Extending a Persian Morphological Analyzer to Blogs

Karine Megerdoomian
University of Maryland, College Park
karinem@umiacs.umd.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 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. The paper also presents an overview of new issues that have arisen since the advent of blogs and the propagation of informal Persian text on the web. This new mode of writing provides the computational system with further challenges. The paper proposes approaches for extending the current morphological system to analyze the material found in Persian blogs.

1 Introduction

Since 2003 there has been an explosion in the number of weblogs, a new form of electronic media, written in the Persian language. A weblog (blog for short) is a form of online journal that is presented in reverse chronological order, can be updated frequently, and typically contains a mixture of hyperlinks, commentary, essays, personal thoughts and opinions. According to the latest statistics, Persian is the 4th largest blog language in the world and there currently exist about 75,000 Persian blogs inside Iran and in the exile community. Linguistically and computationally, an investigation of Persian blogs is of interest since these texts contain features generally observed in contemporary colloquial or informal speech.

Traditionally, Persian text has been written in the formal language. However, with the advent of blogs, colloquial speech has been propagated in written form. In particular, when a blog entry discusses the personal thoughts or daily life of the author, the text tends to be written in an informal language which differs considerably from formal Persian in several domains including lexicon, morphology, and syntactic structure. In addition, since we are dealing with written text, orthographic issues also arise; blogs often contain spelling errors as well as non-standard spacing and punctuation.

Given this “diglossic” situation in Persian blog text, the challenge from a computational perspective is to determine whether today’s algorithms for analyzing Persian newsprint are portable to this new web form. 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), and discusses how the existing rules could be modified to additionally capture informal text.

2 The Language of Blogs

Before the advent of blogs, Persian websites had traditionally been written in formal text. Blogs, however, contain features of both formal and informal speech, often depending on the topic discussed in the web entry. This gives rise to a bigger variation for a computational system processing the online text.

The linguistic features of informal text in Persian blogs differ considerably from the formal counterpart. For instance, informal text makes more frequent use of clitic pronouns (e.g., گرفته‌تم vs. مرا گرفته‌است), has shortened verbal stems and inflectional endings (e.g., میگن vs. می‌گویند), and includes morphemes that do not exist in the formal language such as the definite article (e.g., فروشنده‌هه). The spelling in these texts reflects the pronunciation in colloquial speech (e.g., نظراتون vs. نظرهایتان). In addition, informal text contains more instances of scrambling, idiomatic expressions, loan words, jargons and other non-dictionary words, as well as cultural inferences.

Blogs written in informal text often do not follow a standard set of orthographic rules and may write morphemes attached or unattached as the author sees fit. Blogs contain ellipses, emoticons and hyperlinks, which require special tokenization. In addition, spelling errors are much more common than one may encounter in non-blog websites.

It should also be noted that even blogs written in formal text display a varying range of orthographic patterns when it comes to the printing of derivational and inflectional affixes in Persian, which differ significantly from the standard rules followed by newsprint media online. For instance, while a number of bloggers always write the clitic copula in detached from as in موافقاند, the traditional websites use the attached from موافقند. Authors writing in formal text also use new spellings such as حتا instead of the traditional حتی.

In order to process the material on Persian blogs, the computational system needs to be able to analyze the features encountered in both formal and informal writing, as well as the enormous amount of variation existent within these textual forms. Since tokenization and morphological analysis are at the basis of such a computational system, the next sections present the development of a morphological analyzer for traditional formal text and describe how the existing analyzer could be extended to capture weblog texts in Persian.

3 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. 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 *msafryn* → *msafr*+Noun+Pl
رفت *rft* → *rftn*+Verb+Ind+Pret+3P+Sg
وکیلست
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 formal 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.

[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.

4 Morphological Features of Persian

This section outlines some of the main issues that arise in a computational analysis of Persian text and presents the approach adopted in the finite-state system. Comparisons are made with past work on Persian morphological analyzers when relevant. Some of the issues discussed are related to morphophonology of Persian, while others result from the peculiarities of the writing system and the mixture of informal and formal text in blogs.

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 به, the determiner این, the postposition را, or the relativizer 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) بعقیده شما
to+opinion you
’in your opinion’

(ii) اینکار
this+work
’this work’

In formal blogs, however, the morphemes mentioned above are often transcribed in detached form. For instance, if the preposition به is not considered a part of the lexical element, it is generally written in detached form in formal blog text, with an intervening control character or short space (نیم‌فاصله): به‌هیچ‌وجه, به‌طرز, به‌عقیده‌ی من. On the other hand, complex tokens are very common in informal blog text as in the example below:

(iii) برادرشه
brother+**Pron.3sg+Cop.3sg**
‘he is his/her brother’

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 کتابهایمان and بعقیده:

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) for the pronominal clitic and *bh* (به) 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 در ends in the character ‘r’ which does not distinguish between a final form and an attached form. Sometimes در 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 و 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 tokenization issues appear in both blog and traditional online text. 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

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 (the control character \u200C in Unicode, also known as the zero-width non-joiner) or with an intervening space. The three possible cases are illustrated below for the plural suffix ها and the imperfective prefix می. 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 intervening whitespace. All of these surface forms are attested in various Persian corpora.

Attached	Final Form	Intervening Space
کتابها	کتاب‌ها	کتاب ها
میروند	می‌روند	می روند

Blogs have added to the possible existing forms resulting in more ambiguity in the analysis of affixes. The formal texts in Persian blogs often separate morphemes that were previously treated as attached affixes, as in the following examples: موافق‌ام, دندان‌شان, کافی‌ست, کم‌تر. Although traditionally derivational morphemes are not written separately, in more recent text derivational affixes appear detached as in نماینده‌گان. The informal texts, on the other hand, represent the spoken colloquial form of the morphemes (usually following the standard Tehrani dialect) thus providing a number of new constructions that need to be integrated within the existing morphological rules. Some examples are: plural morpheme reduced to /a/ (غلطای املایی), object marker as /o/ (این سایتو), variant forms of the pronominal clitic (خودتون, نگاه‌های‌شان, ازشون , همسایه‌اشان), هم reduced to /m/ (دوستامونم), and reduced copula forms (مثل‌منن, تیر هواییه, دردناکه). In certain instances, both the stem form and the affix are affected as in مسخرست vs. the formal مسخره است, براتون vs. برایتان , بش vs. بهش/به‌او.

Faced with this variance in forms, the existing morphological rules have to be made more flexible. For instance, in order to capture the personal clitic forms, in addition to the traditional م form, the morphological analyzer needs to also allow for the detached form ام following a consonant. In the finite-state system, the alternate forms of the clitic morpheme are captured at the point that phonological rules apply and any modifications to the rule would need to be performed in phonology, which will be discussed in the following section.

The analysis of detached affixes has been challenging in past morphological systems. In his two-level morphological analyzer, Riazati (1997) decides to treat these elements in syntax. Thus, the two surface realizations of morphemes such as the plural ها 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 is illustrated in the following rules for the *number* feature applied to nouns ending in a consonant:

```
LEXICON NumberCons
%+Sg:0 EzafeCons ;
Number ;

LEXICON Number
+Pl:ha     Ezafe ;
+Pl:_ha    Ezafe ;
```

As shown in the lexicon class named Number, the plural morpheme is recognized either in attached form (i.e., *ha*) or as part of a multiword element (i.e., *_ha*), where the underscore is used to represent a space separating the suffix from the main stem. The continuation class Ezafe processes the form of the *ezafe* morpheme following a vowel. If no plural morpheme is detected in the lexicon class NumberCons, however, the noun is marked by the singular tag and the analysis is continued through the EzafeCons continuation class which processes the *ezafe* form found following a consonant.

This approach allows us to treat both attached and detached forms of a morpheme 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 (see Dehdari 2005 for a similar approach).

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 زنان. If the animate noun ends in a vowel, the glide 'y' is inserted between the stem and the plural morpheme as in گدایان. 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 فرشتگان. The current morphological analyzer is able to apply phonological rules based on the final character of the word.

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, a vowel, or a glide. The system then takes advantage of word boundary tags that are used to determine the relevant phonological alternations. This is particularly useful for characters such as و, ی, ه which can be pronounced both as consonants or vowels. Hence, the word گدا or دانشجو will be marked with a ^WB tag. The words tagged with the boundary marker ^WB undergo phonetic alternations which convert the transliteration of 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. Consider the words بيننده vs. ماه: after the phonetic alternations have applied, the word ماه ending in the consonant ‘h’ is transliterated as [mah] while the word بيننده 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* (بيننده) ‘viewer’ → *bynndgan* (بيندگان) ‘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.

However, the form بينندهگان has now become common in weblogs. Therefore, the rule noted above should be made optional by also including the following pattern, which does not eliminate the silent ‘h’ character before adding the *gân* plural form:

```
define pl2 [%^NB → _ g || e _ a n];
```

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 in morphology to represent the various surface forms of the plural suffix *–ân* (namely, ان-, گان-, and يان–). 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.

As noted earlier, the writing of clitic forms, such as the pronominal clitics or the copula verb, varies enormously depending on the website or the blog. For instance, traditional formal text transcribes the pronominal clitic forms as attached following a consonant, with a glide form added after a vowel, and in detached from following a ‘y’ or silent ‘h’ character. A quick look at blogs clearly shows that these rules are too restrictive and will not be able to capture all possible clitic forms found online. In particular, in formal text found on blogs, the pronominal clitic tends to always be written as a detached morpheme. A similar trend can be observed with the transcription of the copula verb. The following continuation class shows the possible clitic forms for traditional formal text. In each instance, the morphological features of the pronoun are marked on the upper side (i.e., the left hand side of the colon), while the basic clitic form appears on the lower side (i.e., to the right of the colon). Once again, the ^NB tag is used to mark the noun boundary and determines the domain of application of the phonology rule.

```
LEXICON CliticForms
>av%+Pron%+Poss%+1P%+Sg%+Clit:%^NBm Cop;
>av%+Pron%+Poss%+2P%+Sg%+Clit:%^NBt Cop;
>av%+Pron%+Poss%+3P%+Sg%+Clit:%^NBŚ Cop;
>av%+Pron%+Poss%+1P%+Pl%+Clit:%^NBman Cop;
>av%+Pron%+Poss%+2P%+Pl%+Clit:%^NBtan Cop;
>av%+Pron%+Poss%+3P%+Pl%+Clit:%^NBSan Cop;
```

In order to process the various surface forms of the pronominal clitic, the existing phonology rules have to be modified. Hence, to analyze both هراسم observed in traditional formal text and the form هراسام found in formal blog text, the second rule below should be added to the first one thus allowing the optionality of the forms.

```
define Rule1 [%^NB → 0 || Cons _ ];
define Rule2 [%^NB → _ a || Cons _ ];
define Rules [Rule1 | Rule2];
```

Verbal Morphology. 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 tenses. 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:

Infinitival	Present Stem	Past Stem
توانستن	توان	توانست
نوشتن	نویس	نوشت

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 مى توانند. In this instance, we would need to output the past stem form of the verb, namely توانست. 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 می‌توانستند.

```
LEXICON PresentStem
tvanst:tvan VerbReg ;  ! to be able
nvSt:nvys   VerbReg ;  ! to write
aftad:aft   VerbReg ;  ! to fall

LEXICON PastStem
tvanst  InfBoundary ;  ! to be able
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 می‌توانند is thus the infinitival توانستن.

Informal blog text adds another verb form since in spoken Persian, the present stem is often reduced as shown:

نمیذارم vs. نمی‌گذارم , می‌شه vs. می‌شود ,برید vs. بروید
نمیتونه vs. نمی‌تواند , بخواید vs. بخواهید

In order to capture these forms, a third stem form should be listed for this class of verbs in the lexc similar to the PresentStem continuation class, which maps into the past stem form in the upper side as in `tvanst:tvn`.

Long-distance dependencies. Morphological long-distance dependencies have been discussed in the literature (e.g., Sproat, 1992; pages 91-92) since they present a challenge for finite-state models of morphology. In the Persian verbal system, the prefix مى 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 مى is forming e.g., the present tense or the imperfect, the stem and final inflection need to be taken into account. Thus, if مى 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 the imperfective prefix مى is used, as shown with *my gryzd* (مى گريزد) . Although the example shown is from formal text, similar dependencies also exist in informal blog text.

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

Form	Tense	Prefix	Stem	Inflection	Auxiliary
mygryzd مى گريزد	Present	Imperfective *my*	Present *gryz*	Present.3sg *d*	---
mygryxt مى گريخت	Imperfect	Imperfective *my*	Past *gryxt*	Past.3sg ' '	---
mygryxth ast مى گريخته است	Compound Imperfect	Imperfective *my*	Past *gryxt*	Participle *h*	Present be.3sg) *and*
bgryz بگريز	Imperative	Subjunctive *b*	Present *gryz*	Imperative.2sg ' '	---

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

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 in order 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.

Another instance where flag diacritics are used in the verbal paradigm is for determining which verb stems allow the presence of the causative morpheme. The verbs are classified in the lexicon based on whether the causative morpheme may appear on the stem form; as a general rule, unergative and ingestive verbs may appear with the causative morpheme whereas transitives and statives resist causativization. The non-causativizing class of verbs are tagged with the flag diacritic @U.CAUSE.NOTOK@ thus constraining the possible continuation paths for these verbs.

5 Lexicon

In the finite-state system, morphological analysis directly depends on the content of the lexicon. Hence, if the stem form is not listed in the lexical component, the current morphological analyzer will provide all potential output forms and will tag the resulting word stems as unknowns. The lexicon used in the Inxight system consists of 43,154 lemmas including multiword expressions, 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. With informal text, a large number of words have to be added to the *lexc* if the analyzer is to be able to analyze blogs. The following describe some of the issues to consider in extending the lexicon for informal text:

Phonological Alternations. Words are often written as they are pronounced in modern colloquial Persian resulting in new wordforms as in اوضاش, برام, نگا می‌کنم , خونه , همدیگه .

Colloquial Forms. Certain words that have formal counterparts are used frequently in blogs, whereas they do not appear in traditional formal text. واسه (vs. برای) and تو (vs. در) are among this group.

Loan Words. A large number of new loan words (in transliterated form or in the original language) can be found in blogs, in particular words relating to technology and computers. Examples are: چت‌روم , پابلیش, آن‌لاین .

New Words. Bloggers often create new words such as دوستان کامنت‌گذار or لینکدونی. These words can be treated as unknowns of course by the morphological system; however, a more robust system that can perform analysis of compound forms may be able to capture some percentage of these new words.

Interjections. Informal blogs use many forms of interjections and emoticons, such as آآآخ! , والا , وای ,اووه!.

6 Evaluation

The Inxight 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% for formal text. The unanalyzed tokens are often proper nouns or words missing from the lexicon.

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.

There are currently no rules for the analysis of colloquial or informal text.

7 Conclusion

This paper describes some of the challenges encountered in a computational morphological analysis of formal Persian text and discusses the solutions proposed within the finite state system developed at Inxight Software based on the Xerox Finite State Technology. In addition, the paper presents an overview of recent issues that have arisen since the advent of blogs, which contain a mixture of formal and informal speech, as well as online scripting devices such as emoticons and hyperlinks. One of the major difficulties for a computational system is to be able to capture the variation in orthographic forms and standards, the phonological rules arising from colloquial speech patterns, and the preponderance of loan words and non-dictionary items. In each instance, the paper presents possible ways of extending the current morphological analyzer to the new genre within the finite-state system.

8 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 participants at the workshop on 'Computational approaches to Arabic script-based languages' at COLING 2004 for discussion and comments.

References

Bahram Ashraf-Zadeh. زبان فارسی در وبلاگ‌های فارسی http://www.persianfarsi.com/articles/zabaneweblog.htm

Mohammad-Reza Bateni. 1995. توصیف ساختمان دستوری زبان فارسی. Amir Kabir Publishers, Tehran, Iran.

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

Jon Dehdari. 2005. A link Grammar Parser for Persian. Talk presented at the *First International Conference on Aspects of Iranian Linguistics.* Leipzig, Germany.

Alireza Doostdar. 2004. The Vulgar Spirit of Blogging: On Language, Culture, and Power in Persian Weblogestan. In *American Anthropologist,* Vol. 106, No. 4.

Kyumars S. Esmaili, Mohsen Jamali, Mahmood Neshati, Hassan Abolhassani and Yasaman Soltan-Zadeh. 2006. Experiments on Persian Weblogs. Presented at the *Third Annual Workshop on the Weblogging Ecosystem*, hosted by http://blogpulse.com.

ISNA editorial. 2006. زبان فارسي در وبلاگ‌ها؛ فرصت‌ها و تهديدها. http://isna.ir (6 May).

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.

Karine Megerdoomian. 2004. Finite-State Morphological Analysis of Persian. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages, COLING 2004*. Geneva, Switzerland.

Ebrahim Nabavi. 2004. شصت هزار سردبیر. BBC Persian online (15 November).

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.

Mirko Tavosanis. 2006. Linguistic Features of Italian Blogs. In the Proceedings of the Workshop on *NEW TEXT: Wikis and Blogs and Other Dynamic Text Sources, EACL 2006.* Trento, Italy.