Graff, Bouziri, Krouna, & Kulick, 2010). Sarf differs from Buckwalter and SAMA in that it defines agglutinative and fusional affixes using a shorter list of affixes and a list of concatenation compatibility rules that allow prefix-prefix and suffix-suffix concatenations. This allows Sarf to better maintain the morphological tags associated with the affixes.

Buckwalter(Buckwalter, 2002) and SAMA (Maamouri, Bies, Kulick, Zaghouani, et al., 2010) produce a set of segmentation solutions for a word, compute the morphological solutions for each segment, compute the product of the solutions, eliminates the incompatible solutions, and then reports the valid solutions. Sarf traverses the affix and stem structures with the input word character by character and keeps a stack of morpheme nodes. When a morpheme node in a structure is met, it is checked for compatibility with the stack of nodes. Consequently, Sarf generates only the solutions with valid segmentation, and reports only those with compatible stem and affix concatenation.

SAMA was refined to interact with the ATB (Maamouri & Bies, 2004) project after the addition of a large new corpus. The algorithmic changes in SAMA were done manually and worked in integration with the ATB format. Our API approach allows for customizable refinements similar to the refinements of SAMA (Maamouri, Bies, Kulick, Zaghouani, et al., 2010) and allows Sarf to interact with any application on the fly without the modification of the morphological engine itself.

Like ElixirFM (Smrž, 2007), Sarf builds on the lexicon of the Buckwalter analyzer. Sarf also uses deterministic parsing with tries and DAGs to implement the affix and stem structures. We think that the inferential-realizational approach of ElixirFM that is highly compatible with the Arabic linguistic description (Badawi, Carter, & Gully, 2004) can benefit from many features unique to the Arabic language. Sarf leaves implementing that to the developer customization through the API since in several cases the NLP application that uses the morphological analyzer needs only a partial linguistic model of the Arabic language.

MADA+TOKAN (Habash et al., 2009) is a toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming, and lemmatization. Sarf performs all those tasks except for morphological disambiguation where MADA uses SVM. In Sarf, there is no need for a separate segmentor such as TOKAN since each solution keeps a stack of positions that partition text into morphemes.

MADAMIRA is a tool for Arabic morphological analysis and disambiguation that is based on the general design of MADA, an Arabic morphological analyzer and disambiguator, with additional components inspired by AMIRA (Pasha et al., 2014), a language independent SVM based analyzer. MADAMIRA improves upon the two systems and returns information selectively upon the request of the user. Sarf provides means for the

developer to control and adapt the morphological analysis according to application needs. Moreover, the API enables the developer to implement high level applications such as NER which is provided by AMIRA.

Beesley (Beesley & Karttunen, 2003) compiles Xerox rules into specialized finite state machine (FSM) based morphological analyzer. The number of machines generated by a compiler for Xerox rules can not be controlled by the developer of the analyzer, and the composition of the FSMs into a single framework is a difficult task (Beesley, 2001). Consequently the efficiency of the resulting analyzer depends on the way the Xerox rules are written. Writing application specific Xerox grammars and rules, or modifying the existing ones, requires deep knowledge and insight from the NLP application developer in compilation techniques, context free grammars, and morphological analysis. Sarf constructs a framework of efficient structures that encode the stems and the agglutinative and fusional affixes, respectively. Sarf also provides an application customizable API that allows the developer to control the analysis. Doing the equivalent with Beesley requires the modification of the Xerox rules and the recompilation of the analyzer. Unlike Sarf, Beesley provides a root-pattern analysis of the stem.

Fassieh is a commercial Arabic text annotation tool that enables the production of large Arabic text corpora (Attia et al., 2009). The tool supports Arabic text factorization including morphological analysis, POS tagging, full phonetic transcription, and lexical semantics analysis in an automatic mode. Unlike Sarf, Fassieh provides morphological disambiguation and root-pattern analysis. However, Fassieh does not provide segmentation of the affix and reports it as a whole unit. This tool is not directly accessible to the research community and requires commercial licensing. Sarf differs in that it is an open-source application customizable tool that solves the affix segmentation and 'run-on words' problems.

The work in (Attia, Toral, Tounsi, Pecina, & van Genabith, 2010) addresses the detection of Arabic Multi-word Expressions (MWE). They define MWEs as 'idiosyncratic interpretations that cross word boundaries or spaces'. Sarf adopts a similar approach for specific entities such as person names, and place names.

Several researchers stress the importance of correspondence between the input string and the tokens of the morphological solutions. Some work uses POS tags and a syntactic morphological agreement hypothesis to refine syntactic boundaries within words (Lee et al., 2011). The work in (Grefenstette, Semmar, & Elkateb-Gara, 2005)(Nasredine et al., 2008) uses an extensive lexicon with 3,164,000 stems, stem rewrite rules (Darwish, 2002), syntax analysis, proclitics, and enclitics to address the same problem. Parallel traversal of the input string and the tokens of the morphological solution, while accounting for all possible SAMA normalizations, partially solves the problem as reported in (Kulick et al., 2010b; Maamouri et al., 2008). Later notes in the documentation of the ATB (Maamouri, Bies, Kulick, Krouna, et al., 2010) indicate that extensive manual work is still required and that later versions may drop the input tokens. (Lee et al., 2011) uses syntactic analysis to resolve the same problem.

The survey in (Al-Sughaiyer & Al-Kharashi, 2004) compares several morphological analyzers. Analyzers such as (Khoja, 2001)(Darwish, 2002) target specific applications in the analyzer itself or use a specific set of POS tags as their reference. Sarf differs in that it is a general morphological analyzer that reports all possible solutions. It is application

Reason	**ATB**	**TOKAN**	**Frequency** (percent)
Dropping diacritics	أميركياً*āmyrkyāan*	اميركيا*āmyrkyā*	1.456
Hamza normalization	أنقرة*anqrh*	انقرة*ānqrh*	2.799
Other normalizations	مغادرتـه*mġādrt+h*	مغادرة ـه*mġādrt+h*	5.450
Removing letters	لكنـني*lkn+ny*	لكنـي*lkn+y*	0.034
Adding letters	لـلتحقيق*l+lthqyq*	لـالتحقيق*l+ālthqyq*	1.318
Total			**11.058**

Table 8. ATB-TOKAN segmentation disagreement examples

customizable in the sense that the API is used to control and prioritize the analysis, refine the solution features, and associate morphemes with developer-defined categories.

The work in (Attia & Elaraby Ahmed, 2000; Beesley, 2001; Chaâben Kammoun et al., 2010) considers partial diacritics and perform morphological disambiguation by filtering the full morphological solutions and excluding inconsistent ones. This approach constructs several solutions that will be excluded later. Sarf considers diacritic consistency at the morpheme level instead of the final solution level. It checks for diacritic consistency between the input morpheme and the candidate VMF features at every accept node during the traversal of Sarf structures. Sarf analysis proceeds with the consistent VMFs and terminates the inconsistent ones.

(Beesley, 2001), (Chaâben Kammoun et al., 2010), and (Attia & Elaraby Ahmed, 2000) present analyzers that consider partial diacritics for morphological disambiguation. They filter the output morphological analyses based on compatibility with input diacritics if found. Sarf differs in that it considers the diacritics at morpheme boundaries to generate only the diacritic matching solutions, rather than generating all morphological solutions then filtering them.

7 Results

In this section we present and discuss the results of evaluating Sarf and compare it to existing morphological analyzers such as BAMA, SAMA, MADA+TOKAN, and ElixirFM.

We compared the segmentation capabilities of Sarf to that of SAMA and MADA+TOKAN under the ATB v3.1. We also evaluated the presence of the BAMA and SAMA inconsistencies of Tables 4 and 5 in the ATB v3.1 Part 3. Results show that the annotations of Sarf nearly perfectly agree with the manual annotations of the ATB v3.1. The results also show that Sarf can automatically correct 0.76 percent of the ATB annotations due to lexicon inconsistencies.

Moreover, we evaluated the efficiency of Sarf by measuring the size of the lexicons, lexicon augmentation cost, and the runtime and accuracy performance of the analyzer.

Conflict	Δ_{word} (percent)	$\Delta_{analysis}$ (percent)
POS	0.206	0.016
Gloss	8.317	3.226
WaFa	0.251	0.022
Total	**8.774**	**3.264**

Table 9. Effect of lexicon inconsistencies.

We compared the cost of augmenting Sarf with the question clitic (hamza أ) to that of BAMA and SAMA. We also conducted two experiments to evaluate the performance of Sarf compared to SAMA and ElixirFM. The experiments show the advantage of Sarfover the other analyzers in terms of performance.

7.1 Segmentation correspondence

We evaluated the segmentation correspondence capabilities of Sarf under the segmentation guidelines of the ATB v3.1. For each entry in the 'before' section of ATBv3.1 Part 3 with a correct SAMA solution, we automatically computed the segmentation using Sarf and compared the result to the segmentation in the 'after' entries that are manually validated by the LDC.

SAMA had a correct morphological solution for 273,618 words out of 393,201 ATB words. Those required later segmentation and manual validation. Our automatically generated segmentation agree with 99.991 percent with the oracle ATB segmentation. We inspected the 25 entries for which our segmentation disagreed with ATB and found that both segmentations were valid. For example, the entry مِنّا*mnā* is formed of من `min/PREP` (from) and نا `nA/PRON_1P` (us). Our segmentation is مِن+نَا while that of ATB is مِنّا. When the morphemes are concatenated, the two نـ consonants can be fused into a single one with a shadda (ـّـ). Since it is common to omit the shadda, we are left with a single consonant at the boundary, which can correspond to either morpheme.

When we performed the same experiment using TOKAN toolkit of the MADA+TOKAN, we got a total of 88.942 percent agreement with ATB. When analyzing the inconsistent instances, we noticed that TOKAN disregarded input diacritics. It also performed its segmentation based on the POS tags of the morphological solutions in a similar approach to that mentioned in (Maamouri et al., 2008). Table 8 shows examples of the disagreement instances.

Since Sarf preserves correspondence when performing segmentation, it is capable of

	$\|L_p\|$	$\|R_{pp}\|$	$\|L_x\|$	$\|R_{xx}\|$	$\Delta_L^{\text{ء}}$	$\Delta_R^{\text{ء}}$
BAMA	299	–	618	–	295	–
Agglutinative	70	89	181	123	1	32
With fusional	43	89	146	128	1	32
With grouping	41	7	146	32	1	1
SAMA	1325	–	945	–	1,296	–
Agglutinative	107	129	221	188	1	38
With fusional	56	129	188	194	1	38
With grouping	53	18	188	64	1	1

Table 10. Lexicon size comparison

generating a vocalized tag in the ATB 'after' dataset, which carries more information in 15.47 percent of the time than the counterpart POS derived vocalized entry. The vocalized entry in the 'after' dataset was dropped because of maintenance and segmentation issues. With Sarf that entry can be maintained.

7.2 Lexicon consistency

We evaluated the presence of the inconsistencies of Tables tables 4 and 5 in the ATBv3.1 Part 3. The first experiment considered the ATB entries that adopted the SAMA solution. The rest of the entries have manually entered solutions. The gloss inconsistencies affect 0.76 percent of those entries.

The second experiment considered all tokens in the ATB with a SAMA solution. The Δ_{word} column of Table 9 reports the ratio of the affected ATB words, and the $\Delta_{analysis}$ columns reports the ratio of the conflicting morphological analyses. One word might have several morphological solutions which explains the difference. The rows report the effect of the POS and gloss tags, and that of the wrong prefix entry d of Table 5. In total 8.774 percent of the words and 3.264 percent of the morphological solutions are affected. Sarf automatically solves all these conflicts.

7.3 Lexicon size.

The $|L_p|$, $|L_x|$, $|R_{pp}|$, and $|R_{xx}|$ entries in Table 10 report the sizes of the affix lexicons and the number of concatenation rules of BAMA and SAMA. The entries also report the effect of using agglutinative affixes and fusional rules on reducing the size. Sarf only requires 226 and 323 entries to represent the 917 and the 2,270 entries of BAMA and SAMA affixes with inconsistencies corrected, respectively. The transition from SAMA to BAMA required the addition of 1,353 entries to the lexicons of SAMA. Sarf only required the addition of one order of magnitude less entries to accommodate an equivalent change. The 136 entries consisted of 12 more entries in L_p, 42 in L_x, 18 rules in R_{pp}, and 64 in R_{xx}.

Augmentation. The question clitic, denoted by the glottal sign (hamza أ *ʾa*), is missing in BAMA and SAMA as noted by (Attia, 2006). The Δ_L and Δ_R entries in Table 10 show the difference in the number of additional affixes and rules needed to accommodate for the addition of the question clitic. Our method only requires the addition of one atomic affix and one fusional concatenation rule. Whereas BAMA and SAMA need 295 and 1,296 additional entries to their lexicons, respectively, Moreover, the process requires manual intervention with a possibility of inducing inconsistencies in the process. This is evidence that our method is better for the consistency and the maintenance of the lexicons.

7.4 Performance

We evaluated the accuracy and runtime efficiency of Sarf and compared that to SAMA and ElixirFM with one of the applications (Zaraket & Makhlouta, 2012a) that used Sarf as a back-end morphological analyzer. The hadith extraction application (Zaraket & Makhlouta, 2012a) is concerned with analyzing a book of traditions related to prophet Mohammad through a chain of narrators. Narrators are identified by composite proper person names that are connected with family connectors. For example, the chain of narrators حدّثنَا قتيبة بن سعيد عن جرير *ḥddṯnā qtybh bn sʿyd ʿn ǧryr* , starts with the first narrator قتيبة بن سعيد *qtybh bn sʿyd* where قتيبة *qtybh* and سعيد *sʿyd* are `proper person names`. The word بن *bn* (son of) indicates a parental relation that we will refer to as `family-connectors`. The name جرير *ǧryr* is a proper person name denoting the second narrator, and the words حدّثنَا *ḥddṯnā* (told us) and عن *ʿn* (from/about) indicate a narration relation and we refer to them as `tell-connectors`. Key to narrator detection are morphological features that point to places such as names and location prepositions.

Table 11 reports the results of detecting morphological features that define proper names, tell-connectors, and family connectors and compares the results with SAMA and ElixirFM in terms of accuracy and running time. The table considers three books of hadith selected arbitrarily (Al Kulayni, 1996; Al Tousi, 1995; Ibn Hanbal, 2005)[4]. All experiments used a Linux operating system running on a dual core 2.66 Ghz 64-bit processor with 4GB of memory.

Sarf scored higher recall for all the features and approximately similar precision across the three books. The precision and recall measures of the family connectors in Sarf and SAMA are close, unlike ElixirFM which reports lower accuracy measures. After analyzing the results, it turned out that ElixirFM misses some gloss tags such as the 'son' tag associated with the stem 'بن *bn* '. Sarf produces significantly higher proper name recall

[4] We obtained the digitized books from online sources such as http://www.yasoob.com/ and http://www.al-eman.com/.

		Al Kafi			Al Istibsar			Ibn Hanbal		
		Sarf	SAMA	ElixirFM	Sarf	SAMA	ElixirFM	Sarf	SAMA	ElixirFM
Proper Names	precision	0.36	0.36	0.42	0.38	0.35	0.37	0.53	0.55	0.64
	recall	0.95	0.83	0.77	0.96	0.81	0.73	0.98	0.79	0.76
Tell connectors	precision	0.86	0.85	0.84	0.91	0.9	0.92	0.95	0.93	0.95
	recall	0.99	0.99	0.99	0.99	1	1	1	1	1
Family connectors	precision	0.91	0.90	0.78	0.91	0.91	0.77	1	1	0.97
	recall	1	1	0.41	1	1	0.42	1	1	0.69
Total	precision	0.51	0.52	0.56	0.57	0.56	0.55	0.69	0.72	0.78
	recall	0.97	0.92	0.77	0.98	0.92	0.74	0.99	0.90	0.83
Time	(secs)	1.32	6.65	2.78×60^2	1.31	4.55	2.3×60^2	0.096	0.66	29.2×60

Table 11. Comparison of Sarf to SAMA and ElixirF using the hadith application

		Temporal	Hadith	Biography	Genealogy
	Words	125,010	18,047,732	14,710,064	21,385
Without API	Solutions	4.33	5.41	5.61	4.63
	Time (secs)	12.45	45.21×60	133.17×60	2.89
With API	Solutions	1.74	1.82	2.12	2.44
	Time (secs)	2.39	22.64×60	31.77×60	0.41

Table 12. Morphological solutions per word ratio and runtime gains with the customizable Sarf API utility.

measure compared to SAMA and ElixirFM. This mainly due to augmenting the stem lexicon of Sarf with proper names as explained in Section 2.

Sarf outperformed both SAMA and ElixirFM in running time even without the use of the feature priority and the feature selection API. SAMA performed better than ElixirFM.

7.5 Sarf API

We evaluated the application customizable Sarf API with several applications that used Sarf. Table 12 reports the number of morphological solutions per word and the runtime of Sarf for several applications that use the application customizable Sarf API to refine the analysis and compares that with the same applications without the use of the Sarf API. The numbers show the utility of the Sarf API at improving the runtime and the efficiency of NLP applications by an order or magnitude across the four applications.

In what follows, we shortly describe the applications. Detailed results including accuracy measures can be found in the corresponding papers.

The temporal entity extraction application uses finite state machines driven by morphological features that indicate temporal units, intervals, quantities, and prepositions as input (Zaraket & Makhlouta, 2012c). The application processed 43 articles arbitrarily selected from local newspapers. The average number of solutions that Sarf reported per word without the use of the API was 4.33 solutions per word. The application specific refinements of the analysis implemented using the Sarf API eliminated solutions that the application is not interested in and thus the number of solutions per word that Sarf ended up reporting was 1.74 and that in turn resulted in a substantial improvement in runtime.

The hadith segmentation and extraction application extracts chains of narrators from hadith books using finite state machines that take morphological features such as gloss and POS tags that indicate proper names, family relations, tell-connections, places, and possessive nouns as input (Zaraket & Makhlouta, 2012a). The application processed a total of 41 books with a total of 196,171 narrations to build a graph where narrators are nodes and their relation to each other are edges. Similarly to above, the number of solutions per word improved from 5.41 to 1.82 and the runtime improved by more than half.

The biography application matches a narrator extracted from the hadith application to corresponding biographies in biography books and extracts entities such as birth and death dates, location, students, professors, and authentication qualifications (Zaraket & Makhlouta, 2012a). The biography application uses narrator extraction and temporal extraction, and in addition it uses morphological features that indicate qualifying adjectives that relate to authenticity. The application used the graph generated from the hadith application and processed 15 books of biographies with a total of 79,946 biographies to (1) segment the biographies, (2) extract the narrators in the biographies, and (3) annotate the narrator nodes in the graph with qualifiers extracted from the corresponding biographies. The use of the Sarf API improved the solutions per word ratio from 5.61 to 2.12 and improved the run time by almost a factor of 4.

The genealogy extraction application extracts a family tree and learns words that indicate family relations from biblical texts. It uses morphological features that indicate proper names, family relations, places, and professions (Makhlouta et al., 2012). The application processed the book of Genesis with fifty verses. 21,385 words. The use of the Sarf API improved the solutions per word ratio from 4.63 to 2.44 and improved the run time substantially.

8 Conclusion

This paper presents Sarf, an application customizable Arabic morphological analyzer. NLP applications can implement the Sarf API to to refine the morphological analysis on the fly by (1) selecting the interesting features, (2) prioritizing the features, and (3) accepting and rejecting solutions on the fly based on partial features reported so far. Sarf represents affixes using agglutinative and fusional morphemes which (1) significantly reduces the size of the lexicons needed to represent Arabic morphology,(2) fixes inconsistencies in morphological features corresponding to morphemes, (3) simplifies the maintenance and augmentation of the affix morpheme lexicons, and (4) solves the segmentation correspondence problem between the morphological solution and the original text. Sarf also allows the NLP appli-

cation to use partial diacritics for morphological solution disambiguation, and solves the 'run-on words' problem. Sarf is available online as an open source tool.

In the future, we plan to improve Sarf by allowing root analysis of stems, supporting inflectional stems, and providing a graphical user interface to allow the users to edit the affix and stem morpheme lexicons of Sarf.

References

Al Kulayni, M. (1996). *Kitab al-kafi*. Taaruf.

Al Tousi, M. (1995). *Al istibsar*. Taaruf.

AlRajehi, A. (2000a). *The morphological practice/at-tatbiq assarfi* (first ed.). Renaissance (An-nahda).

AlRajehi, A. (2000b). *Semantical practice/at-tatbiq alnahawi* (first ed.). Renaissance (nahda).

Al-Sughaiyer, I., & Al-Kharashi, I. (2004). Arabic morphological analysis techniques: a comprehensive survey. *American Society for Information Science and Technology*, *55*(3), 189–213.

Aoe, J.-i. (1989). An efficient digital search algorithm by using a double-array structure. *IEEE Transactions on Software Engineering*, *15*(9), 1066–1077.

Attia, M. (2006). An ambiguity-controlled morphological analyzer for modern standard arabic modelling finite state networks. In *The challenge of arabic for nlp/mt conference.* The British Computer Society.

Attia, M., & Elaraby Ahmed, M. (2000). *A large-scale computational processor of the arabic morphology* (Unpublished master's thesis). Faculty of Engineering.

Attia, M., Rashwan, M., & Al-Badrashiny, M. (2009). Fassieh, a semi-automatic visual interactive tool for morphological, pos-tags, phonetic, and semantic annotation of arabic text corpora. *Audio, Speech, and Language Processing, IEEE Transactions on*, *17*(5), 916–925.

Attia, M., Toral, A., Tounsi, L., Pecina, P., & van Genabith, J. (2010). Automatic extraction of Arabic multiword expressions. In *Proceedings of the workshop on multiword expressions: from theory to applications (mwe 2010)* (pp. 18–26). Beijing, China: Association for Computational Linguistics.

Badawi, E., Carter, M., & Gully, A. (2004). *Modern written Arabic: A comprehensive grammar*. New York: Routledge.

Beesley, K. (2001). Finite-state morphological analysis and generation of Arabic at xe-

rox research: Status and plans. In *Workshop proceedings on Arabic language processing: Status and prospects* (pp. 1–8). Toulouse, France.

Beesley, K., & Karttunen, L. (2003). *Finite-state morphology: Xerox tools and techniques*. Stanford: CSLI.

Benajiba, Y., Rosso, P., & Benedíruiz, J. (2007). Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational linguistics and intelligent text processing* (pp. 143–153). Springer.

Buckwalter, T. (2002). *Buckwalter Arabic morphological analyzer version 1.0* (Tech. Rep.). LDC catalog number LDC2002L49.

Buckwalter, T. (2004). Issues in Arabic orthography and morphology analysis. In *Semitic '04: Proceedings of the workshop on computational approaches to Arabic script-based languages* (pp. 31–4). Morristown, NJ, USA.

Chaâben Kammoun, N., Hadrich Belguith, L., & Ben Hamadou, A. (2010). The morph2 new version: A robust morphological analyzer for arabic texts. In *Jadt 2010: 10th international conference on statistical analysis of textual data.*

Darwish, K. (2002). Building a shallow Arabic morphological analyzer in one day. In *Proceedings of the acl-02 workshop on computational approaches to semitic languages.*

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., & Alkhalifa, M. (2006). Arabic wordnet and the challenges of arabic. In *Proceedings of arabic nlp/mt conference, london, uk.*

Grefenstette, G., Semmar, N., & Elkateb-Gara, F. (2005). Modifying a natural language processing system for european languages to treat arabic in information processing and information retrieval applications. In *Acl workshop on computational approaches to semitic languages* (pp. 31–37).

Habash, N. (2010). Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1–187.

Habash, N., Rambow, O., & Roth, R. (2009). Mada+tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In K. Choukri & B. Maegaard (Eds.), *Proceedings of the second international conference on Arabic language resources and tools.* Cairo, Egypt: The MEDAR Consortium.

Habash, N., & Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the north american chapter of the association for computational linguistics(naacl)* (pp. 49–52).

Hunspell manual page. (2012). `http://www.linuxcertif.com/man/4/hunspell`.

Ibn Hanbal, A. (2005). *Musnad ahmad*. Noor Foundation.

Jaber, A., & Zaraket, F. (2013). MATAr: Morphology-based tagger for arabic. In *Aiccsa.* Fes, Morocco.

Khoja, S. (2001). Apt: Arabic part of speech tagger. In *Naacl student research workshop.*

Kulick, S., Bies, A., & Maamouri, M. (2010a). Consistent and flexible integration of morphological annotation in the Arabic treebank. In *Proceedings of the seventh conference on international language resources and evaluation (lrec'10).* Valletta, Malta.

Kulick, S., Bies, A., & Maamouri, M. (2010b). Consistent and flexible integration of morphological annotation in the arabic treebank. In *International conference on language resources and evaluation.* European Language Resources Association.

Lee, Y. K., Haghighi, A., & Barzila, R. (2011). Modeling Syntactic Context Improves Morphological Segmentation. In *Conference on computational natural language learning (conll).*

Maamouri, M., & Bies, A. (2004). Developing an Arabic treebank: methods, guidelines, procedures, and tools. In *Semitic '04: Proceedings of the workshop on computational approaches to Arabic script-based languages* (pp. 2–9).

Maamouri, M., Bies, A., Kulick, S., Krouna, S., Gaddeche, F., & Zaghouani, W. (2010). Arabic treebank: Part 3 version 3.2. In *Linguistic data consortium,ldc2010t08.*

Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D., & Ciul, M. (2010). From speech to trees: Applying treebank annotation to Arabic broadcast news. In *Proceedings of the seventh conference on international language resources and evaluation (lrec'10).* Valletta, Malta.

Maamouri, M., Graff, D., Bouziri, B., Krouna, S., & Kulick, S. (2010). Ldc standard arabic morphological analyzer (sama) v. 3.1. *LDC Catalog No. LDC2010L01. ISBN*.

Maamouri, M., Kulick, S., & Bies, A. (2008). Diacritic annotation in the arabic treebank and its impact on parser evaluation. In *International conference on language resources and evaluation.*

Makhlouta, J., Zaraket, F., & Harkous, H. (2012). Arabic entity graph extraction using morphology, finite state machines, and graph transformations. In *Cicling.*

Mosaad, Z. (2009). *The briefing of morphology* (first ed.). As-Sahwa.

Nasredine, S., Laib, M., & Fluhr, C. (2008). Evaluating a natural language processing approach in arabic information retrieval. In *Elra workshop on evaluation.*

Pasha, A., Al-Badrashiny, M., El Kholy, A., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., & Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the ninth international conference on language resources and evaluation.* Reykjavik, Iceland: European Language Resources Association.

Smrž, O. (2007). Elixirfm: implementation of functional Arabic morphology. In *Semitic '07: Proceedings of the 2007 workshop on computational approaches to semitic languages* (pp. 1–8). Prague, Czech Republic.

Spencer, A. (1991). *Morphological theory: An introduction to word structure in generative grammar* (Vol. 2). Basil Blackwell Oxford.

Vajda, E. J. (2001). *Typology.* `http://pandora.cii.wwu.edu/vajda/ling201/test1materials/typology.htm`. Western Washington University.

Zaraket, F., & Makhlouta, J. (2012a). Arabic cross-document NLP for the hadith and biography literature. In *Flairs.*

Zaraket, F., & Makhlouta, J. (2012b). Arabic morphological analyzer with agglutinative affix morphemes and fusional concatenation rules. In *Coling.* Mumbai, India.

Zaraket, F., & Makhlouta, J. (2012c). Arabic temporal entity extraction using morphological analysis. *IJCLA*, *3*, 121-136.