Persian Computational Morphology: A Unification-Based Approach

Karine Megerdoomian

Memoranda in Computer and Cognitive Science
MCCS-00-320

Computing Research Laboratory
New Mexico State University
Las Cruces, New Mexico
April 2000

Abstract

This report provides a complete descriptive analysis of Persian inflectional morphology from a computational perspective. The parts of speech and the morphemes that appear on them as well as their corresponding morphotactics are presented in detail. The verbal paradigm is also described in this document. Since the morphological analyzer designed for this project uses a unification-based grammar with typed feature structures, the morphological information has been defined in terms of features and values. The report describes the current version of the morphological analyzer used in the Shiraz project and discusses any morphological elements that have not been included in this version, mostly due to the colloquial usage of these morphemes. Sample rules of Samba, the grammar specifying the morphological analyzer, as well as the feature specification for the Persian type definitions module are also described.

Contents

1

Introduction

Persian morphology is an affixal system consisting mainly of suffixes and a few prefixes. There is a relatively small number of affixes in the language and they follow a regular morphotactic order. Certain difficulties arise, however, 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*][1] 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, ambiguities arise due to the fact that 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 parser needs to use the information available from the parts of speech that the morpheme appears on, as well as the syntactic context of the lexical element carrying the affix. The morphological analyzer also needs to be able to recognize all the possible surface forms of the affixes, since certain affixes always appear bound to the following or preceding morphemes, while other affixes can be bound as well as free morphemes. So, for instance, the plural morpheme *at* [pronounced *ât*] is always bound to the nominal element it appears on. The plural morpheme *ha* [pronounced *hâ*], on the other hand, can be either attached to the previous morpheme or appear as a free affix. The morphological analyzer obviously needs to recognize both forms of such affixal elements.

This paper presents all the inflectional morphemes needed for a computational analysis of Persian morphology. A basic knowledge of the writing system is assumed. The possible surface forms of the affixes as well as the morphotactics are listed. In addition, any phonological or part of speech information that could help to disambiguate the morphemes is also discussed. This paper also describes all the features and the corresponding values that are needed in order to produce a complete formal specification of the Persian morphological analyzer. The specification language, Samba, utilizes typed feature structures and unification to describe the linguistic information associated with morphemes.

Contemporary Persian distinguishes between the colloquial speech and a more formal written style. Although the morphology employed in colloquial speech is discussed, this report is mainly aimed at the analysis of Persian written text; the writing style of the poetic literature is not included.The morphological analyzer described in this report was tested on online journalistic prose, which consisted mainly of the Hamshahri online newspaper[2].

[1] The romanization used in this report is described in the Appendix. The transliteration used for the pronunciation of the words, given in square brackets, can also be found in the same section.

[2] This newspaper can be found at the following URL: http://www.neda.net

This report is organized as follows: Section 2 presents the Persian parts of speech along with all the morphemes that attach to these categories. Section 3 consists of a list of the inflectional morphemes that appear on the constituents of noun and prepositional phrases. This section also describes the surface realizations and variants of these affixes, and lists the parts of speech they appear on. A thorough description of the verbal inflectional paradigm is given in Section 4. The verbal morphology is quite regular and the complete paradigm can be obtained using a few elements such as present and past stems, aspectual prefixes, and past or present inflection. Section 5 brings together all the features and their corresponding values as used in the specification of the morphological analyzer. Actual examples from the Samba grammar are presented in Section 6, while Section 7 develops the specification of the types and feature structures for each lexical item.

2

Parts of Speech

This section introduces the Persian lexical parts of speech used in Samba, the grammar specifying the morphological analyzer. In Section 2.1, the Non-Verbal lexical categories, and in Section 2.2, the Verbal parts of speech are discussed. The affixes associated with each lexical category as well as their morphotactics are also explained. The morphotactics show the relative ordering of the morphemes on a part of speech, thus indicating the order in which the morphological rules need to apply. The affixes and their morphological functions are represented in the Samba grammar by a list of *features*. This section introduces the features that have been assigned to each lexical part of speech.

The selection of the morphemes that have been included in the system was influenced by several factors: Since the analyzer is to be used on written text, only the affixes which are transcribed were considered; morphemes that are usually indicated by the presence of a short vowel, which is not written in text, can not be detected and are therefore not included. In addition, emphasis was placed on constructions that appear in written text rather than in colloquial speech. Although the colloquial usage of the morphemes are discussed in this paper, they have not always been incorporated within the part of speech structures in the system. The frequency of some affixes in our journalistic corpus has been another factor in determining the necessity of analyzing certain constructions.

All the inflectional morphemes presented in this section are described in detail in Sections 3 and 4.[3]

2.1 Non-Verbal Parts of Speech

2.1.1 Noun

Nouns can be either singular or plural. They may also be accompanied by the *ezafe* (a suffix that connects the elements in a phrase), the indefinite marker, the enclitic particle[4] (a suffix that links words to the following relative clause), the possessive clitic pronoun and the copula. Of these morphemes, the indefinite marker, enclitic element and (possessive) clitics are in complementary distribution since they all indicate the Noun Phrase boundary. This is illustrated in the Morphotactics listed here. In addition, if an *ezafe* is detected, the lexical element could not carry any of the other morphemes since the *ezafe* marks links *within* the Noun Phrase, indicating that the phrase boundary has not yet been reached[5]. All of these morphemes can appear following the

[3] Most of the terminology used in this paper is based on Lazard (1992).

[4] An enclitic is a clitic which attaches to the end of a word.

plural morpheme. The copula may attach itself to any of these morphemes except for the *ezafe*. All morphemes are optional. Note that although the Copula is a morpheme that appears on the Noun, it is a verbal element (see Section 3.7) and in principle, should be analyzed as a separate feature structure[6].

Morphotactics. [Noun + number + ezafe + ...]^{np}
[[Noun + number +indefinite]^{np} + copula]^{vp}
enclitic
clitic

2.1.2 Adjective

Adjectives can appear with suffixes indicating degrees of comparison, namely the comparative and superlative affixes. In addition, the copula can attach to the adjectival element. Persian Adjectives do not show any form of agreement and the purely Adjectival element does not vary in number. But since the adjective appears within a Noun Phrase, it can take any of the morphemes that attach to the last element of that phrase; hence, the adjective can carry the *ezafe*, the indefinite marker, the enclitic, and the possessive clitic morphemes. Again, all morphemes are optional.

When the adjective carries a superlative morpheme, it can behave syntactically as an adjective (e.g., can modify a noun) or it can behave as a noun. For instance, the adjective *bzrg* [*bozorg*] "big" has the superlative form *bzrgtryn* [*bozorgtarin*]"biggest" which can be an adjective as in *bzrgtryn xanh* [*bozorgtarin khAne*]"biggest house", or a noun *bzrgtryn hstm* [*bozorgtarin hastam*]"I am the biggest". In this last case, the adjective + superlative construction can appear with a plural morpheme as in the example

ex.	*q^sng-tryn-ha*	*ra*	*myxvahm*
	[ghashangtarinhâ	*râ*	*mikhâham]*
	prettiest -pl	OBJ	I-want
	'I want the prettiest ones.'		

We will therefore include the feature "number" indicating plurality on the Adjectival structure although it rarely appears in written text.

Certain adjectives in Persian are capable of functioning as a noun. In such constructions, the adjective can take all the morphemes that attach to the noun such as the plural morpheme, the indefinite marker, the enclitic element and the clitics. These lexical elements have been defined as both Noun and Adjective in the dictionary. If the adjective is behaving as a noun, it will be treated as a pure nominal element (see Section 2.1.1). So, for instance, the entry *bykar* [*bikâr*] "unemployed" could be either an adjective or a noun, and it is input as both in the dictionary.

Morphotactics. [Adjective + comparison + number + ezafe + ...]^{np}
[[Adjective + comparison +number+indefinite]^{np} + copula]^{vp}
. enclitic
. clitic

[5] The "+..." following the *ezafe* in the Morphotactics structure indicates that the NP boundary has not yet been reached and other Noun Phrase constituents should follow.

[6] Since the current system does not separate feature structures, the Copula has not been included in the morphological grammar.

2.1.3 Adverb

Adverbs are often invariable in number. However, there is a class of adverbs (mostly adverbs of place and time) that can carry the plural suffix, which then indicates approximation or emphasis (e.g. *az balaha aftad* [*az bâlâhâ oftâd*], from above-PL fell “it fell from somewhere up there”; lit. “it fell from aboves”). Although this construction is colloquial, we have included the “number” feature in the adverbial structure in our morphological analyzer. Time and place adverbs can also take *ezafe*. The pronoun clitic may appear on an adverbial, usually resulting in a partitive construction (see Section 3.5.3 “Partitive Clitics”). Since this clitic is limited to colloquial usage, it has not been included in our current morphological grammar. Certain adverbs, mainly manner adverbs, may appear with the comparative suffix (e.g., *zvdtr*[*zudtar*] “more quickly”, *bhtr*[*behtar*] “better”, *Azadanh~tr*[*âzâdâne-tar*] “more freely”)[7].

Morphotactics. [Adverb + comparison + number + ezafe +...]^{np}
[Adverb + comparison + number + clitic]^{np}

2.1.4 Preposition

The Prepositions in Persian are divided into the two categories of Simple and Compound Prepositions. There are only a few Simple Prepositions in Persian; Compound Prepositions are formed by joining a Simple Preposition to a Noun, but since this Noun has lost its original meaning in these cases, the complete construction can be treated as a Preposition. Examples of Simple Prepositions are: *az*[*az*] “from”, *dr*[*dar*] “in”. Compound Prepositions based on these simple ones include: *az p^st*[*az posht*] “from behind”, *dr p^st*[*dar posht*][8] “behind”, where *p^st* [*posht*] is a noun meaning “back”.

The difference between these two categories of Prepositions lies in the fact that Compound Prepositions can appear with the *ezafe,* which links them to the following element, or with the object clitics. Simple Prepositions, on the other hand, do not carry the *ezafe* morpheme and only a few of them appear with the clitic (namely *bh*[*be*] “to”, *bray*[*barâye*] “for”, *ba*[*bâ*] “with”, *az*[*az*] “from”).

Since the prepositions with a written *ezafe* morpheme seem to be lexicalized forms and their number is anyhow quite limited, we have decided to include them in the lexicon with two forms: one with *ezafe* and one without (e.g, *tv*[*tu*] and *tvy*[*tuye*] “in”). We allow, however, the attachment of the clitic to prepositions. The clitics receive an Object interpretation (accusative Case). (e.g., *brayman* [*barâyemân*] “Lit: for + Clitic/1pl; = for us”) .[9]

Morphotactics. [preposition + clitic]^{pp}

[7] Certain adverbs, mainly manner adverbs, can behave like adjectives. For instance, the word *ea^sqanh [âsheghâne]* can be used as an adverb as in *ea^sqanh ngah^s krd* [*âsheghâne negâhash kard*] ‘He looked at her lovingly’. It can also be used as an adjective as in *ngahy ea^sqanh* [*negâhi âsheghâne*] ‘a loving look’. These elements are marked as both Adjective and Adverb in the dictionary.

[8] Note that the Simple Preposition *dr* is optional in this construction as in many Compound Prepositions. In such cases, the nominal parts of these Compound constructions have been lexicalized and, in colloquial speech, can be used as prepositions independently.

[9] The preposition *bh*[*be*] “to” can itself join to the following element in writing, hence it needs to be included as a morpheme on the Noun Phrase constituents. This prefix is not yet incorporated in our system.

2.1.5 Numerals, Number Units and Classifiers

The cardinal Numerals and the Classifiers (referred to as Numeratives in Lazard (1992)), are used to form numerical expressions. Number Units refer to numbers such as *hzqr [hezâr]* “thousand”, *mylyvn [milion]* “million”. The distinction between Numerals and NumberUnits was made mainly for syntactic reasons. Classifiers are often inserted between the numeral and the noun when speaking of concrete beings or objects. Some examples of these classifiers are:

- *ta* [*tâ*] - used for count nouns
  ex. *dv ta ktab* [*do tâ ketâb*]
  “two CL books”
- *nfr* [*nafar*] - used for people
  ex. *pnJ nfr flsTyny* [*panj nafar felestini*]
  “five CL Palestinians”
- *qladh* [*ghalâde*] “collar” - used for dogs
  ex. *^cnd qladh sg* [*chand ghalâde sag*]
  “a few CL dogs”

Three types of Ordinal constructions exist in Persian, formed by attaching their respective morphemes to the cardinal number or number unit. These three ordinal-forming morphemes are in complementary distribution (i.e., only one of these morphemes can attach to the word at any time). We will represent the ordinals by the feature “numberType”.

Other affixes may occur on the Numerals, Number Units and the Classifiers in colloquial speech, but we have not encountered any of them in our corpus, and thus they will not be included in the morphological analyzer at this point. These morphemes are as follows: The classifer may appear with an indefinite article (which indicates approximation). The numeral, number unit and the classifier can take an *ezafe* that links the numerical expression to a noun and designates a partitive complement (e.g., *dv tay Anha* [*do tâye ânhâ*] “Lit: two CL+ezafe they; = two of them”)[10]. Partitive clitics also attach to the numeral, number unit and the classifier in colloquial constructions (e.g., *dv tay^san*[*do tâyeshân*] “Lit: two CL+clitic/3pl; = two of them”). In addition, the copula can attach itself to a classifier. The morphotactics of these suffixes are shown below, although they are not included in our system (number units and numerals display the same morphotactics).

Morphotactics. [Numeral + numberType + ezafe + ...]^{np}
[Numeral + numberType + clitic]^{np}

Morphotactics. [Classifier + ezafe + ...]^{np}
[Classifier + indefinite]^{np} + copula]^{vp}
enclitic
clitic

2.1.6 Pronouns

We include two types of Pronouns: Personal Pronouns and Quantifier Pronouns (such as *everyone* or *someone*). Pronouns do not appear with inflectional affixes, but they could be inherently marked for Number and Person in the lexicon. These lexical features should be entered in the dictionary for each pronoun. Furthermore, the surface form of the pronoun is the same in Persian regardless

[10] In written text, this construction is usually represented with the preposition *az*[*az*] “from”, as in *dv ta az anha*[*do tâ az ânhâ*] “Lit: two CL from they; = two of them”.

of its Function within the sentence (i.e., accusative, possessive, etc.). The pronoun *mn [man]*, for instance, can be used as the subject, the object, or the possessor in a sentence. In order to distinguish the various English senses, however, the corresponding function of the Pronoun could be included in the dictionary. For the case of *mn [man]*, it could be translated as "I" with function *subject*, "me" as *object* and "my" as the *possessor.* Its Person and Number features will be entered as *first* and *singular* respectively. Certain Pronouns are already in plural form. They are to be marked as "irregulars" in the dictionary using the Regular feature. In the current version of the system, however, the lexical features of the Pronouns are not entered in the dictionary; they are instead treated as non-inflecting parts of speech. The function the Pronouns assume within a sentence is determined during syntactic analysis and marked in the feature structure of the sentence; this information is then used for translating into the correct English pronominal form.

Although the lexical features for both Personal and Quantifier Pronouns are identical, the two Pronoun types are marked as distinct parts of speech due to the differences in their syntactic behavior. The Quantifier Pronoun, for instance, is often used in Partitive constructions that the Personal Pronoun cannot appear in (e.g., *beZy az Anha* [*baazy az ânhâ*] "Lit: some from they; = some of them"). The two Pronoun types are therefore listed as two different feature structures in our system.

2.2 Verbal Parts of Speech

2.2.1 Copula

The Copula represents the enclitic form of the verb *bvdn* [*budan*] "to be" in the Present Indicative. The Copula, by virtue of being a verbal element, also carries verbal features of Tense, Mood, Number and Person. In the specific case of the Copula, the Mood is always marked as *Indicative*, and the Tense is *Present*.

The Copula is itself a morpheme which attaches to nouns and adjectives. However, since it is a verbal element and not a nominal one, in principle the Copula should be separated from the word it is attached to and a new structure should be created to present the properties (features) of the Copula. Hence, unlike other inflectional affixes, the Copula is not to be marked as a feature on the structure of the stem it appears on. In our current system, however, lacking a component that separates a feature structure into two distinct ones, we have represented the Copula as a feature on the Nominal and Adjectival structures.

2.2.2 Verb

The number of simple verbs in Persian is relatively limited. Many verbal predicates, known as Light Verb constructions, are compound verbs consisting of a preverbal element (which can be a noun, a preposition, or an adjective) accompanied by a light verb (see next Section).

The forms of the simple finite verb are characterized by a complete personal inflectional system. They are divided into three moods: indicative, subjunctive, imperative. The indicative consists of a rather large number of tenses that express relations of time or nuances of aspect. The subjunctive has two tenses and the imperative has only one. In addition to the personal forms, the verb can appear in infinitival or participial forms. The infinitival is the citation form of the verb and is listed in the dictionary. Verbs can appear in active or passive voice, and may have a causativized stem.

The negative morpheme appears on the verb in simple conjugated forms and on the auxiliary in compound forms. The verbal conjugation system is discussed in detail in Section 4.

If the verb is transitive, the object clitic could appear on it. In the case of conjugated Light Verb constructions, the object clitic may appear attached either at the end of the light verb or on the preverbal element (preceding the light verb). (see Section 3.5.2 “Object Clitics” for examples). If the verb is intransitive, a subject clitic may appear on the verb.This clitic can not show up on an infinitival (see Section 3.5.4 “Subject Clitics”). In impersonal verbal constructions, the clitic appears on the preverbal element of compound verbs (see Section 3.5.5 “Impersonal Clitics”). Since subject and impersonal clitics are extremely rare in journalistic text, only the object clitics have been included in our current morphological grammar.

2.2.2.1 Infinitival

The infinitival is the citation form of the verb. If the verb is transitive, the infinitival may appear with an object clitic (e.g., *ps az dydn^s*[*pas az didanash*] “Lit: after from see+clitic/3sg; = after seeing him/her”). For the infinitive form of Light Verbs, however, the clitic can only appear at the end of the verbal form and can not attach to the preverbal element as it does with the conjugated verb forms. Syntactically, the infinitive often behaves like a nominal element as the gerunds do in English. In such cases, the clitic that appears on the infinitival is interpreted as a Possessive clitic (e.g., *zn bvdn^s*[*zan budanash*] “Lit: woman be+clitic/3sg; = her being a woman”). The infinitive verb is often a noun phrase constituent, hence the ezafe can appear on it. However, since infinitival verbs always end in the character “n”, and since the ezafe is only written after vowels, the ezafe following an infinitival is never present in Persian text.

2.2.2.2 Auxiliaries

There exist three different verbs that can also be used as auxiliaries in forming the various verbal tenses in Persian. Auxiliaries are marked as verbs in the dictionary and therefore have the same structure as verbs do. They are distinguished by their citation forms, which determine the function of the auxiliary in the verbal conjugation paradigm. The various citation forms of the auxiliary verb distinguish between the passive and active voices as well as the future tense (see Section 4.1.6 “Auxiliaries”). In order to analyze the verbal paradigm, these auxiliary verbs need to be considered as part of the verbal inflectional system. The auxiliary can be fully inflected for mood, tense and aspect; it may also appear with a clitic.

2.2.2.3 Modals

Certain verbs can be used as modals, which are often used in impersonal constructions (i.e, to express possibility or necessity of an action without explicitly mentioning an agent). Modals are tagged as verbs in the dictionary and carry the same structure as verbs. The different modal types are distinguished by their citation forms. Since modals may be separated from the main verb by intervening elements, their relation to the verb is analyzed at the syntactic level. Since modals are verbal elements, they can appear with any of the morphemes that attach to the verbs.

2.2.3 Light Verb

Most verbal constructions in Persian are formed using a light verb. The number of verbs that can be used as light verbs is limited. These structures consist of one or more preverbal elements, which could be a noun, adjective or preposition, followed by a light verb. The light verb in these constructions has partly or completely lost its own meaning and its main function is to derive an expression of a verbal nature when accompanied by a preverbal element. These constructions are known in the literature as Light Verbs or Compound Verbs. Since these verbs are often noncompositional in meaning (i.e., a word by word English translation can not be obtained), they are included in the dictionary.

The elements of a light verb construction can be separated by other lexical elements such as the object of the verbal construction or an adjective-adverb, therefore we will analyze these verbal constructions as Light Verbs only at the syntactic level. The inflectional morphemes, however, that appear on the parts of the Light Verb construction will be analyzed by the morphological analyzer. Verbal inflection can only appear on the light verb itself, but bound morphemes can be attached to the preverbal element as well as to the light verb. Each part of the construction will be analyzed separately (i.e., the verbal part as a verb, the nominal part as a noun, etc.) and the complete construction will be put together at a later level in morphology.

2.3 Non-Inflecting Parts of Speech

The other parts of speech in Persian are the Proper Noun, Title, Postposition, Relativizer, Conjunction, Interrogative, Determiner, Quantifier, and Interjection. Affixes do not appear on these lexical elements, hence these parts of speech will not be analyzed for morphology.

3

The Inflectional Morphemes of Non-Verbal Constituents

This section describes all the non-verbal inflectional morphemes that will be used in the analysis of Persian morphology as well as the features used to represent each morpheme in the Samba specification. Although certain morphemes that appear on the verb are presented, the inflections of the verbal conjugation system are not discussed here; they will be described in full in the following section (see Section 4).

3.1 Comparison Morphemes

The morphemes indicating comparison appear on adjectives. They may also attach to adverbs when such an interpretation is available. These morphemes could mark a comparative or a superlative, and may appear either attached to the word or in a detached form as shown in the examples below[11].

- Comparative
  tr [tar]
  ex. *q^sngtr [ghashangtar] (prettier)*
  q^sng~tr [ghashangtar] (prettier)

- Superlative
  tryn [tarin]
  ex. *q^sngtryn [ghashangtarin] (prettiest)*
  q^sng~tryn [ghashangtarin] (prettiest)

The feature "comparison" is represented on the structure of Adjectives and Adverbs in order to indicate whether a comparative or superlative morpheme was detected on the lexical element.

[11] The symbol ~ indicates that the previous character ends in a final form and thus is not attached to the following letter.

3.2 Plurals

3.2.1 The Plural Morpheme

When a nominal element is singular, it carries no particular mark, but plurality is marked by a plural morpheme which appears immediately following the nominal.[12] The plural can also attach to an adjectival element if the latter is used as a noun (e.g., *sfyd-ha [sefid-hA] (white-plur)* which could be translated as "the white ones"). Infinitivals and certain adverbials may also take the plural morpheme. The plural is indicated by the following suffixes:

- *ha [hâ]*
  This morpheme is the most productive and can appear on any nominal element. This affix can be either detached or attached. The attached form, however, can not appear on words ending in a silent 'h'[13].

- *an [ân]*
  attaches to nouns usually designating animacy. It only appears on words ending in a consonant or 'y'.
  ex. *mrdan [mardân](=men), asban [asbân](=horses), drxtan [derakhtân](=trees),*
  ayranyan [irâniyân](=Iranians)
  Certain nouns indicating temporal notions and nouns representing parts of the human body may also appear with this morpheme.
  ex. ^*sban [shabân](=nights), lbân [labân](=lips)*
  There exist some inanimate nouns which may also take this morpheme.
  ex. *axtran [akhtarân](=stars), sxnan [sokhanân](=words)*
  Variants:
- *yan [yân]*
  an [ân] becomes *yan [yân]* when following the vowels 'a' [*â*] and 'v' [*u*].
  ex. *danayan [dânâyân] (=wise [persons]), dan^sJvyan [dâneshjuyân] (=students)*
- *gan [gân]*[14]
  a*n [ân]* becomes *gan [gân]* when attached to a word ending in silent 'h'. The character 'h' is eliminated when the plural is formed.
  ex. *bndh [bande] (servant)* --> *bndgan [bandegân],*
  mjdh [mozhde] (presage) --> *mjdgan [mozhdegân]*

- *yn [in]*
  usually attaches to nouns that are originally Arabic participles. It does not appear on words ending in a (written) vowel.
  ex. *msafryn [mosâferin] (travelers)*

- *at* [*ât*]
  usually appears on nouns of Arabic origin.

[12] Note that "irregular" plurals are already stored in the lexicon as plurals and are marked as irregular nouns. They need not be analyzed for a plural suffix, hence the morphological analyzer should verify that the nominal element is *regular* before analyzing it for plurality.

[13] The silent 'h' refers to a word-final *h* which is treated as a vowel and is pronounced as *[e]*. See Section 5.1 "Lexical Features" for a more detailed explanation of the silent 'h' in Persian.

[14] According to Mahootian (1997) this morpheme is lexicalized on the nouns. A study of our corpus, however, indicates that plural nouns appearing with *gan[gân]* are frequent enough to justify its inclusion in a morphological rule.

ex. *Hyvanat* [*heyvânât*] *(animals)*, *frmay^sat* [*farmâyeshât*] *(instructions)*
Note: Certain nouns ending with this morpheme are not compositional and since their translation into English can not be obtained by analysis, they should be included in the lexicon as irregular plurals.
ex. a*ntxab* [*entekhâb*] *(choice)* --> *antxabat* [*entekhâbât*]*(elections),*
t_Zahr [*tazâhor*] *(pretention)* --> *t_Zahrat* [*tazâhorât*] *(demonstrations)*
Others do not have a singular form at all and indicate a collective meaning.
ex. *mxlfat* [*mokhalafât*] *(paraphernalia)*

- *at [ât]*
  can also appear on a word ending in silent 'h'. The character 'h' is eliminated when the plural is formed.
  ex. *klmh* [*kalame*] *(word)* --> *klmat* [*kalamât*] *(words),*
  lH_Zh [*lahze*] *(second)* --> *lH_Zat* [*lahazât*] *(seconds)*
  Variants:
- *Jat [jât]*
  when attached to a word ending in silent 'h'. The character 'h' is eliminated when the plural is formed.
  ex. *myvh* [*mive*] *(fruit)* --> *myvJat* [*mivejât*] *(fruits),*
  karxanh [*kârkhâne*] *(factory)* --> *karxanJat* [*kârkhânejât*] *(factories)*
  This affix is not very productive in forming Persian plurals, and seems to be somewhat lexicalized (Mahootian 1997, p.191).

- *vn [un]*
  attaches to certain nouns ending in 'y'. This plural suffix is very rare and the plural nouns formed using this method are usually considered lexicalized.
  ex. *anqlaby* [*enghelâbi*] *(revolutionary)* --> *anqlabyvn* [*enghelâbyun*]*(revolutionaries)*

Plural marking on the nominal elements is represented with the "number" feature. In the case of Nouns, this feature appears under the inflectional feature structure as well as the lexical one. The inflectional "number" feature determines whether an element is marked *singular* or *plural* after morphological analysis (i.e., whether a plural morpheme was detected on the nominal element). The lexical "number" feature is to indicate whether the entry is an irregular plural. So, for instance, if the lexical entry always appears in plural form or if it is an Arabic plural (see next section), the entry will be input in the dictionary as carrying a Plural value. The lexical and inflectional "number" features are linked in the structure of the Noun.

3.2.2 Arabic Loans

Persian contains a large number of Arabic loan words but it also includes a considerable number of morphological borrowings. The main area in inflectional morphology in which the Arabic borrowings are noticeable is in the formation of plural nouns.

There are some plural forms in Persian that follow the Arabic template morphology. Arabic nouns often form plurals based on a consonantal root into which vowels and consonants are inserted. Some examples are illustrated below:

ex. *ktab [ketâb]* --> *ktb [kotob] (books)*
_Zrf [zarf] --> *_Zrvf [zoruf] (dishes)*
fqyr [faghir] --> *fqra [fogharâ] (poor [people])*

These ‘broken’ or ‘discontinuous’ plurals were borrowed into Persian. 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 *Hal [hâl]* “health” and its Arabic plural form *aHval [ahvâl]* used interchangeably, with the meaning “health”. There are cases in which the meaning of the plural has diverged from that of the singular as in *sbb [sabab]* “reason” and *asbab [asbâb]* “goods”. In addition, the word *arbab [arbâb]* is in fact a plural form but is used as a singular in Persian meaning “master”.

Due to the small number of occurrences of these ‘broken’ plurals in Persian texts as well as the fact that this is not a productive word-formation tool in the language, these plural forms need not undergo an analysis according to the Arabic plural morphology. For the purpose of a computational analysis of Persian, the ‘broken’ plurals can be input in the lexicon as irregular plurals. The “number” feature for these words will be marked *plural* in the lexicon. Thus, the morphological analyzer will not analyze these irregular plurals for number when they are encountered.

3.3 Ezafe

The elements within a Nominal Phrase are linked by the enclitic particle called *ezafe*. This construction is used to relate a lexical element to its modifiers or to the possessor. The *ezafe* can also appear on Prepositions linking them to their Noun Phrase complement. Although traditional grammars often attempt to classify the *ezafe* constructions according to their semantics, the role of the *ezafe* is better understood in terms of marking of nominal determination and it indicates nothing as to the nature of the semantic relation between the linked elements. In most cases, this relation can be translated as a genitive structure in English.

The *ezafe* can appear on nouns, infinitives, adjectivals, quantifiers or prepositions. In addition, it could attach to numbers or classifiers in colloquial speech (See “Numerals, Number Units and Classifiers” on page 6). This morpheme is usually an unwritten vowel since short vowels are often not realized overtly in Persian text. The *ezafe* can have an orthographic realization in the following situations:

- *y*
  if the NP element ends in the vowels ‘v’ [*u*] or ‘a’ [*â*]. The *ezafe* suffix ‘y’ is obligatory in written text.

ex. *Sda-y pa-y mn*	ex. *bv-y xv^s*	ex. *rv-y myz*
[sedâ-ye pâ-ye man]	*[bu-ye khosh]*	*[ru-ye miz]*
sound-ez foot-ez my	smell-ez sweet	on-ez table
‘(the) sound of my footsteps’	‘sweet smell’	‘on the table’

  Note: the *ezafe* cannot appear after the vowel ‘y’.

- ;
  the “hamze” character follows an element that ends in silent ‘h’. This suffix is optional in written text.

ex. *xanh-;* *Hsn*
[khâne-ye hasan]
house-ez Hasan
'Hasan's house'

This morpheme is represented with the "ezafe" feature on the structures of the Noun, the Adjective and the Adverb. The orthographic realization of the *ezafe* suffix is marked by this feature. As discussed in Section 2.1.4, the prepositions appearing with the *ezafe* are treated as lexical elements. Since these prepositions are entered in the dictionary, the feature "ezafe" is not included on the prepositional structure in the morphological analyzer.

3.4 Indefinite/Enclitic Morpheme

3.4.1 Indefinite Article

There is no definite marker in Persian. There exists, however, an indefinite article which marks indefiniteness for the Noun Phrase. It usually attaches to the last element included in the construction, which could be a noun, an adjective, a past participle, or a classifier in colloquial speech. It can have the following forms:

- *y [i]*
  the indefinite morpheme takes this form after an element ending with a consonant.
  ex. *mrd-y [mard-i] (man-indef) -- 'a man'*
- *yy [yi]* or *iy [i]*
  if the lexical element ends in the vowels 'v' *[u]* or 'a' *[â]*.
  ex. *Sda-yy [sedâyi] (voice/sound-indef) -- 'a sound' / 'a voice'*
- *ay [i]*
  follows an element that ends in silent 'h' or 'y' (in final form); in these cases, it appears in detached form (indicated by the tilde ~)
  ex. *xanh~ay [khâne-i] (house-indef) -- 'a house'*

Note: The indefinite article can appear following the plural morpheme in which case the translation is 'some'
ex. xanh-ha-yy [khâne-hâ-yi] (house-plur-indef) -- 'some houses'

3.4.2 Enclitic Particle

This morpheme (also known as the demonstrative or referential suffix) joins a noun to a relative clause which determines it. This particle is often immediately followed by the relativizer *ke*, but it could also be separated from the latter by intervening elements. The form and position of the enclitic is very similar to that of the indefinite article; it is, however, different from the Indefinite since the Noun Phrase carrying this enclitic could be interpreted either as a definite or an indefinite. The enclitic attaches to the last element on the Noun Phrase; it can appear on nouns, adjectives, past participles, or classifiers in colloquial speech and it takes the following forms:

- *y [i]*
  after an element ending with a consonant.

- *yy [yi]* or *iy [i]*
  if the NP element ends in the vowels 'v' *[u]* or 'a' *[â]*.
- *ay [i]*
  follows an element that ends in silent 'h' or in 'y' (final form); it appears in detached from in such cases.

3.4.3 Indefinite/Enclitic Feature

Since the Indefinite and the Enclitic morphemes have identical surface forms, the feature "indefiniteEnclitic" was created in conjunction with the two features of "indefinite" and "enclitic". The presence of either one of these suffixes will trigger a positive value on the "indefiniteEnclitic" feature. Since only the structural position of the nominal element on which these morphemes appear can distinguish them, they will be disambiguated at the level of syntactic analysis. At this point, one of the features "indefinite" or "enclitic" will be marked True while the other will receive a False value.

3.5 Clitics

Persian possesses a series of enclitic personal suffixes representing the pronominal morphemes. Although they have the same surface form, clitics can have different functions depending on the part of speech or syntactic context that they appear on. Hence, the clitic could be a Possessive pronominal, an Object or Subject pronoun, or it could be interpreted as a Partitive or Impersonal clitic. Each of these functions is explained in full in this section.

The clitic is itself defined as a feature structure in our system, containing features to determine the "number" and "person" values of the pronoun. It also contains the feature "function", which designates the grammatical function of the pronominal clitic. The clitic is represented on the structures of the Noun, Adjective, Adverb, Preposition and Verb, since it can appear on any of these lexical categories.

The clitics have the following surface realizations:

- following a consonant

personal suffix		***person***	***number***
m	*[am]*	first	singular
t	*[at]*	second	singular
^s	*[ash]*	third	singular
man	*[emân]*	first	plural
tan	*[etân]*	second	plural
^san	*[eshân]*	third	plural

- after the vowels ‘a’ [â] or ‘v’ [u]

personal suffix		***person***	***number***
ym	*[yam]*	first	singular
yt	*[yat]*	second	singular
y^s	*[yash]*	third	singular
yman	*[yemân]*	first	plural
ytan	*[yetân]*	second	plural
y^san	*[yeshân]*	third	plural

- following certain words ending in silent ‘h’ and ‘y’ (in final form only). These clitics appear in detached form.

personal suffix		***person***	***number***
am	*[am]*	first	singular
at	*[at]*	second	singular
a^s	*[ash]*	third	singular
man	*[mân]*	first	plural
tan	*[tân]*	second	plural
^san	*[shân]*	third	plural

3.5.1 Possessive Clitics

The possessive clitics are equivalent to possessive pronouns in English; these constructions can be translated with a genitive formation. The possessive suffixes attach to the last element of the Noun Phrase, hence they can appear on nouns, adjectives, infinitivals, past participles and classifiers. In addition, they can appear on an adjectival which is being used as a definite noun (e.g., *sfyd^s [sefidash] (the white one)*).

ex. *ktab + t (book + Poss.2s)* --> *ktabt [ketâbat] (your book)*
ex. *dva + yman (medication + Poss.1pl)* --> *dvayman [davâyemân] (our medication)*
ex. *xanh + am (house + Poss.1s)* --> *xanh~am [khâneam] (my house)*

3.5.2 Object Clitics

The object clitics are accusative forms of the personal pronouns. Although the usage of these pronouns is usually considered to be limited to colloquial speech and poetry, they often appear in our corpus of journalistic prose. These clitics can be attached to prepositions and to transitive infinitivals as illustrated in the examples below:

- <u>Transitive Infinitival</u>
  zdn + ^san (hitting + Obj.3pl) --> *zdn^san [zadaneshân] (hitting them)*
  tnbyh krdn + tan (punishing + Obj.2pl) --> *tnbyh krdntan [tanbih kardanetân] (punishing you)*
- <u>Preposition</u>
  Jlv + ym (front + Obj.1s) --> *Jlvym [jeloyam] (in front of me)*

In addition, the Object clitics can appear on transitive verbs or light verbs. The clitic often attaches to the preverbal element of a light verb construction, but it could also appear on the verbal part as illustrated in the examples below:

- Transitive Verb
  dyd +m + t (saw + 1s + Obj.2s) --> dydmt [didamat] (I saw you)
- Compound Transitive Verb:
  Clitic on preverbal prepositional element
  br + ^s da^st (up + Obj.3s took) --> br^s da^st [baresh dâsht] (He/She picked it up)
  or Clitic on end of compound verb
  --> *br da^st^s [bar dâshtesh]*[15] *(He/She picked it up)*
- Compound Transitive Verb:
  Clitic on preverbal nominal element
  dvst + t dar +m (like + Obj.2s have +1s) --> dvstt darm [dustat dâram] (I like you / I love you)
  or Clitic on end of compound verb
  --> *dvst darmt [dust dâramat] (I like you / I love you)*

3.5.3 Partitive Clitics

This clitic may appear on adverbials, numerical expressions and interrogative elements with a partitive meaning, as exemplified here:

- Adverbial
  vsT + ^s (middle + Obj.3s) --> vsT^s [vasatash] (in the middle of it)
  bala + yman (up/above + Obj.1pl) --> balayman [bâlâyemân] (above us)
- Numerical Expression
  ^char ta + y^san (four NUM + Obj.3pl) --> ^char tay^san [châhâr tâyeshân] (the four of them)
- Interrogative
  kdam +^s (which + Obj.3s) --> kdam^s [kodâmash] (which one)

These constructions are quite rare in written text. They are currently not included in our morphological system.

3.5.4 Subject Clitics

The subject clitic appears on intransitive verbs (but not infinitivals) and only in the third person singular. The subject clitic is always used with a subject (i.e., it is an instance of clitic doubling). In English, it will not be translated. Since this clitic is usually used in colloquial speech and is quite rare in written text, it has not been included in our morphological analyzer.

ex. *aftad + ^s (fell.3s + Subj.3s) --> aftad^s [oftâdash] (he/she fell)*

[15] The clitic is often pronounced *[esh]* in colloquial speech.

3.5.5 Impersonal Clitics

This clitic is used in impersonal verbal constructions. It attaches to the preverbal element of compound verbs. It is not included in our current grammar.

ex. *xv^s + m Amd (please + Imp.1s came)*
--> *xv^sm Amd [khosham âmad] (I liked it* -- *lit: it pleased me)*

3.6 Ordinals

These morphemes are used to form ordinal numbers from cardinal ones. They are as follows:

- *m [om]*
  The ordinals formed with this morpheme behave as adjectives and adverbs. They indicate numerical rank.
  ex. *pnJm [panjom] (fifth), Sdm [sadom] (hundredth)*

- *my [mi]*
  This formation is used mostly in colloquial speech and for the lower numbers. It gives a definite value to the ordinal. These ordinals behave as adjectives.
  ex. *pnJmy [panjomi] (the fifth)*

- *myn [min]*
  These ordinals indicate the unity which completes a series; they are used especially in order to designate an anniversary. These ordinals behave like an adjective and can not be used as adverbs. By their formation, their syntax and their semantic value, these ordinals are akin to the superlative constructions.
  ex. *pnJmyn [panjomin] (the fifth), Sdmyn [sadomin] (the one hundredth)*

The different ordinal types are represented with the feature “numberType”. This feature determines whether anumber is *Cardinal* or *Ordinal*. Furthermore, it distinguishes between the three types of ordinals, since they each have different syntactic properties.

3.7 The Copula

The present indicative of the verb *bvdn [budan] (to be)* has a series of enclitic forms. These forms are used in the function of verbal copula when attached to Noun Phrases. The copula can appear on nouns, proper names, adjectives, past participles and classifiers. It also serves as auxiliary verb in the formation of the perfect forms of all verbs (See “Auxiliaries” on page 23). The current version of the morphological analyzer is unable to recognize constructions containing a Copula. This is due to the fact that the Copula is a verbal element that appears on nominal constituents. It is, therefore, required to be separated from the nominal feature structure in order to create its independent structure with its own POS value. The splitting fo feature structures is not implemented in the current system. Nevertheless, the enclitic forms of the copula are given below and should be included in a future version of the analyzer.

- following a consonant

copula		***person***	***number***
m	*[am]*	first	singular
y	*[i]*	second	singular
st or *ast*	*[ast]*	third	singular
ym	*[im]*	first	plural
yd	*[id]*	second	plural
nd	*[and]*	third	plural

- after the vowels ‘a’ [â] or ‘v’ [u]

copula		***person***	***number***
ym	*[yam]*	first	singular
yy	*[yi]*	second	singular
st	*[st]*	third	singular
yym	*[yim]*	first	plural
yyd	*[yid]*	second	plural
ynd	*[yand]*	third	plural

- following certain words ending in silent ‘h’ and ‘y’ (in final form only); the copula appears in detached form.

copula		***person***	***number***
am	*[am]*	first	singular
ay	*[i]*	second	singular
ast	*[ast]*	third	singular
aym	*[im]*	first	plural
ayd	*[id]*	second	plural
and	*[and]*	third	plural

Examples

ex. *gl + ym (flower + Cop.1pl)* --> *glym [golim] (we are flowers)*
ex. *zyba + st (beautiful + Cop.3s)* --> *zybast [zibâst] (he/she is beautiful)*
ex. *brdh + and (slave + Cop.3pl)* --> *brdh~and [bardeand] (they are slaves)*

As already mentioned, the Copula is itself separated from the lexical category that it appears on and is placed into a seperate structure. Since it is a verbal element, it needs to have features for “tense”, “mood”, “person” and “number”, in order to represent the various properties of the Copula.

4

Verbal Inflectional Paradigm

The inflectional system for the Persian verbs consists of simple forms and compound forms (forms that require an auxiliary verb). The simple forms are divided into two groups according to the stem they use in their formation. Every Persian verb has two stems: the Present Stem and the Past Stem. The former needs to be specified in the lexicon, while the latter is easily derived from the infinitival form of the verb. The citation form for the verbs is the infinitival.

Among the simple verbs, the tenses that are formed using the Present Stem are the present, the simple subjunctive, the imperative and the present participle. On the Past Stem are formed the preterite, the imperfect, the past participle. Among the compound forms, the future is formed on the Past Stem. All the other compound forms are based on the past participle.

The complete inflectional system is described in this section. To illustrate, the verb *frvxtn (to sell)* has been used throughout. Section 4.1 lists the elements that participate in the formation of the verbal inflectional system in Persian, and the sub-sections that follow explain how these elements combine to obtain the various tenses. Section 4.5 brings all the elements together by proposing a structure based on the surface order of the verbal parts.

4.1 Elements of Verbal Inflection

4.1.1 Stems

The infinitival form of the verb is available from the lexicon. In addition, the Present Stem will be included in the dictionary with its corresponding verb. The Past Stem, however, can be derived by eliminating the final letter ‘n’ from the infinitive.

ex.	***Infinitival:***	*frvxtn [forukhtan]*
	PresentStem:	*frv^s [forush]*
	PastStem:	*frvxt [foruxt]*

In “regular” verbs, the structure of the Past Stem is directly obtained from the Present Stem by adding one of the Past Stem Suffixes: *d, t, yd, ad*. So, for instance, the Past Stem of the verb *xvrdn [khordan] (to eat),* which has the Present Stem *xvr [khor]*, can be constructed by adding the suffix *d*.

ex.		
	Infinitival:	*xvrdn [khordan]*
	PresentStem:	*xvr [khor]*
	PastStem:	*xvrd [khord]*

The Past Stem of the “irregular” verbs can not be obtained so easily from the Present Stem since other variations exist in the latter. An instance of these verbs is the first example given above *frvxtn [forukhtan]* which has a Present Stem ending in *^s [sh]* instead of *x [kh]*. The reason for this is not phonological but rather historical: the Present and Past stems of the “irregular” verbs are in fact derived historically from two different verb roots; the two verbs have later collapsed into one and the two stems have been merged. This obviously makes it difficult to derive the Present Stem from the infinitival (which is the citation form of the verb) and therefore this stem is included in the dictionary[16].

4.1.2 Prefixes

There are two verbal prefixes and one negation particle. These prefixes appear attached to the beginning of the verb or in case of compound verbs, attached to the beginning of the light verb. The prefix ‘b’ and the negation particle are in complementary distribution (i.e., they can’t both appear in the same clause). Similarly, the two verbal prefixes are in complementary distribution. The negative particle appears before the imperfective marker ‘my’.

- *my [mi]*
  imperfective marker. It has a durative value and enters into the formation of the present indicative, the imperfect and the compound imperfect. It could appear in attached or detached form.
- *b [be]*
  a morpheme that characterizes the subjunctive and the imperative. In compound verbs which have the prepositional preverbal elements *dr [dar]* or *br [bar]*, this morpheme is often omitted. In other compound verbs, it is facultative. It never appears in detached form.
  variant:
  by [bi(y)]
  before the vowel ‘a’ *[â]*
  Note: in certain cases, this morpheme may replace the initial vowel ‘|’
  ex. *aftadn [oftâdan] (to fall) /* present stem= *aft [oft]* --> *byftyd [biyoftid] (fall.2pl [subjunctive/imperative])*
- *n [ne/na]*
  the negative morpheme. It is always attached to the following word.
  variant:
  ny [nay]
  before the vowel ‘a’ *[â]*
  Note: in certain cases, this morpheme may replace the initial vowel ‘|’
  ex. *aftadn [oftâdan] (to fall) /* present stem= *aft [oft]* --> *nyftyd [nayoftid] (do not fall)*

[16] In the Shiraz dictionary, we have included the present stem forms for all verbs whether they were regular or irregular.

4.1.3 Personal Inflections

Three types of personal inflections are used for conjugating the Persian verbal forms:
Note: the symbol ‘’ indicates an empty or null value.

present inflection		***person***	***number***
m	*[am]*	first	singular
y	*[i]*	second	singular
d	*[ad]*	third	singular
ym	*[im]*	first	plural
yd	*[id]*	second	plural
nd	*[and]*	third	plural

past inflection		***person***	***number***
m	*[am]*	first	singular
y	*[i]*	second	singular
‘’		third	singular
ym	*[im]*	first	plural
yd	*[id]*	second	plural
nd	*[and]*	third	plural

imperative inflection		***person***	***number***
‘’		second	singular
ym	*[im]*	first	plural
yd	*[id]*	second	plural

4.1.4 Participle-Forming Suffixes

Apart from the personal inflections, there are two verbal suffixes:

- *ndh [ande]*
  used to form the present participle
- *h [e]*
  used to form the past participle

4.1.5 Causation Morpheme

Causatives are obtained by adding the infix *|n [An]* or *|ny [Ani]* to the verb as illustrated below:

Verb Infinitive	Verb Present Stem	English Translation	Causative Verb	English Translation
fhmydn [fahmidan]	*fhm [fahm]*	understand	*fhmandn [fahmâAndan]* *fhmanydn [fahmânidan]*	make understand
trsydn [tarsidan]	*trs [tars]*	fear	*trsandn [tarsândan]* *trsanydn [tarsânidan]*	frighten
dvydn [davidan]	*dv [do (dav)]*	run	*dvandn [davândan]* *dvanydn [davânidan]*	make run
br g^stn [bargashtan]	*br grd [bar gard]*	return, come back	*br grdandn [bargardândan]* *br grdanydn [bargardânidan]*	turn back (someone)
svxtn [sukhtan]	*svz [suz]*	burn	*svzandn [suzândan]* *svzanydn [suzânidan]*	make burn
xvrdn [khordan]	*xvr [khor]*	eat	*xvrandn [khorândan]* *xvranydn [khorânidan]*	make eat

The location for this infix is at the end of the Present Stem of the verb. Hence, given the Infinitival form and the Present Stem of a verb, the causativized inflected verb can be analyzed as follows:

Causative Present Stem = Present Stem + *an* or *any*
Causative Infinitival = Causative Present Stem + *dn*
Causative Past Stem = Causative Infinitival - *n*
= Causative Present Stem + *d*

The inflectional paradigm can now be obtained based on these stems; the causative inflectional system is the same as the one for noncausative verbs.

4.1.6 Auxiliaries

Persian conjugation uses a number of auxiliaries in the compound forms (see Section 4.2.2 "Active Voice: Compound Forms"). Except for the present inflection of the auxiliary *bvdn [budan] (to be)*, all other auxiliary forms are written as independent words.[17]

4.1.6.1 Auxiliary *bvdn [budan] (to be)*

Present Stem = *ba^s [bâsh]*
Past Stem = *bvd [bud]*

This verb has two series of forms, one enclitic, and the other non-enclitic. The enclitic form of this auxiliary (henceforth AuxBe) is the one used in the formation of the perfect forms of all verbs. Its

[17] Modals are not discussed in this report since they are syntactic elements. Their conjugation will be analyzed as that of other verbs but the relation between the modal and the main verb in forming the verb phrase will be parsed at the syntactic level.

inflectional forms are listed below. All the other tenses of this auxiliary are unbound lexical elements[18].

present inflection		***person***	***number***
am	*[am]*	first	singular
ay	*[i]*	second	singular
ast	*[ast]*	third	singular
aym	*[im]*	first	plural
ayd	*[id]*	second	plural
and	*[and]*	third	plural

Although the present inflection of this auxiliary (listed above) is considered an enclitic, it never appears attached to the Past Participle of the verb, since the participle ends in a silent ‘h’[19]. The auxiliary enclitic form for the third person singular *ast [ast]* can sometimes appear separated from the participle by a space. No space precedes the other forms.

4.1.6.2 Auxiliary *xv|stn [khAstan] (to want)*

Present Stem = *xvah [khâh]*
Past Stem = *xvast [khâst]*

This verb is used as an auxiliary in forming the future tenses. It is henceforth referred to as AuxFuture.

4.1.6.3 Auxiliary ^*sdn [shodan] (to become)*

Present Stem = ^*sv [sho]*
Past Stem = ^*sd [shod]*

This auxiliary forms the passive constructions and will be indicated as AuxPassive.
In literary language, the verbs *g^stn [gashtan]* and *grdydn [gardidan]* (both meaning “to turn”, “to become”) are used as the Passive auxiliaries.

4.2 Conjugation

This section provides a description of the Conjugation system for all the tenses in Persian, explaining how each tense can be obtained using the elements given in Section 4.1 “Elements of Verbal Inflection”. Section 4.2.1 and Section 4.2.2 comprise verbs of the Active Voice, and Section 4.2.3 covers the Passive Voice constructions. The descriptions of the tenses as well as their usage is adapted from Lazard (1992).

[18] These tenses are irregular forms and are therefore listed in the morphological rules.

[19] See the section on ‘Silent h’on page 34.

4.2.1 Active Voice: Simple Forms

4.2.1.1 Present Participle

The present participle is mainly used as a nominal by forming an agent noun from the verb.

PresentParticiple **=** PresentStem + *ndh*

ex. *frv^s + ndh* --> *frv^sndh [forushande] (seller, salesperson / selling)*

4.2.1.2 Past Participle

The past participle is used to form compound tenses. It is also used as an adjectival.

PastParticiple **=** PastStem + *h*

ex. *frvxt + 'h'* --> *frvxth [forukhte]* *(sold)*

4.2.1.3 Present Indicative

The present tense is used to refer to the simple present as well as the present continuous in English. It is also used in the case of an action that begun in the past but which still continues in the present (ex. *az dyrvz ta Hala mynvysd [az dyruz tâ hâlâ minevisad] 'He/she has been writing since yesterday'*. In colloquial speech, the present tense can be used for the future.

Present **=** *my* + PresentStem + PresentInflection

ex.	*my + frv^s + m*	-->	*myfrv^sm*	*[miforusham]*	*(I sell / I am selling)*
	my + frv^s + y	-->	*myfrv^sy*	*[miforushi]*	*(you sell / you are selling)*
	my + frv^s + d	-->	*myfrv^sd*	*[miforushad]*	*(he/she/it sells / he/she/it is selling)*
	my + frv^s + ym	-->	*myfrv^sym*	*[miforushim]*	*(we sell / we are selling)*
	my + frv^s + yd	-->	*myfrv^syd*	*[miforushid]*	*(you sell / you are selling)*
	my + frv^s +nd	-->	*myfrv^snd*	*[miforushand]*	*(they sell / they are selling)*

4.2.1.4 Present Subjunctive

The subjunctive is used in instances where the realization of the action is not considered certain. It is used in contexts of doubt, desire, wish, possibility, etc. The subjunctive appears after the modals in a sentence. In addition, the first and third persons of the present subjunctive are used in questions whic are expressed by "shall I/we?" in English. The first person plural can be used in the sense of "lets". The subjunctive is very often used in the subordinate clauses whenever there is a nuance of eventuality, obligation, doubt, wish, etc.

Subjunctive **=** *b* + PresentStem + PresentInflection
The prefix *b* can sometimes be omitted.

ex.	*b + frv^s + m*	-->	*bfrv^sm*	*[beforusham]*	*(that I sell)*
	b + frv^s + y	-->	*bfrv^sy*	*[beforushi]*	*(that you sell)*
	b + frv^s + d	-->	*bfrv^sd*	*[beforushad]*	*(that he/she/it sells)*

b + frv^s + ym	-->	*bfrv^sym*	*[beforushim]*	*(that we sell)*
b + frv^s + yd	-->	*bfrv^syd*	*[beforushid]*	*(that you sell)*
b + frv^s + nd	-->	*bfrv^snd*	*[beforushand]*	*(that they sell)*

4.2.1.5 Imperative

The imperative is used to give an order, invitation, counsel, etc.

Imperative* = *b + PresentStem + ImperativeInflection

ex.	*b + frv^s +”*	-->	*bfrv^s*	*[beforush]*	*(sell)*
	b + frv^s + ym	-->	*bfrv^sym*	*[beforushim]*	*(let’s sell)*
	b + frv^s + yd	-->	*bfrv^syd*	*[beforushid]*	*(sell)*

4.2.1.6 Preterite

Also known as the Simple Past, this tense refers to a completed action. It is also used to describe an action which is about to be completed, so for instance, as an answer to “Are you coming?”, one might say *Amdm [âmadam] ‘I came’* in the meaning of “I am coming.” In a certain subordinate clause, the preterite is used to indicate an action that will be completed at the moment when the action expressed by the main verb will take place; a present would be used in English (ex. *vqty qhvh Amadh ^sd, ma ra Sda kn [vakhti ghahve âmâde shod, mâ râ sedâ kon] ‘When the coffee is ready, call us.’).*

***Preterite* =** PastStem + PastInflection

ex.	*frvxt + m*	-->	*frvxtm*	*[forukhtam]*	*(I sold)*
	frvxt + y	-->	*frvxty*	*[forukhti]*	*(you sold)*
	frvxt + ‘’	-->	*frvxt*	*[forukht]*	*(he/she/it sold)*
	frvxt + ym	-->	*frvxtym*	*[forukhtim]*	*(we sold)*
	frvxt + yd	-->	*frvxtyd*	*[forukhtid]*	*(you sold)*
	frvxt + nd	-->	*frvxtnd*	*[forukhtand]*	*(they sold)*

4.2.1.7 Imperfect

The imperfect expresses continuous, habitual or recurrent actions in the past. It is also used as the conditional tense (ex. *agr mytvanstm, Htma” my xrydm [agar mitavânestam, hatman mikharidam] ‘if I could, I certainly would have bought (it)’.*)

Imperfect* = *my + PastStem + PastInflection

ex.	*my + frvxt + m*	-->	*myfrvxtm*	*[miforukhtam]*	*(I was selling / I used to sell)*
	my + frvxt + y	-->	*myfrvxty*	*[miforukhti]*	*(you were selling / you used to sell)*
	my + frvxt + ‘’	-->	*myfrvxt*	*[miforukht]*	*(he/she/it was selling / he/she/it used to sell)*
	my + frvxt + ym	-->	*myfrvxtym*	*[miforukhtim]*	*(we were selling / we used to sell)*
	my + frvxt + yd	-->	*myfrvxtyd*	*[miforukhtid]*	*(you were selling / you used to sell)*
	my + frvxt + nd	-->	*myfrvxtnd*	*[miforukhtand]*	*(they were selling / they used to sell)*

4.2.2 Active Voice: Compound Forms

The Compound Forms mainly consist of an optional prefix attached to the Past Participle and combined with an auxiliary. The future and Progressive are formed by combining the auxiliary or modal with the verbal element. In the Compound Forms, the tense of the auxiliary determines the tense of the main verb, so, for instance, the auxiliary *bvdn [budan]* (to be) in Present tense is used to construct the Perfect tense while the Preterite tense of the auxiliary forms the Pluperfect. The tense of the auxiliary is formed following the rules described in this section (Section 4.2 “Conjugation”); in that sense, the verbal inflectional system is recursive. Note, however, that the Present tense of the auxiliary *bvdn* is irregular and is given in Section 4.1.6.1[20]

4.2.2.1 Perfect

The perfect refers to an action which has recently been completed or which started in the past but hasn’t yet been completed.

***Perfect* =** PastParticiple + AuxBe[Present]

ex.	*frvxth + am*	-->	*frvxth~am*	*[forukhteam]*	*(I have sold)*
	frvxth + ay	-->	*frvxth~ay*	*[forukhtei]*	*(you have sold)*
	frvxth + ast	-->	*frvxth~ast*	*[forukhteast]*	*(he/she/it has sold)*
	frvxth + aym	-->	*frvxth~aym*	*[forukhteim]*	*(we have sold)*
	frvxth + ayd	-->	*frvxth~ayd*	*[forukhteid]*	*(you have sold)*
	frvxth + and	-->	*frvxth~and*	*[forukhteand]*	*(they have sold)*

4.2.2.2 Compound Imperfect

This tense expresses a past action considered in its duration and which has taken place in a completed past. It is also used when speaking of bygone days.

***CompoundImperfect* =** ‘my’ + PastParticiple + AuxBe[Present]

ex.	*my + frvxth + am*	-->	*myfrvxth~am*	*[miforukhteam]*	*(I have been selling)*
	my + frvxth + ay	-->	*myfrvxth~ay*	*[miforukhtei]*	*(you have been selling)*
	my + frvxth + ast	-->	*myfrvxth~ast*	*[miforukhteast]*	*(he/she/it has been selling)*
	my + frvxth + aym	-->	*myfrvxth~aym*	*[miforukhteim]*	*(we have been selling)*
	my + frvxth + ayd	-->	*myfrvxth~ayd*	*[miforukhteid]*	*(you have been selling)*
	my + frvxth + and	-->	*myfrvxth~and*	*[miforukhteand]*	*(they have been selling)*

4.2.2.3 Pluperfect

The use of the pluperfect is very much the same as in English indicating that at a moment in the past an action was already completed. It is, however, also used as a descriptive tense (ex. *aystadh bvd [istâde bvd] ‘was standing’*).

[20] The final form character ‘~’ following the Past Participle in the examples for the Perfect and Compound Imperfect tenses indicates that the Past Participle ends in a silent ‘h’ and thus the inflectional forms of the Auxiliary are not attached to the ‘h’ character. For an explanation of the properties of the silent ‘h’, see the section on ‘Silent h’on page 34. Also see Section 4.1.6.1 “Auxiliary bvdn [budan] (to be)”

***Pluperfect* =** PastParticiple + AuxBe[Preterite]
AuxBe[Preterite] = AuxBe[PastStem] + PastInflection

ex.	*frvxth bvd + m*	-->	*frvxth bvdm*	*[forukhte budam]*	*(I had sold)*
	frvxth bvd + y	-->	*frvxth bvdy*	*[forukhte budi]*	*(you had sold)*
	frvxth bvd + ‘’	-->	*frvxth bvd*	*[forukhte bud]*	*(he/she/it had sold)*
	frvxth bvd + ym	-->	*frvxth bvdym*	*[forukhte budim]*	*(we had sold)*
	frvxth bvd + yd	-->	*frvxth bvdyd*	*[forukhte budid]*	*(you had sold)*
	frvxth bvd + nd	-->	*frvxth bvdnd*	*[forukhte budand]*	*(they had sold)*

4.2.2.4 Double Compound Past

This tense is the completed past of the perfect, indicating that an action was already completed. It is similar to the pluperfect, except that where the pluperfect refers to an anterior action with respect to the preterite, the double compound past expresses an anterior action in the context of the perfect tense. Sometimes, this tense replaces the pluperfect in order to indicate that a fact is not certain but only presumed.

***DoubleCompound* =** PastParticiple + AuxBe[Perfect]
AuxBe[Perfect] = AuxBe[PastParticiple] + AuxBe[Present]

ex.	*frvxth bvdh + am*	-->	*frvxth bvdh~am*	*[forukhte budeam]*	*(I had sold)*
	frvxth bvdh + ay	-->	*frvxth bvdh~ay*	*[forukhte budei]*	*(you had sold)*
	frvxth bvdh + ast	-->	*frvxth bvdh~ast*	*[forukhte budeast]*	*(he/she/it had sold)*
	frvxth bvdh + aym	-->	*frvxth bvdh~aym*	*[forukhte budeim]*	*(we had sold)*
	frvxth bvdh + ayd	-->	*frvxth bvdh~ayd*	*[forukhte budeid]*	*(you had sold)*
	frvxth bvdh + and	-->	*frvxth bvdh~and*	*[forukhte budeand]*	*(they had sold)*

4.2.2.5 Compound Subjunctive

Also known as the Past Subjunctive, this tense has the same modal value as the simple subjunctive but it expresses actions referring to the past. (ex. *mmkn ast gm ^sdh ba^sd [momken ast gom shode bâshad] ‘he/she may have gotten lost.’*)

***CompoundSubjunctive* =** PastParticiple + AuxBe[Subjunctive]
AuxBe[Subjunctive] = AuxBe[PresentStem] + PresentInflection

ex.	*frvxth ba^s +m*	-->	*frvxth ba^sm*	*[forukhte bâsham]*	*(that I have sold)*
	frvxth ba^s + y	-->	*frvxth ba^sy*	*[forukhte bâshi]*	*(that you have sold)*
	frvxth ba^s + d	-->	*frvxth ba^sd*	*[forukhte bâshad]*	*(that he/she/it has sold)*
	frvxth ba^s + ym	-->	*frvxth ba^sym*	*[forukhte bâshim]*	*(that we have sold)*
	frvxth ba^s + yd	-->	*frvxth ba^syd*	*[forukhte bâshid]*	*(that you have sold)*
	frvxth ba^s + nd	-->	*frvxth ba^snd*	*[forukhte bâshand]*	*(that they have sold)*

4.2.2.6 Future

The future tense expresses an action in the future. It is mainly used in formal and written speech; in colloquial speech the present tense is used to express the future.

Future **=** AuxFuture + PastStem
where AuxFuture refers to the present tense of the auxiliary without the prefix ‘my’.

ex.	*xvah + m frvxt*	-->	*xvahm frvxt*	*[khâham forukht]*	*(I will sell)*
	xvah + y frvxt	-->	*xvahy frvxt*	*[khâhi forukht]*	*(you will sell)*
	xvah + d frvxt	-->	*xvahd frvxt*	*[khâhad forukht]*	*(he/she/it will sell)*
	xvah + ym frvxt	-->	*xvahym frvxt*	*[khâhim forukht]*	*(we will sell)*
	xvah + yd frvxt	-->	*xvahyd frvxt*	*[khâhid forukht]*	*(you will sell)*
	xvah +nd frvxt	-->	*xvahnd frvxt*	*[khâhand forukht]*	*(they will sell)*

4.2.3 Passive Voice

A complete passive conjugation is formed with the help of the past participle followed by the AuxPassive auxiliary verb, regularly conjugated. The formation of all the tenses is listed below for the passive. Only the third person singular inflection is illustrated for each case.

The use of the passive is restricted in Persian, although it is more common in written and journalistic text. It is usually not used when the sentence can be expressed by the active voice. The passive is used particularly when the agent of the action is not expressed.

<u>Indicative</u>

Present **=** PastParticiple + AuxPassive[Present]
frvxth my^svd [forukhte mishavad] (is being sold)
Preterite **=** PastParticiple + AuxPassive[Preterite]
frvxth ^sd [forukhte shod] (was sold)
Imperfect **=** PastParticiple + AuxPassive[Imperfect]
frvxth my^sd [forukhte mishod] (was being sold)
Perfect **=** PastParticiple + AuxPassive[Perfect]
frvxth ^sdh ast [forukhte shode ast] (has been sold)
CompoundImperfect **=** PastParticiple + AuxPassive[CompoundImperfect]
frvxth my^sdh ast [forukhte mishode ast] (was being sold)
Pluperfect**=** PastParticiple + AuxPassive[Pluperfect]
frvxth ^sdh bvd [forukhte shode bud] (had been sold)
DoubleCompound **=** PastParticiple + AuxPassive[DoubleCompound]
frvxth ^sdh bvdh ast [forukhte shode bude ast] (had been sold)
Future **=** PastParticiple + AuxPassive[Future]
frvxth xvahd ^sd [forukhte khâhad shod] (will be sold)

<u>Subjunctive</u>

Subjunctive **=** PastParticiple + AuxPassive[subjunctive]
frvxth ^svd [forukhte shavad] (that he/she/it be sold)
CompoundSubjunctive **=** PastParticiple + AuxPassive[CompoundSubjunctive]
frvxth ^sdh ba^sd [forukhte shode bâshad] (that he/she/it has been sold)

<u>Imperative</u>

Imperative **=** PastParticiple + AuxPassive[imperative]
frvxth ^sv [forukhte sho](be sold)

4.3 Irregular Verbs

A few verbs do not follow the conjugation patterns described above and should be listed within the morphological grammar with their irregular forms. These verbs are the isolated present forms of *bvdn* [*budan*] “to be, to exist” in both positive and negative polarity, and the present forms of *da^stn* [*dâshtan*] “to have” which appear without the present tense prefix *my* [*mi*]. They are listed below.

present form of “bvdn”

hstm	*[hastam]*	*(I am, I exist)*
hsty	*[hasti]*	*(you are, you exist)*
hst	*[hast]*	*(he/she/it is, he/she/it exists)*
hstym	*[hastim]*	*(we are, we exist)*
hstyd	*[hastid]*	*(you are, you exist)*
hstnd	*[hastand]*	*(they are, they exist)*

present form of “bvdn” - negative

nystm	*[nistam]*	*(I am not)*
nysty	*[nisti]*	*(you are not)*
nyst	*[nist]*	*(he/she/it is not)*
nystym	*[nistim]*	*(we are not)*
nystyd	*[nistid]*	*(you are not)*
nystnd	*[nistand]*	*(they are not)*

present form of “da^stn”

darm	*[dâram]*	*(I have)*
dary	*[dâri]*	*(you have)*
dard	*[dârad]*	*(he/she/it has)*
darym	*[dârim]*	*(we have)*
daryd	*[dârid]*	*(you have)*
darnd	*[dârand]*	*(they have)*

4.4 Negation

In the Active voice, the negative morpheme appears on the beginning of simple verbs[21]. In the compound forms, the negative prefix also appears on the beginning of the conjugated verbal element since it attaches to the past participle of the main verb. The only exception is the Double Compound Past, in which the negative attaches to the past participle form of the auxiliary and not on the main verb. In addition, if a modal is present in the sentence, the negation can appear on the modal element. The Present Participle does not carry negation since it is often used as a nominal element. The negative affix can appear on Past Participles and Infinitivals. In the Passive voice, with the exception of the Future Passive, the negative prefix always appears on the passive auxiliary.

Examples are given below for each tense in the Active Voice. The negative prefix can be seen on the beginning of all the verb forms with the exception of the Double Compound Past. In this case,

[21] The negative prefix is in complementary distribution with the prefix used in the Subjunctive and Imperative moods.

the *n* morpheme is attached to the past participle of the auxiliary *bvdn*[*budan*] (to be). The negative prefix has the form *n* before consonants and the form *ny* before the vowels *a [â]* and *v [u]*.

Active Voice:

Past Participle	*nfrvxth [naforukhte]*	*(not sold)*
Present Indicative	*nmyfrv^sm [nemiforusham]*	*(I don't sell/ I am not selling)*
Preterite	*nfrvxtm [naforukhtam]*	*(I didn't sell)*
Imperfect	*nmyfrvxtm [nemiforukhtam]*	*(I wasn't selling/I didn't use to sell)*
Perfect	*nfrvxth~am [naforukhteam]*	*(I have not sold)*
Compound Imperfect	*nmyfrvxth~am [nemiforukhteam]*	*(I have not been selling)*
Pluperfect	*nfrvxth bvdm [naforukhte budam]*	*(I had not sold)*
Future[22]	*nxvahm frvxt [nakhâham forukht]*	*(I will not sell)*
Present Subjunctive	*nfrv^sm [naforusham]*	*(that I do not sell)*
Compound Subjunctive	*nfrvxth ba^sm [naforukhte bâsham]*	*(that I have not sold)*
Imperative	*nfrv^s [naforush]*	*(do not sell)*
Double Compound Past	*frvxth nbvdh~am [forukhte nabudeam]*	*(I had not sold)*

The Passive forms are listed below. In all tenses, the negative prefix attaches to the passive auxiliary ^*sdn*[*shodan*] (to become) and never on the verb. The only exception is the Future tense in the passive; in this case, the negative prefix attaches to the future auxiliary.

Passive Voice:

Present Indicative	*frvxth nmy^svd [forukhte nemishavad]*	*(is not being sold)*
Preterite	*frvxth n^sd [forukhte nashod]*	*(was not sold)*
Imperfect	*frvxth nmy^sd [forukhte nemishod]*	*(wasn't being sold)*
Perfect	*frvxth n^sdh ast [forukhte nashode ast]*	*(has not been sold)*
CompoundImperfect	*frvxth nmy^sdh ast[forukhth nemishode ast]*	*(was not being sold)*
Pluperfect	*frvxth n^sdh bvd [forukhte nashode bud]*	*(had not been sold)*
DoubleCompound	*frvxth n^sdh budh ast [forukhte nashode bude ast]*	*(had not been sold)*
Present Subjunctive	*frvxth n^svd [forukhte nashavad]*	*(that it not be sold)*
Compound Subjunctive	*frvxth n^sdh ba^sd [forukhte nashode bâshad]*	*(that it has not been sold)*
Imperative	*frvxth n^sv [forukhte nasho]*	*(do not be sold)*
Future	*frvxth nxvahd ^sd [forukhte nakhâhad shod]*	*(will not be sold)*

4.5 Structure of the Verbal Paradigm

The complete inflectional system can be obtained by the combination of the main elements of the verbal paradigm, which were introduced and discussed in Section 4.1. The tables below demonstrate how the various tenses can be formed by different combinations of prefixes, stems, inflections and auxiliaries[23]. Table 1 illustrates the verbal paradigm for the active voice and Table 2 the paradigm for the passive.

[22] Note that the auxiliary in the Future tense precedes the main verb stem.

[23] Since we analyze negation and causatives as properties external to the tense paradigm, these elements are not included in this table.

Table 1: Verbal Paradigm for the tenses in the 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
	CompoundImperfect	*my*	Past	*h*	Aux: Present
	DoubleCompound	--	Past	*h*	Aux: Imperfect
	Future	--	Past	--	AuxFuture: Present[a]
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.

Table 2: Verbal Paradigm for the tenses in the Passive voice

Mood	Tense	Prefix	Stem	Inflection	Auxiliary
Indicative	**Present**	--	Past	*h*	AuxPassive: Present
	Preterite	--	Past	*h*	AuxPassive: Preterite
	Imperfect	--	Past	*h*	AuxPassive: Imperfect
	Perfect	--	Past	*h*	AuxPassive: Perfect
	Pluperfect	--	Past	*h*	AuxPassive: Pluperfect
	CompoundImperfect	--	Past	*h*	AuxPassive: CompoudImperfect
	DoubleCompound	--	Past	*h*	AuxPassive: DoubleCompound
	Future	--	Past	*h*	AuxPassive: Future
Subjunctive	**Present**	--	Past	*h*	AuxPassive: Subjunctive
	CompoundPast	--	Past	*h*	AuxPassive: Subjunctive CompoundPast
Imperative	**Present**	--	Past	*h*	AuxPassive: Imperative

Causatives can be analyzed in all of the tenses shown above. The only difference is in the fact that the Present and Past Stems of the verb are modified to include the Causative morpheme. These stems can be obtained as follows:

Causative Present Stem = Regular Present Stem + *an* or *any*
Causative Past Stem = Causative Present Stem + *d*

After the tenses have been analyzed according to Tables 1 and 2, rules can be written to recognize the negative morpheme on the conjugated verbs.

5

Feature Structures

This section provides a complete summary by listing all the features and their corresponding values used in the Samba grammar. These features represent the lexical and inflectional morphemes used in the specification of the morphological analyzer, and the values assigned to each feature determine the function of the morpheme.

5.1 Lexical Features

The Lexical Features are the features that are stored in the dictionary for each entry. These features consist of properties of the lexical element needed by the morphological analyzer while applying the Samba grammar for the analysis of text.

In the dictionary, all lexical entries have been marked with a part of speech (POS); these parts of speech are compiled as Types in the system (e.g., POS=Noun is treated as a lexical element of type Noun). Some nouns may carry a number feature. The present stem of simple verbs is also given by a feature in the lexicon. Irregularity, such as for the case of irregular plurals, is indicated by a feature as well. The possible values for each feature are given below:

POS. Part of Speech of the lexical entry.

- pos: Noun, Adjective, Adverb, Preposition, Numeral, Classifier (or Numerative), Pronoun, QuantifierPronoun, Copula, Verb, LightVerb, ProperNoun, Title, Postposition, Relativizer, Conjunction, Interrogative, Determiner, Quantifier, NumberUnit, Interjection.[24]

Number. If the lexical entry always appears in plural form or if it is an Arabic plural (see Section 3.2.2), the entry will be input in the dictionary as carrying a Plural value. In this case, the feature Number is set to the value *plural*; this value is linked to the Number feature of the inflectional features described in Section 5.2. In such cases, the word is also marked as an "irregular" (see feature **Regular** below). If the lexical entry is not defined as a plural item, the value is set to *singular* as a default.

- number: Singular, Plural.

Regular. If a lexical item is to be treated as an *irregular* element, it is marked as such in the dictionary. For Persian, if Nouns or Pronouns have an irregular plural form, the value of this feature is set to *False*. These words then do not undergo the morphological rule determining plurality.

- regular: True, False.

[24] Some of these POS had to be rearranged in order to be used in syntactic analysis. For instance, pronouns and proper names had to be classified under a common type in order to capture certain identical syntactic behavior displayed by these parts of speech. The main classifications of the POS, however, remain as described here.

Present Stem. The present stem of simple verbs is input in the dictionary. This stem is used in conjugation. Its value can be any string.

- presentStem: String.

There exist additional lexical features that have not been included in the current specification of the morphological analyzer, but which could be very helpful in eliminating ambiguities in analysis. These lexical features could be added to the Samba grammar at a later date. They are described below:

SilentH. If a word ends in ‘h’, two possibilities arise. The final ‘h’ could be pronounced (e.g., *dh [deh] (village), bh [beh] (quince), kargah [kârgâh] (studio, workshop)*), in which case it is considered a consonant. If the ‘h’ is “silent”, it is pronounced as /e/ (e.g., *xanh [khâne] (house), synh [sine] (chest), prdh [parde] (curtain)*). The silent ‘h’ is always written in final form and, unlike the pronounced version, never attaches to the next character. The presence of the silent ‘h’ can therefore determine the surface form of the affixes appearing on the word (i.e., whether the affixes should be in attached or detached form).
The silent and pronounced ‘h’ at the end of a word can not be distinguished in written form, only in speech. In order to take this distinction into account in written form, this property has to be included as a feature in the lexicon. If the word ends in a silent ‘h’, the value of this feature could be set to *True*, otherwise the value is *False*.

- silentH: True, False.

Category. This feature determines whether the verbal element is transitive, intransitive or impersonal. If the verb carries a pronoun clitic, this knowledge can be used to determine its function (i.e., whether the clitic is accusative, a subject clitic, or a pronominal form used in an impersonal construction). This feature will also be very useful in syntactic analysis of Verb Phrases.

- category: Transitive, Intransitive, Impersonal.

Animacy. Persian morphological and syntactic behavior is often sensitive to the animacy of the lexical element. Certain plural morphemes, for instance, appear only on animate nouns. In syntax, the verb does not always agree with inanimate subjects, whereas subject-verb agreement always holds for animate elements.

- animacy: True, False.

Person. Pronouns are often inherently marked for number and person. If this information were included in the lexicon, translation quality would improve. The Person feature will indicate the person value of the pronoun.

- person: First, Second, Third.

5.2 Inflectional Features

The Inflectional Features are the features computed by the morphological analyzer.

Comparison. The Comparison feature indicates whether an adjective or adverb carries a morpheme denoting a *Comparative* or *Superlative* value. The *Null* value is used to indicate the absence of this morpheme on the lexical element.

- comparison: Comparative, Superlative, Null.

Number. This feature determines whether an element is marked *singular* or *plural*. Plurality is marked by the presence of the Plural morpheme on the lexical items.

- number: Singular, Plural.

NumberAgreement. This feature marks the number agreement on the conjugated verbal forms. It indicates whether the conjugation is in *singular* or *plural* form. The value *undefined* was used to explicitly set the number value of infinitival and participial forms of the verb.

- numberAgreement: Singular, Plural, Undefined.

Person. This feature appears on verbal elements to mark the person on the conjugated for. It could also appear on the structure for Clitics (see below). The possible values for this feature are *first, second* and *third*. The value *undefined* allows us to explicitly set the person value for tenses in which the person is not determined, such as in infinitives and participles.

- person: First, Second, Third, Undefined.

Function. The part of speech that the pronoun clitic appears on or the syntactic context where it occurs determines the grammatical function the clitic fulfills. The *null* value marks the absence of the clitic.

- function: Possessive, Object, Subject, Partitive, Impersonal, Null.

Tense. This feature determines the tense of the verb. If the verb is not conjugated, it will be marked as *Infinitive*. The participles are indicated as *participial* tense.The tenses listed below include nuances of aspect; they are described in full in Section 4.

- tense: Present, Preterite, Imperfect, Perfect, Pluperfect, CompoundImperfect, CompoundPast, DoubleCompound, Future, Participial, Infinitive.

Voice. Verbs can be active or passive. Infinitives and participles are treated as having an *undefined* voice.

- voice: Active, Passive, Undefined.

Mood. The mood of the conjugated verb is determined by this feature. The *undefined* value refers to infinitives and participles.

- mood: Indicative, Subjunctive, Imperative, Undefined.

Ezafe. The presence of this feature is marked by the Ezafe morpheme when present on the lexical item[25]. This feature will be used in syntax to indicate whether an element is attached to the following one, within a Noun Phrase construction. This feature can be marked absent if the word ends in a vowel yet no Ezafe was detected. The presence or absence of this morpheme is indicated by *EzTrue* or *EzFalse*, respectively. If the lexical word ends in a character following which the Ezafe is not overtly written, the morphological analyzer can not determine for certain whether the Ezafe is present or absent and should allow for both possibilities. In such cases, the value of the feature is set to *undefined*.

- ezafe: EzTrue, EzFalse, Undefined.

Indefinite/Enclitic. Since the Indefinite marker and the Enclitic morpheme have identical surface realization, this feature was created to capture the ambiguity. If the surface form for this feature is detected, the value is set to *True*, otherwise the value is marked as *False*. If the value is positive, the morpheme will be disambiguated in syntax based on the structural position of the element within the clause.

- indefiniteEnclitic: True, False.

When the morpheme has been disambiguated at the stage of syntactic analysis, either the Indefinite or the Enclitic feature will have to be marked positive. These features are described below:

Indefinite. Definiteness is not marked morphologically in Persian. There exists, however, an Indefinite morpheme. The values are *True* or *False*.

[25] By “present” we mean present in the written text. Since the ezafe morpheme is a short vowel, it is usually not transcribed in written text. It appears only after certain letters (see Section 3.3 “Ezafe”).

- indefinite: True, False.

Enclitic. The Enclitic, when attached to a word, relates the latter to the following relative clause. The values are *True* or *False*.
- enclitic: True, False.

Clitic. The pronominal Clitics have different forms based on Person and Number as well as their phonological context. Depending on which part of speech it appears on, each clitic fulfills a different grammatical Function. If no clitic has been detected, a *Null* value is used. (The Person, Number, Function values are as defined above for those features.)
- clitic:person: First, Second, Third.
  number: Singular, Plural.
  function: Possessive, Object, Subject, Partitive, Impersonal, Null.

NumberType. This feature determines whether a number is *Cardinal* or *Ordinal*. In addition, it distinguishes between the three types of ordinal numbers available in Persian. Since the cardinal form of a number does not carry an attached morpheme and only the ordinals appear with an affix attached to them, the cardinal value is chosen to represent the unmarked case.
- numberType: Cardinal, Ordinal1, Ordinal2, Ordinal3.

Participle. If a verbal element appears in Participial form, this feature is filled accordingly. If a Present Participle is detected, the feature is marked as *prsnt*, and a Past Participle is indicated as *pst*. In the cases of tenses that are not based on the Participle for their formation, the value of this feature is set to *False*.
- participle: Prsnt, Pst, False[26].

Causative. This feature determines whether the causation morpheme has been detected on the verb. The possible values are *True* or *False*.
- causative: True. False.

Negation. If the verb carries a negative morpheme, the value of the Negative feature is *True*. Otherwise, the feature is marked *False*.
- negation: True, False.

The formal specification of the feature structures for each lexical part of speech item, indicating all the possible lexical and inflectional features and values for that specific element, as specified in the ttype definitions module (Tango) and the morphological grammar (Samba), can be found in Section 7.

[26] The values *prsnt* and *pst* were chosen instead of the more normal *present* and *past* in order to avoid confusion with the values for the “tense” feature during unification.

6

Samba 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 within the Samba language.

6.1 Simple Rules

Consider the rules[27] for analyzing the plural morphemes on nouns given in this section. Recall from Section 3.2.1 “The Plural Morpheme”, 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 the Shiraz 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"> |
        < <$stem \ NonVowel> <"an" | "at">> |
        < "~"? "ha" >
    >
    per.Noun[exp: "$stem$",             // the morphological features
             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">` says that the stem ends in a vowel and it is followed by the plural morpheme *yan [yân]*, and the

[27] This rule is in the General Rule format, which is slightly different from the original Samba rules described in Zajac(1998).

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 [ân]* or the morpheme *at [ât]*. The final form character is represented by “~”.

If any of the possible plural endings have been recognized, the feature structure describing the morphological features is unified. The citation form is stored under the path `exp`; and any lexical feature available from the dictionary 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. The Plural rule below 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” or “at” or “Jat”
    per.Noun[
        exp: "$stem$h",         // citation form = stem + “h”
        lex: [regular: True],   // entry is a regular Noun
        infl: [number: Plural]] //number feature is assigned the value Plural
>;
```

6.2 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 or enclitic[28], 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 indefinite or enclitic morpheme, the result of the `Number` rule, which recognizes the plural morphemes, is used as the input to the `IndefiniteEnclitic` rule, and is indicated by the variable `$base`.

[28] Recall that the surface form of the indefiite and of the enclitic is identical, hence they cannot be distinguished at the morphological level. The feature *indefiniteEnclitic* is used to indicate that the morpheme is present; it will be disambiguated at the syntactic level.

Take, for instance, the input word *stargany [setârегâni] (some stars)*, which contains both a plural morpheme (*gan*) and an indefinte/enclitic 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 `IndefiniteEnclitic` 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., *stargan = starh (*‘star’*)* *+ gan (*plural*)*). The feature structure formed by this rule is transferred to the `IndefiniteEnclitic` rule given below, which locates the indefinite morpheme “*y*”.

```
IndefiniteEnclitic = <
   <$base = <Number>>            // for all these base is Number rule
   <
     <<$base \ Consonant> "y"> |     //base ends in consonant;followed by “y”,or
     <<$base \ Vowel> <"yy" | "iy">> | // base ends in vowel + “yy” or “iy”,or
       <<$base \ "h y"> "~ay">       // base ends in “h” or “y”;followed by “~ay”
   >
   <
    per.Noun[                         // word is a Noun
       infl: [indefiniteEnclitic: True, //indefiniteEnclitic feature set to True
              ezafe: per.EzFalse,
              clitic.function: Null]]
    |
    per.Adjective[                    // or word is an Adjective
       infl: [indefiniteEnclitic: True,// indefiniteEnclitic feature set to True
              ezafe: EzFalse,
              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 indefinite or enclitic suffix is detected, the value for the *IndefiniteEnclitic* feature is set to True. Since the ezafe and the pronominal clitic are in complementary distribution with the indefinite/enclitic 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. Since the morphological analyzer is unification-based, if the value of a feature is left undefined, it will cause ambiguous results in later components. For instance, a syntax rule might only operate on [ezafe: True], but if the value of the ezafe feature has been left undefined, the syntax rule would always unify and will therefore always apply. It is therefore better practice to explicitly assign the values for all the features during morphological analysis.

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. This will be explained in the following section.

6.3 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 table 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 `numberAgreement` is set to `Singular`, and the value for `person` is set to `First`.

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

The Past Inflection is used in forming several of the Past tenses in Persian. In the grammar, it is described separately and is then used in conjugation rules that refer to it. 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 unification 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 the rules thus forming an ordered set of the rules (i.e., the rule `PastStem` applies before `PastInfl`). Any common features (*paradigm* according to Zajac(1998)) 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 and
    infl:[                           // Preterite tenses.
        voice: per.Active,
        mood: per.Indicative,
        participle: per.PartFalse]]
    <
      < <"my" "~"?>                  // if there is a prefix, tense is Imperfect.
        per.Verbal[infl.tense: per.Imperfect]
      > |                            // if there is no prefix, tense is Preterite
      < per.Verbal[infl.tense: per.Preterite]
      >
   >
   PastStem                      // the past stem as defined by PastStem rule.
   PastInfl                      // past inflection as defined by PastInfl rule
                                 //  described above.
>;
```

7

Specification of Feature Structures

This section develops the type definitions for Persian, called per.Tango, which were discussed in previous sections. The first section provides a summary of the features for each part-of-speech category. These feature structures are used to encode lexical and morphological properties of lexical items. The following section provides the lexical and inflectional knowledge as specified in the types in the Tango module.

7.1 Summary of Features

The table below shows the Persian part of speech categories that appear with affixes. It also introduces the inflectional features that correspond to each part of speech. These results can be used to design the feature structures and type hierarchies.

Table 3: POS and corresponding inflectional features for Persian

	Noun	**Adjective**	**Adverb**	**Preposition**	**Numeral**	**Verb**
number	*yes*	*yes*	*yes*	*no*	*no*	*no*
ezafe	*yes*	*yes*	*yes*	*no*	*no*	*no*
indefiniteEnclitic	*yes*	*yes*	*no*	*no*	*no*	*no*
clitic*	*yes (Poss)*	*yes (Poss)*	*no*	*yes (Obj)*	*no*	*yes*
comparison	*no*	*yes*	*yes*	*no*	*no*	*no*
number Type	*no*	*no*	*no*	*no*	*yes*	*no*
person	*no*	*no*	*no*	*no*	*no*	*yes*
numberAgreement	*no*	*no*	*no*	*no*	*no*	*yes*
mood	*no*	*no*	*no*	*no*	*no*	*yes*
tense	*no*	*no*	*no*	*no*	*no*	*yes*
voice	*no*	*no*	*no*	*no*	*no*	*yes*
participle	*no*	*no*	*no*	*no*	*no*	*yes*
negation	*no*	*no*	*no*	*no*	*no*	*yes*
causative	*no*	*no*	*no*	*no*	*no*	*yes*

*The clitic is itself a feature structure. It carries the features for *number*, *person* and *function*.

The other parts of speech in Persian are the Proper Noun, Personal and Quantifier Pronouns, Title, Postposition, Relativizer, Conjunction, Interrogative, Determiner, Quantifier and Interjection. No inflectional morpheme appears on these lexical categories; they are classified as NonInflecting categories.

Light Verbs are considered as a part-of-speech category as well; however, they are constructed in post-morphology by putting the structures of the preverbal element(s) and the verb together. A light verb has three possible structures: the light verb may be formed by joining the verbal element to a noun (*NominalLV*), to an adjective (*AdjectivalLV*), or to a preposition (*PrepositionalLV*). These three subcategories are defined as subtypes of the Light Verb. Note that a NominalLV should have all the inflectional features that appear on a verb as well as all the nominal features; an AdjectivalLV has both adjectival and verbal features; a PrepositionalLV contains both prepositional and verbal features.

7.2 Feature Structures

The structure for each part-of-speech category is given below. The lexical information (*lex*) includes the features available in the dictionary for the part of speech category; the inflectional part (*infl*) describes the features obtained through morphological analysis.

```
NounMorphology = per.Noun[
                      lex: [number: Singular, Plural,
                           regular: Yes, No],
                      infl: [number: Singular, Plural,
                           ezafe: True, False, Undefined,
                           indefiniteEnclitic: True, False,
                           indefinite: True, False,
                           enclitic: True, False,
                           clitic: [number: Singular, Plural,
                                  person: First, Second, Third,
                                  function: Possessive, Object, Subject, Partitive, Impersonal, Null]]];

AdjectiveMorphology = per.Adjective[
                           infl:   [number: Singular, Plural,
                                  ezafe: True, False, Undefined,
                                  indefiniteEnclitic: True, False,
                                  indefinite: True, False,
                                  enclitic: True, False,
                                  clitic: [number: Singular, Plural,
                                         person: First, Second,Third,
                                         function: Possessive, Object, Subject, Partitive, Impersonal, Null],
                                  comparison: Comparative, Superlative, Null]];

AdverbMorphology = per.Adverb[
                         infl:     [number: Singular, Plural,
                                  ezafe: True, False, Undefined,
                                  comparison: Comparative, Superlative, Null]];

PrepositionMorphology = per.Preposition[
                           infl:  [clitic: [number: Singular, Plural,
                                        person: First, Second, Third,
                                        function: Possessive, Object, Subject, Partitive, Impersonal,Null]]];

NumeralMorphology = per.Numeral[
                               infl:[numberType: Cardinal, Ordinal1, Ordinal2, Ordinal3]];


VerbMorphology = per.Verb[
                    lex:   [presentStem: String],
                    infl:   [voice: Active, Passive, Undefined,
                           mood: Indicative, Subjunctive, Imperative,
                           tense: Present, Preterite, Imperfect, Perfect, Pluperfect, CompoundImperfect,
                                  CompoundPast, DoubleCompound, Future, Infinitive, Participial;
                           participle: Present, Past, Null,
                           numberAgreement: Singular, Plural, Undefined,
                           person: First, Second, Third, Undefined,
                           negation: True, False,
                           causative: True, False;
                           clitic: [number: Singular, Plural,
                                  person: First, Second, Third,
                                  function: Possessive, Object, Subject, Partitive, Impersonal, Null]]];
```

LightVerbMorphology = per.LightVerb[

```
lex:    [presentStem: String],
infl:   [voice: Active, Passive, Undefined,
         mood: Indicative, Subjunctive, Imperative,
         tense: Present, Preterite, Imperfect, Perfect, Pluperfect,
                CompoundImperfect, CompoundPast, DoubleCompound, Future,
                Infinitive, Participial;
         participle: Present, Past, Null,
         numberAgreement: Singular, Plural, Undefined,
         person: First, Second, Third, Undefined,
         negation: True, False,
         causative: True, False;
         clitic: [number: Singular, Plural,
                  person: First, Second, Third,
                  function: Possessive, Object, Subject, Partitive, Impersonal,
                                  Null]]];
```

ProperNounMorphology = per.ProperNoun;
PronounMorphology = per.Pronoun;
QuantifierPronounMorphology = per.QuantifierPronoun;
TitleMorphology = per.Title;
PostpositionMorphology = per.Postposition;
RelativizerMorphology = per.Relativizer;
ConjunctionMorphology = per.Conjunction;
InterrogativeMorphology = per.Interrogative;
DeterminerMorphology = per.Determiner;
QuantifierMorphology = per.Quantifier;
InterjectionMorphology = per.Interjection;

Note: NominalLightVerbs are obtained by the unification of the Noun and LightVerb structures. PrepositionalLightVerb structure is the unification of the Preposition and LightVerb structures. AdjectivalLightVerbs are obtained by unfying the structures for the Adjective and the LightVerb.

7.3 Tango Specification

Tango defines the feature structures used to encode lexical, morphological and syntactic properties of a lexical item for each part of speech types used in the morphological analyzer.

7.3.1 Specializing the Generic Entry for Persian Morphology

A set of language-independent definitions specify the generic structure of a lexical item. This structure also contains fields for encoding morphological and syntactic information. The definitions for Persian use the generic structure and add Persian-specific definitions for Persian morphology and syntax.

The *lex* feature includes the lexical knowledge which is available from the dictionary. The *infl* feature represents the inflectional data obtained by the morphological analyzer. Each Part of Speech inherits this two-part morphological structure. Note that some Part of Speech categories could actually be subtypes of other more basic categories.

The types for each part of speech are shown in the figure below. This hierarchy depicts the lexical type for each category, as well as the corresponding inflectional type given under the feature *infl*. The inflectional types are defined in the next section. Note that, apart from the Nominal and Verbal entries, which carry special lexical features (*number* and *regular* for the Nominal and *presentStem* for the Verbal), all other types use the generic lexical structure.

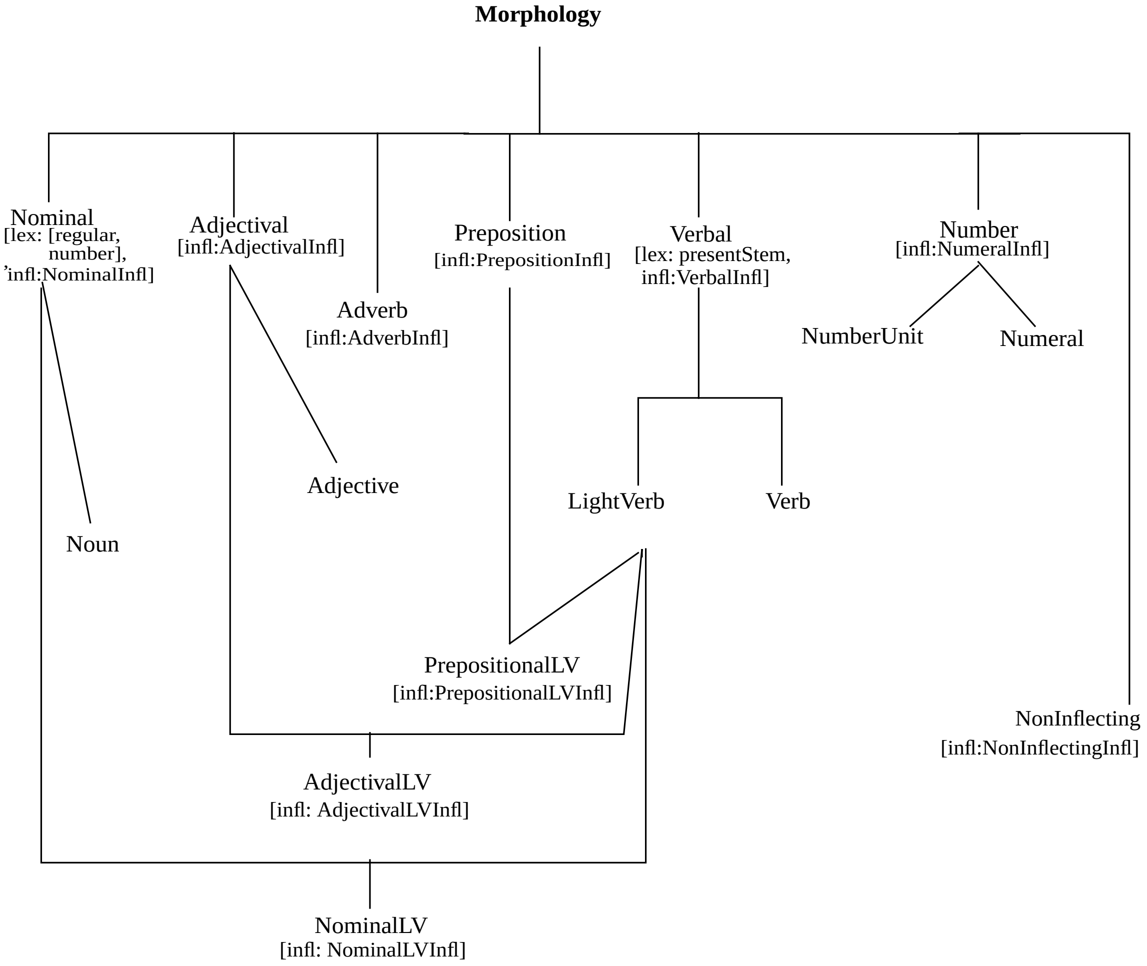

Figure 1: Morphology Types for the Persian Tango Module

7.3.2 Inflectional Knowledge

The inflectional information is again inherited from the language-independent module. The type *InflectionalMorphology* defines the inflectional knowledge obtained from a morphological analysis of the word.

The following tree structure shows the building of the final inflectional feature structures for each POS type. For instance, the Numeral entry has, in addition to the generic inflection, the feature *numberType*. Base2 is an intermediary type, which carries the *clitic*, *ezafe*, *indefinite* and *enclitic* features. This type joins (unifies) with another intermediary type, HasNumber, to form the type for Nominal entries (e.g., Noun) which carry all the features in types Base2 and HasNumber combined.

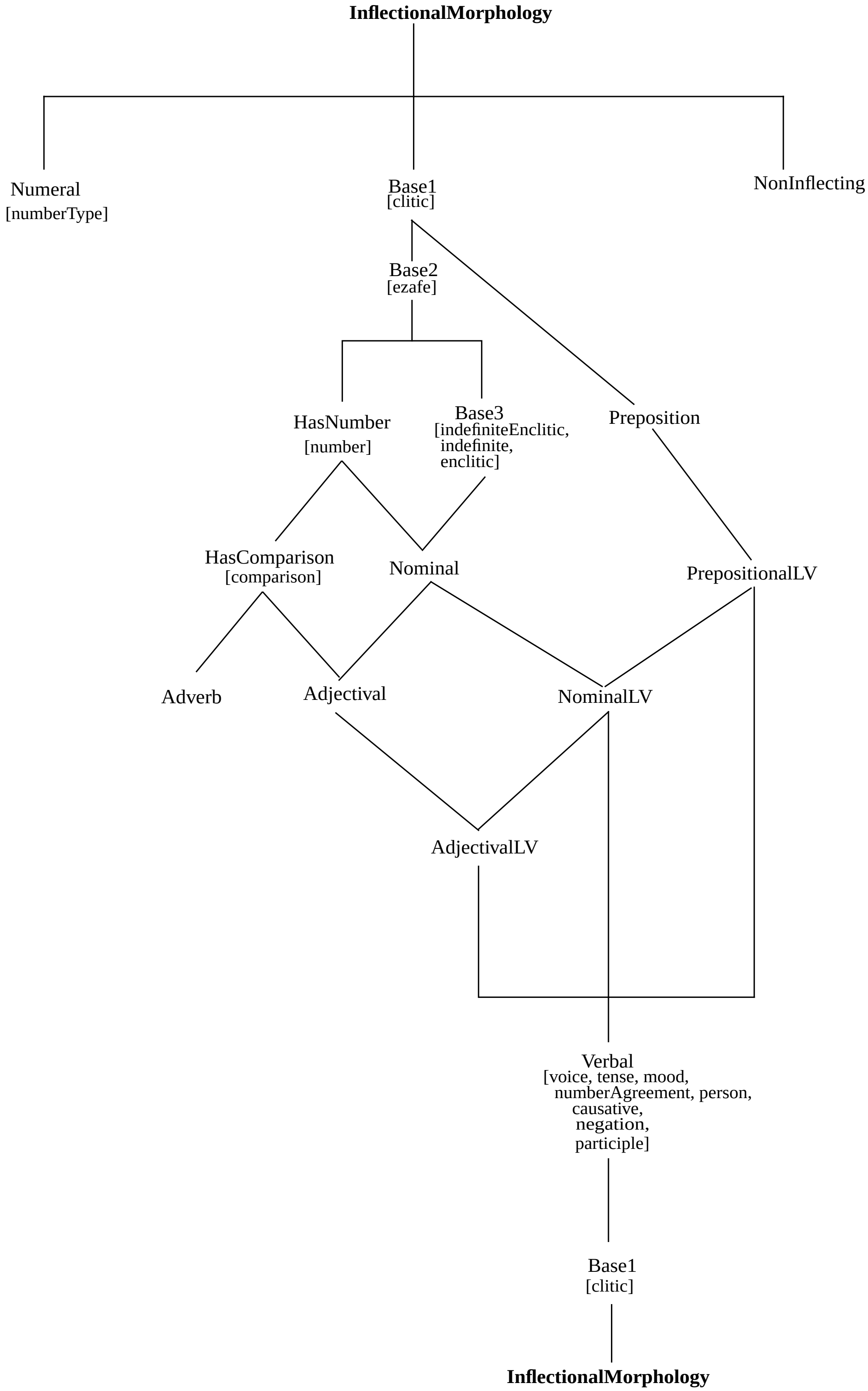

Figure 2: Inflectional Types for Persian Tango Module

8

Conclusion

This report provides a complete descriptive analysis of Persian inflectional morphology from a computational perspective. The parts of speech and the morphemes that appear on them as well as their corresponding morphotactics are presented in detail. The verbal paradigm is also described in this document. Since the morphological analyzer designed for this project uses a unification-based grammar with typed feature structures, the morphological information has been defined in terms of features and values. The report describes the current version of the morphological analyzer used in the Shiraz project and discusses any morphological elements that have not been included in this version, mostly due to the colloquial usage of these morphemes. Sample rules of Samba, the grammar specifying the morphological analyzer, as well as the feature specification for the Persian type definitions module are also described.

Appendix: Romanization

Table 4: Shiraz romanization

Persian Letters	Romanization
alef with madd	A
alef	a
be	b
pe	p
te	t
se	_s
jim	J
che	^c
he	H
khe	x
dal	d
zal	_z
re	r
ze	z
zhe	j
sin	s
shin	^s
sat	S
zat	Z
ta	T
za	_Z
eyn	e
gheyn	Q
fe	f
ghaf	q
kaf	k
gaf	g
lam	l
mim	m
nun	n
vav	v
he	h
ye with hamze	i
ye	y
hamze	;
tanvin	^
short space (marking a final form character)	~

Table 5: Pronunciation guide

Transliteration Letter	Pronounced as...
b	boy
d	dog
f	fun
g	great
h	horse
j	Joe
k	clock
l	love
m	Mary
n	nun
p	pool
r	*similar to* Spanish “r”
s	sun
t	toy
v	very
y	you
z	Zorro
kh	*similar to* German buch
gh	*similar to* French “r”
ch	church
sh	shoe
zh	mirage
a	and
â	father
e	bed
i	sea
o	so
u	food

References

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

Beesley, K. R. (1997). Arabic Morphological Analysis on the Internet. Xerox Research Centre Europe.

Comrie, B. (1990). *The World's Major Languages.* Oxford University Press.

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

Kiraz, G. (1994). Computational Analyses of Arabic Morphology. Computer Laboratory, University of Cambridge.

Lazard, G. (1992). *A Grammar of Contemporary Persian.* Mazda Publishers.

Mahootian, Sh. (1997). *Persian.* Routledge.

Moshiri, L. (1988). *Colloquial Persian.* Routledge.

Rezaei, S. (1998). Persian Morphology. Report.

Riazati, D. (1997). Computational Analysis of Persian Morphology. MSc thesis, Department of Computer Science, RMIT.

Yasamy, R. et al. (1994). *Dastur-e Zaban-e Panj Ostad [Grammar of Five Masters].* Ferdows Publications, Tehran, Iran.

Zajac, R. (1992). Inheritance and Constraint-Based Grammar Formalisms. Association for Computation Linguistics.

Zajac, R. (1998). Feature Structures, Unification and Finite-State Transducers. *Paper presented to FSMNLP'98: International Workshop, on Finite State Methods in Natural Language Processing.*