14

Arabic Morphological Representations for Machine Translation

Nizar Habash
Center for Computational Learning Systems, Columbia University
habash@cs.columbia.edu

Abstract: Arabic has a very rich morphology characterized by a combination of templatic and affixational morphemes, complex morphological rules, and a rich feature system. This complexity makes working with Arabic as a source of target language in machine translation (MT) a challenge for two reasons. First, it is not clear what the right representation is for two reasons. First, it is not clear what the right representation is for Arabic words given a specific MT approach or system. And secondly, there are many MT-relevant resources for Arabic morphology, lexicography and syntax (e.g., morphological analyzers, dictionaries and treebanks) that adopt various representations that are not necessarily compatible with each other. The result is that for MT researchers, there is a need to experiment with and to relate multiple representations used by different resources or components to each other within a single system. In this chapter, we describe different Arabic morphological representations used by MT-relevant natural language processing resources and tools and we discuss their usability in different MT approaches. We also present a common framework for relating different levels of representations to each other

14.1 Introduction

Arabic has a very rich morphology characterized by a combination of templatic and affixational morphemes, complex morphological rules, and a rich feature system. This complexity makes working with Arabic as a source or target language in Machine Translation (MT) a challenge for two reasons. First, it is not clear what the right representation is for Arabic words given a specific MT approach or system. It is not even clear whether the same representation is optimal for every component in an MT system, e.g., word alignment versus decoding in statistical MT or parsing versus structural transfer in symbolic MT. Secondly, there are many MT-relevant resources for Arabic morphology, lexicography and syntax (e.g., morphological analyzers, dictionaries and treebanks) that adopt various representations that are not necessarily compatible with each other. For example, dictionaries use the notion of a *lexeme* that is different from the root/pattern/vocalism and stem/affix representations used by many morphological analyzers. And statistical parsers can be content with a

A. Soudi, A. van den Bosch and G. Neumann (eds.), Arabic Computational Morphology, 263–285.

minimally tokenized inflected undiacritized word as the proper level of representation for Arabic, which is different from input text and also potentially different from later processing steps. The result is that for MT researchers, there is a need to experiment with and to relate multiple representations used by different resources or components to each other within a single system. This challenge has different implications for research in statistical MT, symbolic MT or hybrid approaches to MT.

In this chapter, we describe different Arabic morphological representations used by MT-relevant Natural Language Processing (NLP) resources and tools and we discuss their usability in different MT approaches. We also present a common framework for relating different levels of representations to each other. We motivate the lexeme-and-feature level of representation as a common representation to analyze to. From that representation, we can regenerate to other desirable shallower representation. This framework allows for easy navigation between representations used by different resources. It also allows for exploring the effect of using different representations in MT. The interaction between analysis and generation makes this framework direction-independent, i.e., useful for working with Arabic as a source or target MT language. Finally, we describe and evaluate ALMORGEANA, a large-scale system for analysis and generation from/to the lexeme-and-feature representation. We also discuss how to use it to relate different morphological representations for Arabic.

Section 14.2 introduces different representations in Arabic morphology.[1] Section 14.3 discusses the role of morphological representations in different approaches to MT. Section 14.4 and Section 14.5 describe ALMORGEANA and how it can be used for navigating among different morphological representations, respectively.

14.2 Representations of Arabic Morphology

In discussing representations of Arabic morphology, it is important to separate two different aspects of morphemes: *type* versus *function*. Morpheme *type* refers to the different formal kinds of morphemes and their interactions with each other. A distinguishing feature of Arabic (in fact, Semitic) morphology is the presence of templatic morphemes in addition to affixational morphemes. Morpheme *function* refers to the distinction between derivational morphology and inflectional morphology. These two aspects, type and function, are independent, i.e., a morpheme type does not determine its function and vice versa. This independence complicates the task of deciding on the proper representation of morphology in different NLP resources and tools. This section introduces these two aspects and their interactions in more detail.

[1] Additional discussions of Arabic morphological phenomena are presented in Chapter 3 and in the four chapters in Part 2 of this book. See Chapter 15 for a discussion of Arabic generation in the context of MT.

14.2.1 Morpheme Type: Templatic vs. Affixational

Arabic has three categories of morphemes: templatic morphemes, affixational morphemes, and Non-Templatic Word Stems (NTWS). Templatic morphemes come in three types that are equally needed to create a templatic word stem: roots, patterns and vocalisms. The root morpheme is a sequence of three, four or five consonants (termed *radicals*) that signifies some abstract meaning shared by all its derivations. For example, the words كَتَب *katab* 'to write', كَاتِب *kaAtib* 'writer', and مَكْتُوب *maktuwb* 'written' all share the root morpheme (كتب) *ktb* 'writing-related'. The pattern morpheme is an abstract template in which roots and vocalisms are inserted.[2] For example, the verbal pattern *tV1V22V3* indicates that a non-root consonant (t) is added and that the second root radical is doubled. The vocalism morpheme specifies which vowels to use with a pattern. A word stem is constructed by interleaving a root, a pattern and a vocalism. For example, the word stem كَتَب *katab* 'to write' is constructed from the root كتب *ktb*, the pattern *1V2V3* and the vocalism *aa*.

Arabic affixes can be prefixes such as +سَـ *sa+* 'will/[future]', suffixes such as +ُونَ *+uwna* '[masculine plural]' or circumfixes such as تَـ++ـنَ *ta++na* '[subject imperfective 2nd person feminine plural]'. Some of the affixes are clitics, such as the conjunction +وَ *wa+* 'and', the preposition (+لِـ *li+* 'to/for', and the pronominal object/possessive clitics (e.g. +هَا *+haA* 'her/it/its'). Others are bound morphemes.

Finally, NTWS are word stems that are not derivable from templatic morphemes. They tend to be foreign names (e.g., وَاشِنطُن *waAšinTun* 'Washington').

An Arabic word is constructed by first creating a word stem from templatic morphemes or using a NTWS, to which affixational morphemes are then added. For example, the word وَسَيَكْتُبُونَهَا *wasayaktubuwnahaA* has two prefixes, one circumfix and one suffix in addition to a root, a pattern and a vocalism:

(1) wa+ sa+ y+ [ktb+V12V3+au] +uwna +haA
and+ will+ 3rd+ write +plural +it
'And they will write it'

The process of combining morphemes can involve a number of phonological, morphological and orthographic rules that modify the form of the created word; it is not always a simple interleaving and concatenation of its morphemic components. One example is the feminine morpheme, ة+ *+ħ (ta marbuta)*, which is turned into ت+ *+t* when followed by a possessive clitic: أَمِيرَةُ+هُم *Âamiyraħu+hum* 'princess+their' is realized as أَمِيرَتُهُم *Âamiyratuhum* 'their princess'. Another example is the deletion of the Alif (ا) of the definite article +الـ *Al+* when preceded

[2] In this chapter, numbers, (1, 2, 3, 4, or 5), are used in a pattern to indicate radical position as opposed to the common practice in the literature of using the symbol *C*. The symbol *V* is used to indicate a vocalism position.

by the preposition +لِـ *l*+ 'for'. For example, لِـ+الـ+بَيت *li+Al+bayt* 'for+the+house' is realized as لِلبَيت *lilbayt* 'for the house'. These rules clearly complicate the process of analyzing and generating Arabic words.

14.2.2 Morpheme Function: Derivational vs. Inflectional

The distinction between derivational and inflectional morphology in Arabic is similar to that in other languages. Derivational morphology is concerned with creating new words from other words, a process in which the core meaning of the word is modified. For example, the Arabic كَاتِب *kaAtib* 'writer' can be seen as derived from the root كتب *ktb* the same way the English *writer* can be seen as a derivation from *write*. Although compositional aspects of derivations do exist, the derived meaning is often idiosyncratic. For example, the *masculine* noun مَكْتَب *maktab* 'office/bureau/agency' and the *feminine* noun مَكْتَبَة *maktabaħ* 'library/bookstore' are derived from the root كتب *ktb* 'writing-related' with the pattern+vocalism *ma12a3*, which indicates location. The exact type of the location is thus idiosyncratic, and it is not clear how the nominal gender difference can account for the semantic difference.

On the other hand, in inflectional morphology, the core meaning of the word remains intact and the extensions are always predictable. For example, the semantic relationship between كَاتِب *kaAtib* 'writer' and كُتَّاب *kut~aAb* 'writers' maintains the sense of the kind of person described, but only varies the number. The change in number in this example is accomplished using templatic morphemes (pattern and vocalism change). This form of plural construction in Arabic is often called "broken plural" to distinguish it from the strictly affixational "sound plural" (e.g. كَاتِب+ات *kaAtib+aAt* 'writers [fem]').

Broken plurals are one example highlighting the independence of morpheme type from morpheme function: templatic morphemes can be derivational or inflectional, with the exception of the roots, which are always derivational. Similarly, the majority of affixational morphemes are inflectional but there are some affixational derivational morphemes: the adjective كُتُبِيّ *kutubiy~* 'book-related' is derived from the noun كُتُب *kutub* 'books' using the affix +يِيّ *+iy~*.

14.2.3 Arabic Morphological Representations

Given the variability in the relationship between morpheme type and function in addition to the presence of phonological, morphological, and orthographic adjustment phenomena, there are many ways to represent Arabic words in terms of their morphological units. Table 14.1 illustrates some of these possible representations using the example وَلِكَتَبَتِهِم؟ *walikatabatihim?* 'and for their writers?'.

There are many variations among these different representations: (a.) whether they address inflectional/derivational phenomena, templatic/affixational phenomena

Table 14.1. Morphological representations of Arabic words

Representation	Example	Found where?
Natural Token	wlktbthm?	naturally occurring text
Simple Token	wlktbthm ?	common preprocessing for NLP [29]
Segmentation	wl+ ktb +thm ?	[11, 12]
	w+ l+ ktbt +hm ?	[51, 21]
	w+ l+ ktb +t +hm ?	[40, 39]
Normalized Segmentation	w+ l+ ktbħ +hm ?	Penn Arab Treebank [41, 29]
	w+ l+ ktb +ħ +hm ?	[59, 29]
Templatic Segmentation	w+ l+ ktb+1V2V3aħ+aa +hm ?	[33]
Morphemes and Features	w+/CONJ l+/PREP kataba*ħ* +hm/P:3MP ? ktb&CaCaCa*ħ* w+ l+ +P:3MP ? ktb +PL w+ l+ +GEN +P:3MP ?	[6, 11, 12, 29]
Lexeme and Features	[kAtib w+ l+ PL P:3MP] [?]	ALMORGEANA, [27], dictionaries (lexeme only)

or both, (b.) whether they preserve or resolve ambiguity,[3] and (c.) which degree of abstraction from allomorphs (actual form of morpheme after applying various adjustment rules) they use. And since any subset (or all) of the morphemes can be separated from the word and/or be normalized, there is a very large space of possible specific representations to select from.

The *natural token* refers to the way Arabic words appear in actual text where they are undiacritized and segmented only using white space. Punctuations, for example, could be attached to the word string in this representation. All naturally occurring Arabic text is in this representation. *Simple tokenization* separates punctuation but maintains the morphological complexity of the Arabic word tokens. There is no change in ambiguity compared to the natural token.

Segmentation is the simplest way to dissect an Arabic word. It is strictly defined here to exclude any form of orthographic, morphological or phonological normalization. Segmentation splits up the letters into segments that correspond to clusters of a stem plus one or more affixational morphemes. There are many ways to segment an Arabic word as Table 14.1 shows. Segmentation can select a subset of analyses of a word. For example, segmenting للجنة *lljnħ* into *l+l+jnħ* (*li+l+jan~aħ* 'to Paradise'

[3] This discussion does not address the issue of morphological disambiguation, which is outside the scope of this chapter [26, 54, 30].

or *li+l+jin~aħ* 'to insanity/mania') is selecting a subset of analyses excluding *l+ljnħ* (*li+lajnaħ* 'to a committee' or *li+l~ajnaħ* 'to the committee').

Normalized segmentation abstracts away from some of the adjustment phenomena discussed earlier. In the example in Table 14.1, the form of the segmented word stem is *ktbħ* not *ktbt*. Normalization disambiguates the unnormalized segmented form *ktbt* ('he/she/you[sg.] wrote' or 'writers'). The Penn Arabic Treebank [41] uses a normalized segmentation that breaks up a word into four regions: conjunction, particle, normalized word stem and pronominal clitic.

Templatic segmentation is a deeper level of segmentation that involves normalization by definition. Here, the root, pattern and vocalism are separated. Up to this level of representation, the tokens are driven by a templatic/affixational view of morphology rather than a derivational/inflectional view. The introduction of features at the next level of representation, *morphemes and features*, abstracts away from different morphemes that at an underlying level signify the same feature. For example, The affixational morphemes يـ++ـُونَ *y++uwna*, يـ++ـُوا *y++uwA* and ـُوا+ *+uwA* all realize the third person masculine plural subject for different verb aspect/mood combinations. There are many different degrees to the transition from morphemes to features. A combination of both is often used.

The final representation is *lexeme and features*. The lexeme can be defined as an abstraction over a set of word forms differing only in inflectional morphology. The lexeme itself captures a specific meaning that does not change with inflectional variations. The traditional citation form of a lexeme used in dictionaries is the perfective third person masculine singular for verbs and the singular masculine form for nouns and adjectives. If there is no masculine form, the feminine singular is used. As such, the Lexeme [كَاتِب] [kaAtib] 'writer' normalizes over all the different inflectional forms of كَاتِب *kaAtib* such as كَاتِبَان *kaAtibaAn* 'two writers', كَتَبَة *katabaħ* 'writers', and كَاتِبَة *kaAtibaħ* 'female writer'. Lexemes as opposed to stems provide a desirable level of abstraction that is to a certain degree language independent for applications such as MT. Lexemes are also less abstract than roots and patterns which tend to be too vague semantically and derivationally unpredictable, making them less useful in practice for MT.

The next section discusses how these different levels of representation interact with different MT approaches.

14.3 Morphological Representations for Machine Translation

In statistical approaches to MT, a translation model is trained on word-aligned [46] parallel text of source and target languages [9, 10, 35, 37, 36]. The translation model is then used to generate multiple target language hypotheses from the source language input. The target hypotheses are typically ranked using a log-linear combination of a variety of features [45]. Statistical MT has been quite successful in

producing good quality[4] MT on the genre it is trained on in much faster time than symbolic approaches. For statistical MT, in principle, it doesn't matter what level of morphological representation is used as long as the input is on the same level as the data used in training. Practically however there are certain concerns with issues such as sparsity, ambiguity, language-pair differences in morphological complexity, and training-data size. Shallower representations such as simple tokenization tend to maintain distinctions among morphological forms that might not be relevant for translation, thus increasing the sparsity of the data. This point interacts with the MT language pair: for example, normalizing subject inflections of Arabic verbs when translating to a morphologically poor language like English might be desirable since it reduces sparsity without potentially affecting translation quality. If the target language is morphologically rich, such as French, that would not be the case. This, of course, may not be a problem when large amounts of training data are available. Additionally, transforming the training text to deeper representations comes at a cost since selecting a deeper representation involves some degree of morphological disambiguation, a task that is typically neither cheap nor foolproof [26].

The anecdotal intuition in the field of statistical MT is that reduction of morphological sparsity often improves translation quality. This reduction can be achieved by increasing training data or via morphologically-driven preprocessing [22]. Recent investigations of the effect of morphology on MT quality focused on morphologically rich languages such as Catalan [49], Czech [22], German [43], Serbian [49] and Spanish [34, 49]. These studies examined the effects of various kinds of tokenization, lemmatization and part-of-speech (POS) tagging and showed a positive effect on MT quality.

Specifically for Arabic, Lee [39] investigated the use of automatic alignment of POS-tagged English and affix-stem segmented Arabic to determine appropriate tokenizations of Arabic. Her results show that morphological preprocessing helps but only for the smaller corpora sizes she investigated. As size increases, the benefits diminish. Habash and Sadat [29, 52] reached similar conclusions on a much larger set of experiments including multiple preprocessing schemes reflecting different levels of morphological representation and multiple techniques for disambiguation/tokenization. Two of their techniques used the ALMORGEANA system described later in this chapter. They showed that specific preprocessing decisions can have a positive effect when decoding text with a different genre than that of the training data (in essence another form of data sparsity). They also demonstrated gains in MT quality through combination of different preprocessing schemes. Additional similar results were reported using specific preprocessing schemes and techniques [59, 51, 21, 44].

Research in the use of different morphological representations of Arabic in Example-based MT, a corpus-based approach related to statistical MT [55, 14], is promising, at least in terms of improved coverage of training examples [48].

[4] The question of how to judge the quality of MT, i.e., MT Evaluation, is outside the scope of this chapter. Currently, the most accepted yet still controversial approaches are automatic, e.g., BLEU [47, 13] and METEOR [5].

Finally, the newest addition to research on morphology within phrase-based statistical MT is Moses, a decoder for factored [8] phrase-based translation models. Moses allows using a mix of different levels of morphological representation.[5] At the time of writing this chapter, no work on Arabic factored translation models have been done.

In symbolic approaches to MT, such as transfer-based or interlingual MT, linguistically motivated rules (morphological, syntactic and/or semantic) are manually or semi-automatically constructed to create a system that translates the source language into the target language [20]. Symbolic MT approaches tend to capture more abstract generalities about the languages they translate between compared to statistical MT. This comes at a cost of being more complex than statistical MT, involving more human effort, and depending on already existing resources for morphological analysis and parsing. This dependence on already existing resources highlights the problem of variation in morphological representations for Arabic. In a typical situation, the input/output text of an MT system is in natural or simplified tokenization. But, a statistical parser (such as [16] or [7]) trained out-of-the-box on the Penn Arabic Treebank assumes the same kind of tokenization (4-way normalized segments) used by the treebank. This means that a separate tokenizer is needed to convert input text to this representation [19, 26]. Moreover, the output of such a parser, being in normalized segmentation, will not contain morphological information such as features or lexemes that are important for translation: Arabic-English dictionaries use lexemes and proper translation of features, such as number and tense, requires access to these features in both source and target languages. As a result, additional conversion is needed to relate the normalized segmentation to the lexeme and feature levels. Of course, in principle, the treebank and parser could be modified to be at the desired level of representation (i.e., lexeme and features). But this can be a rather involved task for researchers interested in MT. We are aware of the following published research on Arabic symbolic MT: [4, 53] (within the transfer approach) and [58, 56, 1] (within the interlingua approach). Given the inhibiting costs of building large scale symbolic MT system, they tend to be developed by commercial institutions, which are less inclined to publicize their trade secrets.[6]

Finally, the current hybridization direction in the field of MT is interested in exploring statistical and symbolic combinations of resources and tools within and beyond the level of morphology. Some hybrids rooted in statistical MT include syntactic information as part of the preprocessing phase [17], the decoding phase [50] or the n-best rescoring phase [45]. Such approaches will share challenges relevant to both statistical and symbolic MT when extended to Arabic. A detailed discussion of such challenges are presented in the context of extending a Generation Heavy MT system, a hybrid approach rooted in symbolic MT [23], to Arabic [25].

[5] Moses was developed during the 2006 summer workshop at Johns Hopkins University as an enhancement to Pharaoh [36]. See http://www.clsp.jhu.edu/ws2006/groups/ossmt/ and http://www.statmt.org/moses/ for more details.

[6] Two of the top Arabic MT companies using rule-based MT systems are Apptek (http://www.apptek.com/) and Sakhr (http://www.sakhr.com/).

In the next section, we describe ALMORGEANA (Arabic Lexeme-base MORphological GEnerator/ANAlyzer). ALMORGEANA is a morphological analysis and generation system built on top of the Buckwalter analyzer databases, which are at a different level of representation (3-way segmentation). Being an analysis and generation system, it can be used with MT systems analyzing or generating Arabic. ALMORGEANA relates the deepest level of representation (lexeme and features) to the shallowest (simple tokenization).[7] This wide range together with bidirectionality (analysis/generation) allows using ALMORGEANA to navigate between different levels of representations as will be discussed in Section 14.5. Morphological disambiguation, or the selection of an analysis from a list of possible analyses, is a different task that is out of the scope of this chapter although it is quite relevant to MT [26, 54, 30].

14.4 ALMORGEANA

ALMORGEANA is a large-scale lexeme-based Arabic morphological analysis and generation system.[8] ALMORGEANA uses the databases of the Buckwalter Arabic morphological analysis system with a different engine focused on generation from and analysis to the lexeme-and-feature level of representation. The building of ALMORGEANA didn't just involve the reversal of the Buckwalter analyzer engine, which only focuses on analysis, but also extending it and its databases to be used in a lexeme-and-feature level of representation for both analysis and generation.

The next section reviews other efforts on morphological analysis and generation in Arabic. Section 14.4.2 introduces the Buckwalter analyzer's database and engine. Section 14.4.3 describes the different components of ALMORGEANA. An evaluation of ALMORGEANA is discussed in Section 14.4.4.

14.4.1 Morphological Analysis and Generation

Arabic morphological analysis has been the focus of researchers in natural language processing for a long time. This is due to features of Arabic Semitic morphology such as optional diacritization and templatic morphology. Numerous forms of morphological analyzers have been built for a wide range of application areas from Information Retrieval (IR) to MT in a variety of linguistic theoretical contexts [3, 2, 6, 11, 12, 18, 33, 27].

Arabic morphological generation, by comparison, has received little attention although the types of problems in generation can be as complex as in analysis.

[7] Going to natural tokenization is a trivial step where, for example, punctuation marks are attached to preceding words.

[8] A previous publication about ALMORGEANA focused on the generation component of the system which was named Aragen [24].

Finite-State Transducer (FST) approaches to morphology [38] and their extensions for Arabic such as the Xerox Arabic analyzer [6] are attractive for being generative models. However, a major hurdle to their usability is that lexical and surface levels are very close [32]. Thus, generation from the lexical level is not useful to many applications such as symbolic MT where the input to a generation component is typically a lexeme with a feature list. A solution to this problem was proposed by [32], which involved composition of multiple FSTs that convert input from a deep level of representation to the lexical level. However, there are still many restrictions on the order of elements presented as input and their compatibility.[9] The MAGEAD (Morphological Analysis and Generation for Arabic and its Dialects) system attempts to design an end-to-end lexeme-and-features to surface FST-based system for Arabic [28]. As of the time of the writing of this chapter, MAGEAD's coverage is limited to verbs in Modern Standard Arabic and Levantine Arabic. The only work on Arabic morphological generation that focuses on generation issues within a lexeme-based approach is done by [15, 57]. Their work uses transformational rules to address the issue of stem change in various prefix/suffix contexts. Their system is a prototype that lacks in large-scale coverage.

There are certain desiderata that are expected from a morphological analysis/generation system for any language. These include (1) coverage of the language of interest in terms of both lexical coverage (large scale) and coverage of morphological and orthographic phenomena (robustness); (2) the surface forms are mapped to/from a deep level of representation that abstracts over language-specific morphological and orthographic features; (3) full reversibility of the system so it can be used as an analyzer or a generator; (4) usability in a wide range of natural language processing applications such as MT or IR; and finally, (5) availability for the research community. These issues are essential in the design of ALMORGEANA for Arabic morphological analysis and generation. ALMORGEANA[10] is a lexeme-based system built on top of a publicly available large-scale database, Buckwalter's lexicon for morphological analysis.

14.4.2 Buckwalter Morphological Analyzer

The Buckwalter morphological analyzer uses a concatenative lexicon-driven approach where morphotactics and orthographic rules are built directly into the lexicon itself instead of being specified in terms of general rules that interact to realize the output [11, 12]. The system has three components: the lexicon, the compatibility tables and the analysis engine. An Arabic word is viewed as a concatenation of three regions, a prefix region, a stem region and a suffix region. The prefix

[9] Other work on using FSTs designed for analysis in generation is discussed in [42].

[10] The ALMORGEANA engine can be freely downloaded under an OpenSource license for research purposes from http://www.ccls.columbia.edu/cadim/resources.html. The lexical databases need to be acquired independently from the Linguistic Data Consortium (LDC) as part of the Buckwalter Arabic Morphological Analyzer [11, 12].

وَ/wa	Pref-Wa	*and*
بِـ/bi	NPref-Bi	*by/with*
وَبِـ/wabi	NPref-Bi	*and + by/with*
الـ/Al	NPref-Al	*the*
بِالـ/biAl	NPref-BiAl	*with/by + the*
وَبِالـ/wabiAl	NPref-BiAl	*and + with/bythe*
ة/ap	NSuff-ap	*[fem.sg.]*
تَانِ/atAni	NSuff-atAn	*two*
تَيْنِ/atayoni	NSuff-tayn	*two*
تَاهُ/atAhu	NSuff-atAh	*his/its two*
ات/At	NSuff-At	*[fem.pl.]*

;;1_ كَتَب/katab-u_1		
كَتَب/katab	PV	*write*
كُتُب/kotub	IV	*write*
كُتِب/kutib	PV_Pass	*be written*
كُتَب/kotab	IV_Pass_yu	*be written*
;;1_ كِتَاب/kitAb_1		
كِتَاب/kitAb	Ndu	*book*
كُتُب/kutub	N	*books*
;;1_ كِتَابَة/kitAbap_1		
كِتَاب/kitAb	Nap	*writing*

Fig. 14.1. Some Buckwalter lexical entries

and suffix regions can be null. Prefix and suffix lexicon entries cover all possible concatenations of Arabic prefixes and suffixes, respectively. For every lexicon entry, a morphological compatibility category, an English gloss and occasional Part-Of-Speech (POS) data are specified. Stem lexicon entries are clustered around their specific lexeme, which is not used in the analysis process. Figure 14.1[11] shows sample entries: the first six in the left column are prefixes; the rest in that column are suffixes; the right column contains seven stems belonging to three lexemes. The stem entries also include English glosses which allows the lexicon to function as a dictionary. However, the presence of inflected forms, such as passives and plurals among these glosses makes them less usable as lexemic translations.

Compatibility tables specify which morphological categories are allowed to co-occur. For example, the morphological category for the prefix conjunction وَ/`wa` *wa+* 'and', `Pref-Wa`, is compatible with all noun stem categories and perfect verb stem categories. However, `Pref-Wa` is not compatible with imperfective verb stems because they must contain a subject prefix. Similarly, the stem كِتَاب/`kitAb` *kitaAb* of the the lexeme 1_كِتَاب/`kitAb_1` *kitaAb* 'book' has the category (`Ndu`), which is not compatible with the category of the feminine marker ة/`ap` *aħ*: `NSuff-ap`. The same stem, كِتَاب/`kitAb` *kitaAb*, appears as one of the stems of the lexeme 1_كِتَابَة/`kitAbap_1` *kitaAbaħ* 'writing' with a category that *requires* a suffix with the feminine marker. Cases such as these are quite common and pose a challenge to the use of stems as tokens since they add unnecessary ambiguity.

The analysis algorithm is rather simple since all of the hard decisions are coded in the lexicon and the compatibility tables: Arabic words are segmented into all possible sets of prefix, stem and suffix strings. In a valid segmentation, the three strings exist in the lexicon and are three-way compatible (prefix-stem, stem-suffix and prefix-suffix).

[11] The Buckwalter transliteration is preserved in examples of Buckwalter lexicon entries (see Chapter 2).

14.4.3 ALMORGEANA Components

14.4.3.1 Input/Output

In generation mode, the input to ALMORGEANA is a *feature-set*, a set of lexeme and features from a closed class of inflectional phenomena. The output of generation is one or more word strings in simple tokenization. In analysis mode, the input is the string and the output a set of possible feature-sets. The features in a feature-set include number, gender and case inflections, which do appear in other languages, but also prefix conjunctions and prepositions that are written as part of the word in Arabic orthography. Table 14.2 lists the different features and their possible values.

The first column includes the names of the features. The second and third column list the possible values they can have and their definitions, respectively. The last column lists the default value assigned during generation in case a feature is unspecified based on its type. There are two types of features: obligatory and optional. Obligatory features, such as verb subject or noun number, require a value to be specified. Therefore, in case of under-specification, all possible values are generated. Optional features, such as conjunction, preposition or pronominal object/possessive clitics, on the other hand can be absent. The pronominal features, subject, object and possessive, are defined in terms of sub-features specifying person, gender and number. In case any of these sub-features is under-specified, they are expanded to all their possible values. For example, the subject feature `S:2`, as in the case of the English pronoun 'you' (which is under-specified for gender and number), is expanded to `(S:2MS S:2FS S:2D S:2MP S:2FP)`. If no POS is specified, it is automatically determined by the lexeme and/or features. For example, the presence of a definite article implies the lexeme is a noun or an adjective; whereas a verbal particle or a subject/object implies the lexeme is a verb.[12]

The following is an example of an Arabic word and its lexeme-and-feature representation in ALMORGEANA.

(2) `[kitAb_1 POS:N PL Al+ l+]`
لِلكُتُبِ *lilkutubi*
'for the books'

The feature-set in this example consists of the nominal lexeme `kitAb_1` 'book' with the feature `PL` '*plural*', the definite article `Al+` 'the' and the prefix preposition `l+` 'to/for'.

14.4.3.2 Preprocessing Buckwalter Lexicons

ALMORGEANA uses the Buckwalter lexicon described in Section 14.4.2 *as is*. The lexicon is processed in ALMORGEANA to index entries based on inferred sets of

[12] Other POS not included in Table 14.2 are D *Determiner*, C *Conjunction*, NEG *Negative particle*, NUM *Number*, AB *Abbreviation*, IJ *Interjection*, and PX *Punctuation*.

Table 14.2. ALMORGEANA features

Feature	Value	Definition	Default
Part-of-Speech	POS:N	*Noun*	automatically
	POS:PN	*Proper Noun*	determined
	POS:V	*Verb*	
	POS:AJ	*Adjective*	
	POS:AV	*Adverb*	
	POS:PRO	*Pronoun*	
	POS:P *and others*	*Preposition*	
Conjunction	w+	*'and'*	none
	f+	*'and, so'*	
Preposition	b+	*'by, with'*	none
	k+	*'like'*	
	l+	*'for, to'*	
Verbal Particle	s+	*'will'*	none
	l+	*so as to*	
Definite Article	Al+	*the*	none
Verb Aspect	PV	*Perfective*	all
	IV	*Imperfective*	
	CV	*Imperative*	
Voice	PASS	*Passive*	all
Gender	FEM	*Feminine*	all
	MASC	*Masculine*	
Subject	S:PerGenNum	**Per***son = {1,2,3}*	all
Object	O:PerGenNum	**Gen***der = {M,F}*	none
Possessive	P:PerGenNum	**Num***ber = {S,D,P}*	none
Mood	MOOD:I	*Indicative*	all
	MOOD:S	*Subjunctive*	
	MOOD:J	*Jussive*	
Number	SG	*Singular*	all
	DU	*Dual*	
	PL	*Plural*	
Case	NOM	*Nominative*	all
	ACC	*Accusative*	
	GEN	*Genetive*	
Definiteness	INDEF	Indefinite	all
Possession	POSS	Possessed	all

features values (or *feature-keys*) that are used to map features in the input feature-sets to proper lexicon entries. This task is trivial for cases where the lexicon entry provides all necessary information. For example, verb voice and aspect are always part of the stem: the feature-key for *kutib*, the stem of the passive perfective form of the verb كُتِبَ/`katab` is `katab+PV+PASS`.

Many lexicon entries, however, lack feature specifications. One example is broken plurals, which appear under their lexeme cluster, but are not marked in any way for

plurality (see the entry for كُتُب/kutub in Figure 14.1). Detecting when a stem is plural is necessary to include the feature *plural* in the feature-key for that stem. Using the English gloss to detect the presence of a broken plural is a possible solution. However, it fails for adjectival entries since English adjectives do not inflect for plurality, e.g. كَبِير/kabiyr (SG) and كِبَار/kibAr (PL) are both glossed as '*big*'. Additionally, some sound plural stems in the lexicon are glossed as plurals. The Buckwalter categories are not helpful on their own for this task. For example, the presence of a stem with morphological category N is ambiguous as to being a broken plural or a singular nominalization of a form I verb [11]. The solution for this problem stems from the observation that a singular verbal nominalization is its own *lexeme*, whereas a broken plural is always listed under a lexeme that is in a singular base form. A broken plural is by definition a major change in the form of the lexeme. Therefore, if a stem under a lexeme has the morphological category N, Ndip, or Nap (all of which can mark a broken plural) AND it is **not** a subset string of the lexeme, it is considered a broken plural. This technique works for entries considered part of the same lexeme in the Buckwalter lexicon. Entries that treat a broken plural as a separate lexeme will not be processed correctly, e.g. the lexeme إِخْوَة_ *Ăixwaħ* 'brothers'.

14.4.3.3 Analysis and Generation

Analysis in ALMORGEANA is similar to Buckwalter's analyzer (Section 14.4.2). The difference lies in an extra step that uses feature-keys associated with stem, prefix and suffix to construct a feature-set for the lexeme-and-feature output. In the case of failed analysis, a back-off step is explored where prefix and suffix substrings are sought. If a compatible pair is found, the stem is used as a degenerate lexeme and the features are constructed from the feature-keys associated with the prefix and suffix.

The process of generating from feature-sets is also similar to Buckwalter analysis except that feature-keys are used instead of string sequences. First, the feature-set is expanded to include all forms of underspecified obligatory features, such as case, gender, number, etc. Next, all feature-keys in the ALMORGEANA lexicon that fully match any subset of the expanded feature-set are selected. All combinations of feature-keys that completely cover the features in the expanded feature-set are matched up in prefix-stem-suffix triples. Then, each feature-key is converted to its corresponding prefix, stem or suffix. The same compatibility tables used in Buckwalter analysis are used to accept or reject prefix-stem-suffix triples. Finally, all unique accepted triples are concatenated and output. In the case that no surface form is found, a back-off solution that attempts to regenerate after discarding one of the input features is explored. If the back-off fails, typically due to a missing lexical entry, a baseline Arabic morphological generator is used.

The baseline generator uses a simple concatenative word structure rule and a small lexicon. The lexicon contains 70 entries that map all features to most common surface realizations. For example, FEM maps to (ـة /ap *aħ*, ـت /at, and ϕ) and PL

maps to (ات /At, ِينَ /iyna, ِي /iy, ُونَ /uwna and ُو /uw). Subtleties of feature interaction are generally ignored except for the case of subject and verb aspect since the circumfix realization of subjects in the imperfective/imperative form is rather complex to model concatenatively. The only word structure rule used in the baseline generator is the following:

```
<WORD> ::= (w|f) (s|l|b|k) Al <SubjectAspect>
<Lexeme>
<AspectSubject> <Gender> <Number> <Object> <Possessive>
```

14.4.4 Evaluation

ALMORGEANA uses the databases of the Buckwalter analyzer; therefore, its coverage is equivalent to the coverage of these lexicons. In this section, we evaluate ALMORGEANA engine for analysis and generation only.[13]

A sample text of over one million Arabic words from the UN Arabic-English corpus [31] was used in this evaluation. For each unique word in the text, ALMORGEANA is used in analysis mode to produce feature-sets. The resulting feature-sets are then input to two systems: the complete ALMORGEANA as described earlier *and* the baseline generator used as back-off to ALMORGEANA generation. For each feature-set, there are two sets of words: (a) words that analyze into the feature-set (A words) and (b) words that are generated from the feature-set (G words) (see Figure 14.2). The bigger the intersection between the two sets (C words), the better the performance of a system. Generated words that are not part of the intersection (C words) are Overgenerated words (O words). Words that analyze into the feature-set but are not generated are Undergenerated words (U words). In principle, U words are definite signs of problems in the generation system; whereas, O words can be correct but unseen in the analyzed text.

A system's Undergeneration Error (UnderErr) is defined as the ratio of U words to A words. Overgeneration Error (OverErr) is defined as the ratio of O words to G words. These two measure are equivalent to (1 - Recall) and (1 - Precision) respectively, if the set of A words paired with a feature-set is considered a gold standard to

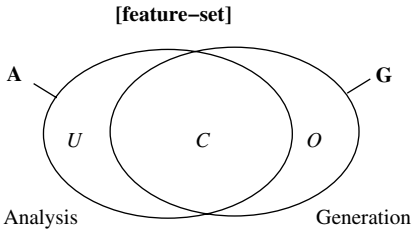

Fig. 14.2. ALMORGEANA evaluation

[13] The evaluation described here was run over the Buckwalter lexicons (version 1) [11].

be replicated in reverse by a generation system. The Combined Undergeneration and Overgeneration Error (CombErr) is calculated as (1 - *the corresponding F-score*):[14]

$$UnderErr = \frac{U}{A} = \frac{A - C}{A}, OverErr = \frac{O}{G} = \frac{G - C}{G},$$
$$CombErr = 1 - (\frac{2 \times (1 - UnderErr) \times (1 - OverErr)}{(1 - UnderErr) + (1 - OverErr)})$$

The evaluation text contained 63,066 undiacritized unique words, which were analyzed into 118,835 unique feature-sets corresponding to 14,883 unique lexemes. The number of unique diacritized words corresponding to the text words is 104,117. The evaluation was run in two modes controlling for the type of matching between A words and G words: diacritized (or diacritization-sensitive) and undiacritized. Evaluation results comparing ALMORGEANA to the baseline are presented in Table 14.3. The baseline system is almost six times faster than ALMORGEANA[15], but it had high undergeneration and overgeneration error rates. Both error rates were reduced in the undiacritized mode, where some erroneous output became ambiguous with correct output. ALMORGEANA, by comparison, reduced the combined error rate from the baseline by over 84%.

Many of the overgeneration errors are false alarms. They include cases of overgeneration of broken plurals, some of which are archaic or genre-specific but correct. For example, the word for 'sheik', شَيخ *šayx*, has three uncommon broken plurals in addition to the common شُيُوخ *šuyuwx*: أَشيَاخ *ÂšyaAx*, مَشَايِخ *mašaAyix*, and مَشَائِخ *mašaAŷix*. Another very common overgeneration error resulted from the underspecification of some mood-specific vocalic verbal suffixes in the Buckwalter lexicon. Arabic hollow verbs, for example, undergo a stem change in the jussive mood (from يَقُول *yaquwl* to يَقُل *yaqul*), which is indistinguishable in the analysis.

Table 14.3. Evaluation results

System	UnderErr	OverErr	CombErr	Time (secs)
ALMORGEANA *diacritized*	0.39%	12.22%	6.68%	1,769
ALMORGEANA *undiacritized*	0.38%	12.42%	6.79%	1,745
Baseline *diacritized*	43.90%	60.99%	53.98%	281
Baseline *undiacritized*	32.84%	47.93%	41.34%	293

[14] I would like to thank Christian Monson for suggesting this formula to computing CombErr. A previously published formula was biased toward underestimating the combined error [24].

[15] The experiments were run on a Dell Inspiron machine with Pentium 4 CPU, 512 MB RAM and 2.66 GHz.

Undergeneration errors stem exclusively from lexicon errors. These are not many and they can be expected in a manually created database. One example is caused by a missing lexeme comment in the Buckwalter lexicon which resulted in pairing all the forms of the verb رَأَى *raÂaý* 'to see' to the lexeme that appears just before it, رَاوَند *raAwand* 'rhubarb'. Such cases suggest a valuable use of ALMORGEANA as a debugging tool for the Buckwalter lexicon.[16]

14.5 Interoperability of Morphological Representations

This section describes how ALMORGEANA can be used to navigate between different levels of morphological representation. An Arabic word in simple tokenization can be analyzed using ALMORGEANA to multiple possible lexeme-and-feature analyses. This automatically gives us access to the lexeme-and-feature level and also the three-way segmentation used by Buckwalter's lexicons. To generate an intermediate representation such as the normalized segmentation used by the Penn Arabic Treebank [41], the features for conjunction, preposition and pronominal object/possessive can be stripped from the lexeme-and-feature analyses. The remaining features and lexeme are then used to generate the word stem using ALMORGEANA to guarantee a normalized form. The stripped features are also trivially generated and positioned relative to the word stem: [conjunction] [preposition] [word-stem] [pronoun]. Table 14.4 shows the different analyses for each word in the sentence. وقد كَاتبته فتحية لمدة سنتين *wqd kAtbth ftHyħ lmdħ sntyn. 'and Fathia continued to correspond with him for two years'*. The correct Penn Arabic Treebank tokenization for this example is و قد كَاتبت ه فتحية . ل مدة سنتين *w qd kAtbt h ftHyħ l mdħ sntyn.*

The ambiguity inherent in both the analysis and generation processes results in multiple possibilities (column 3 in Table 14.4). To select a specific segmentation, any of a set of possible techniques can be used such as rule-based heuristics or language models trained on text in the correct tokenization. For example, in the case of the Penn Arabic Treebank, the already tokenized text of the treebank can be used to build a language model for ranking/selecting among options produced by this technique (similar to [40]). Alternatively, machine learning over the features of the annotated words in the Penn Arabic Treebank can be used to select among the different analyses (similar to [26, 54, 30]).[17]

We developed a general tokenizer, TOKAN, as an implementation of this analyze-then-regenerate approach to tokenization. TOKAN is built on top of ALMORGEANA. TOKAN takes as input (a.) disambiguated ALMORGEANA analyses and (b.) a token

[16] All of the errors described here are for version 1 of the Buckwalter analyzer only [11]. We did not conduct a similar study on version 2 of the Buckwalter analyzer [12].

[17] The Morphological Analysis and Disambiguation for Arabic (MADA) tool [26] is a disambiguation system fully integrated with ALMORGEANA. More information on MADA is available at http://www.ccls.columbia.edu/cadim/resources.html.

Table 14.4. Normalized segmentation example

Word	Analysis	Segments
wqd	[qad~_1 POS:N w+ +SG +MASC gloss:size/physique] [qad_2 POS:F w+ gloss:may/might] [qad_1 POS:F w+ gloss:has/have] [qid~_1 POS:N w+ +SG +MASC gloss:thong/strap]	wqd
	[waq~ad_1 POS:V +PV +S:3MS gloss:kindle/ignite] [waqod_1 POS:N +SG +MASC gloss:fuel/burning] [waqadi_1 POS:V +PV +S:3MS gloss:ignite/burn]	wqd
kAtbth	[kAtib_1 POS:N +FEM +SG +P:3MS gloss:author/writer/clerk] [kAtib_2 POS:AJ +FEM +SG +P:3MS gloss:writing]	kAtb*ħ* h
	[kAtab_1 POS:V +PV +S:3FS +O:3MS gloss:correspond_with] [kAtab_1 POS:V +PV +S:1S +O:3MS gloss:correspond_with] [kAtab_1 POS:V +PV +S:2FS +O:3MS gloss:correspond_with] [kAtab_1 POS:V +PV +S:2MS +O:3MS gloss:correspond_with]	kAtbt h
ftHy*ħ*	[taHiy~ap_1 POS:N +FEM +SG f+ gloss:greeting/salute]	f tHy*ħ*
	[fatHiy~ap_1 POS:PN gloss:Fathia]	ftHy*ħ*
lmd*ħ*	[mud~ap_1 POS:N +FEM +SG l+ gloss:interval/period]	l md*ħ*
sntyn	[sinot_1 POS:N +MASC +DU +ACCGEN gloss:cent] [sanap_1 POS:N +FEM +DU +ACCGEN gloss:year]	sntyn
.	[. POS:PX gloss:.]	.

definition sequence that specifies which features are to be extracted from the word and where they should be placed. For example, the token definition for splitting off the conjunction *w*+ only is `"w+ REST"`. This token definition specifies that the conjunction *w*+ is split from the word and whatever is left (REST) is regenerated after the conjunction *w*+. Similarly, the token definition for the Penn Arab Treebank tokenization is `"w+ f+ l+ k+ b+ REST +O: +P:"`.[18] ALMORGEANA and TOKAN have been used in both statistical and symbolic MT systems [29, 25].

14.6 Conclusions

In this chapter, we described obstacles facing MT researchers when working with Arabic resources in differing morphological representations. The lexeme-and-feature level of representation has been motivated and, ALMORGEANA, a large-scale system for analysis and generation from/to that level has been described and evaluated. We presented a framework using ALMORGEANA for navigating between Arabic

[18] More information on TOKAN is available at http://www.ccls.columbia.edu/cadim/resources.html.

morphological representations. This framework is useful for research exploring the effects of using different Arabic representations in MT.

Acknowledgments

This work has been supported, in part, by Army Research Lab Cooperative Agreement DAAD190320020, NSF CISE Research Infrastructure Award EIA0130422, Office of Naval Research MURI Contract FCPO.810548265, NSF Award #0329163 and Defense Advanced Research Projects Agency Contract No. HR0011-06-C-0023. I would like to thank Owen Rambow, Mona Diab, Bonnie Dorr, Tim Buckwalter, Michael Subotin and Christian Monson for helpful discussions.

References

[1] Azza Abdel-Monem, Khaled Shaalan, Ahmed Rafea, and Hoda Baraka. A Proposed Approach for Generating Arabic from Interlingua in a Multilingual Machine Translation System. In *Proceedings of the 4th Conference on Language Engineering*, pp. 197–206, 2003. Cairo, Egypt.

[2] Imad Al-Sughaiyer and Ibrahim Al-Kharashi. Arabic Morphological Analysis Techniques: A Comprehensive Survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213, 2004.

[3] Muhammed Aljlayl and Ophir Frieder. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. In *Proceedings of ACM Eleventh Conference on Information and Knowledge Management, Mclean, VA*, pp. 340–347, 2002.

[4] Haytham Alsharaf, Sylviane Cardey, Peter Greenfield, and Yihui Shen. Problems and Solutions in Machine Translation Involving Arabic, Chinese and French. In *Proceedings of the International Conference on Information Technology*, pp. 293–297, Las Vegas, Nevada, 2004.

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[6] Kenneth Beesley. Arabic Finite-State Morphological Analysis and Generation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pp. 89–94, Copenhagen, Denmark, 1996.

[7] Daniel Bikel. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *Proceedings of International Conference on Human Language Technology Research (HLT)*, pp. 24–27, 2002.

[8] Jeff A. Bilmes and Katrin Kirchhoff. Factored Language Models and Generalized Parallel Backoff. In *Proceedings of the Human Language Technology Conference/North American Chapter of Association for Computational Linguistics (HLT/NAACL-03)*, pp. 4–6, Edmonton, Canada, 2003.

[9] Peter Brown, John Cocke, Stephen Della-Pietra, Vincent Della-Pietra, Fredrick Jelinek, John Lafferty, Robert Mercer, and Paul Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16:79–85, June 1990.

[10] Peter Brown, Stephen Della-Pietra, Vincent Della-Pietra, and Robert Mercer. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[11] Tim Buckwalter. Buckwalter Arabic Morphological Analyzer Version 1.0, 2002. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2002L49.

[12] Tim Buckwalter. Buckwalter Arabic Morphological Analyzer Version 2.0, 2004. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02, ISBN 1-58563-324-0.

[13] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 249–256, Trento, Italy, 2006.

[14] Michael Carl and Andy Way. *Recent Advances in Example-Based Machine Translation*. Kluwer Academic Publishers, Dordrecht, Holland, 1988.

[15] Violetta Cavalli-Sforza, Abdelhadi Soudi, and Teruko Mitamura. Arabic Morphology Generation Using a Concatenative Strategy. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP 2000)*, pp. 86–93, Seattle, Washington, USA, 2000.

[16] Michael Collins. Three Generative, Lexicalised Models for Statistical Parsing. In *Proceedings of the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL)*, pp. 16–23, Madrid, Spain, 1997.

[17] Michael Collins, Philipp Koehn, and Ivona Kucerova. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 531–540, Ann Arbor, Michigan, 2005.

[18] Kareem Darwish. Building a Shallow Morphological Analyzer in One Day. In *Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pp. 47–54, Philadelphia, PA, USA, 2002.

[19] Mona Diab, Kadri Hacioglu, and Daniel Jurafsky. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *Proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*, pp. 149–152, Boston, MA, 2004.

[20] Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A Survey of Current Research in Machine Translation. In M. Zelkowitz, editor, *Advances in Computers, Vol. 49*, pp. 1–68. Academic Press, London, 1999.

[21] Anas El Isbihani, Shahram Khadivi, Oliver Bender, and Hermann Ney. Morpho-syntactic arabic preprocessing for arabic to english statistical machine translation. In *Proceedings on the Workshop on Statistical Machine Translation*, pp. 15–22, New York City, June 2006. Association for Computational Linguistics.

[22] Sharon Goldwater and David McClosky. Improving Statistical MT Through Morphological Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 676–683, Vancouver, Canada, 2005.

[23] Nizar Habash. *Generation Heavy Hybrid Machine Translation*. PhD thesis, University of Maryland College Park, 2003.

[24] Nizar Habash. Large Scale Lexeme Based Arabic Morphological Generation. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN-04)*, pp. 271–276, 2004. Fez, Morocco.

[25] Nizar Habash, Bonnie Dorr, and Christof Monz. Challenges in Building an Arabic-English GHMT System with SMT Components. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA06)*, pp. 56–65, Cambridge,MA, 2006.

[26] Nizar Habash and Owen Rambow. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 573–580, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[27] Nizar Habash and Owen Rambow. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 681–688, Sydney, Australia, July 2006. Association for Computational Linguistics.

[28] Nizar Habash, Owen Rambow, and George Kiraz. Morphological Analysis and Generation for Arabic Dialects. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages at 43rd Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 17–24, Ann Arbor, Michigan, 2005.

[29] Nizar Habash and Fatiha Sadat. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the 7th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL06)*, pp. 49–52, New York, NY, 2006.

[30] Jan Hajič, Otakar Smrž, Tim Buckwalter, and Hubert Jin. Feature-based Tagger of Approximations of Functional Arabic Morphology. In Ma. Antonia Martí Montserrat Civit, Sandra Kübler, editor, *Proceedings of Treebanks and Linguistic Theories (TLT)*, pp. 53–64, Barcelona, Spain, 2005.

[31] Xu Jinxi. UN Parallel Text (Arabic-English), LDC Catalog No.: LDC2002E15, 2002. Linguistic Data Consortium, University of Pennsylvania.

[32] Lauri Karttunen, Ronald Kaplan, and Annie Zaenen. Two-level Morphology with Composition. In *Proceedings of Fourteenth International Conference on Computational Linguistics (COLING-92)*, pp. 141–148, Nantes, France, July 20–28 1992.

[33] George Kiraz. Multi-tape Two-level Morphology: A Case study in Semitic Non-Linear Morphology. In *Proceedings of Fifteenth International Conference on Computational Linguistics (COLING-94)*, pp. 180–186, Kyoto, Japan, 1994.

[34] Katrin Kirchhoff, Mei Yang, and Kevin Duh. Statistical Machine Translation of Parliamentary Proceedings Using Morpho-Syntactic Knowledge. In *TC-STAR Workshop on Speech-to-Speech Translation*, pp. 57–62, Barcelona, Spain, 2006.

[35] Kevin Knight. A Statistical MT Tutorial Workbook, April 30 1999. http://www.clsp.jhu.edu/ws99/projects/mt/mt-workbook.htm.

[36] Philipp Koehn. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. In *Proceedings of the Association for Machine Translation in the Americas*, pp. 115–124, 2004.

[37] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-based Translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, pp. 127–133, Edmonton, Canada, 2003.

[38] Kimmo Koskenniemi. Two-Level Model for Morphological Analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pp. 683–685, 1983.

[39] Young-Suk Lee. Morphological Analysis for Statistical Machine Translation. In *Proceedings of the 5th Meeting of the North American Chapter of the*

Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04), pp. 57–60, Boston, MA, 2004.

[40] Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. Language Model Based Arabic Word Segmentation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL'03)*, pp. 399–406, Sapporo, Japan, 2003.

[41] Mohamed Maamouri, Ann Bies, and Tim Buckwalter. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2004.

[42] Guido Minnen, John Carroll, and Darren Pearce. Robust, Applied Morphological Generation. In *Proceedings of the 1st International Conference on Natural Language Generation (INLG 2000)*, pp. 201–208, Mitzpe Ramon, Israel, 2000.

[43] Sonja Nieıen and Hermann Ney. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2), 2004.

[44] Franz Josef Och. Google System Description for the 2005 NIST MT Evaluation. In *MT Eval Workshop (unpublished talk)*, 2005.

[45] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. A Smorgasbord of Features for Statistical Machine Translation. In *Proceedings of the Human Language Technology / North American Association of Computational Linguistics Conference*, pp. 161–168, Boston, Massachusetts, 2004.

[46] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–52, 2003.

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, PA, 2002.

[48] Aaron Phillips and Violetta Cavalli-Sforza. Arabic-to-English Example Based Machine Translation Using Context-Insensitive Morphological Analysis. In *Journées dŠEtudes sur le Traitement Automatique de la Langue Arabe (JETALA)*, Rabat, Morocco, 2006.

[49] Maja Popović and Hermann Ney. Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pp. 1585–1588, Lisbon, Portugal, May 2004.

[50] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency Treelet Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 271–279, Ann Arbor, Michigan, 2005.

[51] Jason Riesa and David Yarowsky. Minimally Supervised Morphological Segmentation with Applications to Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA06)*, pp. 185–192, Cambridge, MA, 2006.

[52] Fatiha Sadat and Nizar Habash. Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics.

[53] Mohammed Sharaf. Implications of the Agreement Features in (English to Arabic) Machine Translation. Master's thesis, Al-Azhar University, 2002.

[54] Noah Smith, David Smith, and Roy Tromble. Context-Based Morphological Disambiguation with Random Fields. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP05)*, pp. 475–482, Vancouver, Canada, 2005.

[55] Harold Somers. Review Article: Example-based Machine Translation. *Machine Translation*, 14(2):113–157, 1999.

[56] Abdelhadi Soudi. Challenges in the Generation of Arabic from Interlingua. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN-04)*, pp. 343–350, 2004. Fez, Morocco.

[57] Abdelhadi Soudi, Violetta Cavalli-Sforza, and Abderrahim Jamari. A Computational Lexeme-Based Treatment of Arabic Morphology. In *Proceedings of the Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001)*, pp. 50–57, Toulouse, France, 2001.

[58] Abdelhadi Soudi, Violetta Cavalli-Sforza, and Abderrahim Jamari. A Prototype English-to-Arabic Interlingua-based MT system. In *Proceedings of the Third International Conference on Language Resources and Evaluation: Workshop on Arabic language resources and evaluation*, Las Palmas, Spain, 2002.

[59] Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. Bridging the inflection morphology gap for arabic statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 201–204, New York City, USA, 2006. Association for Computational Linguistics.