Unification-Based Persian Morphology

Karine Megerdoomian

We present a complete formalization of Persian inflectional morphology using a unification-based framework. The morphological analyzer was developed for use in a Persian-English machine translation system; it computes the part of speech categories and returns all syntactically relevant inflectional features for a word. The morphological analyses are represented as feature structures, which can easily be used by a syntactic parser. The morphological formalism consists of a declarative description of rules utilizing typed feature structures. Persian morphotactics include a few prefixes and sequences of suffixes with co-occurrence constraints between non-adjacent morphemes. The verbal inflectional morphology is rich and is characterized by a complex system of conjugations. A morphological rule associates a regular expression describing a set of character strings to a typed feature structure. Rules can be combined using regular expression operators and they can be factorized in conjugation tables. The morphological engine is implemented as a finite-state transducer where the left projection is the input string and the right projection is a typed feature structure.

1 INTRODUCTION

In this paper, we describe the implementation of an inflectional morphological analyzer for Persian, which is based on finite state transducers and typed feature structures with unification. The analyzer was designed to provide an interface to the syntactic parser in the Shiraz Persian-English machine translation system (http://crl.nmsu.edu/shiraz) and was tested on online newspaper articles. The system includes a dictionary with 50,000 entries which is used for lookup after morphological analysis has been performed.

This paper also provides a detailed description of Persian inflectional morphology. Persian is an affixal system consisting mainly of suffixes and a few prefixes. The nominal paradigm consists of a relatively small number of affixes but the language has a complete verbal inflectional system, which can be obtained by the combination of prefixes, stems inflections and auxiliaries. The affixes in the language follow a strict morphotactic order. One of the main problems for the analysis of Persian written text is discontinuity in the word structure. Certain affixes in the language are always bound to the stem, while others may appear as either bound or free morphemes. For instance, the plural

morpheme *ât* is always bound to the nominal element it appears on. The plural morpheme *hâ*, on the other hand, can be either attached to the previous morpheme or appear as a free affix. Hence, in order to recognize both forms of such affixal elements, the analyzer should be used in conjunction with a tokenization component. Certain ambiguities also arise in a computational analysis of Persian text since the same surface form can represent different morphemes. For instance, the suffix *-y* on *mrdy* [pronounced *mardi*] can be analyzed as the indefinite article (i.e., "a man"), the enclitic particle which links the noun to a relativizer, or the copula for the second person singular (i.e., "you are a man"). In addition, short vowels are often not marked in written text which results in different possibilities of analysis. The previous example, for instance, could also be pronounced with the vowel 'o' [*mordi*] in which case the suffix could be analyzed as the verbal inflectional ending for the second person singular (i.e., "you died"). Furthermore, the morpheme *-y* is also used in forming various lexical elements in derivational morphology. In order to disambiguate such cases, the morphological analyzer needs to use the information available from the parts of speech that the morpheme appears on.

Although there has 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. The only thorough research of Persian morphology from a computational perspective is [Riazati 1997], which uses two-level morphology to analyze both derivational and inflectional affixation in Persian. Riazati's analyzer, Perslex, is modeled on the basis of Englex [Antworth 1990].

In the framework presented here, the linguistic information associated with the morphemes is described using Samba, a morphological formalism combining typed feature structures with a declarative unification-based language [Zajac 1998]. The morphological rule describes the concatenation of stems and morphemes (using regular expressions) and the combination of morphological features of words and morphemes (using feature structures and unification). A morphological rule associates a surface form, representing a sequence of morphemes, to a set of morphological features, and describes how the features of the stem and the morpheme are combined. The specification formalism is language-independent and can be used in multilingual environments. One important advantage of using typed feature structures for the description of morphological knowledge is the uniformity obtained, since modules for the analyses of words and further linguistic processing can be integrated seamlessly.

2 DISCONTINUOUS ELEMENTS

Persian uses the Arabic alphabet and texts are written from right to left. Letters in a word are often connected to each other, but most characters have a different

form depending on their position within the word. The initial form indicates that no element is attached to the element from the right (i.e., there is no "attaching" character before it, but there is one following the character). Characters are in medial form if they have an attaching character both before and after them. The final form denotes that the character is at the end of a word. The final forms can therefore be used to mark word boundaries.

final	*medial*	*initial*	
ب	ـبـ	بـ	**“b”**
گ	ـگـ	گـ	**“g”**
ج	ـجـ	جـ	**“J”**

Figure 1: Sample Persian character forms

The Persian writing system allows certain morphemes to appear either as bound to the following or preceding morpheme or as free affixes. When a morpheme is attached, it uses the initial or medial form of the character. But when, for instance, a prefix appears detached, its last character is in final form. An instance of this is the imperfective marker *my* which can be written attached or detached as shown in (1). Note that the morpheme is not separated from the stem by a space; in Unicode, the final forms are indicated by a control character which, in our transliteration, is represented by tilda /~/. Other examples of such morphemes are the superlative affix, several plural forms, indefinite and enclitic suffixes, as well as certain auxiliary forms.

a. *myrvm* (I am going) (1)
b. *my~rvm* (I am going)

In his two-level morphological analyzer, [Riazati 1997] is unable to analyze the detached affixes and decides to treat these discontinuous elements in syntax. Thus, the two surface realizations of morphemes such as the imperfective *my* are analyzed in different levels of the system (the attached version in the morphological analyzer and the detached form in the syntactic parser). We have opted, instead, to use a preprocessing component which joins these morphemes to the stem separated by the control character. The morphological grammar is then designed to recognize both surface forms. This allows us to treat both forms uniformly in the morphological analyzer and we have no need to delay the analysis of the detached morphemes to the syntactic level.

Most verbal constructions in Persian are formed using a light verb such as *kardan* (do, make), *dâdan* (give), *zadan* (hit, strike). The number of verbs that can be used as light verbs is limited, but these constructions are extremely productive in Persian. These structures consist of a preverbal element, which could be a noun, adjective or preposition, followed by a light verb, which has partly or completely lost its original meaning. Since these constructions are noncompositional in meaning, they are included in the dictionary as compounds. In these Light Verb or Compound Verb constructions, verbal inflection can only appear on the light verb itself, but bound morphemes can be attached to the preverbal element as well as the light verb. An example of this construction is given in (2).

tshvigh- shan *krd-nd* [pronounced *tashvigheshân kardand]* (2)
encourage-them did-3pl
'They encouraged them.'

The morphological grammar analyzes these inflectional morphemes separately on each element. A later syntactic component unifies the two parts of the light verb and combines the morphemes into a single feature structure.

3 NON-VERBAL MORPHOLOGY

There are no gender distinctions in Persian and the language has only one case form. Person, number and sometimes animacy, however, are distinguished. There exist several morphemes to mark plurality, some of which are borrowings from Arabic; these suffixes vary based on the animacy or phonological properties of the stem. There are also some plural forms in Persian that follow the Arabic template morphology (also known as "broken" plurals) such as *ketâb --> kotob* (books) or *faghir --> fogharâ* ([the] poor). But the rules for forming these plurals are not used productively in Persian; instead, the forms derived from the Arabic morphological paradigm have been lexicalized and are being used as fixed vocabulary. Furthermore, in certain instances, the singular/plural distinction between the distinct Arabic word forms has not been maintained in Persian. This is the case for the word *hâl* 'health' and its Arabic plural form *ahvâl*, which are used interchangeably. There are cases in which the meaning of the plural has diverged from that of the singular as in *sabab* 'reason' and *asbâb* 'goods'. In addition, the word *arbâb* is in fact a plural form but is used as a singular in Persian meaning 'master'. These loan words are listed as irregular plurals in the lexicon and need not undergo morphological analysis.

Although there is no overt definite marker, a suffix is used on nouns and adjectives to indicate indefiniteness. The enclitic suffix which links nominal elements to a relative clause has the same surface form as the indefinite; the two morphemes can not be disambiguated at the morphological level. Hence, *xâne~i* could mean either 'a house' as an indefinite or 'a/the house' with the enclitic

since the latter does not supply any information on the definiteness of the noun. We have created a hybrid feature *indefEncl* representing the presence of the indefinite or the enclitic in morphology, which is succesfully disambiguated in the syntactic parser.

The elements within a noun phrase are linked by the enclitic particle called *ezafe*. This morpheme is usually an unwritten vowel, but it could also have an orthographic realization in certain phonological environments. In most cases, this relation can be translated as a genitive structure. Examples of this construction are given below:

a. *sedâ-ye pâ-ye man* (3)
sound-ez foot-ez my
'(the) sound of my footsteps'
b. *ru-ye miz*
on-ez table
'on the table'

Adjectives follow the same morphological patterns as nouns. They can also appear with comparative and superlative morphemes, e.g., *bozorg-tarin* (biggest). Certain adverbs, mainly manner adverbs, can behave like adjectives and can appear with all the adjectival affixes.

Personal pronouns can appear either as free forms or as clitics. Although these cliticized pronouns have identical surface forms, they can have different functions depending on the part of speech they appear on or their syntactic context: On the last element of a noun phrase, the clitic is interpreted as a possessive pronoun *ketâb-at* [book + 2sg] (your book). Attached to transitive verbs and prepositions, the clitic is the accusative form of the personal pronoun *did-am-at* [see(past) + 1sg infl. + 2sg] (I saw you). The clitic may appear on adverbials, numerical expressions and interrogative elements with a partitive meaning, *vasat-ash* [middle + 3sg] (in the middle of it). On intransitive verbs, it could be used as the subject clitic. It is also used in impersonal verbal constructions. Some of these usages, however, are limited to colloquial speech and apart from the possessive and accusative clitics, they are rarely used in written text.

There are three types of ordinal constructions in Persian, which are formed by attaching their respective morphemes to the cardinal number illustrated in (4).

panj (five)--> *panjom* (fifth), *panjomi* (the fifth), *panjomin* (the fifth) (4)

There exist other lexical elements, such as the preposition *be*, the postposition *râ*, or the relativizer *ke*, that usually appear as separate words in written text, but which can also be found as attached morphemes.

The lexical categories may carry several morphemes appearing in a strict ordering. For example, the adjective *enqelâbi* (revolutionary) can have a plural suffix, a superlative affix and a pronominal clitic, but there exists only one possible ordering for these elements: *enqelâbi+tarin+hâ+yeshân* [revolutionary+*superlative + plural + clitic/3pl*] which could be translated as 'the revolutionariest ones among them.' The morphological rules, of course, incorporate the morphotactics in order to constrain the number of analyses produced.

The present indicative of the verb *budan* (to be) behaves as a copula. It has a series of enclitic forms which can attach to the constituents of a noun phrase. This morpheme is a verbal element but it can attach to nouns, adjectives and classifiers, e.g., *zabânshenâs-im* [linguist + copula/1pl] (we are linguists). In the current implementation, the morphological analyzer can not handle the copula, since it requires the analyzer to split the word structure into two distinct feature structures (one for the nominal element and for the verb).

4 VERBAL MORPHOLOGY

The inflectional system for the Persian verbs consists of simple forms and compound forms; the latter are forms that require an auxiliary verb. The simple forms are divided into two groups according to the stem they use in their formation: the tenses that use the Present Stem and those formed on the Past (or Aorist) Stem. The Present Stem needs to be specified in the lexicon since it cannot be derived, while the Past Stem is easily derivable from the infinitival form of the verb as exemplified in (5). The citation form for the verb is the infinitive.

Infinitival:	*foruxtan*(to sell)	*kardan* (to do; to make)	(5)
Present Stem:	*forush*	*kon*	
Past Stem:	*foruxt*	*kard*	

In addition to the verb stems, the following elements also participate in the formation of the verbal inflectional system in Persian:

- **Prefixes:** the imperfective prefix *my* and the morpheme *b* or *by*, which characterizes the subjunctive and the imperative. Negation is marked by the *n* or *ny* prefix.
- **Personal Inflections:** present, past and imperative personal inflections are used in conjugating the Persian verb. All verb forms are marked for person and number.
- **Suffixes:** the suffix *ande* marks the present participle ending and *e* (written *h*) is used to form the past participle.

- **Causation morpheme:** causatives are obtained by adding the affix *ân* or *âni* to the end of the Present Stem of the verb (Table 1). Personal inflections and suffixes can then be attached to the Causative Present Stem to derive all verbal forms for the causative construction.

Table 1: Causative formation

Verb Infinitive	Verb Present Stem	English Translation	Causative Verb	English Translation
fahmidan	*fahm*	understand	*fahmândan* *fahmânidan*	make understand
tarsidan	*tars*	fear	*tarsândan* *tarsânidan*	frighten
bar gashtan	*bar gard*	return, come back	*bar gardândan* *bar gardânidan*	turn back (someone)

- **Auxiliaries:** Persian conjugation uses a number of auxiliaries in the compound forms. The enclitic form of the auxiliary *budan* (be) is the one used in the formation of the perfect forms of all verbs. The verb *xâstan* (want) is used as an auxiliary in forming the future tenses. The auxiliary *shodan* (become) forms the passive constructions.

The complete inflectional system can be obtained by the various combinations of these elements as illustrated for the Active voice in the table below.

Table 2: Conjugation paradigm (Active voice)

Mood	Tense	Prefix	Stem	Inflection	Auxiliary
	Infinitival	--	Past	*n*	--
	PresentParticiple	--	Present	*ndh*	--
	PastParticiple	--	Past	*h*	--
Indicative	**Present**	*my*	Present	Present	--
	Preterite	--	Past	Past	--
	Imperfect	*my*	Past	Past	--
	Perfect	--	Past	*h*	Aux: Present
	Pluperfect	--	Past	*h*	Aux: Preterite
	Compound Imperfect	*my*	Past	*h*	Aux: Present
	DoubleCompound	--	Past	*h*	Aux: Imperfect
	Future	--	Past	--	AuxFuture: Present[a]

Mood	Tense	Prefix	Stem	Inflection	Auxiliary
Subjunctive	**Present**	*b/by*	Present	Present	--
	CompoundPast	--	Past	*h*	Aux: Subjunctive Present
Imperative	**Present**	*b/by*	Present	Imperative	--

a. AuxFuture is placed *before* the Past Stem in forming the Future tense.

Thus, to obtain the Indicative Present, the imperfective prefix *my* is combined with the present stem of the verb followed by the present inflection for person and number. The Indicative Pluperfect, however, results from the combination of the past participle (past stem of the verb + *h*) and the auxiliary *budan* (to be) in the preterite tense. Similarly, the complete conjugation paradigm for the Passive voice can be obtained by combining the past participle with the passive auxiliary verb *shodan* in the corresponding tense form, e.g., ***Indicative Preterite*** = Past stem + *h* + *shodan* (in Preterite).

5 MORPHOLOGICAL GRAMMAR

The Samba language uses typed feature structures and a unification-based declarative framework to describe morphology. The basic element of a morphological description is a *morphological rule* which associates a surface form, representing a sequence of morphemes, to a set of morphological features. The surface form is formally represented as a regular expression on characters. The morphological features are specified as a feature structure that contains the lexical and inflectional information provided by the rule. These feature structures describe how the stem and the morphological features of the affixes are combined. The examples discussed in this section demonstrate how certain morphological properties of Persian are represented in the specification language.

5.1 Simple rules

As an illustration, consider the rules for analyzing the plural morphemes on nouns. Recall that Persian includes several plural morphemes, based on the phonology, the animacy value or the etymological origin of the word. The first set of plural morphemes is described in the rule `NominalPlural1`. In Samba, string variables are prefixed with the dollar sign; regular expressions are enclosed between angle brackets and a transliteration is used in order to represent the morphemes. Concatenation is represented by space, and optionality by `?`.

```
NominalPlural1= <
    $stem = <Character Character+>//surface string has at least
                                  //two characters
    <                             //the possible plural endings
        < <$stem \ Vowel> "yan"> |
```

```
        < <$stem \ Consonant> "yn"> |
        < <$stem \ "y"> "vn"> |
        < <$stem \ NonVowel> <"an" | "at">> |
        < "~"? "ha" >
    >
    per.Noun[                    //the morphological features
        exp: “$stem$”,
        lex.regular: True,
        infl.number: per.Plural]
>;
```

In this example, `$stem` represents the surface form of the word, which can contain two or more characters. The format `<$stem \ VALUE >` indicates that the stem ends in the character group represented by `VALUE`. Hence, in the case of the first morpheme, the rule `<$stem \ Vowel> "yan">` indicates that the stem ends in a vowel and it is followed by the plural morpheme *yan*, and the rule `<$stem \ NonVowel> <"an" | "at">>` indicates that the stem ends in a non-vowel (a consonant or a “y”) and is followed by either the morpheme *an* or the morpheme *at*. The final form character is represented by “~”. Its optionality indicates that the morphological analyzer will be able to recognize the plural morpheme *ha* whether it appears in attached or detached form.

If any of the possible plural endings have been recognized, the feature structure describing the morphological features is unified. The part of speech categories are defined as types in this system [cf. Carpenter 1992]. The citation form is stored under the path `exp`; and any lexical feature available from the dictionary, such as the regularity of the noun, can be found under the path `lex`. The features added through unification by the morphological analyzer are stored under the path `infl`. In this example, the stem `$stem` is stored as the citation form; this string is used later to look up the words in the dictionary. This plural rule requires that the word be a regular Noun, and it assigns the value Plural to the *number* feature.

Now consider the Plural rule below which analyzes the second set of plural morphemes that can appear on a noun. These morphemes (“*gan*”, “*at*”, “*Jat*”) appear only after consonants and “y” (NonVowels) and replace the word-final “h” character of the singular form. Hence, in order to obtain the correct citation form, we need to add the character “h” to the stem that has been recognized.

```
NominalPlural2 = <
    $stem = <Character Character+>//surface string has at least
                                  //two characters
    <$stem \ NonVowel>           //stem ends in consonant or “y”
    <"gan" | "at" | "Jat">>      //followed by gan,at or Jat
    per.Noun[
        exp: "$stem$h",          // citation form = stem + “h”
        lex.regular: True,
```

```
        infl.number: per.Plural]//number feature is assigned the
                                 //value Plural
>;
```

5.2 Morphotactics

Since the morphemes that can appear on a word are ordered, the rules need to capture the relative ordering of the affixes. As an example, consider the Noun, on which may appear the plural suffix, the indefinite marker, the enclitic linking the noun to a relative clause, the pronominal clitic, the ezafe and the copula. The order in which these morphemes appear is quite constrained. The first suffix on a Noun is the plural morpheme. The second position is occupied by the ezafe, the indefinite, the enclitic, or the possessive pronominal clitic; these morphemes are in complementary distribution. The last morpheme to appear on the Noun is the Copula verb, but it may not follow the ezafe. The morphotactics of the Noun are shown below:

Morphotactics. [Noun + plural + ezafe + ...]^{np}

[[Noun + plural + indefinite]^{np} + copula]^{vp}

enclitic

clitic

Morphotactics are portrayed in the Samba grammar by making the output of a rule the input of the following one. For instance, in order to process the presence of the *ezafe* morpheme, the result of the `Number` rule, which recognizes the plural morphemes, is used as the input to the `Ezafe` rule, and is indicated by the variable `$base`.

Take, for instance, the input word *ketâbhây* (books (of)), which contains both a plural morpheme (*hâ*) and an ezafe suffix (*y*). The morphological analyzer first recognizes the plural morpheme when it processes the `Number` rule, and then the second suffix (*y*) when it passes through the `Ezafe` rule. In addition, the morphological analyzer needs to provide the correct citation form for the entry. In this case, the `Number` rule locates the plural morpheme and forms the correct citation form for the singular Noun (e.g., *ketabhay = ketab (*'book'*) + ha (*plural*)*). The feature structure formed by this rule is unified with the output of the `Ezafe` rule given below, which locates the ezafe morpheme "*y*".

```
Ezafe = <
    <$base = <Number>>      // base is Number rule
    <<$base \ Vowel> "y">   // base ends in vowel; followed by "y"
    <
    per.Noun[
        infl: [ezafe: True, // ezafe feature is set to True
               indefEncl: False,
               indefinite: False,
```

```
                enclitic: False,
                clitic.function: Null]]
    >
>;
```

Note that since the citation form has already been set during the recognition of the plural morpheme, it need not be set here. However, if the ezafe suffix is detected, the value for the *Ezafe* feature is set to True. Since the indefinite morpheme, the enclitic and the pronominal clitic are in complementary distribution with the ezafe morpheme, these other features can already be set to False or Null. Explicitly setting the values for the features eliminates ambiguities that might arise at later stages.

Using a complete morphological rule as the base input to another rule is one way of capturing the morphotactics in a natural language. A second way to represent morphotactics is simply by concatenation of rules, described in the following section.

5.3 Paradigmatic morphology

The conjugation or declension for a given verbal paradigm can be grouped together in a format that describes forms that belong to the same paradigm. These rules specify a disjunction of rules that share a common information (these rules are similar to *tables* as described in [Zajac 1998]). The following example describes the endings for the past tenses in Persian represented in a disjunction of rules. The first rule, for instance, looks for the first, singular morpheme “m” and if it is recognized, the structure is unified with the morphological features described under the path `infl.` In this particular case, the value for the feature `numberAgr` (number agreement) is set to `Singular`, and the value for `person` is set to `First`.

```
PastInfl= <
  <"m" per.Verbal[infl: [numberAgr:per.Singular, person:per.First]]> |
  <"y" per.Verbal[infl: [numberAgr:per.Singular, person:per.Second]]>|
  < per.Verbal[infl: [numberAgr:per.Singular, person:per.Third]]> |
  <"ym" per.Verbal[infl: [numberAgr:per.Plural, person:per.First]]> |
  <"yd" per.Verbal[infl: [numberAgr:per.Plural, person:per.Second]]> |
  <"nd" per.Verbal[infl: [numberAgr:per.Plural, person:per.Third]]>
>;
```

The Past Inflection is used in forming several of the past tenses in Persian. In the morphological grammar, the inflection forms are described separately as shown and are then used in conjugation rules that refer to the `PastInfl`. This is exemplified in the `SimplePast` rule below. The `SimplePast` rule analyzes the *Imperfect* and *Preterite* tenses in Persian. The *Imperfect* is formed by the

concatenation of the *my* imperfective prefix, followed by the past stem of the verb and the past inflection. The *Preterite* lacks the prefix; it is obtained by concatenating the past stem and the past inflection. As shown in the rule below, the concatenation of these elements can be specified by the concatenation of regular expressions. The rules for analyzing the Past stem (`PastStem`) and Past Inflection (`PastInfl`) are simply called by referring to the name of the rule. This allows for describing the morphotactics of rules by concatenating them, thus forming an ordered set of the rules (i.e., the rule `PastStem` applies before `PastInfl`). Any common features are "factored out" and specified in the form of a feature structure in the beginning of this rule. The successful application of the rule will add (unify) this structure to the output feature structure.

```
SimplePast= <
    per.Verbal[        //common features for Imperfect & Preterite
      infl:[voice: per.Active,
            mood: per.Indicative,
            participle: per.PartFalse]]
    <
      < <"my" "~"?>   //if there is a prefix, tense is Imperfect.
        per.Verbal[infl.tense: per.Imperfect]
      > |
      per.Verbal[infl.tense: per.Preterite] //if no prefix, tense
                                            //is Preterite.
    >
    PastStem       // the past stem as defined by PastStem rule.
    PastInfl       // past inflection as defined by PastInfl rule
                   // above
>;
```

6 SYSTEM ARCHITECTURE

The morphological analyzer described here is part of a large automatic translation system for Persian text. Like the morphological analyzer, all other components use typed feature structures with unification to represent linguistic objects and linguistic knowledge in the form of grammars, etc. The information flow between components is modeled by viewing hypotheses of various kinds as partial knowledge about a certain interval of the input. The underlying model for this is a layered chart, capable of representing heterogeneous types of hypotheses in an integrated way [Amtrup 1999].

The morphological analyzer computes the part of speech categories and returns the word's inflectional features in the form of a feature structure. Since the analyzer is not equipped with an internal dictionary, it has to produce all possible results, regardless of whether the proposed citation form can be found in a

dictionary or not. Usually, the morphological analysis of a word yields around ten structurally different analyses. This architecture shows two main advantages: First, the analyzer is not closely related to the dictionary in the system and does not have to be recompiled each time the dictionary is modified. Second, the handling of unknown words is greatly improved. Instead of supplying two different analyzers (one for words for which the citation form is known, and one for unknown words), or at least two passes through roughly the same analyzer (the second pass would ignore dictionary information), our system analyzes all words (known and unknown) in one integrated pass.

After morphological analysis, dictionary lookup removes unwanted analyses and augments the morphological information with lexical knowledge, such as regularity and English translations. The dictionary contains 50,000 entries which include single words, compounds, proper names and phrases. After dictionary lookup, on the average two valid analyses remain. These results function as the input to the syntactic parser, which creates phrase structure hypotheses, which in turn undergo transfer and generation (cf. [Amtrup et al. 1999] for a more detailed description of the system architecture).

Two examples of complete morphological analysis are shown below. The first example describes the word *keshvarhâye* (countries (of)), a noun appearing with a plural morpheme followed by the ezafe suffix.

```
Noun[
    lex : LexMorph[
        regular : True],
    infl : NominalInfl[
        number : Plural,
        clitic : Clitic[
            function : Null],
        ezafe : True,
        indefEncl : False,
        indefinite : False,
        enclitic : False],
    exp : "k^svr"]
```

The second example represents a verb with a detached auxiliary. The surface form of the verb is *neveshte~ast* (has written) and the analysis is given below.

```
Verb[
    infl : VerbalInfl[
        voice : Active,
        clitic : Clitic[
            function : Null],
        tense : Perfect,
        causative : False,
        negation : False,
        mood : Indicative,
```

```
        person : Third,
        participle : Pst,
        numberAgr : Singular],
    exp : "nv^stn"]
```

The system was tested on online Persian newspaper text. In its current preliminary implementation, the morphological analyzer processes approximately 100 words per second on a Pentium II PC. The system is also available for Solaris (Sun) and Linux (PC). The complete system takes about 3 seconds to analyze a sentence of average length (23 words).

7 CONCLUSION

In this paper, we provided a detailed descriptive analysis of Persian morphology, presenting the morphemes that appear on non-verbal parts of speech as well as the complete verbal paradigm and all corresponding morphotactics. We have also described the implementation of a morphological analyzer for Persian, which utilizes finite-state transducers, typed feature structures and a unification-based language. By using a combination of feature structures and unification, the formalism can handle long-distance dependencies in the word structure. It is also able to provide an elegant account of the morphotactics in the language. The grammars are easy to develop since the language uses a declarative framework for writing the rules. Furthermore, the morphological analyzer is a language independent tool and can easily be used in multilingual environments. In addition to Persian, a morphology of Arabic has also been implemented and tested in this formalism. The formalism has also been used to process morphological generation in English on the target language side.

The morphological analyzer for Persian could be further improved if the preprocessor responsible for joining detached morphemes to the stems were to be included inside the analyzer itself. In addition, the analyzer can be equipped with the ability of providing sequences of analyses for single words (which would be needed to handle attached copulas in Persian).

The morphological analyzer is an integral component of the Shiraz Persian-English machine translation system developed at the Computing Research Laboratory (CRL). In the future, we plan to use the same architecture for the morphological analysis and machine translation of other languages, such as Serbo-Croatian and Korean.

REFERENCES

Amtrup, Jan W., 1999. *Incremental Speech Translation*. Lecture Notes in Artificial Intelligence 1735, Springer Verlag, Berlin, Heidelberg.

Amtrup, Jan W., Karine Megerdoomian and Remi Zajac, 1999. "Rapid Development of Translation Tools". In *Proceedings of Machine Translation Summit VII*, Singapore, pp.385-389, September 1999.

Antworth, Evan L., 1990. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*. Occasional Publications in Academic Computing No. 16. Dallas, TX:Summer Institute of Linguistics.

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

Carpenter, Bob, 1992. *The Logic of Typed Feature Structures.* New York, NY: CambridgeUniversity Press.

Comrie, Bernard, 1990. *The World's Major Languages.* Oxford University Press.

Gholamali Zadeh, Khosrow, 1995. *Sakht-e Zaban-e Farsi [Structure of Persian Language].* Ahya ketab Publishers, Tehran, Iran.

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

Mahootian, Shahrzad, 1997. *Persian.* Routledge.

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

Zajac, Remi, 1998. "Feature Structures, Unification and Finite-State Transducers". In *FSMNLP'98: International Workshop, on Finite State Methods in Natural Language Processing*.

Karine Megerdoomian is a doctoral student in Linguistics at the University of Southern California in Los Angeles. She is currently working as a computational linguist at Computing Research Laboratory, New Mexico State University, PO Box 30001, Las Cruces, NM 88003, USA. She can be reached at karine@crl.nmsu.edu.

The research described in this paper was funded in part by DoD, Maryland Procurement Office, MDA904-96-C-1040. The Samba formalism was designed by Remi Zajac; it was developed and implemented by Jan Amtrup.