Finite-State Morphological Analysis of Persian

Karine Megerdoomian

Inxight Software, Inc.
500 Macara Avenue
Sunnyvale, CA 94085, USA
karinem@inxight.com

University of California, San Diego
Linguistics Department
9500 Gilman Drive, #0108
La Jolla, CA 92093, USA
karinem@ling.ucsd.edu

Abstract

This paper describes a two-level morphological analyzer for Persian using a system based on the Xerox finite state tools. Persian language presents certain challenges to computational analysis: There is a complex verbal conjugation paradigm which includes long-distance morphological dependencies; phonological alternations apply at morpheme boundaries; word and noun phrase boundaries are difficult to define since morphemes may be detached from their stems and distinct words can appear without an intervening space. In this work, we develop these problems and provide solutions in a finite-state morphology system.

1 Introduction

This paper describes the design of a two-level morphological analyzer for Persian developed at Inxight Software, based on Xerox finite-state technology (Beesley and Karttunen, 2001), by focusing on some of the issues that arise in a computational analysis of the language.

Persian morphology raises some interesting issues for a computational analysis. One of the main challenges of Persian resides in the tokenization of the input text, since word boundaries are not always respected in written text. Hence, morphemes may appear detached from their stems while distinct tokens may be written without an intervening space. Furthermore, the use of the Arabic script and the fact that short vowels are not written and capitalization is not used create ambiguities that impede computational analysis of text. Persian includes *complex tokens* whereby two distinct part of speech items may be joined; these attaching elements (e.g., prepositions, pronominal clitics or verbs) should be treated as inflectional morphemes in the morphological analyzer. Persian does not have the problems that have been observed in Semitic languages such as the template-based morphology of Arabic, and is in general more concatenative. However, the verbal conjugation consists of a complex paradigm, which includes long-distance dependencies that may be problematic for a linear approach depending solely on surface forms. Finally, the phonetic representation of Persian nominals directly affects the phonological alternations applying at morpheme boundaries; however, the orthographic realization of certain words may not reflect their phonetics and require special manipulations to eliminate the ambiguities.

Although there have been some significant studies in the area of parsing and syntactic analysis for Persian, very little work has been done on computational morphology in this language. In this paper, we elaborate on some of the challenges presented by a morphological analysis of Persian and discuss the solutions provided with a two-level finite-state formalism.

2 System Description

The Persian system is developed using Xerox Finite-State Technology. The lexicons and morphological rules are written in the format of *lexc*, which is the lexicon compiler (Karttunen and Beesley, 1992). The lexicon and grammar are compiled into a finite-state transducer (fst) where the lower side consists of the input string and the upper side provides the baseform of the word with associated morphosyntactic features. In this system, the fsts for each part of speech category are created separately and then composed. Similarly, phonological rules are composed on the relevant fst, thus performing the required phonetic and phonological alternations on the word forms. The composition of all the part of speech transducers with the rules results in the final lexical transducer used for morphological analysis. Since all intermediate levels disappear during a composition, the final transducer consists of a single two-level fst with surface strings in the bottom and the morphological output on the top.

Consider the simple lexc example below. This lexc consists of three small LEXICONs, beginning with the one named Root, which marks the start of the network. The lexicon class named Root

includes three entries and each entry consists of a *form* and a *continuation class.*

```
LEXICON Root
dog  Noun ;
cat  Noun ;
laugh      Verb ;

LEXICON Noun
+Plural:s   # ;
+Singular:0 # ;

LEXICON Verb
+Present:s  # ;
+Past:ed    # ;
+Gerund:ing # ;
            # ;  !empty string
```

The forms, such as 'dog', are interpreted by the lexc as a regular expression as in {d o g}. Continuation classes are used to account for word-formation by capturing morphotactic rules. In the example under consideration, the string 'dog' is followed by the continuation class Noun. As the Noun lexicon shows, the rule allows 'dog' to be followed either by the morpheme 's' or by a null morpheme represented as '0'. The Noun continuation class maps the lower string 's' to the +Plural tag on the upper side of the two-level transducer. Similarly, the Verb continuation class allows the concatenation of the verbal stem 'laugh' with the various inflectional morphemes.

The Persian morphological analyzer at Inxight currently consists of about 55,000 stem forms, including multiword tokens, and a system of rules that identify the baseform of each token. Examples of the output of the morphological analyzer are shown below where the left hand side represents the lower input string and the right hand side is the upper side output[1]:

مسافرین 'travelers'
msâfryn → *msâfr*+Noun+Pl
رفت 'he/she left'
rft → *rftn*+Verb+Ind+Pret+3P+Sg
وکیلست 'he/she is a lawyer'
vkylst→*vkyl*+Noun>*bvdn*+Verb+Ind+Pres+3P+Sg

The rules are written as regular expressions and are represented as continuation paths within the lexc grammar. The morphological analyzer covers all main features of the Persian language with full verbal conjugation and nonverbal inflection, including irregular morphology. In addition, about twenty phonological rules are used to capture the various surface word forms and alternations that occur in the language. Common Proper Nouns are also recognized and tagged.

3 Challenges of the Persian System

This section outlines some of the main issues that arise in a computational analysis of Persian text and presents the approach adopted in the current finite-state system. Comparisons are made with past work on Persian morphological analyzers when relevant.

Persian is an affixal system consisting mainly of suffixes and a number of prefixes appearing in strict morphotactic order. The nonverbal paradigm consists of a relatively small number of affixes marking number, indefiniteness or comparatives, but the language has a complete verbal inflectional system, which can be obtained by the various combinations of prefixes, stems, person and number inflections and auxiliaries.

3.1 Nonverbal Morphology

The Arabic script used in Persian distinguishes between the attached and unattached (or final) forms of the characters. Thus, letters in a word are often connected to each other, whereas all but six characters have a final form if they appear at the end of a word or token. Thus, most characters have a different form depending on their position within the word and the final forms can therefore be used to mark word boundaries. But as we will see in this section, these boundaries are not without ambiguity.

Detached inflectional morphemes. The Persian writing system allows certain morphemes to appear either as bound to the host or as free affixes – free affixes could be separated by a final form character or with an intervening space. The three possible cases are illustrated for the plural suffix *hâ* (ها) in *flsTyny hâ* (فلسطینی ها) 'Palestinians' and the imperfective prefix *my* (می) in *my rvnd* (می روند) 'they are going'. In these examples, the tilde (~) is used to indicate the final form marker which is represented as the control character \u200C in Unicode (also known as the zero-width non-joiner). As shown, the affixes may be attached to the stem, they may be separated with the final form control marker, or they can be detached and appear with the intervening control marker as well as a whitespace. All of these surface forms are attested in various Persian corpora.

[1] Unless otherwise specified, the Persian examples are direct transliterations of the Persian script and do not include short vowels, since that would require disambiguation of word senses and is beyond the scope of the current application. For issues in automatic diacritization of Arabic script-based tests see (Vergyri and Kirchhoff, 2004) in this volume.

Attached	Final Form	Intervening Space
flsTynyhâ	*flsTyny~hâ*	*flsTyny~ hâ*
myrvnd	*my~rvnd*	*my~ rvnd*

In his two-level morphological analyzer, (Riazati, 1997) is unable to analyze the detached affixes and decides to treat these elements in syntax. Thus, the two surface realizations of morphemes such as the plural *hâ* are analyzed in different levels of the system (the attached version in the morphological analyzer and the detached form in the syntactic parser). In the unification-based system developed at CRL (Megerdoomian, 2000), a post-tokenization component is used to join the detached morpheme to the stem, separated by the control character. The morphological grammar is then designed to recognize both surface forms.

The advantage of the finite-state system described here is the ability to process multiword tokens in the analyzer. Thus, by treating the final form character (the zero-width non-joiner) as a space in the tokenization rules, we are able to analyze the detached morphemes in Persian as part of multiword tokens within the lexc grammar module. This allows us to treat both forms uniformly in the morphological analyzer and there is no need for a preprocessing module or for delaying the analysis of the detached morphemes to the syntactic level.

Complex tokens. "Complex tokens" refer to multi-element forms, which consist of affixes that represent a separate lexical category or part of speech than the one they attach to. As in languages such as Arabic and Hebrew, Persian also allows attached word-like morphemes such as the preposition *bh* (به) (*b-* in attached form), the determiner *ayn* (این), the postposition *râ* (را), or the relativizer *kh* (که), that form such complex tokens and need to be analyzed within the morphological analyzer. Similarly, a number of pronominal or verbal clitic elements may appear on various parts of speech categories, giving rise to complex tokens. The examples below illustrate some of these complex constructions where two distinct part of speech items appear attached. The word-like affixes are shown in bold in the examples below.

(i) ***b****eqydh* *Smâ* بعقیده شما
to+opinion you
'in your opinion'

(ii) ***ayn****kâr* اینکار
this+work
'this work'

(iii) anqlaby-tryn-ha-**ySan-nd** انقلابیترینهایشانند
revolutionary+Sup+Plur+**Pron.3pl+Cop.3pl**
'they are the most revolutionary ones'

To account for these cases in the Persian system, the different part of speech items are analyzed within the morphological analyzer and they are separated with an angle bracket as shown below for *ktabhayman* (کتابهایمان) 'our books' and *beqydh* (بعقیده) 'to+opinion'.

ktabhayman
→*ktab*+Noun+Pl>*av*+Pron+Pers+Poss+1P+Pl+Clit
beqydh
→ *bh*+Prep< *eqydh* +Noun+Sg

The angle brackets are used to distinguish these elements from regular inflectional morphemes since the distinct part of speech information may be needed at a later stage of processing, e.g., for parsing or machine translation. Each word-like prefix is presented by its stem form: *av* (او) 'he/she' for the pronominal clitic and *bh* (به) 'to' for the baseform of the preposition. This stem form is then followed by the relevant morphosyntactic tags. If the information is not required, as in the case of certain information retrieval applications, the elements separated by the angle brackets can easily be stripped off without losing the information of the content carrying category, namely the noun in these examples.

In certain cases, two distinct syntactic categories may appear without an intervening space even though they are not attached. For instance, the preposition *dr* (در) 'in' ends in the character 'r' which does not distinguish between a final form and an attached form. Sometimes *dr* appears without a space separating it from the following word and the tokenizer is not able to segment the two words since there is no final form to mark the word boundary. Similarly, in many online corpora sources, the coordination marker *v* (و) 'and' appears juxtaposed with the following word without an intervening space; and since the letter 'v' does not distinguish between a final and attached form, the tokenizer cannot determine the word boundary. These common words that often appear written without an intervening space, though not actually inflectional morphemes, are treated as prefixes in the system as illustrated below:

vgft → *v*+Coord< *gftn*+Verb+Pret+3P+Sg وگفت
drdftr → *dr*+Prep< *dftr*+Noun+Sg دردفتر

Phonetics & Phonological Rules. In Persian, the form of morphological affixes varies based on

the ending character of the stem. Hence, if an animate noun ends in a consonant, it receives the plural morpheme *–ân* as in *znân* (زنان) 'women'. If the animate noun ends in a vowel, the glide 'y' is inserted between the stem and the plural morpheme as in *gdâyân* (گدایان) 'the poor'. Similarly, for animate nouns that end in a silent 'h' (i.e., the letter 'h' which is pronounced as *é*), they take the morpheme *–gân* as in *frSth* (فرشته) → *frStgân* (فرشتگان) 'angels'.

A problem arises in Persian with characters that may be either vowels or consonants and cannot be analyzed correctly simply based on the orthography. For instance, the character 'v' is a consonant in *gâv* (گاو) 'cow' (pronounced 'gaav') but a vowel in *dânSJv* (دانشجو) 'university student' (pronounced 'daneshjoo'). The character 'h' is pronounced as a consonant in *mâh* (ماه) 'moon' but as a vowel in *bynndh* (بیننده) 'viewer' (pronounced 'binandé'). Similarly, 'y' is a glide in *r'ay* 'vote' but a vowel in *mâhy* (ماهی) 'fish' (pronounced 'maahee'). Hence, it is clear that in Persian, the orthographic realization of a character does not necessarily correspond to the phonetic pronunciation, yet phonological alternations of morphemes are sensitive to the phonetics of stems.

In the finite-state lexicon, the nonverbal and closed class lexical items are separated based on their final character, i.e., whether they end in a consonant or a vowel, and word boundary tags are used to determine the relevant phonological alternations. In particular, the words ending in a vowel sound are marked with a word boundary tag ^WB. Hence, the words *dânSJv*, *bynndh* and *mâhy* will be marked with a ^WB tag but not those ending in the consonant pronunciation of the same characters, namely *gâv*, *mâh* and *r'ay*. This allows us to convert the nominal endings of these words to their phonetic pronunciation rather than maintaining their orthographic realization, helping us disambiguate phonological rules for nominal affixes.

The words tagged with the boundary marker ^WB undergo phonetic alternations which convert the ending characters 'v', 'h' and 'y' to 'u', 'e' and 'i', respectively, in order to distinguish vowels and consonants when the phonological rules apply. Thus, after the phonetic alternations have applied, the word *mâh* ending in the consonant 'h' is transliterated as [mah] while the word *bynndh* ending in the vowel or silent 'h' is represented as [bynnde].

Once the ending vowel and consonant characters have been differentiated orthographically, the phonological alternation rules can apply correctly. We mark morpheme boundaries in the lexc with the tag ^NB. This permits the analysis routine to easily locate the area of application of the phonological alternations when the rules are composed with the lexicon transducer. One such phonological rule for the animate plural marker *-ân* is exemplified below:

```
define plural [e %^NB → g || _ a n];
```

This regular expression rule indicates that the word ending in the vowel 'e' and followed by a morpheme boundary marker is to be replaced by 'g', in the context of the plural morpheme 'an'. This rule captures the phonological alternation for *bynndh* (بیننده) 'viewer' → *bynndgân* (بینندگان) 'viewers'.

Thus, since the phonetic representation of Persian nouns and adjectives plays a crucial role in the type of phonological rule that should apply to morpheme boundaries, we manipulate the orthographic realization of certain words in order to eliminate the ambiguities that may arise otherwise.

Past morphological analysis systems have either not captured the pronunciation-orthography discrepancy in Persian thus not constraining the analyses allowed, or they have preclassified the form of the morpheme that can appear on each token. The advantage of the current system is that, by using phonological rules that apply across the board at all morpheme boundaries, we can capture important linguistic generalizations. For instance, there is no need to write three distinct plural rules to represent the various surface forms of the plural suffix *–ân* (namely, *-ân*, *-gân*, and *–yân*). Instead, we can write one single rule adding the *–ân* morpheme and apply phonological rules that can also apply to the boundaries for the pronoun clitic, indefinite, 'ezafe' and relativizing enclitic morphemes, providing a very effective linguistic generalization.

3.2 Verbal Paradigm

The inflectional system for Persian verbs is quite complex and consists of simple forms and compound forms; the latter are forms that require an auxiliary verb. There are two stems used in the formation of the verbal conjugation, which may combine with prefixes marking the imperfective, negation or subjunctive, person and number inflections, suffixes for marking participle forms, and the causative infix. Certain tenses also use auxiliaries to form the perfect forms, the future tense or the passive constructions.

Two stems. One of the intricacies of the Persian verbal system (and of Indo-Aryan verbal systems in general) is the existence of two distinct stem types used in the formation of different

Form	Tense	Prefix	Stem	Inflection	Auxiliary
mygryzd می گریزد	Present	Imperfective *my*	Present *gryz*	Present.3sg *d*	---
mygryxt می گریخت	Imperfect	Imperfective *my*	Past *gryxt*	Past.3sg ‘ ‘	---
mygryxth ast می گریخته است	Compound Imperfect	Imperfective *my*	Past *gryxt*	Participle *h*	Present be.3sg) *and*
bgryz بگریز	Imperative	Subjunctive *b*	Present *gryz*	Imperative.2sg ‘ ‘	---

Table 1: Long-distance dependency between prefix and personal inflection

tenses: The *present stem* is used in the creation of the present tense, the simple subjunctive, the imperative and the present participle. On what is known as the *past stem* are formed the preterite, the imperfect, the past participle and past compounds. Furthermore, all infinitives and future tenses are built on the past stem while all causatives, regardless of tense, are created on the present stem. For computational purposes, the two stems are treated as distinct entities because they often have different surface forms and cannot be derived from each other. Two examples are given below for *krdn* (کردن) and *gryxtn* (گریختن) in the actual pronunciation[2]:

Infinitival	Present Stem	Past Stem	
kardan	*kon*	*kard*	‘to do/make’
gorixtan	*goriz*	*gorixt*	‘to flee’

Since the infinitival or citation form of the verbs is built on the past stem, the verbal finite-state transducer has to produce the past stem on the upper side, allowing the derivation of the infinitive. A problem arises when the input string is the present stem form as in the present tense *my gryznd* (می گریزند) ‘they are fleeing’. In this instance, we would need to output the past stem form of the verb, namely *gryxt* (گریخت). In order to capture the association between the present and past stems in Persian, we link these forms in the verbal lexicon by allowing all present stems to map to the past stem form in the upper side of the transducer, as illustrated in the first continuation class below. In addition, the same verbs have to be listed in a different lexical continuation class with the past stems alone (i.e., past stem on both lower and upper sides) in order to analyze the tenses formed on the past stem of the verb such as the imperfect *my gryxtnd* (می گریختند) ‘they were fleeing’.

```
LEXICON PresentStem
gryxt:gryz  VerbReg ;  ! to flee
nvSt:nvys   VerbReg ;  ! to write
aftad:aft   VerbReg ;  ! to fall

LEXICON PastStem
gryxt   InfBoundary ;  ! to flee
nvSt    InfBoundary ;  ! to write
aftad   InfBoundary ;  ! to fall
```

In both cases the upper side past stem string is marked with a delimiter tag ^INF which is later mapped to ‘n’, forming the surface form of the infinitive. The resulting stem form for the finite verb *my gryznd* (می گریزند) ‘they are fleeing’ is thus the infinitival *gryxtn* (گریختن) ‘to flee’.

Long-distance dependencies[3]. As can be seen in the examples given above for the verb *gryxtn* (گریختن) ‘to flee’, the prefix *my-* (می) cannot be used to distinguish the tense of the verbal entry since it is used in the formation of the present, the imperfect or the compound imperfect. In order to decide whether *my* is forming e.g., the present tense or the past imperfect, the stem and final inflection need to be taken into account. Thus, if *my* is attached to the present stem, it forms the regular present tense forms but if it is attached to the past stem, then it gives rise to either the simple imperfect or the compound imperfect, depending on the final inflection forms (see Table 1). Similarly, the imperative inflection can only appear on a present stem with the subjunctive prefix ‘b’, as shown in *bgryz* (بگریز) in Table 1, whereas only the present inflection can be used if

[2] Note that in Persian, the short vowels such as *o,a,e* are not generally transcribed, hence the direct transliteration of the examples would be

krdn kn krd ‘to do, to make’
gryxtn gryz gryxt ‘to flee’

[3] See for instance (Sproat, 1992; pages 91-92) for a description of the issue raised by “morphological long-distance dependencies” in finite-state models of morphology.

the imperfective prefix ‘my’ is used, as shown with *my gryzd* (می گریزد) .

Accounting for the long-distance dependency between the prefix and the personal inflection in Persian in a finite-state two-level morphological analyzer leads to very complex paths and continuation class structures in the lexical grammar. Also, using filters to capture long-distance dependencies can sometimes largely increase the size of the transducer. Since there exist several cases of interdependencies between non-adjacent morphemes in Persian verb formation, we have opted to keep a simpler continuation class structure in the lexc grammars and to instead take advantage of *flag diacritics* and their unification process.

Flag diacritics are multicharacter symbols and can be used within the lexc grammar to permit the analysis routines to use the information provided in terms of feature-value settings to constrain subsequent paths. Hence, whether a transition to the following path would apply depends on the success of the operation defined by the flag diacritic. In essence, the flag diacritic allows the system to perform a unification of the features set in the analysis process. Xerox finite state technology includes a number of different flag diacritic operators but the only one used in this Persian system is the U-type or the Unification flag diacritic. The template for the format of these flags is as follows: @U.feature.value@. Flag diacritics are used to keep the fst small and yet be able to apply certain constraints, in particular when dealing with interdependencies between non-adjacent morphemes within a word.

For example, to capture the choice of the imperative vs. the present tense inflection based on the prefix that appears on the present stem of the verb, we use a flag diacritic with the attribute PFXTYP (PrefixType) which is then set to IMP (for imperfective) or SUB (for subjunctive). This flag diacritic is set when the prefixes are read and they are unified with the PFXTYP flags at the lexical class defining the personal inflectional paradigm. If the values of the PFXTYP flag diacritic match at this point, unification takes place allowing the concatenation of the prefix and present stem combination with the personal inflection.

Similarly, the agentive, infinitive and participial forms can be formed only if there is no prefix at all on the verbal stem. This is captured by the flag diacritic attribute PFX, which has the two possible values PRESENT and ABSENT. Thus, the lexc rule for the Infinitive, for instance, requires that the PFX flag’s value be set to ABSENT. This, in effect, captures the fact that *mygryxtn* (*my* ‘imperfective’ + *gryxt* ‘past stem’ + *n* ‘infinitive marker’) is not a valid form since the infinitive marker *–n* can only appear on a past stem that lacks an overt prefix.

4 Evaluation

The lexicon used in the Inxight system currently consists of 43,154 lemmas, which include nouns, adjectives, verbs, adverbs and closed class items. In addition, there are about 12,000 comomon proper noun entities listed in the lexicon. The system also recognizes date, number and internet expressions.

The current Persian morphological analyzer has a coverage of 97.5% on a 7MB corpus collected mostly from online news sources. The accuracy of the system is about 95%. The unanalyzed tokens are often proper nouns or words missing from the lexicon. In addition, colloquial forms of speech are not covered in the current system.

The finite state transducer consists of 178,452 states and 928,982 arcs before optimization. And the speed of the analyzer is 20.84 CPU time in seconds for processing a 10MB file executed on a modern Sun SparcStation.

5 Conclusion

This paper describes some of the challenges encountered in a computational morphological analysis of Persian and discusses the solutions proposed within the finite state system developed at Inxight Software based on the Xerox Finite State Technology. The approaches adopted are compared with past systems of Persian whenever relevant. The paper presents the problems arising from detached inflectional morphemes, as well as attached word-like elements forming complex tokens, the discrepancy between orthography and phonetics in application of phonological rules, and the interdependency between non-adjacent morphemes in a word. In each case, it was argued that methods adopted from the finite-state calculus can capture linguistic generalizations and reduce the transducer to a manageable and commercially viable size.

6 Acknowledgements

We gratefully acknowledge the help and support provided by the development team at Inxight Software and the insightful suggestions of the members of the Lingware group. I would also like to thank the anonymous reviewers for their detailed comments.

References

Mohammad-Reza Bateni. 1995. *Towsif-e Sakhteman-e Dastury-e Zaban-e Farsi* [Description of the Linguistic Structure of Persian Language]. Amir Kabir Publishers, Tehran, Iran.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite-State Morphology: Xerox Tools and Techniques*. CSLI Publications, Palo Alto.

Lauri Karttunen and Kenneth R. Beesley. 1992. *Two-Level Rule Compiler*. Technical Report. ISTL-1992-2. Xerox Palo Alto Research Center. Palo Alto, California.

Gilbert Lazard. 1992. *A Grammar of Contemporary Persian*. Mazda Publishers.

Shahrzad Mahootian. 1997. *Persian*.Routledge.

Karine Megerdoomian. 2000. Unification-Based Persian Morphology. In *Proceedings of CICLing 2000*. Alexander Gelbukh, ed. Centro de Investigación en Computación-IPN, Mexico.

Dariush Riazati.1997. *Computational Analysis of Persian Morphology*. MSc thesis, Department of Computer Science, RMIT.

Richard Sproat. 1992. *Morphology and Computation*.MIT Press, Cambridge, Massachusetts.