Persian-English Machine Translation: An Overview of the Shiraz Project

Jan W. Amtrup, Hamid Mansouri Rad,
Karine Megerdoomian and Rémi Zajac

Memoranda in Computer and Cognitive Science
MCCS-00-319

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

Abstract

This report describes the Shiraz project MT prototype for a Persian to English machine translation system using typed feature structures and unification. An overview of the linguistic properties of Persian is presented and the morphological and syntactic grammars developed within the Shiraz project are discussed. The underlying model for the system is a layered chart, capable of representing heterogeneous types of hypotheses in an integrated way.

Contents

1

Introduction

The goal of the Shiraz machine translation project (http://crl.nmsu.edu/shiraz) was to build a prototype system that translates Persian text into English. The project began in October 1997 and the final version was delivered on August 1999. The system uses typed feature structures and an underlying unification-based formalism to describe Persian linguistic phenomena. It is able to run on Unix as well as Windows machines. The Shiraz system uses an electronic bilingual Persian to English dictionary consisting of approximately 50,000 terms, a complete morphological analyzer and a syntactic parser. The system components were tested on a bilingual tagged corpus developed from a large Persian corpus of on-line material (approximately 10MB). The machine translation system is mainly targeted at translating news material. The current system performs tokenization and full morphological analysis. Compounds and light verb constructions are also recognized. The syntactic parser can analyze noun phrases, preposition phrases, relative clauses and basic sentential constructions.

This report is a gathering of various introductory papers written during the course of the project, thus providing an overview of the linguistic and computational research performed in developing the system. This document includes descriptions of Persian language, the structure of the Shiraz dictionary, the system architecture and a discussion of several morphological and syntactic rules. Since this report presents an overview of the various modules in the translation system, it does not discuss any one component in depth. For a detailed description of the Persian language, the architecture of the system and the grammars, the reader is referred to the Related Publications listed under References (page 41).

An overall description of Persian is presented in Section 2, which discusses the history, writing system, morphology and syntax of the language. Section 3 is a description of Persian inflectional morphology. This section covers the nominal and verbal inflection, morphotactics and a brief description of the morphological formalism. Section 4 describes the structure of the Shiraz Persian-English dictionary which was developed to be used both as a machine translation tool and as a stand-alone dictionary. This section also proposes ways to enhance the current lexicon. Section 5 presents the basic syntactic structure of Persian sentences. This section discusses certain properties that are interesting for a computational analysis of Persian, rather than a complete description of syntactic grammar. The system architecture is described in the final section. It provides an overview of the various modules used in the Shiraz machine translation system, as well as a discussion of the underlying structure.

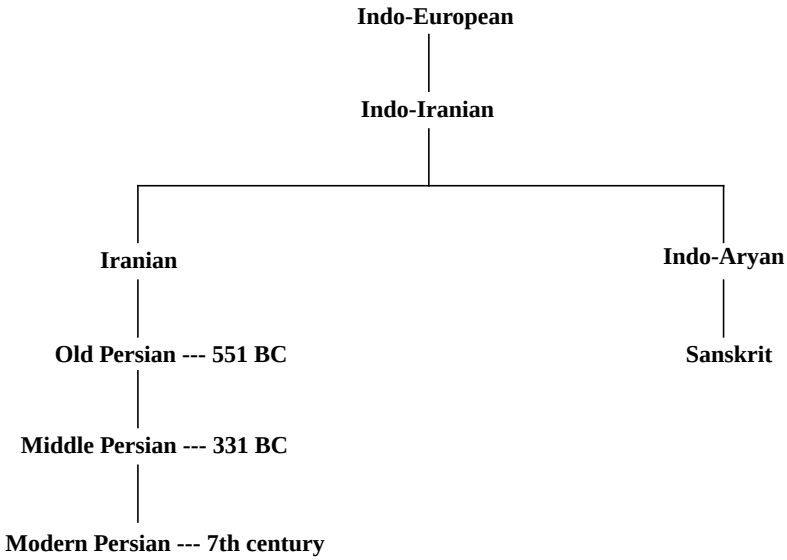
2

Introduction to Persian Linguistics

Persian, also known as Farsi, is the official language of Iran. It is also one of the two main languages spoken in Afghanistan, and the main language in Tajikestan, a formal central Asian republic of Soviet Union. The Persian spoken in these three countries has been influenced by the local environments. This is especially true in Tajikestan since it was isolated from the other Persian speaking countries during the Soviet era. The Persian in this country has many Russian borrowings and also uses the Russian alphabet. The language described here is mainly the Persian spoken in Iran. This section is an overview of Persian, with an emphasis on some of the interesting aspects for a computational analysis of written text in this language.

2.1 Background

Persian is derived from Indo-Iranian, one of the branches of the Indo-European languages. Indo-Iranian split into the Iranian languages and the Indo-Aryan (Indic) languages, from which most languages of India are derived. This split is estimated to have taken place around 1500 BC. The major Iranian languages are Persian, Kurdish, Pashto and Baluchi.



Although Persia was inhabited by the first millennium BC, the first inscriptions of Old Persian were obtained at approximately 551 BC, at the beginning of the Achemenides Empire. Old Persian had a complex morphology with a rich case and agreement system. By 331 BC (at the time of the

conquest of Persia by Alexander the Macedonian) the language had simplified and had lost most of its cases and agreements. At this level of development, the language is known as Middle Persian. Modern Persian dates from the 7th century, marking the Arabic conquest of Persia.

Old Persian was inscribed in Cuneiform. During the time of Middle Persian, an official alphabet did not exist; instead, each religious group had its own alphabet. So, for instance, the Zoroastrians (followers of Zoroaster or Zarathustra) wrote their religious texts in a special alphabet while the Manicheans (followers of Mani) and the Christians had their own alphabets. Administrative and government text was written in a yet another alphabet. It was only after the Arabic conquest that the Arabic script was adopted for writing Persian and the writing system was thus unified. Four sounds that did not exist in Arabic were added to the alphabet for Persian.

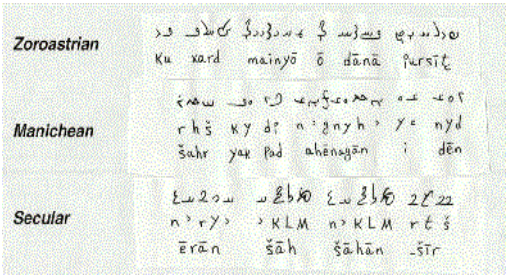


Figure 1: Alphabet samples from *Abolghasemi (1995)*

The Arabic conquest of Persia lasted for four centuries, from the 7th to the 11th AD. Arabic became the language of the intellectuals:Writers, poets and philosophers, as well as people in the administration, spoke and wrote in Arabic. During this period, many Arabic words were imported into the Persian language. More recently, there has been an Islamic resurgence in Iran since the revolution of 1979, and a considerable number of new Arabic borrowings are being used in Persian writing, which have also been added to the lexicon of the language. For a computational analysis of Persian, all these "new" Arabic loans have to be included in the dictionary.

Arabic has had an extensive influence on the Persian lexicon, but it has not really affected the structure of the language. Although a considerable portion of the lexicon is derived from Arabic roots, including the Arabic plural patterns, the morphological process used to obtain these lexical elements has not been imported into Persian and it is not productive in the language. The examples below show how the Arabic root system is used to derive nouns by inserting certain vowel patterns in the blank slots in the root template. (The transliteration used is described in the Appendix).

- ***Root Form***

  k_t_b

- ***Some words derived from the Root Form***

ketâb	"book"
kotob	"books"
katbi	"written"

katib	"scribe"
maktab	"(primary) school"

These Arabic words have been imported and lexicalized in Persian. So, for instance, the Arabic plural form for *ketâb* is *kotob* obtained by the root derivation system. In Persian, the plural for the lexical word *ketâb* can be given as in Arabic (*kotob*), or it can be obtained by just adding the Persian plural morpheme (*ketâb+hâ --> ketâbhâ*). Any new Persian words, however, can only be pluralized by the addition of the plural morpheme since the Arabic root system is not a productive process in Persian. In addition, since the plurals formed by the Arabic morphological system constitute a small portion of the Persian vocabulary (about 5% in the Shiraz corpus), it is not necessary to include them in the morphology; they are instead listed in the dictionary as irregular forms.

2.2 Writing System

Persian uses the Arabic alphabet. Texts are written from right to left. Vowels generally known as short vowels (*a, e, o*) are usually not written; only the long vowels (*y, u, â*) are represented in the text. This, of course, creates certain ambiguities. Since the short vowels are not inscribed, the word *krm*, for instance, can be pronounced with different vowel combinations resulting in five possible lexical elements. A reader uses the context to determine the word in the sentence.

kerm "worm", *karam* "generosity", *kerem* "cream", *krom* "chrome", *karm* "vine"

In the Persian writing system, letters in a word are often connected to each other. Most characters have a different form depending on their position within the word. The initial form indicates that no element is attached to the element from the right (i.e., there is no "attaching" character before it, but there is one following the character). Note that an initial form does not mean that the character is in the beginning of a word, it only indicates that the character is not at the end of the word. Characters are in medial form if they have an attaching character both before and after them. The final form denotes that the character is at the end of a word. The final forms can therefore be used to mark the word boundaries.

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

Figure 2: Sample Persian character forms

Certain characters (*alef, dâl, zâl, re, ze, zhe, vâv*) have only one form regardless of their position within the word.

In written text, words are usually separated by a space. Compounds and detachable morphemes (i.e., morphemes following a word ending in final form character), however, are written without a space separating them. In other words, the two parts of a compound appear next to each other but

the first element in the compound will usually end in a final form character, hence it would be possible to recognize the two parts of the compound. This format is not very consistent, however, and sometimes words can appear without a space between them. If the first word ends in a character that has a final form, then we can easily distinguish the word boundary. But if the first word ends in one of the characters that have only one form, the end of the word is not clear. Although this latter case is usually avoided in written text, it is not rare. Furthermore, a space is sometimes inserted between a word and the morpheme. In such cases, the morpheme needs to be reattached (or the space eliminated) before proceeding to the morphological analysis of the text.

2.3 Ambiguities in Morphology

Persian morphology is affixal, consisting of a few prefixes and mostly suffixes. There are no case morphemes or definite markers, but the language has an indefinite marker, several plural morphemes and pronoun clitics, and an enclitic form for the copula. Certain ambiguities arise in a morphological analysis of written text because different morphemes have the same surface form. This, combined with the fact that the short vowels are not written, give rise to a few parallel analyses of inflected words.

This is illustrated in the example below. The word written as *mrdy* could be pronounced with either the /a/ or the /o/ vowels as shown. Additionally, the suffix *y* can fulfill different functions. The resulting interpretation for *mrdy* is five-way ambiguous as shown. (The "enclitic" mentioned in the second interpretation is a clitic that links the noun to a relative clause and is not translatable.)

example:

the suffix y

- *mrdy*

1. *mardy* ..... "a man" ..... Noun+Indefinite marker
2. *mardy* ..... "a/the man" ..... Noun+Relative clause linking enclitic
3. *mardy* ..... "you are a man" ..... Noun+Copula (2nd, singular)
4. *mordy* ..... "you died" ..... Verb(past)+2nd, sing. inflection
5. *mardy* ..... "manliness" ..... Noun+derivational morpheme

2.4 Light Verb Constructions

These constructions consist of an element (noun, adjective or preposition) followed by a light verb such as the verbs "do", "give" or "hit". In these structures, the verb has lost its original meaning. Instead, it joins to the preverbal element to form a new verb. The meaning of a light verb construction is noncompositional; in other words, it can not be obtained by translating each element separately as the examples illustrate:

zamyn khordan	"floor eat"	to fall
fekr kardan	"thought do"	to think
dust dâshtan	"friend have"	to like/love
gush dâdan	"ear give"	to listen
jâru zadan	“broom hit"	to sweep

This is reminiscent of the light verb constructions such as "to give an ear" in Old English, and "to make an announcement" or "to catch a cold" in contemporary English. These structures, however, are extremely productive in Persian. New verbs are formed following this pattern such as

email zadan	"email hit"	to (send) email
klik kardan	"click do"	to click (on a mouse)

In addition, verbs in simple form have been and currently are in the process of dying out and are being transformed into the light verb constructions.

In analysis of written text, if these verbs always appeared as one single unit, they could have easily been recognized. In other words, it would have been more straightforward if the preverbal element and the light verb were to be considered as a light verb construction each time they appeared next to each other, and if they were to be treated as two distinct units whenever they appeared separated in the sentence. This is not the case, however. These verbal constructions can be separated from each other by other intervening elements. Syntactic analysis is needed to determine whether the element and the light verb should be treated as a unit or as two separate entities. This is shown in the examples below:

jâru	*khub*	*mizad*
broom	good	was hitting
	"he/she used to sweep well"	

jâru	*râ*	*zad*
broom	OBJ	hit
	"he/she hit the broom"	

In the first case, *jâru* and the light verb form a light verb construction with the meaning "to sweep", even though the two elements are separated by an intervening word. In the second instance, *jâru* and *zad* are separated by the specific object marker *râ* but remain as distinct units, and the verb maintains its original meaning of "hitting"; no light verb construction is formed.

2.5 Syntax

One of the main problems in a computational analysis of written text in Persian is determining phrase boundaries, especially the boundaries of Noun Phrases (henceforth NP).

- **No Cases**

Determining phrase boundaries is difficult since Persian is a verb final language but there are no markers or cases to distinguish the Subject or the Objects in a sentence (with the exception of the specific object marker *râ*). The resulting structure is then

Subject Object Verb or *Subject Predicate Copula*,

but there are no obvious markers to determine where the Subject ends and the Object or the Predicate begins.

- **No Phrase-Internal Markers**

Often, there are no distinct markers to link the constituents within the NP. The only such element is the ‘ezafe’ morpheme, pronounced /e/, which is a short vowel and is therefore not written in text. The ezafe only appears in written text after the vowels /â/ and /u/; in these cases it is written as *y*. For instance, in English, the *’s* in *John’s book* indicates that the word *John* is linked to *book*. In the example shown, the ezafe is not written between the words (although it is pronounced in spoken language) and there is no indication that the elements are related to each other within this NP.

ketâb syâh rezâ [pronounced: *ketâb-e syâh-e rezâ*]
book black Reza
‘Reza’s black book’

- **Boundary Ambiguity**

The example below illustrates a sentence with a Prepositional Phrase, followed by a Subject, a Predicate and the Copula. The syntactic parser will have to determine the phrase boundary for the NP within the Prepositional Phrase, as well as the boundaries for the Subject and the Predicate NPs.

به گفته رئیس ستاد حوادث غیرمترقبه استانداری خراسان مرکز زلزله
دیروز شهر کرمانشاه بود.

Translation:
"According to the chief of the Center of Unexpected Events of the provincial government of Khorasan, the center of yesterday’s earthquake was the city of Kermanshah."

In spoken language, the ezafe (the short vowel) is present to link the various constituents within the Noun Phrase. When there is no ezafe pronounced, the phrase boundary is defined; the absence of the ezafe indicates the end of an NP. In this written example, however, the ezafe is not marked at all. The result is a set of 12 elements that occur between the first and last words in the sentence, and which could all belong to an NP. A word by word translation of these 12 nouns and adjectives into English gives the following:

saying chief center events unexpected provincial government Khorasan center earthquake yesterday city Kermanshah

And the phrase boundaries should be as indicated below with a slash:

saying chief center events unexpected provincial government Khorasan /center earthquake yesterday / city Kermanshah /

Without any clear markers to determine the phrase boundaries, and without the ezafe to link the phrase constituents, how can we distinguish the end of a phrase?

- **NP Structure**

If we study the Noun Phrase structure in Persian, we can see that the Pronouns and the Indefinite Morphemes are always the last elements in the NP. In addition, the Proper Names are almost always at the end of an NP as well. Going back to the previous example, we can see that two of the phrase boundaries do in fact occur right after the two Proper Names in the sentence: *Khorasan* and *Kermanshah*.

If the ezafe morpheme is available in the text (e.g., after the vowels /a/ and /u/), it will indicate that a boundary can not occur between the element on which the ezafe appears and the following word, since the ezafe shows that these two elements are linked to each other. So, in the example shown below, no boundary can appear between *niru* and *havâ*, or *havâ* and *artesh*.

niru-ye	*havâ-ye*	*artesh*	*irân*
force-EZ	air-EZ	army	Iran

"the air force of the army of Iran"

To sum up, a boundary should be inserted after pronouns, indefinite markers, and proper names, and no boundary is inserted following an element with an ezafe marker. If none of these elements appear in a text, however, means other than syntactic analysis or morphology should be used in order to determine the phrase boundaries.

3

Persian Morphology

Persian morphology is an affixal system consisting mainly of suffixes and a few prefixes. The nominal paradigm consists of a relatively small number of affixes. The verbal inflectional system is quite regular and can be obtained by the combination of prefixes, stems, inflections and auxiliaries.

The following is a brief description of Persian morphology as well as the morphological grammar used in the Shiraz project. Since we are only dealing with written text, the difficulties encountered in the analysis of the written language are briefly discussed. The following sections describe the nominal and verbal morphology of Persian. The last section presents examples of the rules used in the morphological analyzer.

3.1 Ambiguities in Written Text

Certain ambiguities arise in a computational analysis of Persian text since the same surface form can represent different morphemes. In addition, short vowels are not marked in written text, which results in different possibilities of analysis. For instance, the word *mrdm* could be analyzed, among other possibilities, either as the noun *mardom* (people) or as the past tense of the verb *mordan* (to die): *mordam* (I died).

Furthermore, certain affixes always appear bound whereas others can also appear as free morphemes. The morphological analyzer is able to recognize all the possible surface forms of the affixes; it also uses the information available from the parts of speech that the morpheme appears on in order to disambiguate.

3.2 Nominal Morphology

There are no case forms and no gender distinctions in Persian. Person, number and sometimes animacy, however, are distinguished. Although there is no overt definite marker, a suffix is used on nouns and adjectives to indicate indefiniteness. The enclitic suffix which links nominal elements to a relative clause has the same surface form as the indefinite. There exist several morphemes to mark plurality, some of which are borrowings from Arabic. There are also some plural forms in Persian that follow the Arabic template morphology (also known as "broken" plurals) as shown below.

ketâb --> *kotob* (books)
faghir --> *foghará* (poor [people])

But the rules for forming these plurals are not used productively in Persian. These loan words are listed in the Shiraz lexicon and need not undergo morphological analysis.

The elements within a Noun Phrase are linked by the enclitic particle called ezafe. This morpheme is usually an unwritten vowel, but it could also have an orthographic realization in certain phonological environments. The role of the ezafe is to mark 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 (or possessive) structure. Examples of this construction are given below:

sedâ-ye pâ-ye man
sound-ez foot-ez my
‘(the) sound of my footsteps’

ru-ye miz
on-ez table
‘on the table’

Adjectives follow the same morphological patterns as nouns. They can also appear with comparative and superlative morphemes. Certain adverbs, mainly manner adverbs, can behave like adjectives and can appear with all the adjectival affixes. There are three types of ordinal constructions in Persian, which are formed by attaching their respective morphemes to the cardinal number.

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

The present indicative of the verb *budan* (to be) has a series of enclitic forms which can attach to the elements within a Noun Phrase. This morpheme is a verbal element but it can attach to nouns, adjectives and classifiers. The morphological analyzer needs to recognize this copula morpheme and separate it into a distinct lexical structure.

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

3.3 Verbal Morphology

3.3.1 Inflectional Paradigm

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

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

- **Prefixes:** the imperfective prefix written as *my* and the morpheme *b* or *by,* which characterizes the subjunctive and the imperative. Negation is marked by the *n* or *ny* prefix.
- **Personal Inflections:** present, past and imperative personal inflections are used in conjugating the Persian verb. All verb forms are marked for person and number.
- **Suffixes:** the suffix *ande* marks the present participle ending and *e* (written *h*) is used to form the past participle.
- **Causation morpheme:** causatives are obtained by adding the affix *ân* or *âni* to the end of the Present Stem of the verb. Personal inflections and suffixes can then be attached to the Causative Present Stem to derive all verbal forms for the causative construction.
- **Auxiliaries:** Persian conjugation uses a number of auxiliaries in the compound forms. The enclitic form of the auxiliary *budan* (be) is the one used in the formation of the perfect forms of all verbs. The verb *khâstan* (want) is used as an auxiliary in forming the future tenses. The auxiliary *shodan* (become) forms the passive constructions.

The complete inflectional system can be obtained by the various combinations of these elements.

3.3.2 Light Verbs

Most verbal constructions in Persian are formed using a light verb such as *kardan* (do, make), *dâda*n (give), *zadan* (hit, strike). The number of verbs that can be used as light verbs is limited, but these constructions are extremely productive in Persian. These structures consist of a preverbal element, which could be a noun, adjective or preposition, followed by a light verb, which has partly or completely lost its original meaning. Since these Light Verb or Compound Verb constructions are noncompositional in meaning, they are included in the dictionary as compounds.

Verbal inflection can only appear on the light verb itself, but bound morphemes can be attached to the preverbal element as well as the light verb. These inflectional morphemes are analyzed in the morphological component.

3.4 Morphological Grammar

The linguistic information associated with the morphemes is described using a unification-based morphological formalism. The morphological rule describes the concatenation of stems and

morphemes (using regular expressions) and the combination of morphological features of words and morphemes (using feature structures and unification). Stems and their features are stored in the lexicon as feature structures. A morphological rule associates a surface form, representing a sequence of morphemes, to a set of morphological features, and describes how the features of the stem and the morpheme are combined.

As an example, consider the Plural rule for Persian given below (string variables are prefixed with the dollar sign, regular expressions are enclosed between angle brackets):

```
Plural = <
          <$stem “hA”>
          Noun[exp: “$stem$”,
               lex: [regular: True],
               infl: [number: Plural]]
>;
```

The regular expression in angled brackets describes the surface form of the morpheme (the suffix *hA (= hâ)* in this example). The feature structure on the next line gives a partial description of the entry. The type is defined as `Noun` which indicates that the plural morpheme appears on a lexical element with type Noun. `exp` is the orthographic form (or citation form) of the entry as it is input in the lexicon. The lexical information available in the dictionary is presented under `lex` and the inflectional information is given under the path `infl`. The feature structure unifies the given inflection with the morpheme information. In this specific example, the morphological rule marks the `number` feature as *Plural* in case the lexical element is of type *Noun* and it is marked as *regular* in the dictionary.

The morphotactics of the inflections (i.e., the relative order of the morphemes) can be captured by applying the result of a rule as the input of the following one. For instance, the indefinite marker in Persian can follow the plural morpheme but the reverse is not true. This rule can be written in the following manner:

```
Indefinite = <
               <  <<$base \ Vowel> “yy”>
                  Noun[infl.indefinite: True] > |
               <  $base
                  Noun[infl.indefinite: False] >
>;
```

The plural rule can be used within the Indefinite rule in order to account for more complex morphological phenomena. The string analyzed by the Plural rule is bound to the variable base. This variable can thus be used in the Indefinite rule for checking, for instance, the character that it ends in. In other words, after the plural morpheme *hA (=hâ)* has been detected on the word, the Indefinite rule applies. The first alternative checks if the surface form of the base application ends in a vowel; this is true since *hA* ends with the vowel "*A*". The following feature structure requires this entry to be of type Noun. The successful application of this rule will add (unify) the corresponding structure to the output feature structure. So, in this example, if the suffix *yy* has been recognized following the plural morpheme *hA*, the *indefinite* feature in the structure is marked *True*, otherwise it’s marked *False*.

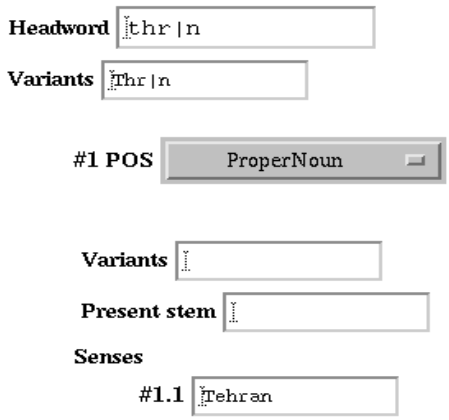
4

Shiraz Dictionary Structure

The Shiraz Persian to English dictionary was built by a team of Persian lexicographers and consists of approximately 50,000 entries including single words, phrases and proper names. The dictionary contains information about the orthography, morphosyntactic category and syntactic properties of lexical items as well as the English word-sense equivalents. It also contains morphological features for irregular entries. This information is stored as feature structures describing each dictionary entry. In what follows each element constructing the structure of the dictionary is discussed in more detail.

4.1 Citation Form

Each entry is input in the dictionary in citation form using the Shiraz romanization method[1]. The citation form entered for nonverbal elements is the non-inflected form or stem, and the citation form entered for verbal elements is the infinitival. In case entries have other orthographic variants, they are also included in a specific field called “variants.” The following example illustrates the dictionary entry *thran* [pronounced *tehrân*] and its variant *Thran* (which is written with the letter *tâ* instead of *te*) shown below in Shiraz romanization as seen on the dictionary interface..



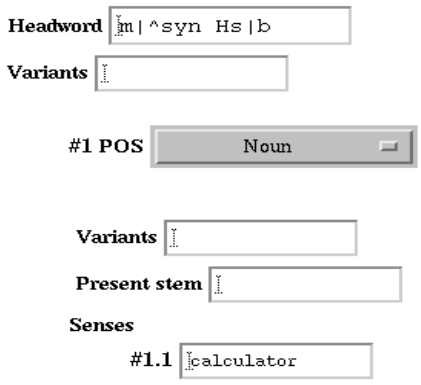
1. This romanization was developed specifically for the Shiraz project. It was designed to be bijective, so that the text could be converted to the original Persian format automatically without losing any information. The attempt was to keep the romanization easily readable for the language acquirers, hence the forms are based on their pronunciation in Persian; the distinction between the various characters is provided by different diacritics. The romanization does not transcribe the short vowels since they are absent from written text.

Vowels generally known as short vowels (a, e, o) are usually not written in Persian; only the long vowels (y, u, â) are represented in text. Therefore, words with different short vowels are input as one entry in the dictionary. This, of course, creates certain ambiguities. Since the short vowels are not inscribed, the word *krm*, for instance, can be pronounced with different vowel combinations resulting in five possible lexical elements. A reader uses the context to determine the word in the sentence.

kerm "worm", *karam* "generosity", *kerem* "lotion", *krom* "chrome", *karm* "vine"

The following shows the actual dictionary entry for the word *krm.* Note the different translations for the word.

A great number of words in Persian language lexicon exist as compounds. Persian Compounds such as LightVerbs, Compound Nouns, and a number of Prepositions are input with a space between their constituent elements. For example, the word *mâshin hesâb* "calculator":



4.2 Part of Speech

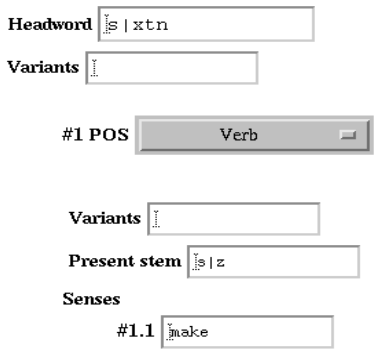
The next field for the dictionary entry holds the Part-of-Speech (POS) of the word. The main POS's in the dictionary could be divided into two categories of Open Class, and Closed Class. The main Open Class POS's in Persian are: Noun, Adjective, Proper Name, Verb, Light Verb. Among them, Light Verbs are more specific to Persian language. They consist of one or more preverbal elements, which could be a noun, adjective or preposition, followed by a light verb. For example, *esrâr kardan*, meaning "insist", which consists of the noun *esrâ*r "insistence", and the verb *kardan* "do". These constructions are categorized as LightVerb in our dictionary.

Closed Class items in the Shiraz dictionary are: Prepositions, Postposition (the object marker *râ*), Conjunctions, Relativizers, Numerals (numbers and digits), Determiners, Interrogatives, Interjections, Titles, Phrases, Numeratives (classifiers used to form numeral expressions), Number Units (which refer to numbers such as *hezâr* "thousand", *milyon* "million"). Pronouns are also among the Closed Class items, which fall under two categories: Personal Pronouns, such as *man* "I,me,my" or *shomâ* "you,your", and Quantifier Pronouns like *hame* "everyone".

All the above mentioned POS's are input for each dictionary entry. In the current version of the dictionary there are POS's included (such as POSNotAvailable) for the entries whose POS's were not clear for the acquirers. These entries need to be edited.

4.3 Present Stem

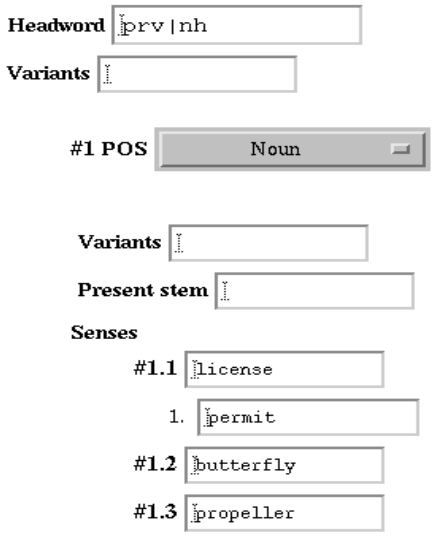
The verbs are input in the dictionary in their infinitival form. Every Persian verb has two stems, Present and Past. The Past Stem could be derived from the infinitival form of the verb, but the Present Stem is not easily obtained from the surface structure of the infinitival. As a result, the Present Stem of verbs are entered into the dictionary in a designated field.



As illustrated above, the present stem for the verb *sâkhtan* is *sâz*, which is used in forming the present tense and subjunctive inflections.

4.4 Sense

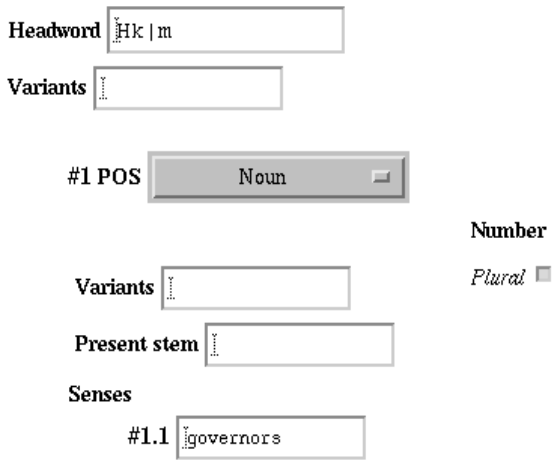
For machine translation purposes, the most generic meaning of every Persian entry is input in the dictionary. However, since the dictionary also needs to be used as a stand-alone tool, the synonyms and infrequent meanings of the entries are also included. Note the meanings for the word *parvâne* listed in the dictionary from its most to least frequent meaning in written text:



4.5 Features

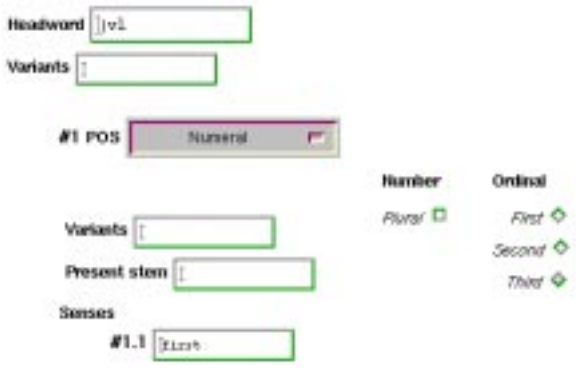
4.5.1 Number

There are a number of Arabic loans in Persian, one of which is the Arabic plural form for certain nominal elements. The plural form for these elements does not follow Persian plural-formation rules, and as a result, they can not be recognized by the system. Therefore, these nominal elements are considered irregular and entered in the dictionary with the feature Number set to Plural. For example, the plural form of the noun *hâkem* (governor), is *hokâm* shown below:



4.5.2 Number Type

There are also a number of irregular ordinal numbers that can not be derived from their cardinal forms. These ordinal numbers that do not follow the rules for ordinal-formation are input in the dictionary as irregular with the feature Ordinal indicating the number type. Further, there exist three types of ordinal values in Persian depending on their morphological structure. For example, one of the ordinal forms of the cardinal number *yek* (one) is *avval* (first). In this case, *aval* is input in the dictionary and the number type is set to *Ordinal Third*.



4.5.3 Regular Feature

The Regular feature is not set manually. Instead, if an entry is to be treated as irregular (i.e., if the Number feature is set to Plural, or if it is marked as an Ordinal), the value of the Regular feature is automatically set to False.

4.6 A Richer Lexicon

The dictionary developed in the Shiraz project was built from scratch. The short time-frame did not allow for the creation of an elaborate lexicon, thus the information entered in the dictionary was kept to a bare minimum. In addition to Number and Number Type, for instance, there are other features that have not been added to the current version of the dictionary. Including these features in the dictionary will help with preventing ambiguities in the output of morphological and syntactic analysis and therefore improving the quality of translation.

4.6.1 Silent H

The analysis for Persian entries ending in "h" can be very ambiguous. The reason is that the final "h" could be considered as both a consonant (pronounced /h/) or a vowel (pronounced /e/; also known as "silent h"). The final "h" as a consonant could attach to the morpheme following it. For instance, if the word *karagah* [pronounced *kârâgâh*] (=detective), which ends in the consonant "h" is followed by the indefinite marker "y", it will produce *karagahy* [*kârâgâhi*] (= a detective). In this case, the indefinite article "y" attaches to the word-final "h". On the other hand, when the word *xanh* [*khâne*] (= house), ending in silent "h" is followed by the indefinite marker "y", it will produce *xanh~ay* [*khâne i*], since morphemes can not attch to the silent "h"[1]. It forms, instead, a detached morpheme which follows the noun *xanh* without connecting to it. Since there is no way of distinguishing between the two types of final "h" by the surface structure of the words, the ambiguity in the analysis of the words with final "h" is inevitable. The method to prevent this ambiguity could be including a Silent H feature in the lexicon.

4.6.2 Vowel Feature

Similarly, there are two types of "v" in Persian: Vowel [pronounced *u* or *o*] and Consonant [pronounced *v}*. They are both written the same way and there is no way of distinguishing them by the surface structure of the words in which they appear. This does not cause problems in translation if the "v" occurs at the beginning or in the middle of Persian words, since the words are present as lexical elements in the dictionary with their English equivalents. However, the morphological analyses for words *ending* in "v" could be problematic. It is due to the fact that the inflectional rules for words ending with "v" as vowel are different from the ones ending in consonant "v". For example, in Persian, the word *Jjarv [jâru]* meaning "broom" ends in the vowel form of "v". If the word is followed by the ezafe morpheme, a "y" will appear at the end of the word forming *Jarvy.* On the other hand, the word *gav [gâv]* meaning "cow" ends in the consonant "v". When the ezafe morpheme appears on this word, it will be represented as the short vowel /e/, which is not transcribed in written text. In cases like this, as in *gav hmsayh [gâv-e hamsâye]* meaning "the cow of the neighbor," the morphological analyzer will mark it as ezafe=Undefined. In general, when a word ends in "v" the system produces both analyses of ezafe=False (i.e., analyzes "v" as a vowel) and ezafe=Undefined (i.e., analyzes "v" as a consonant). Both analyses are carried through to syntax resulting in further ambiguities (since the ezafe is used in determining Noun Phrase boundaries). Marking the words ending in "v" in the dictionary based on their being a vowel or a consonant will disambiguate these morphological and syntactic analyses.

1. The tilda ~ in *xanh~ay* marks the boundary between the word and the detached morpheme.

4.6.3 Person and Number

Persian does not include gender specific pronouns, but pronouns are inherently marked for number and person. For example:

- *to* (you, your) refers to singular second person
- *ânhâ* (they, them) is used to refer to plural third person

As a result, there exists ambiguity in analysis and translation due to pronoun references. Marking these in the dictionary will make pronoun references clear. It will also be helpful in subject-verb agreement in syntactic grammar and will improve the quality of translations.

4.6.4 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.

4.6.5 Verb Category

This feature determines whether the verbal element is transitive, intransitive or impersonal. In morphological analysis, 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, since it helps the parser to determine whether the verb requires an object or not.

4.6.6 Compound Head

As mentioned earlier, a great number of lexical elements are compounds. In Persian the head of the compound could be any of the elements forming the compound. In the Shiraz system, morphological analysis first applies to single words. The parts of the compounds are joined in a later component. At this point, the system needs to recognize which inflectional information should be transferred to the compound feature structure.

Consider the following examples. In the compound in (1), the plural morpheme attaches to the word *mâshin,* the first element of the compound *mâshin hesab* (calculator) forming *mâshin-**hâ-ye** hesâb.* However, in example (2), the plural morpheme attaches to the last element of the compound *az khod gozashtegi*, forming as *khod gozashtegi-**hâ***, meaning “self sacrifices”.

(1)

	ماشین حساب		ماشین های حساب
	mâshin hesâb		*mâshin~hâye hesâb*
Gloss	machine count	=>	machine+pl+ez count
Translation	“calculator”		“calculators”

(2)

	az khod gozashtegi		*az khod gozashtegi~hâ*
Gloss	from self passing	=>	from self passing+pl
Translation	“self-sacrifice”		“self-sacrifices“

Currently, the result of the process of recognition and analysis of compounds is very ambiguous. Explicitly marking the head of compounds in the dictionary will enormously reduce the ambiguities in the translation.

4.6.7 The Causative

There are a number of causative verbs in Persian that require more consideration. In Persian, causative verbs are marked by the morpheme *ân* or *âni*. In most cases, their translations in English can be compositionally derived using the verb “make”. For instance, *khanndândan*, the causative form of the verb *khandidan* (to laugh), could be marked by the morphological analyzer as a causative construction which, in turn, can be translated into the English “to make laugh”. Thus, we only need to include the basic form of the verb *khandidan* (=laugh) in the dictionary and allow the causative translation to be built in English generation. However, in certain cases, the translation using “make” is not as felicitous since English uses lexical items to translate these causative verbs. For instance, for the verb *tarsidan* (to fear), the causative form *tarsândan* can not be compositionally translated in English as “make fear”. Instead, it has to be lexicalized as “to frighten”. For these cases, we input the Persian causative verb in the dictionary with the corresponding English sense. A feature needs to be included in the lexicon for such verbs so that the morphological analyzer does not analyze them for causative constructions.

5

Persian Syntax

Persian syntax is quite ambiguous in written form and causes immense difficulties in automatic parsing of written text. Several factors contribute to the ambiguity: Although Persian is a verb-final language, it does not adhere to a strict word order and the sentential constituents may occur in various positions in the clause. This is especially the case for preposition phrases and adverbials. In addition, there are no overt markers, such as case morphology, to indicate the function of a noun phrase or its boundary; in Persian, only specific direct objects receive an overt marker. Furthermore, subjects are optional in Persian which increases the ambiguity obtained. Verbs in Persian agree with the subject in number and person; however, if the subject is inanimate, agreement may be omitted and instead a default third person singular agreement is used on the verb. Determining the boundaries of noun phrases is very difficult since there are very few overt morphemes to mark such boundaries. Although in spoken language, the *ezafe* morpheme is used to link the elements within the noun phrase, this morpheme, being a short vowel, is absent in written text. Since short vowels are not transcribed, lexical ambiguity is also another problem in automatic parsing of Persian text.

1 Persian preposition phrases, however, are easily recognized and can be used to mark phrasal boundaries in the sentence. Additionally, the verb almost always occurs in the sentence-final position in written text which facilitates parsing. This section provides a description of Persian syntax, especially concerning issues that may arise in a computational analysis of written text.[1] Syntactic disambiguation methods, if available, are also discussed.

5.1 Word Order

Persian is an SOV language: the sentences appear in the word order Subject-Object-Verb. The verb is marked for tense and aspect and usually agrees with the subject in person and number. Persian is a pro-drop language, thus the subject is optional. The object marker *râ* is used to indicate specific direct objects in simple sentences.[2]

(1) a. bache-hâ panjare râ shekast-and
child-Plur window Obj break-Past-3pl
'The children broke the window.'

1. Although selectional restrictions or subcategorization marked on the lexical items could also be helpful in disambiguating phrasal boundaries, they are not discussed in this report. Since, the current Shiraz parser and lexicon do not carry any subcategorization information.
2. Brackets indicate optionality in the examples. Imp=imperfective marker, Neg=negation, Plur=plural, Obj=object marker, Past=past tense, Present=present tense, Subj=subjunctive. Person is marked by 1, 2 or 3; number is either pl=plural or sg=singular.

b. (mâ) kabâb mi-khor-im
(we) kabob Imp-eat-Present/2pl
'We are eating kabob.'

If there is an oblique object or a Prepositional Phrase in the clause, it precedes the indefinite direct object as shown in (2), but usually follows the specific or definite object as in (3).

(2) Hushang be bache-hâ nân dâd
Hushang to child-Plur bread gave/3sg
'Hushang gave (some) bread to the children.'

(3) Hushang nân râ be bache-hâ dâd
Hushang bread Obj to child-Plur gave/3sg
'Hushang gave the bread to the children.'

Although these examples describe the canonical word order, Persian is a free word order language and the sentential constituents can be moved around in the clause as illustrated below:[1]

(4) pirzan vâse bache-hâ dâstân-e "mâhi syâh-e kuchulu" râ ta'rif kard
old woman for child-Plur story-Ez fish black-Ez little Obj telling did/3sg
'The old woman told the children the story of the "little black fish".

(5) a. dâstân-e "mâhi syâh-e kuchulu" râ pirzan vâse bache-hâ ta'rif kard
story-Ez fish black-Ez little Obj old woman for child-Plur telling did/3sg
b. vâse bache-hâ dâstân-e "mâhi syâh-e kuchulu" râ pirzan ta'rif kard
for child-Plur story-Ez fish black-Ez little Obj old woman telling did/3sg
c. dâstân-e "mâhi syâh-e kuchulu" râ pirzan ta'rif kard vâse bache-hâ
story-Ez fish black-Ez little Obj old woman telling did/3sg for child-Plur

These "scrambled" clauses often give rise to focused or topicalized readings. In the written language, although most elements may appear in relatively free word order, the sentences often remain verb-final. Adverbs and preposition phrases, however, can appear in various positions quite freely. Apart from manner adverbs, which occur within the verb phrase, other adverbs may appear almost anywhere in the clause, in between the various constituents. Adverbs usually can not occur following the verb.

Although Persian is verb-final at the sentential level, it behaves like head-initial languages in noun phrases (NP) and preposition phrases (PP). Thus, the head noun in a NP is often followed by the modifiers and possessors (6), and the preposition precedes the complement NP (7).

(6) a. ketâb-e man
book-Ez my
'my book'

b. yek khâne-ye bozorg
one house-Ez big
'a big house'

(7) mardom *dar* *khyâbân-hâ* tazâhorât mi-kard-and
people *in* *street-Plur* demonstrations Imp-do-3pl
'People were demonstrating *in the streets*.'

Certain preposition phrases, such as locative and directional PPs, can follow the verb as shown in the following examples. The preposition is sometimes optional in these cases. These constructions,

1. Ez is the *ezafe* morpheme, discussed in Section 5.2.

however, do not often occur in written text.

(8) a. bache-hâ raft-and (be) manzel
child-Plur go-Past-3pl (to) home
'The children went home.'
b. pâkat râ gozâsht ru-ye miz
envelope Obj put-Past-3sg on-Ez table
'He/she put the envelope on the table.'

Subordinate clauses follow the main clause as illustrated in (9). Persian has the complementizer *ke* (that) which marks both subordinate constructions and relative clauses; it is often optional.

(9) mardom ne-mi-khâst-and (ke) rafsanjâni dar in entekhâbât barande shavad
people neg-Imp-want-Past/3pl (that) Rafsanjani in these elections winner become-Subj/3sg
'People didn't want Rafsanjani to win in these elections.'

Questions are usually formed *in-situ*, i.e., the element being questioned is replaced by the interrogative form without changing the word order.

(10) a. bache-hâ *chi* râ shekast-and?
child-Plur *what* Obj break-Past-3pl
'What did the children break?'
b. *ki* panjare râ shekast?
who window Obj break-Past-3sg
'Who broke the window?'

5.2 Noun Phrases

5.2.1 Simple Noun Phrase

The head of a noun phrase could be a noun or an infinitival verb. Pronouns and proper names may also head noun phrases and they function as *possessors* in forming complex noun phrases (i.e., possessive constructions such as *ketâb-e hushang (Hushang's book)).*

The head noun is preceded by the determiner, the numeral constructions and the quantifiers, and it is followed by the *modifiers*, which usually consist of an adjectival phrase (AP). Superlative adjectives, however, do not appear in the AP; instead, they precede the head noun. Numeral constructions, quantifiers and superlative adjectives are in complementary distribution, i.e., if one of these elements is present, the others cannot occur within the NP.

The relative ordering of the constituents of the simple NP is given below:

NP = *determiner specifier head modifier*

where the *head* is a Noun and the parts of speech or phrases that can appear in each of the other categories are as shown below. Note that all the constituents, with the exception of the head noun, are optional.

determiner:	Determiner	ex. *in* (this, these)
specifier:	Numeral (Unit) (Classifier)	ex. *do milion nafar* (two million person*)*

	Numeral [Ordinal]	ex. *dovomin* ((the) second)
	Adjective [Superlative]	ex. *bozorgtarin* ((the) biggest)
	Quantifier	ex. *ba'zi* (some)
modifier:	(Adverb) Adjective	ex. *kheyli kohne* (very old)
	Note: Modifiers may be recursive	

The modifiers are linked to the head noun with the *ezafe* morpheme. The example in (11) represents a simple Noun Phrase where CL stands for Classifier and Ez for the ezafe morpheme. Classifiers indicate the class or type of the noun. Thus, for instance, *tâ* is used with count inanimate nouns, *nafar* indicates people, *qalâde* (=collar) can be used when giving a count for dogs, etc.

(11) in do tâ ketâb-e kohne
this two CL book-Ez old
'These two old books'

The infinitival constructions are very similar to the English gerundive. The infinitive head can appear in a predicate construction or with an adverbial. The objects of the verb become arguments of a possessive construction as exemplified in (13).

(12) zan budan-ash
woman be-her
'her being a woman'

(13) koshtan-e shir
kill-Ez lion
'the killing of a lion'

5.2.2 Possessive Constructions

These constructions are the equivalent of the genitive or possessive constructions in English, such as "Mao's red book", "her mother's hat" or "the syntax of noun phrases". In English, the link between the two nouns is marked by " 's" (e.g., Mao's) or the preposition "of". In the case of pronouns, the latter appear in their genitive form (e.g., her). The element joining the Persian noun phrase constituents to each other is the *ezafe* suffix. The ezafe, however, is usually pronounced as the short vowel /e/ and is therefore not marked in written text. The result, in Persian written text, is a series of consecutive nouns without any overt links or boundaries as shown in the example in (14) transcribed as it appears in Persian text (i.e., without short vowels). The actual pronunciation for this example is given in (15); the ezafe morpheme is represented by the *-e* following the first three nouns, linking each one to the following constituent. Note that the last constituent in the NP does not carry the ezafe suffix, thus marking the end boundary of the noun phrase.

(14) ktab dvst pdr daryvsh
book friend father Dariush
'Dariush's father's friend's book'

(15) ketâb-e dust-e pedar-e dâryush

In this example, each noun forms a simple NP which then join together to form the complex NP given in (14). The sentence in (16) illustrates another complex noun phrase; the pronunciation is

given in (17).

(16) jlsh kmysyvn amvr daxly mjls ayn kshvr
meeting committee affairs internal parliament this country
'The meeting of the Committee on Internal affairs of this country's Parliament'

(17) jalase-ye komision-e omur-e dâkheli-e majles-e in keshvar

When pronouns are used as the possessor, the constructions are identical:

(18) ketâb-e man
book-Ez 1sg-pronoun[1]
'My book'

5.2.3 NP Boundaries

Certain morphological and syntactic elements can help resolve some of the ambiguities arising in parsing of Persian written text. As already mentioned, the *ezafe* suffix can mark boundaries within Noun Phrases. Unfortunately, this morpheme is often absent from written form. It does occur, however, after the vowels *â* and *u* as exemplified below.

(19) zn zybay daryvsh vard shod *[zan-e zibâ-ye dâryush vâred shod]*
wife beautiful-Ez Dariush entered
'Dariush's beautiful wife entered.'

(20) [wife beautiful Dariush]*NP* entered.

In (19), the adjective *zyba* is followed by the ezafe suffix *y*, which indicates that the adjective is linked to the following element *daryvsh*. Thus, the absence of the ezafe after the adjective *zyba* will mark a noun phrase boundary as illustrated in (21)and (22).

(21) zn zyba daryvsh ra shnaxt *[zan-e zibâ dâryush râ shenâkht]*
woman beautiful Dariush OBJ recognized
'The beautiful woman recognized Dariush.'

(22) [woman beautiful]*NP* [Dariush OBJ]*NP* recognized.

Certain morphemes, such as the pronominal clitics, the indefinite article and the enclitic used to link NPs to relative clauses, can only occur as the last element in the NP. The detection of any of these morphemes indicates that the boundary of the noun phrase has been reached. In addition, proper names and pronouns often mark the boundary of the noun phrase.

In the current Shiraz grammar, these boundary markers have been incorporated within the NP rules. Thus, if a simple noun phrase carries a boundary marker, it is not allowed to join with another NP to form a more complex phrase. As a simple illustration, consider the two N'-forming rules, `NounBarClitic` and `NounBarEzafe`. These rules contain a left-hand side (lhs) and a right-hand side (rhs) as in rewrite rules. In the first rule, the right-hand side is satisfied if a clitic is detected (indicated by `clitic.function: True`). As can be seen in the left-hand side of this rule,

1. Since there is no Case in Persian, the surface form of the pronoun is always the same whether it is used in a subject, object or possessive context.

this nominal element is tagged as the head of the N' and the value of the *boundary* feature is set to True. This boundary value is transferred up when the higher NP level is formed; this NP will not be allowed to join to another noun phrase following it since the boundary has already been set to True.

```
// N' --> N   carrying a boundary marker
NounBarClitic = per.Rule[
    lhs: per.NounBar[
        head: #head,
        boundary: True],
    rhs: <:
        #head= per.Noun[infl.clitic.function: True]
    :>
];
```

In the case of the `NounBarEzafe` rule, however, when an ezafe feature is detected (shown in the right-hand side of the rule as `infl.ezafe:per.EzTrue`), the *boundary* feature in the left-hand side is set to False. This allows the N' and the higher NP to join to the following noun phrase construction.

```
// N' --> N   carrying ezafe - no boundary set
NounBarEzafe = per.Rule[
    lhs: per.NounBar[
        head: #head,
        boundary: False],
    rhs: <:
        #head= per.Noun[infl.ezafe: per.EzTrue]
    :>
];
```

5.3 Relative Clauses

Persian relative clauses are usually introduced by the relativizer *ke* (that), which is used regardless of the animacy, gender or function of the head noun. In nonrestrictive relative clauses, the head noun often carries an enclitic morpheme (Encl) which links the noun to the following relative clause. If the relativized noun is the object of the main sentence, then it may appear with the object marker *râ* as illustrated in (24).

(23) zan-**i** ke injâ neshaste ast hamsar-e Nâder ast
woman-Encl that here sit-Part is spouse-Ez Nader is
'The woman that is sitting here is Nader's wife.'

(24) ktâb-**i** râ ke diruz kharide budam emruz sobh tamâm kard-am
book-Encl Obj that yesterday bought was today morning finish did-1sg
'This morning, I finished the book that I had bought yesterday.'

The relative clause may be separated from the head noun by the main verb as illustrated below. In addition, several relative clauses could follow a head noun.

(25) mâ pesar-ân-i râ *entekhâb mi-kon-im* ke dar jang sherkat na-karde-and
we boy-Plur-Encl Obj choosing Imp-do-1pl that in war participation neg-done-3pl
'We choose (the) boys that have not participated in the war.'

If the head noun is the subject or direct object of the relative clause, it is often left as a gap as was shown in the examples in (23) and (24). However, even in such cases, the relativized noun may be replaced by a resumptive pronoun in the clause it originated from. Thus, in (26), the head noun *plâk-e kuchak* (small plaque) is the subject of the relative clause; it is substituted by the resumptive pronoun *ân* (it). The use of the resumptive pronoun usually occurs when the head noun is separated from the relative clause by an intervening verb. In this example, the verb *pey borde-and* (have found) precedes the relative clause.

(26) dâneshmandân be plâk-e kuchak-i dar maqz pey borde-and ke **ân** niz
scientist-Plur to plaque-Ez small-Encl in brain found-3pl that it also

tâkonun nâshenâxte mânde bud.
until now unknown remained was
'Scientists have found a small plaque in the brain that until now had remained undiscovered.'

When the head noun is the indirect object or is extracted from a Prepositional Phrase adjunct in the clause, a resumptive pronoun is used. In other words, the position from which the head noun originates is substituted by a pronoun that agrees with the head noun. This is exemplified in the three NP cases below:

(27) in bache-hâ ke *az* ***ânhâ*** âdres mi-porsid-i ...
this kid-Plur that from them address Imp-ask-2sg
'These kids from whom you asked for the address...'

(28) shahr-i ke *dar* ***ân*** tazâhorât shode bud ...
city-Encl that in it demonstrations become was
'The city in which demonstrations took place...'

(29) zan-i ke *barây-**ash*** ketâb kharid-i ...
woman-Encl that for-Clitic(3sg) book buy-Past-2sg
'The woman for whom you bought a book...' or
'The woman that you bought a book for...'

5.4 Verb Phrases

As already discussed in Section 5.1, the verb in Persian usually occurs in the sentence-final position, with objects, adverbials and adjuncts preceding it. The relative order of the direct object and the indirect object or PP may be modified based on the specificity of the direct object.

(30) Hushang be bache-hâ nân dâd
Hushang to child-Plur bread gave/3sg
'Hushang gave (some) bread to the children.'

(31) Hushang nân râ be bache-hâ dâd
Hushang bread Obj to child-Plur gave/3sg
'Hushang gave the bread to the children.'

The verb agrees in number and person with the subject of the clause. However, if the subject is inanimate, the agreement may default to the third person singular as illustrated in the contrast in the examples below, taken from the same newspaper article, both containing an inanimate plural

subject but giving rise to different agreements on the verb:

(32) *jangande-hâ-ye* esraili jonub-e lobnân râ bombârân kard-***and***
fighter plane-Plur-Ez Israeli south-Ez Lebanon Obj bombing did-***3pl***
'Israeli fighter planes bombed the south of Lebanon.'

(33) *bombârân-hâ-ye* esrail ma'mulan motevajeh-e manâteq-e
bombardment-Plur-Ez Israel usually directed-Ez regions-Ez

maskuni-e jonub-e lobnân ***ast***
residential-Ez south-Ez Lebanon is***(3sg)***
'Israel's bombardments are usually directed towards the residential regions of the south of Lebanon.'

5.5 Light Verb Constructions

Persian simple verbs are quite rare compared to the number of light verb constructions, also known as complex predicates, in the language. These constructions consist of a noun, adjective or preposition followed by a light verb such as the verbs "do", "give" or "hit", forming non-compositional units of meaning. In other words, the meaning of these light verb constructions can not be obtained by translating each element separately as the examples illustrate:

zamin xordan	"floor eat"	to fall
zendegi kardan	"life do"	to live
gul zadan	"deception hit"	to deceive
shekast dâdan	"defeat give"	to defeat
e'lâm kardan	"announcement do"	to announce
âsib didan	"damage see"	to be damaged
pâyân yâftan	"ending find"	to end
na're keshidan	"yelling pull"	to yell, to roar
e'teqâd dâshtan	"belief have"	to believe
be donyâ âmadan	"to world come"	to be born
az dast dâdan	"from hand give"	to lose

These constructions can also be used as purely idiomatic expressions:

del be daryâ zadan	"heart to sea hit"	to take a risk

In any case, these complex predicates are extremely productive in Persian. New verbs are formed following this pattern, by joining a nominal or adjectival word (possibly a loan word) to a light verb as shown:

email zadan	"email hit"	to (send) email
klik kardan	"click do"	to click (on a mouse)

In addition, verbs in simple form have been and currently are in the process of dying out and are being transformed into the light verb constructions.

The light verbs used in these complex predicates are not always semantically vacuous. In fact, these verbs may contribute to the aspectual readings of the predicate or provide a causation interpretation to the verb. They may also contribute to the transitivity of the verb phrase as shown in (34). The first sentence consists of the light verb construction *shekanje dâdan* (torture give) and gives rise to a transitive sentence. (34b), on the other hand, is formed with the light verb construction *shekanje didan* (torture see) and the result is a passive reading.

(34) a. sarbâz-hâ golesorxi râ dar zendân shekanje dâdand
soldier-Pl Golesorkhi Obj in prison torture gave-3pl
'The soldiers tortured Golesorkhi in prison.'
b. golesorxi dar zendân (be dast-e sarbâz-hâ) shekanje did
Golesorkhi in prison (to hand-Ez soldier-Pl) torture saw/3sg
'Golesorkhi was tortured in jail (by the soldiers).'

For the purposes of the Shiraz project, however, light verb constructions were input into the dictionary as lexical units with their corresponding translations into English. In other words, light verbs are treated as compounds in the Shiraz machine translation system: Each element of the construction undergoes morphological analysis and the results are joined together when the light verb construction is recognized. Consider the example in (35) representing a light verb construction, in which both the nominal and the verbal parts carry morphemes.

(35) bache-hâ pirmard râ *kotak-**esh*** *zad-**and***
child-Plur old man Obj beating-him hit-3pl
'The children beat up the old man.'

In this examples, the light verb *zadand* carries information on tense, aspect, number and person. The nominal part *kotak* (beating) carries the clitic pronoun for third person singular. This clitic is analyzed as an object (i.e., accusative) on verbs. The result of morphological analysis and lexical lookup for each part is shown in (36) for the nominal part and in (37) for the verbal part, where *lex* represents the lexical information and *infl* is the inflectional information computed by the morphological analyzer. Note that in (36), the noun has been analyzed as a singular, carrying a clitic pronoun (third person singular). In (37), the verb is analyzed as active voice, preterite tense, third person plural agreement; there are no clitic pronouns on the verb.

```
(36)  Noun[
        lex : LexMorph[ number : Singular, regular : True],
        infl : NominalInfl[
           number : Singular,
           clitic : Clitic[ person : Third,
                            number : Singular,
                            function : Possessive],
           ezafe : EzFalse,
           indefEncl : False,
           indefinite : False,
           enclitic : False],
        exp : "ktk",
        trans : <: LSign[exp : "beating"] :>]


(37)  Verb[
        lex : LexMorph[
           number : Singular,
           presentStem : "zn",
```

```
    regular : True],
  infl : VerbalInfl[
    voice : Active,
    clitic : Clitic[function : Null],
    tense : Preterite,
    causative : False,
    negation : False,
    mood : Indicative,
    person : Third,
    participle : PartFalse,
    numberAgr : Plural],
  exp : "zdn",
  trans : <:
   LSign[
    exp : "hit"]LSign[
    exp : "play"]:>]
```

The simple rule below shows how the two parts of such a light verb construction are unified in the Shiraz grammar, and how their morphological information is percolated from the right-hand side, up to the left-hand side, in order to form the single light verb construction NominalLVEntry.

```
(38)  NominalLV = per.Rule[
        lhs: per.NominalLVEntry[
         infl: [mood: #mood,    //verbal morphemes
             tense: #tense,
             voice: #voice,
             person: #person,
             numberAgr: #numberAgr,
             causative: #caus,
             negation: #neg,
             participle: #part,
             clitic: #clitic,  //nominal morphemes
             number: #number,
             ezafe: #ezafe,
             indefEncl: #indencl,
             indefinite: #indef,
             enclitic: #encl]],
        rhs: <:
         per.Noun[infl: [
             number: #number = Top,
             ezafe: #ezafe = Top,
             indefEncl: #indencl = Top,
             indefinite: #indef = Top,
             enclitic: #encl = Top,
             clitic: #clitic = Top]]
         per.Verb[infl: [mood: #mood = Top,
             tense: #tense = Top,
             voice: #voice = Top,
             person: #person = Top,
             numberAgr: #numberAgr = Top,
             causative: #caus = Top,
             negation: #neg = Top,
             participle: #part = Top]]
        :>,
        recursive: True,       // This rule can apply recursively
```

```
        lookup: True,           // Perform dictionary lookup after creating lhs
        remove: True];          // Remove all edges used by this rule after pars-
ing
```

The final structure for the light verb construction, after the rule in (38) has applied and the final structure is looked up in the dictionary, is shown in (39). We now have a light verb construction (NominalLVEntry) resulting from the unification of the two parts.[1]

```
(39)   NominalLVEntry[
         lex : LexMorph[
           number : Singular,
           regular : True],
         infl : NominalLVInfl[
           voice : Active,
           number : Singular,
           causative : False,
           ezafe : EzFalse,
           tense : Preterite,
           person : Third,
           clitic : Clitic[ person : Third,
                            number : Singular,
                            function : Possessive],
           participle : PartFalse,
           mood : Indicative,
           negation : False,
           numberAgr : Plural,
           indefinite : False,
           indefEncl : False,
           enclitic : False],
         exp : "ktk zdn",
         trans : <: LSign[ exp : "beat up"]:>]
```

Thus, if light verb constructions always occurred as one single unit, they could easily be recognized. This is not the case, however. These verbal constructions can be separated from each other by other intervening elements. The object of the light verb, for instance, may appear between the two parts of the construction as shown in (40) for the light verb construction *âsheq shodan* (fall in love). In (41), the light verb predicate *afzâyesh yâftan* (increase) has been separated by the adjective *shadid*, which is behaving as an adverb. (42) represents the light verb construction *xâstâr shodan* (request) with an intervening object, which itself consists of a complex noun phrase composed of a NP and a PP.

(40) majnun **âsheq**-e leyli **shod**
Majnun lover-Ez Leyli became
'Majnun fell in love with Leyli.'

(41) shomâr-e bikâr-ân **afzâyesh**-e shadid-i **yâfte ast**
number-Ez unemployed-Plur increase-Ez intense-Indef found is
'The number of the unemployed has increased intensely.'

(42) englis **xâstâr**-e moshârekat-e rusiye dar hall-e bohrân-e kozovo **shod**
England requester-Ez cooperation-Ez Russia in solving crisis-Ez Kosovo became
'England requested Russia's cooperation in solving the Kosovo crisis.

1. In the current version of the grammar, the function of the clitic pronoun is not changed from Possessive to Object; this is taken care of at the transfer level.

In all of these examples, the separated parts of the light verb are still to be recognized as one unit. However, in certain cases, the separated constituents lose the light verb construction meaning. Compare the two sentences in (43). In (43a), the light verb construction is interpreted as a unit, whereas in (43b), the intervening object marker splits the light verb construction. In this case, the nominal part *jâru* (broom) has become the direct object of the verb *zadan* (to hit). A similar effect is obtained by the relativization of the nominal part in (44).

(43) a. vaqti vâred shodam nader dâsht **jâru** **mi-zad**
when enter became-1sg Nader had-3sg broom Imp-hit
'When I entered, Nader was (in the process of) sweeping.'
b. vaqti vâred shodam nader dâsht **jâru** râ **mi-zad**
when enter became-1sg Nader had-3sg broom Obj Imp-hit
'When I entered, Nader was (in the process of) hitting the broom.'

(44) a. nader dishab **zamin** **khord**
Nader last night floor eat-Past-3sg
'Nader fell last night.'
b. **zamin**-i râ ke nader **khord** bâvar nakardani bud
floor-Encl Obj that Nader ate-3sg unbelievable was
'The floor that Nader ate was unbelievable.'
[Not 'Nader's fall was unbelievable.']

Compare (44), however, to the construction in (45) with the light verb predicate *latme zadan* (damage). In this instance, even when the nominal element is relativized, the light verb construction still obtains.

(45) a. tagarg dishab be baq-e man **latme** **zad**
hail last night to garden-Ez my damage hit-Past-3sg
'The hail damaged my garden last night.'
b. **latme**-i râ ke tagarg be baq-e man **zad** bâvar nakardani bud
damage-Encl Obj that hail to garden-Ez my hit-3sg unbelievable was
'The damage that the hail caused to my garden was unbelievable.'

The examples discussed in this section show that light verb constructions do not form a unified category. Some research is required, however, to be able to better classify the various light verb predicates based on their properties. The current Shiraz dictionary contains more than 8000 light verb constructions and the syntactic parser can correctly recognize them as well as any inflection that appears on them. The parser, however, is unable at this point to recognize light verb constructions with intervening elements.

6

System Architecture

The architecture of the Shiraz machine translation system is centered around the notion of a chart, capable of storing partial and complete results on a multitude of description levels, ranging from simple tokens, which appear in the source text, to syntactic analyses and target language output strings. The system operates on Unicode strings, complex typed feature structures are used to encode linguistic knowledge and intermediate results. The system consists of a number of modules which can be configured to perform different tasks, from glossing a text to full machine translation.

This report describes some aspects of the architecture and gives an outline of the modules involved in the translation from Persian to English. The system is completely written in C++ and can be used both on Unix machines or on PCs. In different configurations, the system is also used for several other translation projects at CRL, e.g., to translate Turkish, Korean, Arabic and others.

6.1 Charts for Shiraz

The central data structure within Shiraz is a chart, which is used to store partial and completed results on all levels of linguistic description. A Chart [Kay:80] is an acyclic, directed graph of hypotheses about parts of a document. Vertices correspond to points between words, edges denote words or descriptions of a sequence of words. Charts are extremely suitable for the representation of results within a natural language processing system. They allow to separate the description of what needs to be processed from the exact order in which actions are carried out, thus allowing for a wide range of search and processing strategies. Moreover, they remove redundancy since not only complete results are stored, but also all partial results that arise during a computation. These partial results can be reused in a larger context.

Shiraz uses several types of edges to distinguish between different types and levels of description. Thus, the chart can not only be used for a single purpose (say, syntactic parsing or generation), but it stores all hypotheses on all levels. Internally, so-called tags are used to mark edges as to what module they belong. In fact, the chart used for Shiraz is a weaker version of the layered chart used in [Amtrup:97], in that it does not support hypergraphs or the distribution of modules to employ parallel processing.

Edges in the chart are annotated with complex typed feature structures following [Carpenter:92]. Different types of feature structures can be used to encode different aspects of liguistic knowledge conveniently. We use an efficient implementation based on a vector-oriented representation for feature structures. Figure 1 shows an image of a chart and some of its edges. In the lower part of the image, part of a feature structure is shown.

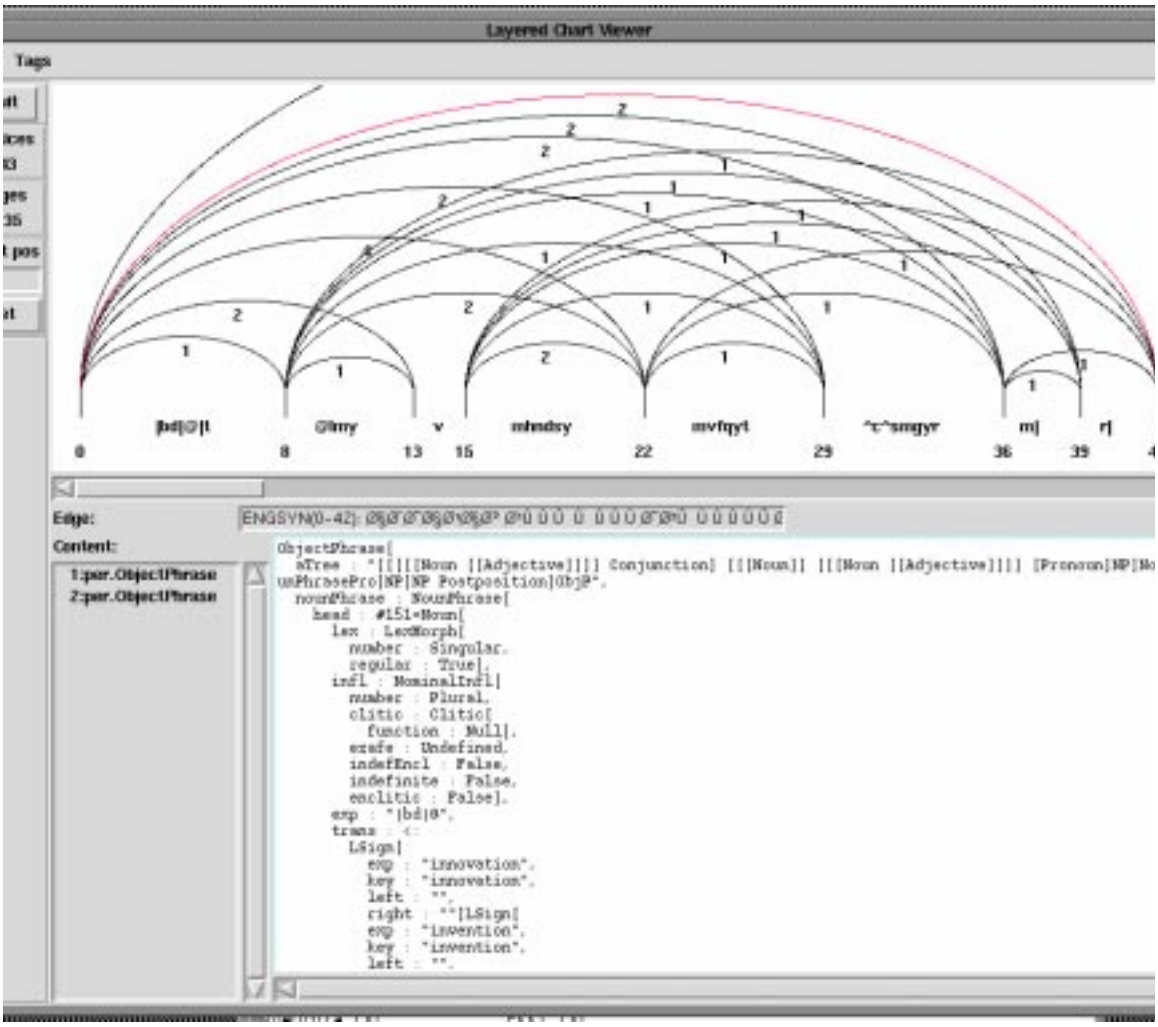


Figure 1: A Chart and some edges

6.2 Components and the application definition

The Shiraz system is designed to fulfill different functions within a natural language processing scenario. Two main requirements have to be met:

- The core system must be capable of processing different languages, probably using different scripts. Thus, Shiraz uses Unicode characters to represent user data throughout. Strings within feature structures are also Unicode.
- The system has to handle differents tasks. In the Persian case, at least a glosser and a full translation system have to be supported.

The approach we chose in order to realize a configurable, flexible system is a combination of extreme modularization and user-defined application. Shiraz consists of currently 27 different modules. The user is able to compose a sequence of modules in order to build a complete application. Upon runtime, the system interprets the application definition and executes the modules needed.

An application definition file defines

- A set of variable definitions, which can be used later on to save on typing and to group things,
- A set of application definitions, which define which modules to execute for which application, and
- A set of module definitions, which define the parameters for individual modules.

A small excerpt from the Persian application definition file is shown in figure 2. It exemplifies the composition of modules to form a complete application, as well as the definition of parameters, variables, and the incorporation of command-line parameters.

```
// Variable definitions
$RES=/home/mcm2/meat/per

// Global parameters
tangoModule = $(RES)/shiraz.mod

// An application
application lookup = Tokenizer($File=$1):PostTokenizer:MorphAnalyzer:
                     DictionaryLookup:DictionaryCompoundLookup:ChartViewer

// Sample module definitions
module Tokenizer {
    class = Tokenizer
    inputFile = /home/mcm/$File
    encoding = UTF8
}

module MorphAnalyzer {
    class = MorphAnalyzer
    grammar = $(RES)/GenMorph.samba
    rule = Morphology
    type = chart
    sourceTag = TOKEN
    targetTag = MATOKEN
}
```

Figure 2: Application definition file

6.3 Components of the Shiraz system

In this section, we give a short overview of the main components that are involved in constructing an English translation from a Persian document (see Figure 3). Using the mechanism just mentioned, an application is defined as a sequence of modules which are executed one after the other. The results of each component are gathered in the central chart and can be used by any other component. The translation process can be divided into five major steps:

1. Reading and preparing the input text
2. Morphological analysis and dictionary lookup
3. Syntactic parsing
4. Transfer
5. Generation and preparation of the target language output

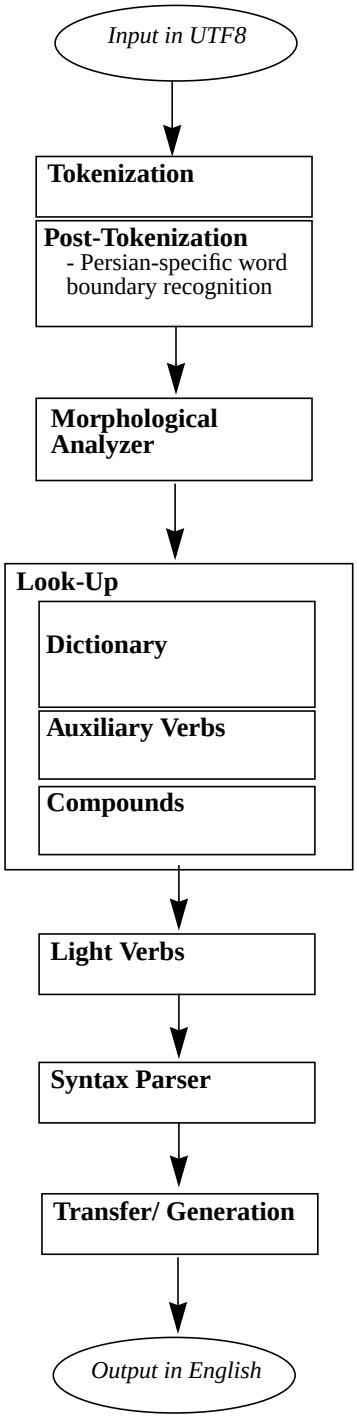


Figure 3: Shiraz system components

6.4 Preparing the input text

The first step in preparing the input text for a translation is performed by a Tokenizer, which reads an input file and splits it up into separate items such as words, punctuation, numbers etc. The input file is usually not in ASCII format, but rather a code conversion from some encoding to Unicode has to be performed. The tokenizer is a generic Unicode tokenizer, it is not specialized for any language.

For the Persian case, we also added a Posttokenizer. The task of this component is to postprocess the Tokenizer output with respect to some peculiarities of Persian. In particular, detached affixes are again attached to their kernels.

6.5 Morphological analysis and dictionary lookup

Then, in order to be able to perform dictionary lookup, the inflected surface words need to be processed by a Morphological Analyzer. We use a finite state transducer with feature structures formalism called Samba [Zajac:98] to describe morphological properties of words. Figure 3 shows a simple rule that describes the suffix which marks the causative form of Persian verbs. For more information see the web report on Persian Morphology.

```
CausativePastStem < GeneralRule;
CausativePastStem =
    <   RegularPresentStem
        <"|n" "d">
        [infl: per.VerbalInflection[causative: True]]
>;
```

Figure 4: A simple morphological rule

The dictionary itself is based on citation forms. It contains approx. 50000 entries. Dictionary Lookup takes the citation forms generated by the morphological analyzer and uses them to access lemma definitions in the lexicon. The inflectional information gained by morphology is then unified with the dictionary entry, rendering a rich description of the input word. For a more detailed description of the structure of the lexicon, see Section 4.

Compounding is taken care of in the Compound lookup component. Here, we are not looking for individual words in the dictionary, but rather take any sequence of words to find compounds. The compound lookup procedure is based both on citation forms and surface forms, since some compound parts are not words on their own right. We do not record the internal structure of compounds in the dictionary, but since Persian is a head-final language, we assume that the last element in a compound carries the most important inflectional information. The compound inherits this inflectional information, if possible.

6.6 Syntactic Parsing

The parser employed in Shiraz is a unification-based, bidirectional Chart parser. Figure 4 shows a simple syntax rule for the composition of complex noun phrases. The rules are phrase structure rules, and consist of a left hand side, which describes the constituent being formed, and a right hand side, which describes which subconstituents are used for the construction. Feature structures

on both sides allow to formulate restrictions and to build up structure. The rules can be parametrized to allow for certain special situations. First, they can be marked as non-recursive, in which case they are not used to propose new categories more than once at the same position. Second, they can be marked to perform dictionary lookup. If this happens, the left hand side is considered to refer to a dictionary entry and it is only constructed if there is an entry in the dictionary which matches the citation form built.

```
complexNP = per.Rule[
    lhs: per.NounPhrase[
            head: #np1,
            possessor: #np2],
    rhs: <:
        #np1= per.NounPhraseZero[
                    boundary: per.FalseOrUndefined]
        #np2= per.NounPhrase[
                    head:per.Nominal]
    :>
];
```

Figure 5: A sample syntax rule

In the Shiraz system, we use three incarnations of the parser to perform different tasks. You can think of this as having a grammar with different levels, each of which is applied in sequence. These incarnations are:

- The Auxiliary Verb Parser which is used to attach auxiliary verbs to their main verb counterparts. In doing that, morphological information is merged.
- The Light Verb Parser. Persian shows a large number of light verbs, which are combinations of non-verbal words (Nouns, Adjectives, etc.) with semantically poor verbs (most often ‘‘do’’). These combinations have a non-compositional semantics, so they need to be lexicalized. The light verb parser combines the individual parts of a light verb and looks it up in the dictionary.
- Finally, the Syntax Parser is used to construct syntactic constituents from individual words.

6.7 Transfer

The Transfer component is used to transform Persian syntactic structures to their English counterparts. Currently, we are only performing lexical transfer, i.e. the Persian morphological information is mapped to English inflectional features. Like all components within the Shiraz system, transfer is based on the chart notion. Incorporating syntactic transfer will allow to reuse partial translations within larger constructs (cf. [Amtrup:95]).

6.8 Generation and Surface Construction

Two components are involved in the final construction of English surface strings: The Syntactic Generator creates English fragments, and the Surface Generator searches for a suitable path through the fragments.

Syntactic generation currently uses a simple method of linearization of English words. There is no complex mechanism to generate surface strings from syntactic descriptions. A sample rule for the generator is shown in Figure 5.

```
np1 = [ structure: per.NounPhrase[
            head: #1= Top,
```

```
            relClause: #2= Top],
          order: <: #1 #2 :>,
          trigger: "relClause"
];
```

Figure 6: A sample generation rule

The rule demonstrates the three elements present in a generation rule:

1. The structure defines what kind of syntactic structure can be handled by the rule.
2. The order defines in which surface order the underlying parts should be generated. Fixed strings can be inserted here as well, e.g. to mark the possessor in English with an additional ‘‘of’’.
3. The trigger restricts the application of rules. If present, then the feature marked by the trigger path has to be non-empty in order for the rule to be applicable.

Apart from constructing surface strings from syntactic descriptions, a morphological generation procedure is performed during this phase. Thus, English words are generated with correct inflection. The surface generation, finally, chooses the best path through the graph of generated English surface fragments and issues these as output. In the future, we plan to use an English language model to choose among the many possible surface strings. The string which is ranked best by the model will be issued.

6.9 System Statistics

The system is completely written in C++ (with the exception of a small Java applet used to render Persian script for the glosser). It consists of approx. 27000 lines of code. It can be run both on Unix platforms (using the Gnu compiler) and PCs running Windows 98/NT (using Visual C++). Translating a sentence of medium length and complexity (i.e., ambiguity) takes between 3 and 5 seconds.

6.10 Conclusion

Shiraz is a machine translation system for translating Persian written text into English. It is based on two main architectural foundations: The use of a chart throughout the system, which allows an integrated view on results created on all levels of linguistic description, and the use of a complex typed feature structure formalism, which unifies the view on the descriptions itself.

Acknowledgments

We wish to thank Mike Freider, Jane Freider, Nigel Sharples, Mohammad Reza Aidinejad, Javad Moghaddas and Saadat Pour Mozafari for their participation in the Shiraz project.

Appendix

Transliteration (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
q	*similar to* French “r”
ch	church
sh	shoe
zh	mirage
a	and
â	father
e	bed
i	sea
o	so
u	food

References

Abolghasemi, Mohsen. 1995. *Rishe Shenasi [Etymology]*. Tehran, Iran: Teghnus Press.

Amtrup, Jan W. 1997. Layered Charts for Speech Translation. In Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation, TMI '97, Sante Fe, NM, Jul. 1997, pp. 192-199.

Amtrup, Jan W. 1995. Chart-based Incremental Transfer in Machine Translation. In Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation. KU Leuven, July 1995, pp. 188-195.

Bateni, Mohamad-Reza. 1995. *Towsif-e Sakhteman-e Dastury-e Zaban-e Farsi* [Description of Persian Syntax].Tehran,Iran: Amir Kabir Publishers.

Carpenter, Bob. 1992. The Logic of Typed Feature Structures. Tracts in Theoretical Computer Science, Cambridge University Press, Cambridge, MA.

Comrie, Bernard. 1987. *The World's Major Languages*. New York: Oxford University Press.

Karimi-Doostan, Mohamad-Reza (1997). Light Verb Constructions in Persian. Doctoral dissertation, University of Essex.

Kay, Martin. 1980. Algorithmic Schemata and Data Structures in Syntactic Processing. Technical Report CSL-80-12, Xerox Palo Alto Research Center.

Khanlari, Parviz N. 1995. *Tarikh-e Zaban-e Farsi* [History of the Persian Language]. Tehran, Iran: Simorgh Press.

Lazard, Gerard. 1992. *A Grammar of Contemporary Persian*. Costa Mesa, California: Mazda Publishers.

Mahootian, Shahrzad. 1997. *Persian*. Routledge, New York, NY.

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

Related Publications

Amtrup, Jan W., Karine Megerdoomian and Remi Zajac, 1999. "Rapid Development of Translation Tools". In *Proceedings of Machine Translation Summit VII*, Singapore, pp.385-389, September1999.

Megerdoomian, Karine, 2000. "Unification-Based Persian Morphology". In *Proceedings of CICLing 2000*. Alexander Gelbukh (ed.). Centro de Investigacion en Computacion-IPN, Mexico.

Megerdoomian, Karine. 2000. "Persian Computational Morphology: A Unification-Based Approach". NMSU, CRL, Memoranda in Computer and Cognitive Science Report (MCCS-00-320).

Megerdoomian, Karine. 2000. "A Computational Analysis of the Persian Noun Phrase". NMSU, CRL, Memoranda in Computer and Cognitive Science Report (MCCS-00-321).

Megerdoomian, Karine and Hamid Mansouri Rad. 2000. “Acquisition of Persian Resources: Corpora and Dictionary Development in the Shiraz Project”. NMSU, CRL, Memoranda in Computer and Cognitive Science Report (MCCS-00-323).
Megerdoomian, Karine and Remi Zajac. 2000. “Processing Persian Text: Tokenization in the Shiraz Project”. NMSU, CRL, Memoranda in Computer and Cognitive Science Report (MCCS-00-322).