FACULTY OF HUMAN AND SOCIAL SCIENCES

An Electronic Dictionary of Persian Verbs

Bahareh Kakanaeeni

Dissertation
Master of Language Sciences

Supervisor Prof. Doutor Jorge Baptista

Faro, 2014

Statement

An Electronic Dictionary of Persian Verbs

Declaração de autoria do trabalho

Declaro ser a autora deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

An Electronic Dictionary of Persian Verbs

Declaration of authorship of the work

I hereby declare to be the author of this work, which is original and unpublished. Writers and works consulted are properly quoted in the text and listed in the list of references here included.

Acknowledgements

Foremost, I would like to express my honest appreciation to my supervisor Prof. Jorge Baptista for support of my Master study and thesis, for his patience, motivation, enthusiasm, and enormous knowledge. His supervision helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and teacher for my Master study.

My deepest heartfelt appreciation goes to my dear husband, Vahid, for his passion, kindness and sustenance, without him this project would simply stay as a nice idea.

Last but not the least, I express my love and gratitude to my families, for their understanding and endless love, through the duration of my studies.

Faro, July 3rd 2014

Bahareh Kakanaeeni

Abstract

There are more than 110 million Persian speakers in the world, but the lexical resources for Natural Language Processing (NLP) of the Persian language are still too scarce. Though Persian (or Fārsi) uses an adapted form of the Arabic script, it is an Indo-European language and its verbal inflection is based on stems and affixes.

This project is to be considered as a first step towards the construction of large-scale lexical resources for Persian, and the development of a Persian module to be distributed with the Unitex linguistic development platform. The specific goal of this dissertation is to build a morphologic, machine-readable dictionary of Persian verbs, using a dictionary of lemmas and a set of morphologic finite-state transducers (FST) to generate all the inflected forms associated to each lemma, and encode them with all the relevant morphosyntactic information (tense, person-number, etc.).

This task is complicated in Persian verbal morphology by the fact that each verb has two stems (past and present), different inflection paradigms are used for written (formal) and oral (informal) language uses, and several compound tenses can be formed through combining prefixes and suffixes with base inflected forms, the same tense being able to constitute one, two or more different tokens (separate written forms).

A small dictionary of lemmas (145) and their respective written (292) and spoken (127) stems was built, each stem was provided with the appropriate inflection conventional code, which correspond to an inflection paradigm. The list of lemmas was compiled based on frequency data from a large Persian corpus, the TEP (Tehran English-Persian Parallel Corpus), containing around 4.5 million words, and selecting the most frequent verb forms.

At its current state, the dictionary of inflected forms contains 1,536 entries.

In Persian, in average, each verb lemma yields 28 simple inflected forms, 14 simple 'written' inflected forms and 14 'spoken' inflected forms, and relatively, for each stem there are 7 inflected forms.

The recognition of compound tenses is carried out by a set of 22 FST using the system morphological mode and the previous lexical annotation of simple verb forms. These FSTs allowed for the retrieval of 3,953 compound verb tenses from the corpus.

For the evaluation of this language resource, a sample text, retrieved from the www.persian.euronews.com, and containing around 1000 words was used to assess the lexical coverage of the simple words' dictionary and the compound tenses' lexical graphs. The evaluation was done manually, based on the recognized words and those that were not identified by lexical resources built here.

Keywords

Persian (Farsi), verb, electronic (or machine-readable) dictionary, inflectional morphology, finite-state automata and transducers

Resumo

Há mais de 110 milhões de falantes da língua Persa (ou Farsi) no mundo, mas os recursos lexicais para Processamento de Língua Natural (NLP) disponíveis para esta língua ainda são muito escassos. Embora o Persa use uma forma adaptada do alfabeto árabe, é uma língua indo europeia e sua flexão verbal é baseada em morfemas radicais e afixos.

Este projeto deve ser considerado como um primeiro passo para a construção de recursos lexicais em larga escala para o Persa, e o desenvolvimento de um módulo do Persa a ser distribuído com a plataforma de desenvolvimento linguístico UNITEX. O objetivo específico deste trabalho é construir um dicionário eletrónico, legível por máquinas, de verbos do persa, usando um dicionário de lemas e um conjunto de transdutores de estados finitos (FST) morfológicos para gerar todas as formas flexionadas associadas a cada lema, codificando-as com toda a informação morfossintática relevante (pessoa-número, tempo-modo, etc.). Esta tarefa é complicada na morfologia verbal persa pelo facto de cada verbo tem dois estemas (ing. *stem*), um para os tempos do passado e outro para os do presente; são também empregues diferentes paradigmas de flexão na linguagem escrita (mais formal) e na oralidade (informal); e, finalmente, pelo facto de vários tempos compostos poderem ser formados através da combinação a prefixos e sufixos com formas flexionadas de base, ao mesmo tempo que podem constituir-se num único, em dois ou até mais diferentes palavras gráficas (*tokens*).

Um pequeno dicionário de lemas (145) e seus respectivos estemas da forma 'escrita' (292) e 'falada' (127) foi, então, construído, e cada estema acompanhado pelo código de flexão convencional de flexão adequado, e que corresponde a um paradigma de inflexão. A lista dos lemas foi compilada com base em dados de frequência de um grande corpus persa, o TEP (Tehran English –Persian Parallel Corpus), contendo cerca de 4,5 milhões de palavras, e selecionando as formas verbais mais frequentes. No seu estado atual, o dicionário de formas flexionadas contém 1.536 entradas. Em Persa, em média, cada lema verbal produz 28 formes simples: 14 formas simples 'escritas' flexionadas e 14 formas 'faladas', dado que cada stema.

O reconhecimento de tempos compostos é realizado por um conjunto de 22 FSTs utilizando o modo morfológico do sistema e a prévia anotação lexical de formas verbais simples. Estes grafos permitiram reconhecer 3.953 formas compostas no corpus utilizado.

Para a avaliação deste recurso linguistico, utilizou-se um texto obtido a partir da www.persian.euronews.com e contendo cerca de 1.000 palavras, a fim de estimar a cobertura lexical dos recursos construídos. A avaliação foi feita manualmente com base nas palavras reconhecidas e nas não identificadas pelos recursos lexicais aqui construídos.

Palavras-chave

Persa (Farsi), verbo, dicionário eletrónico, morfologia flexional, autómatos e transdutores de estados finitos

Extended Abstract

This project is to be considered as a first step towards the construction of large-scale lexical resources for Persian, and the development of a Persian module to be distributed with the Unitex linguistic development platform.

The Unitex 3.0 has resources for English, Finnish, French (France, Quebec), Georgian (Ancient), German, Greek (Modern and Ancient), Italian, Korean, Norwegian, Polish, Portuguese (Brazil, Portugal), Russian, Serbian (Cyrillic and Latin alphabets), Spanish and Thai.

The specific goal of this dissertation is to build a morphologic, machine-readable dictionary of Persian verbs, using a dictionary of lemmas and a set of morphologic finite-state transducers (FST) to generate all the inflected forms associated to each lemma, and encode them with all the relevant morphosyntactic information (tense, person-number, etc.) because, there are more than 110 million Persian speakers in the world, but the lexical resources for Natural Language Processing (NLP) of the Persian language are still too scarce.

Though Persian (or Fārsi) uses an adapted form of the Arabic script, it is an Indo-European language and its verbal inflection is based on stems and affixes.

This task is complicated in Persian verbal morphology by the fact that each verb has two stems (past and present), different inflection paradigms are used for written (formal) and oral (informal) language uses, and several compound tenses can be formed through combining prefixes and suffixes with base inflected forms, the same tense being able to constitute one, two or more different tokens (separate written forms).

A small dictionary of lemmas (145) and their respective written (292) and spoken (127) stems was built, each stem was provided with the appropriate inflection conventional code, which correspond to an inflection paradigm.

Since the inflected forms of a given verb are produced from its stems, the lemma proper was encoded as a supplementary feature in the verbal entry, which allows for the retrieval of any inflected verb form from texts using that feature.

The list of lemmas was compiled based on frequency data from a large Persian corpus, the TEP (Tehran English-Persian Parallel Corpus), containing around 4.5 million words, and selecting the most frequent verb forms.

For each inflection paradigm, a morphologic FST was built to generate the corresponding inflected forms.

At its current state, the dictionary of inflected forms contains 1,536 entries.

In Persian, in average, each verb lemma yields 28 simple inflected form, 14 simple 'written' inflected forms and 14 'spoken' inflected forms, and relatively, for each stem has 7 inflected forms.

The recognition of compound tenses is carried out by a set of 22 FST using the system morphological mode and the previous lexical annotation of simple verb forms. These FSTs allowed for the retrieval of 3,953 compound verb tenses from the corpus.

Two Bakus-Baur form (BNF), one for simple and another for compound inflected forms' entries were built in the form of finite-state graphs and used to automatically validate the format of the lexical entries.

For the evaluation of this language resource, a sample text, retrieved from the www.persian.euronews.com, and containing around 1000 words was used to assess the lexical coverage of the simple words' dictionary and the compound tenses' lexical graphs.

The evaluation was done manually, based on the recognized words and those that were not identified by lexical resources built here.

Only one simple verb form, from the 43 verb forms found (107 occurrences), was not in the dictionary by a lapse in the construction of the respective graph flexion, which has been corrected.

Regarding compound verb forms, the lexical graphs produced 31 lexical entries corresponding to 20 different forms, occurring 46 times in the text. Only one form, involving the contraction of a negation morpheme with a verb (a single instance) was not identified.

Resumo Alargado

Há mais de 110 milhões de falantes da língua Persa (ou Farsi) no mundo, mas os recursos lexicais para Processamento de Língua Natural (NLP) disponíveis para esta língua ainda são muito escassos. Embora o Persa use uma forma adaptada do alfabeto árabe, é uma língua indo europeia e sua flexão verbal é baseada em morfemas radicais e afixos.

Este projeto deve ser considerado como um primeiro passo para a construção de recursos lexicais em larga escala para o Persa, e o desenvolvimento de um módulo do Persa a ser distribuído com a plataforma de desenvolvimento linguístico UNITEX.

O objetivo específico deste trabalho é construir um dicionário eletrónico, legível por máquinas, de verbos do persa, usando um dicionário de lemas e um conjunto de transdutores de estados finitos (FST) morfológicos para gerar todas as formas flexionadas associadas a cada lema, codificando-as com toda a informação morfossintática relevante (pessoa-número, tempo-modo, etc.).

Esta tarefa é complicada na morfologia verbal persa pelo facto de cada verbo tem dois estemas (ing. *stem*), um para os tempos do passado e outro para os do presente; são também empregues diferentes paradigmas de flexão na linguagem escrita (mais formal) e na oralidade (informal); e, finalmente, pelo facto de vários tempos compostos poderem ser formados através da combinação de prefixos e sufixos com formas flexionadas de base, ao mesmo tempo que podem constituir-se num único, em dois ou até mais diferentes palavras gráficas (*tokens*).

Um pequeno dicionário de lemas (145) e seus respectivos estemas da forma 'escrita' (292) e 'falada' (127) foi, então, construído, e cada estema acompanhado pelo código de flexão convencional de flexão adequado, e que corresponde a um paradigma de inflexão.

Uma vez que a flexão é produzida a partir dos estemas, o lema do verbo é tratado como um traço suplementar nas entradas do dicionário, o que permite pesquisar as ocorrências de um dado verbo num texto a partir do respetivo lema.

. A lista dos lemas foi compilada com base em dados de frequência de um grande corpus persa, o TEP (Tehran English –Persian Parallel Corpus), contendo cerca de 4,5 milhões de palavras, e selecionando as formas verbais mais frequentes.

No seu estado atual, o dicionário de formas flexionadas contém 1.536 entradas. Em Persa, em média, cada lema verbal produz 28 formes simples: 14 formas simples 'escritas' flexionadas e 14 formas 'faladas', dado que cada stema.

O reconhecimento de tempos compostos é realizado por um conjunto de 22 FSTs utilizando o modo morfológico do sistema e a prévia anotação lexical de formas verbais simples. Estes grafos permitiram reconhecer 3.953 formas compostas no corpus utilizado.

Duas formas Bakus-Naur (BNF), uma para as entradas das palavras simples e outra para as entradas de formas compostas flexionadas, foram construídas sos a forma de gráficos de estados finitos e utilizadas para validar automaticamente o formato das entradas lexicais.

Para a avaliação deste recurso linguistico, utilizou-se um texto obtido a partir da www.persian.euronews.com e contendo cerca de 1.000 palavras, a fim de estimar a cobertura lexical dos recursos construídos.

A avaliação foi feita manualmente com base nas palavras reconhecidas e nas não identificadas pelos recursos lexicais aqui construídos.

Apenas uma forma simples de um verbo, de entre as 43 encontradas (107 ocorrências), não estava registada no dicionário, por lapso na construção do respetivo grafo de flexão, o que foi corrigido.

Quanto às formas verbais compostas, os grafos lexicais produziram 31 entradas lexicais, correspondendo a diferentes flexões de 20 formas, que ocorrem 46 vezes no texto.

Apenas uma forma, correspondendo à contração de um morfema de negação com um verbo (uma única ocorrência) não foi identificada.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

There are approximately 110 million Persian (natively known as فارسی **Farsi** [fɒːrˈsiː] or پارسی **Parsi**) speakers worldwide[1], with the language property official status in Iran, Afghanistan and Tajikistan. For centuries, Persian has likewise been a prestigious cultural language in Central Asia, South Asia, and Western Asia. Persian is used as a liturgical language of Islam not only in Iran, Afghanistan, and Tajikistan, but also in Pakistan and North India.

In spite of its cultural and demographic importance, studies in Natural Language Processing (NLP) on Persian are scarce. One of the main problems is the fact that lexical resources for NLP are almost non-existed in Persian, as well as, no literature review existed for this theme.

We intend to start the building of a new module for the Unitex[2] linguistic development platform (Paumier 2003, 2013) by describing the inflection of Persian verbs.

This project is to be measured as a first step towards the construction of large-scale lexical resources for Persian, and the development of a Persian module to be distributed with the Unitex linguistic development platform.

The project consists in formalizing the main inflectional paradigms of Persian verbs and produce automatically all the inflected forms, simple and compound, both formal and informal forms[3], associated with them from a dictionary of lemmas.

[1] http://en.wikipedia.org/wiki/Persian_language [URL <access in 10-05-2013>]

[2] http://www-igm.univ-mlv.fr/~unitex/

[3] In Persian, verbs can be used both in formal and informal speech and these informal uses occur most of the times in speaking but they can appear in writing also; formal and informal verb forms are morphologically different so they have to be taken into account in inflectional dictionary. In following pages, we call these verbs as a *spoken* (informal) and *written* (formal) forms. For example, آمد *amad* is the written stem of the verb "come" but the spoken stem is اومد *omad*.

Unitex is a collection of programs developed for the analysis of texts in natural language by using linguistic resources and tools. These resources consist of electronic dictionaries, local grammars and the lexicon-grammar tables with lexicon-grammatical information about the predicative element's properties of a language (M. Gross 1996).

These languages already exist in Unitex: English, Finnish, French (France, Quebec), Georgian (Ancient), German, Greek (Modern and Ancient), Italian, Korean, Norwegian, Polish, Portuguese (Brazil, Portugal), Russian, Serbian (Cyrillic and Latin alphabets), Spanish and Thai.

In such electronic dictionary of inflected forms, each verb form is linked to the corresponding lemma, its part of speech (POS) and all grammatically relevant information pertaining to that word.

1.1. Objectives

For achieving the aim of this project, we must perform these following steps:

1. Build a sufficiently large lexicon of Persian verbs, in parliamentary procedure to reach a reasonable lexical coverage;
2. Produce a Persian verbs dictionary of lemmas, associated with their corresponding inflectional paradigms;
3. Build inflectional graphs to generate automatically the both simple and compound forms of Persian verbs, in order to produce a dictionary of inflected forms, aiming at a highly granular and systematic description of all the relevant grammatical values pertaining to those forms.
4. Apply the BNFs to validate automatically the inflect words' dictionary that retrieved from the corpus to evaluate the formal correction of the result.
5. Apply the Persian dictionary to a sample of text in order to assess the lexical coverage and grammatical adequacy of the dictionary.

2. Methodology

2.1. Synopsis of Persian verb morphology

Persian has a complex verbal tense system[4] (Tabatabayi, A. (2010))[5] (its Instructional Technology Services. (2007))[6]. Besides having specific endings for some tenses, for instance:

- Simple Past (GS[7]) رفتم, *raftam*[8], (lit. "went I"), means "I went"
- Simple Present (HS) روم, *ravam*, (lit. "go I"), means "I go";

Certain tenses are produced by adding prefixes to some inflected forms, for example:

- Past Continuous (GC), می رفتم (میرفتم) *miraftam [mi+raftam]*, (lit. "ing=duration went I"), means "I was going" which is formed by the prefix *mi–* (می) and the simple past forms followed by the suffix *–am* (ام) ;
- Present Continuous (HC) می روم (میروم), *miravam [mi ravam]*, and (lit. "ing=duration go I"), means "I am going" that is also shaped by the prefix *mi–* (می) and the simple present forms followed by the suffix *–am* (ام). These two tenses can be composed with the prefixes both joined and parted;

[4] http://en.wikipedia.org/wiki/Persian_verbs [URL <access in 15-06-2013>]
[5] http://dlib.ical.ir/site/catalogue/789489 [URL <access in 15-06-2013>]
[6] http://sites.la.utexas.edu/persian_online_resources/verbs/ [URL <access in 01-07-2013>]
[7] The codes here used for tenses are conventional, but they were derived from the Persian terminology, for easy mnemonic; **GS**= Gozashteye Sade (گذشته ساده) Simple Past, **HS**= Hale Sade (حال ساده) Simple Present, **GC**= Gozashteye estemrari (گذشته استمراری) Past Continuous, **HC**= Hale estemrari (حال استمراری) Present Continuous, **HP**= Hale kamel (حال کامل) Present Perfect, **HU**= Hale eltezami (حال التزامی) Present Subjunctive, **I**= amri (امری) Imperative, **HF**= Hale Kamel estemrari (حال کامل استمراری) Present Perfect Continuous, **GP**= Gozashteye kamel (گذشته کامل) Past Perfect, **GF**= Gozashteye Kamel estemrari (گذشته کامل استمراری) Past Perfect Continuous, **HK**= hale kamele eltezami (حال کامل التزامی) Perfect Subjunctive), **AS**= Ayande (آینده) Simple Future.
[8] All example provided in a first person singular.

- Present Subjunctive (HU) بـرومِ, *berravam*, (lit. “be=subjunctive particle go I”), means, “(that) I go” that is formed by the prefix *be–* (بِـ) and the simple present forms followed by the suffix *–am* (اَم).
- Imperative (I) بـرو, *boro*, (lit. “b=imperative particle go you”), means “(you) GO!” which is made by the prefix *b–* (بـ) followed by the root of the present form *ro* (رو);

In addition, Persian has one tense, which is created by appending suffixes to the simple past form of the verb like as:

- Present Perfect (HP), رفـتـه ام *rafteh am*, (lit. “gone I”), means “I have gone” which is made by the simple past forms followed by the suffix *–ham* (هَ اَم);

Furthermore, compound tenses are created by adding a prefix and a suffix to some inflected forms and they can be written as two tokens or three tokens compound, such every bit:

- Present Perfect Continuous (HF) مـیـرفـتـه ام (مـی رفـتـه ام), *mirafteh am [mi+rafteh+am*], and (lit. “ing=duration gone I”), means “I had been going” is formed by the prefix *mi–* (مـی) and the simple past forms followed by the suffix *–ham* (هَ اَم).

There are also some compound tenses made by adding suffix or prefix to some inflected forms, but in this case, the suffix or prefix is not meaningless words, it is a verb and also for changing, the person of the verb the suffix or prefix verb changed not a basic verb, such as:

- Past Perfect (GP) رفـتـه بـودم, *rafteh bodam* (lit. “gone been I”), means “I had gone” is formed by the simple past third singular person followed by *–h* (ه) and inflected form of the simple past forms of *bod* (بـود) *been;*

- Past Perfect (GF) رفـتـه بـود ه ام, *rafteh bode am [rafteh+bode+am]*, (lit. "gone had been I"), means "I had gone" is formed by the simple past third singular person followed by *–h* (ه) and the simple past forms of *bod* (بـود) *been* followed by the suffix *–ham* (ه ام);
- Present Perfect Subjunctive (HK) رفـتـه بـا شم *rafteh basham* (lit. "gone be I"), means "I have gone" is formed by the simple past third singular person followed by *–h* (ه) and inflected forms of the simple present forms of *bash* (بـا ش) *be;*
- Simple Future (AS) خـو ا هم رفـت, *khaham raft* (lit. "Want I go"), means "I will go" is formed by inflected forms of the simple present forms of *Khah* (خـو ا ه) *want* followed by the simple past third singular person form of the verb.

2.2. A lexicon of Persian verbs

In order to select the most frequent verbs for the lexicon, we used the TEP corpus: Tehran English-Persian Parallel Corpus (Pilevar *et al*. (n.d.)) [9] obtained from the Natural Language and Text Processing Laboratory of Tehran University[10]. The corpus consists of 4 million tokens and, after being tokenized with Unitex 3.0, it features 556,234 sentences, 15,166,987 (64,492 different) tokens, 4,485,147 (64.365 different) simple word forms and 3,239,250 (10 different) digits.

Using the frequency list of the corpus' words ranked by decreasing frequency and automatically listed by the Unitex, we estimated the cumulative frequency of its word forms.

We then annotated the verb forms, indicating their lemma, starting from the most frequent words until we attained a rank corresponding to a cumulative percentage of 90% of the total corpus word forms (around 3 million and a half) corresponding to 1,466 different verb forms.

[9] http://ece.ut.ac.ir/nlp/resources.htm
[10] http://ece.ut.ac.ir/nlp/

Count	Token	Cum.count	Cum. %	Lemma	Stem	Transliteration	Translation	V. Class	DELAS entry (Stem)
16877	بود	16877	0.015707	بودن	بود	bod	Been	V002	بود,V002
12935	باشه	29812	0.027746	بودن	باش	bash	Be	V001	باش,V008
11564	شده	41376	0.038509	شدن	شد	shod	became	V002	شد,V002
10835	کنم	52211	0.048593	کردن	کن	kon	Do	V001	کن,V001
10295	داره	73057	0.067994	داشتن	دار	dar	have	V001	دار,V008
9428	کن	82485	0.076769	کردن	کن	kon	do	V001	کن,V001
9344	دارم	91829	0.085466	داشتن	دار	dar	have	V001	دار,V001
8614	است	100443	0.093483	بودن	است	ast	Is	V003	است,V003
7104	هست	107547	0.100094	بودن	هست	hast	Is	V004	هست,V004
6932	داري	114479	0.106546	داشتن	دار	dar	have	V001	دار,V001
6756	میشه	121235	0.112834	شدن	شو	sho	become	V001	ش,V008

Table 1. Lexicon table of Verbs

Table 1 indicates a fragment of the lexicon of verbs, built from the list of simple row and their frequencies, where just the verb classes are presented, in decreasing frequency in decreasing frequency.

The Count column shows the frequency of the token in the corpus and the Cumulative count column presents the total number of verb forms up to that word.

The Cumulative percentage column indicates the percentage of the corpus that the previous Cumulative count represents.

For each token, the corresponding lemma was added. Since, in Persian, the simple tenses are derived from two distinct stems, each stem constitutes a different entree, but the same lemma is provided for both stems, so that the two sets of inflected forms will be linked up to the same lemma.

For example: رفـتـن *raftan* “to go” has two stems; رو *ro* “go” is the stem of the present form and رفـت *raft* "gone” is the stem of past form of the same verb. In Persian, lemma builds by adding *“*ـَن *an”* to the end of past stem, for example: رفـت *raft* + ـَن *an* = رفـتـن *raftan*.

Also for more consideration about each verb stem and the lemma, we use two websites (Redirected from Persian in Texas)[11] and (فـعـل سـا د ه feele sade “Simple Verb”[12]) that these websites are contain of Persian verb stems and lemmas.

2.3. A Persian verbs dictionary of inflected forms

For building a dictionary of lemmas and their respective written and spoken stems, for each stem, we must provide the appropriate inflection conventional code, which correspond to an inflection paradigm.

Since the inflected forms of a given verb are produced from its stems, the lemma proper was encoded as a supplementary feature in the verbal entry, which allows for the retrieval of any inflected verb form from texts using that feature.

[11] http://persian.nmelrc.org/pvc/simpleverbs.php
[12]http://fa.wikipedia.org/wiki/%D9%81%D8%B9%D9%84_%D8%B3%D8%A7%D8%AF%D9%87

2.3.1. BNFs

In order to provide formal guidelines for the description of the verb tense inflection[13], we have built a BNF (Backus Normal Form or Backus–Naur Form)[14] with all the relevant grammatical information pertaining to the verb inflection. In computer science, a BNF is an annotation technique for context-free grammars, often used to identify the syntax of languages used in computing, often used to describe the syntax of languages used in computing. Two Finite-State Automata (FSA), shown in **Figure 1** and **Figure 2**, were built to be used as BNFs for the validation of the simple and compound inflected words' dictionaries.

The goal is to use these FSA in order to validate automatically the inflect words' dictionary entries.

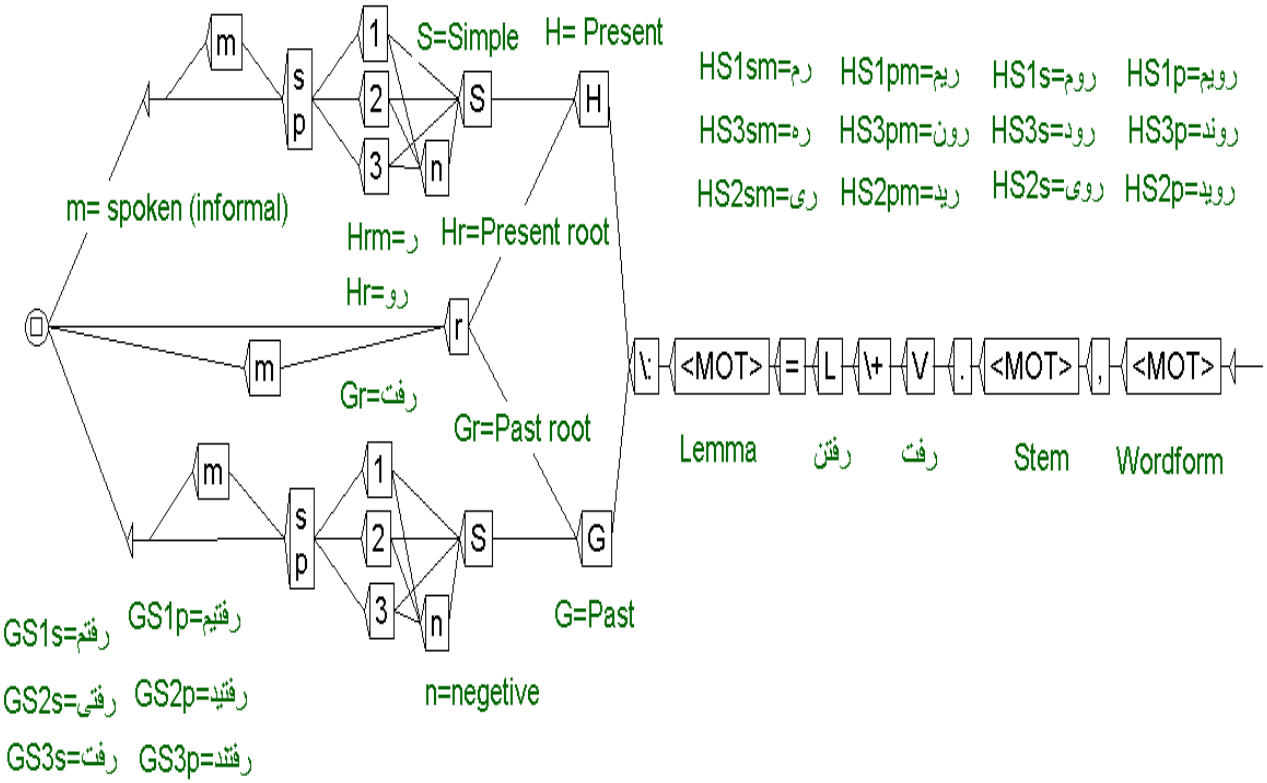


Figure 1. BNF1 for simple tenses

[13] http://en.wikipedia.org/wiki/Persian_verbs
[14] http://en.wikipedia.org/wiki/Backus%E2%80%93Naur_Form

The graph (**Figure 1**) describes the format of an inflected word entry in the DELAF (dictionary of inflected form) and it reads (right to left[15]) as follows: the first box with the in-built mask <MOT> stands for any simple word (one token); it is followed by a comma ',' and the second word stands for the stem; then a dot '.' separates it from the part-of-speech and here, only "V" for verb, followed by '+' and "L" is a code for lemma, "=" shows that the "L" is equal to the corresponding Lemma that shows as a third word (MOT); in this stage, the verb entry has two major paths issuing from that point onward: one for "H", representing the present forms, and another for "G", which represents the past forms; a third path goes to "r" for the roots of both the present and the past forms. As this is a simple word entries' BNF, the past and the present forms have only simple tenses; thus, each main path is followed by the "S" for simple tense, the code "n" is used to show the negation, the "1", "2", and "3" symbols for 1st, 2nd, 3rd person respectively, "s" and "p" for singular and plural forms which, "m" stands for spoken (informal) verb forms and also the line without "m" shows the written (formal), for example: *HS1s* is written simple present first singular person and *HS1sm* is spoken simple present first singular person. The same is done for the past forms.

The graph in the following figure (**Figure 2**) is read in the similar way as the previous one, but it describes the entries of compound tenses. The first box shows a loop for the compound form entry (more than one token). For the present tenses "H", we find the codes[16] for the Continuous "C", Subjunctive "U", Perfect Continuous "F", Perfect "P" and Perfect Subjunctive "K" tenses; for the past tense "G", we find the Continuous "C", Perfect continuous "F" and Perfect "P". For the Imperative mode the code "I" is used, but only the second person value has been considered, both

[15] Normally, graphs in Unitex are from left to right but the direction here is opposite because writing order in Persian is from right to left.

[16] We do not use the same letters (codes) as in Unitex manual, because it does not have the all letters which we need, also we wanted that the present (H) and the past (G) be divided to be more understandable.

for singular and plural. For the Future tense “A”, there is only the Simple “S”, which in fact is a compound form (e.g. خواهم رفت, *khaham raft* (lit. “Want I go”) means, “I will go”), the “simple” being the traditional grammatical term used to describe it.

Examples of all inflections are provided for the verb *رفتنraftan “to go”.*

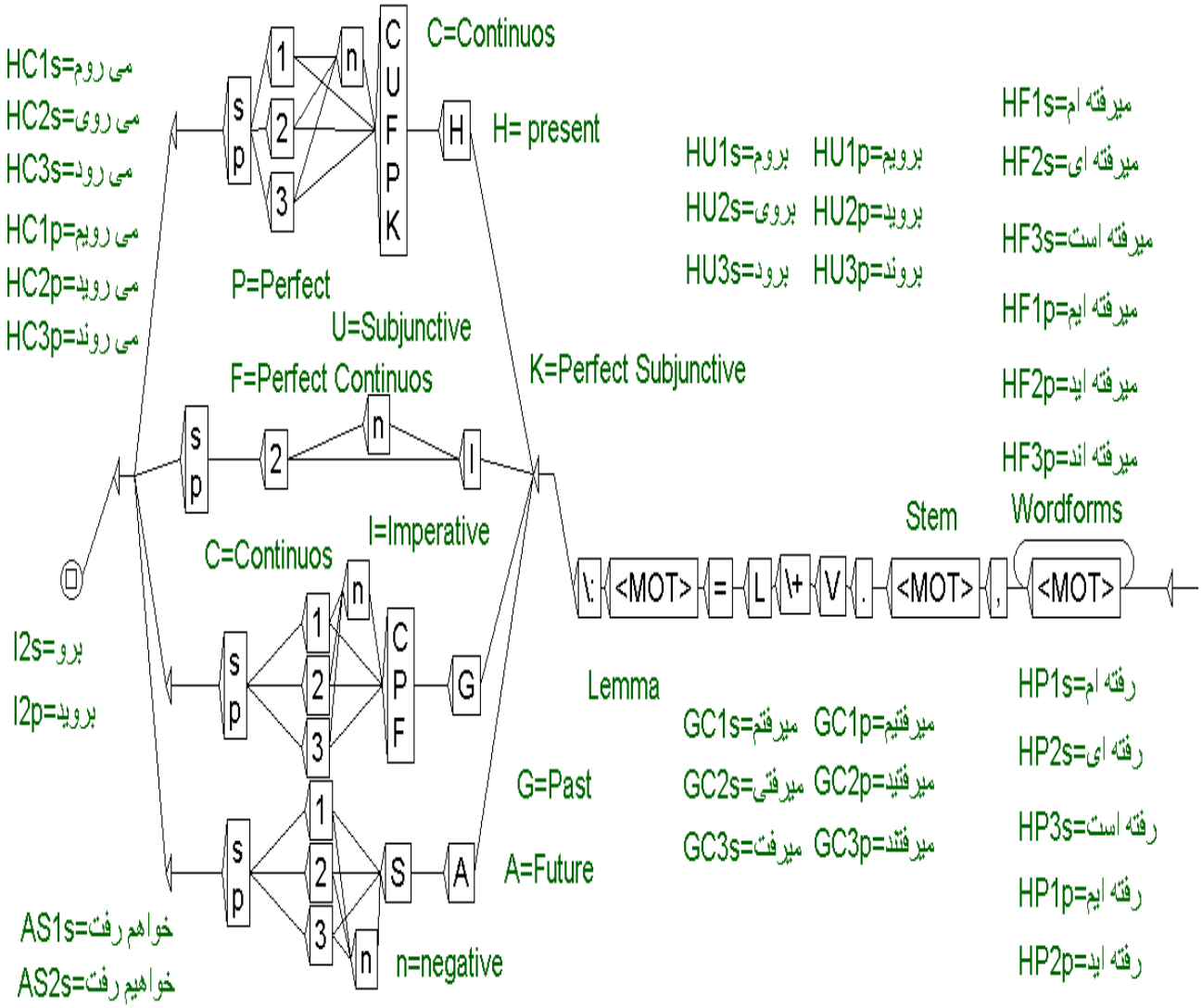


Figure 2. BNF2 for compound tenses

2.3.2. Inflection FSTs for Written simple words

In order to generate the inflected forms associated to each lemma, we built four graphs for written simple forms of verbs. These graphs are shown in **Figure 3** and **Figure 4**.

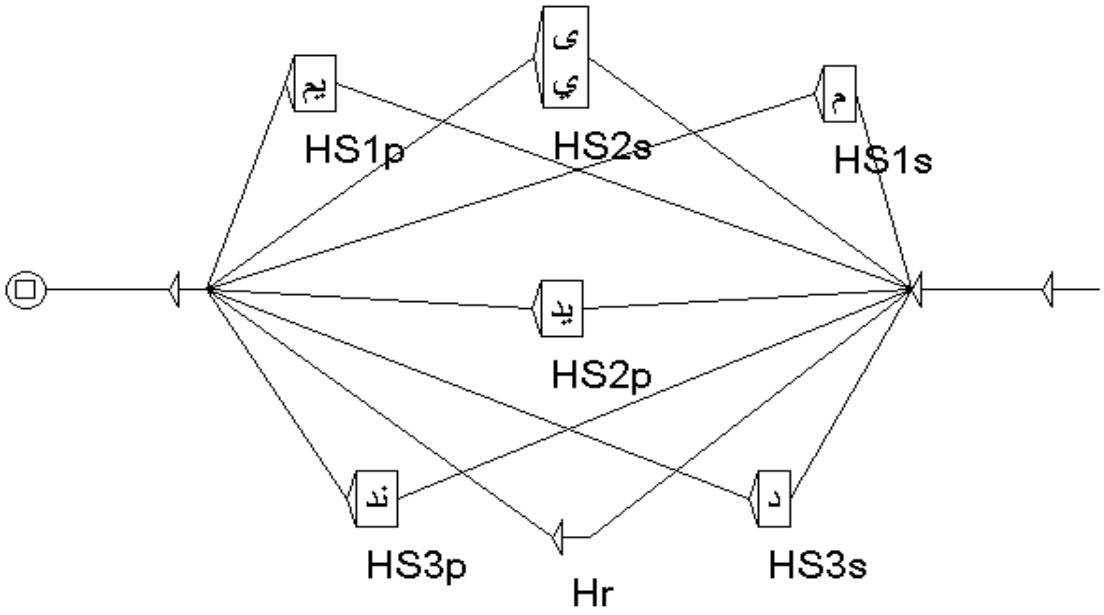


Figure 3. V001: written simple present dictionary graph

This graph describes changes the stem of the word undergoes in order to generate all inflected forms of written simple present. Starting with the box in the upper-right side, we see that the simple present first person singular (HS1s) is produced by adding "م *m"* to the stem, for example: for the verb رفـتـن *raftan "to go"* whose stem of the present is رو *rav*[17], this will yield the form روم *ravam*.

[17] When the present stem of the verb ends up with –av as in رو *rav* "go", it changes to –o as in the present root (Hr) رو *ro*. In the second person singular of imperative form prefix بِـــ be- also becomes بُـــ bo-.

The same is done for the remaining inflections in **Figure 4**; the graph shows the inflection of the written simple past.

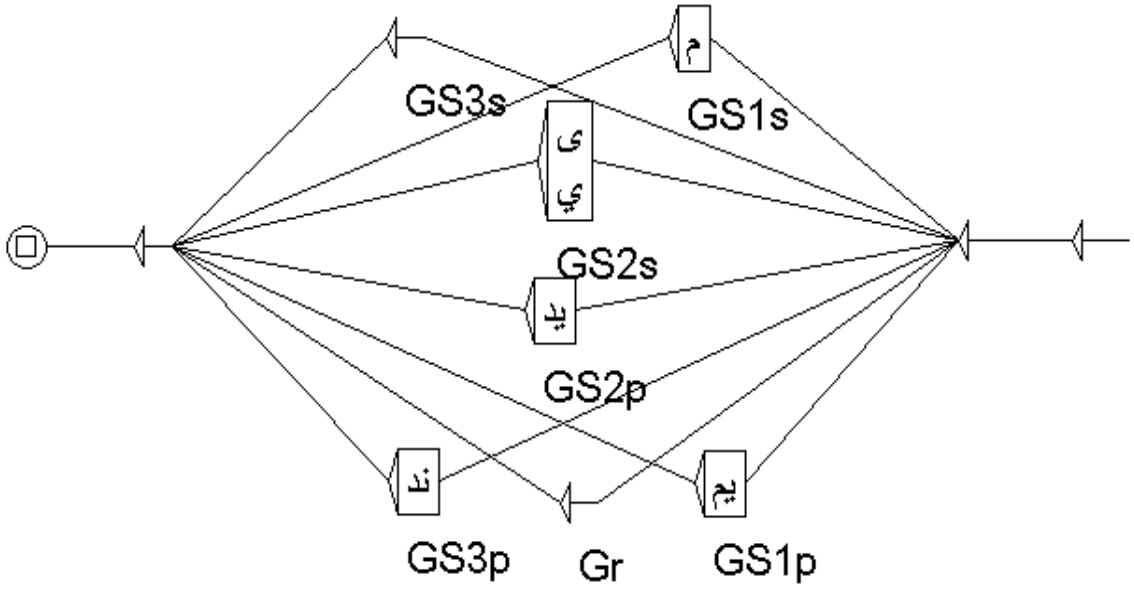


Figure 4. V002: written simple past dictionary graph

In Persian, we have two verbs ست/*ast "to be"* and هست *hast "to be"* that derived from the verb of بـودن *bodan "to be"* which do not follow the same structure as simple present. Because of this, for *ast "be"* and هست *hast "be"*, a specific set of graphs must be built. These are shown in **Figure 5** and **Figure 6**.

In the case of ست/*ast "be"*, only the present simple third singular person (HS3s) and the present root (Hr) exist.

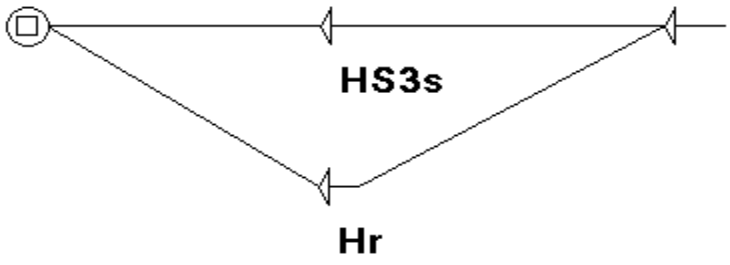


Figure 5. V003: است ast (be) dictionary graph

In the case of هست *hast "be"*, only the simple present (HS) tenses (all forms), and the present root (Hr) exist, the difference between this verb inflections from the other simple present verb inflections is that, this verb followed the simple past inflections rule.

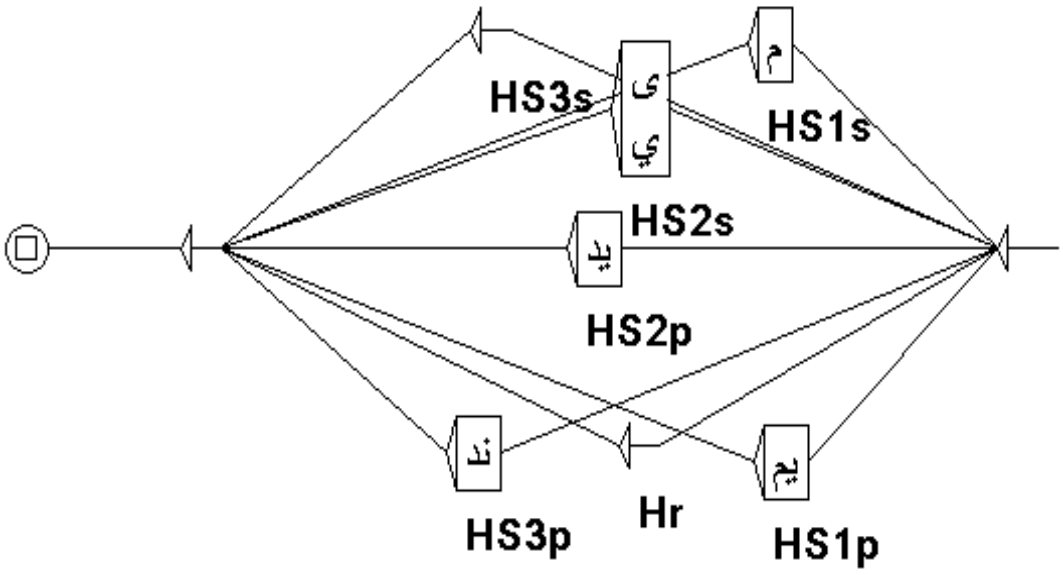


Figure 6. V004: هست hast (be) dictionary graph

2.3.3. Inflection FSTs for Spoken simple words

As noted before, Persian uses different inflections for written (formal) and spoken (informal) forms.

Nevertheless, these forms also appear in written texts. So, they also have to be taken into account. In order to generate the inflected forms associated to each lemma of the spoken inflections, we built five graphs. These graphs are shown in **Figure 7**, **Figure 8**, **Figure 9**, **Figure 10** and **Figure 11**, which explain the differences between these five graphs below. The "spoken" forms are signaled by the code 'm'. The term محاوره ای *mohavereyi "spoken"*, used in Persian grammar to donate the spoken form.

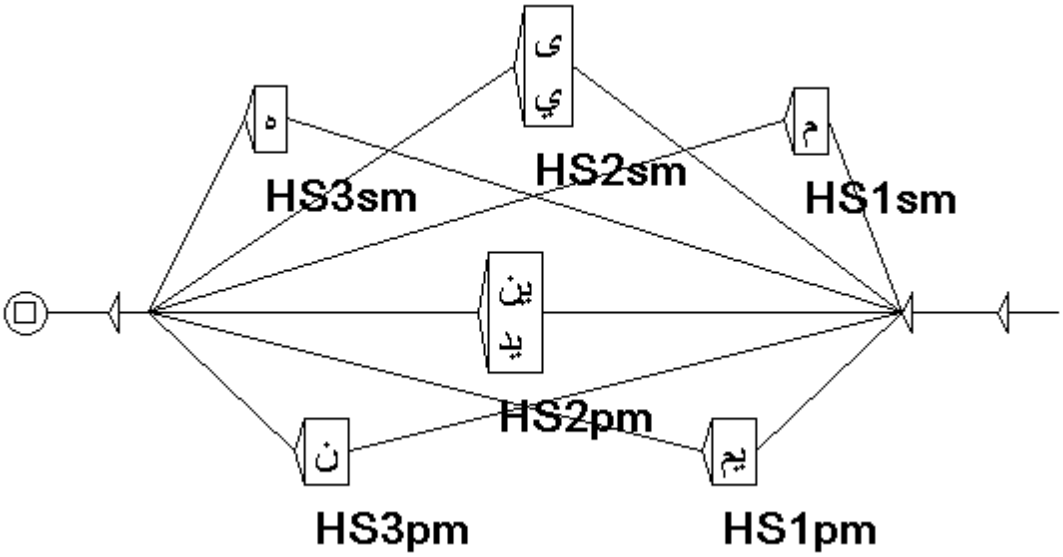


Figure 7. V005: spoken simple present dictionary graph without root

The graph in the **Figure 7** is read in the similar way as the written simple present graph, but only suffixes in plural forms are different and also it does not have a root.

We made another graph for a spoken simple present, which are presented in the graph of **Figure 8**; simply, the difference of this graph from previous one is that this graph has a root form (Hrm).

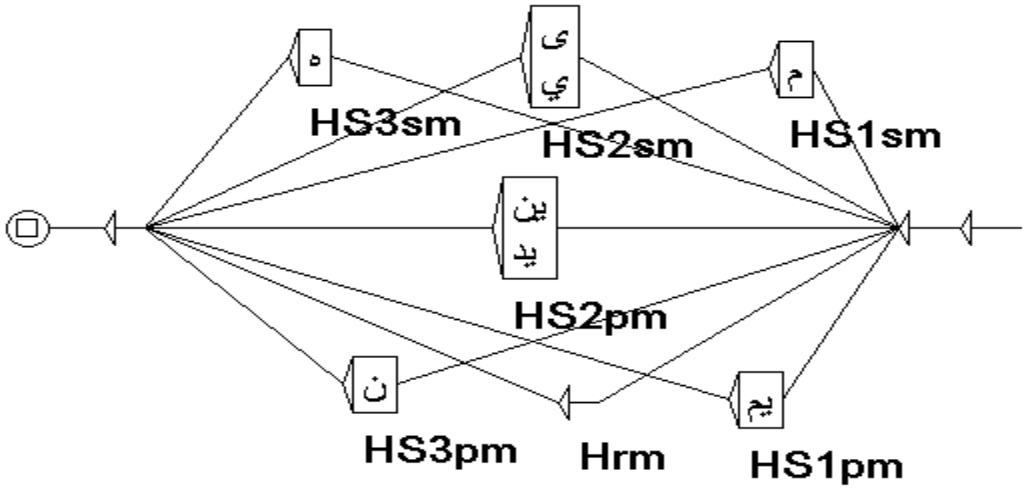


Figure 8. V008: spoken simple present dictionary graph with root

In the other cases, some verbs get د *d* instead of ه *h* as a suffix in the third singular person (HS3sm), for example: the verb خواستن *khastan* "want" with the spoken stem of خوا *kha* "want" in a third singular person from gets د *d* (خوا *kha* + د *d* = *khad* خواد).

This graph is shown in **Figure 9**.

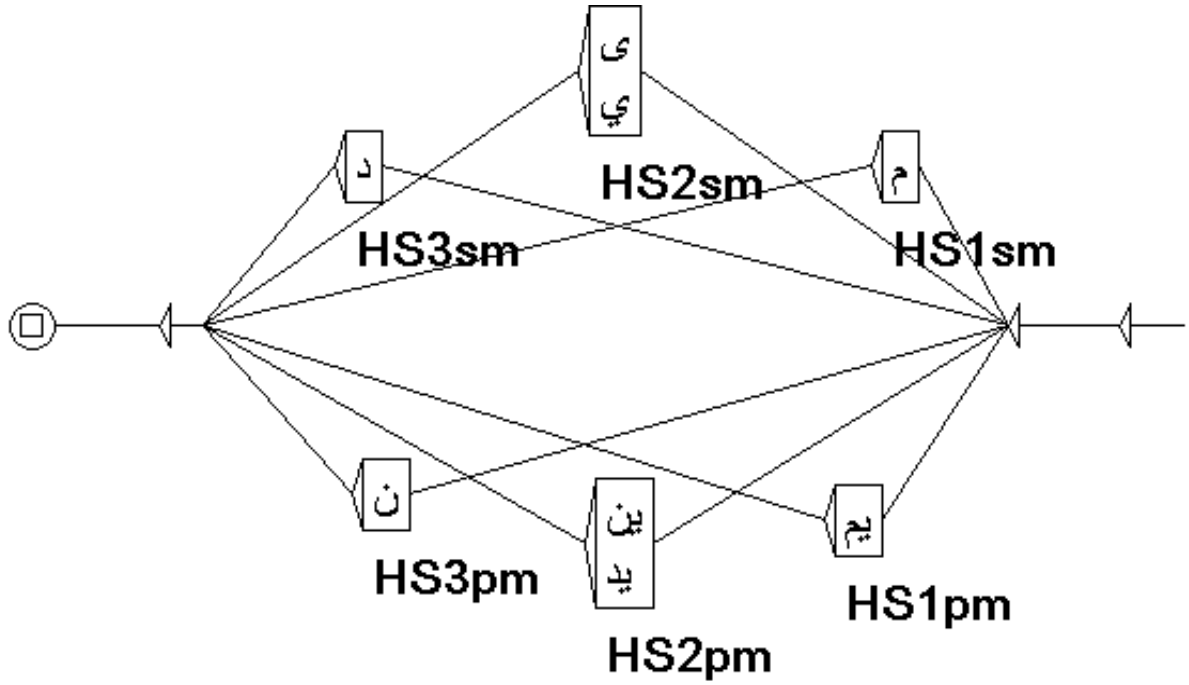


Figure 9. V007: spoken simple present dictionary graph with changes in HS3sm

The graph in the **Figure 10** is read in the similar way as the written simple past graph, but only suffixes in plural forms are changed, this graph made and used only for a spoken (informal) stem that never use these stems in formal writing.

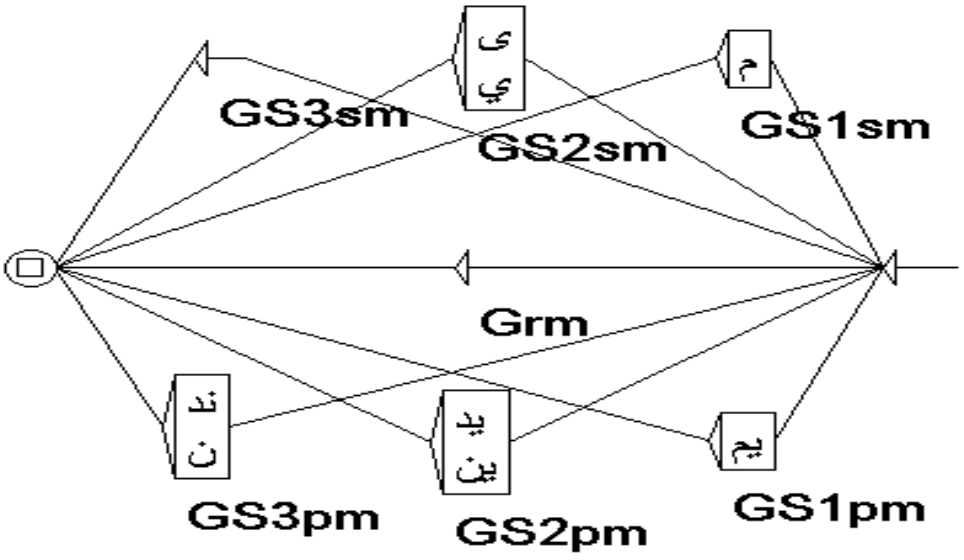


Figure 10. V009: spoken simple past dictionary graph

We added one more graph in **Figure 11** for spoken simple past, sometimes the past written (formal) stem of the verb by getting a spoken (informal) suffixes become an informal verb and this case only happened in second and plural person of simple past. For example: the verb پـختـن *pokhtan "cook"* with stem of پـخت *pokht "cook"* in written (formal) third plural person get ـَنـد *and* (پـخت *pokht* + ـَنـد *and* = پـختـند *pokhtand*) but if we add ـَن *an* (informal suffix) instead of ـَنـد *and*, it has become an informal verb (پـخت *pokht* + ـَن *an* = پـختـن *pokhtan*).

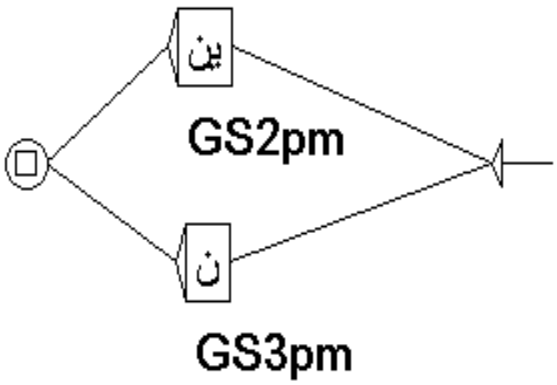


Figure 11. V006: spoken simple past dictionary graph with changes in GS2pm and GS3pm

2.3.4. Compound inflection FSTs

Persian adopts a complex system of compound tenses to express other mode, tense, and aspect variation (See 2.1).

We built several graphs for compound tenses as sub-graphs[18] in one main graph, which intends to capture all correspond inflected forms. This is shown in **Figure 12**.

The sub-graph on the left consist of the compound tenses without negation, while the sub-graphs on the right describe the negative compound tenses. This is indicated in the graphs names by an "n".

For example, **HC** stands for the Present Continuous, while, **HCn** designates the negative Present Continuous.

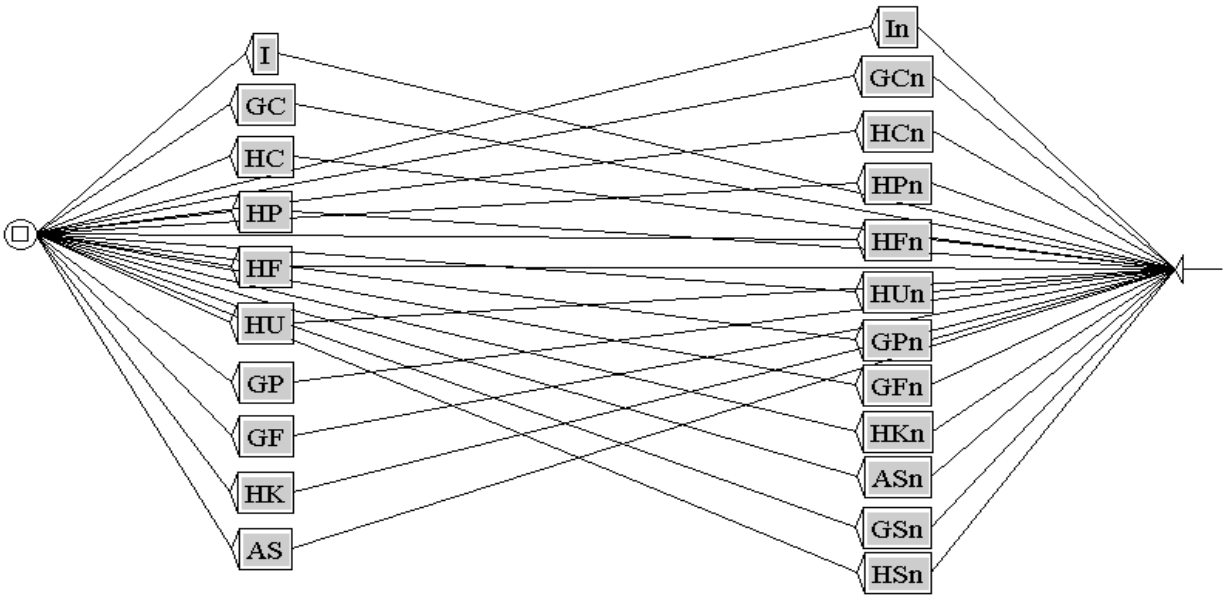


Figure 12. Compound-tenses dictionary graph

To illustrate a sub-graph, consider **Figure 13**.

[18] Each sub-graph is one graph, that we put it in one main graph by using a ":" and the name of the graph, for example: we write ":HC" in one box and Unitex automatically find the graph, you can open a sub-graph by clicking on the grey line while pressing the Alt key.

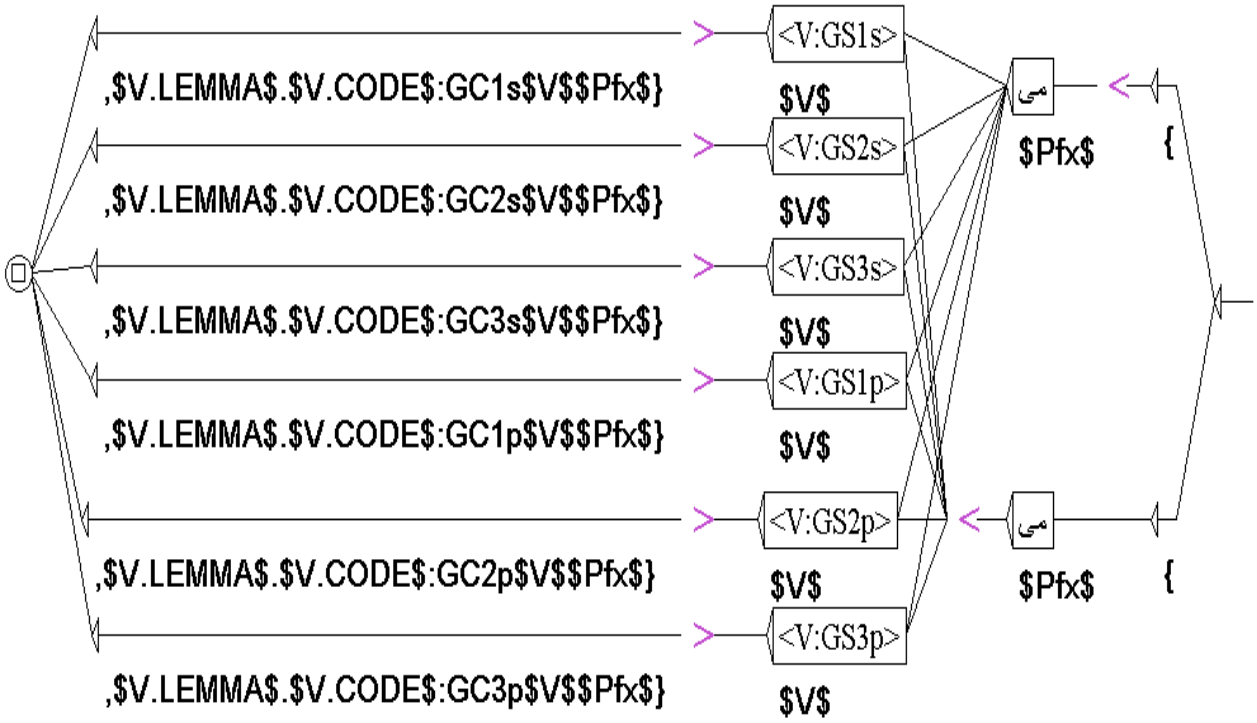


Figure 13. GC sub-graph: Past continuous (GC)

This graph describes the past continuous compound forms and it has two main paths that correspond to the possibility of the prefix می *mi–* to be attached (upper path) or split up (lower path) from the base word. In the first path, the simple form that is to be recognized is marked between > < to indicate that the graph is working in the morphological mode (Paumier 2013 page 67 ff.; 129 ff.). The string می *mi–* is associated to the variable Pfx because it is a prefix, which is then followed by the form that has already been tagged as <V: GS1s>, that is a verb in the past simple first person singular. This form is associated to variable V. In the output, the system produces the compound entry which consist of the content of the variables Pfx and V, the lemma of V, the codes of V and the inflection code for the past continuous first singular person {GC1s}. A similar procedure is carried out for the remaining simple inflected forms. Along

the second path, the only difference is that مـی *mi–* is taken care of as a separate word but the output is the same.

Moreover, we write an output in this order because the graph is applied from right to left, in the correct order, so that the output must be something like: {**$Pfx$$V$, $V. LEMMA$. $V.CODE$:GC1s}**.

As an analyzed form, these compound inflected forms are delimited by the meta-symbols "{" and "}".

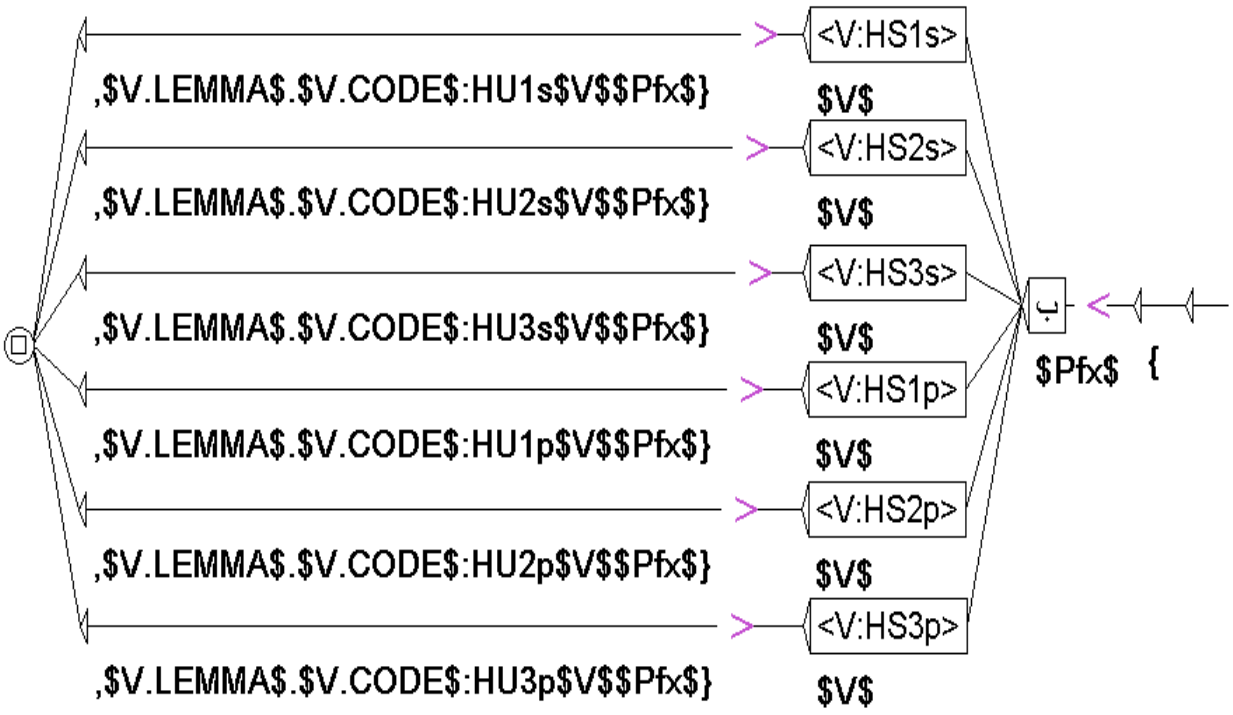


Figure 14. HU sub-graph: Present Subjunctive dictionary graph

Figure 14 is similar to the previous one, only there is just a simple word (one token), regularly derived by adding the prefix بـ *be–* to the simple present forms in order to produce the present subjunctive inflected forms.

2.4. Generating the dictionary of inflected simple forms

In order to construct the dictionary of inflected forms, we first have to build a dictionary of lemmas, which has the following structure:

<stem>,<inflection-code>+L=<LEMMA>

where <stem> represents the form that will be used to generate the inflected forms; and <inflection-codes> is composed of the word part-of-speech (POS) in this case "L" for the lemma, and conventional number for indicating the inflectional FST describing that inflection paradigm.

For example, here are some of the entries of the Persian verbs DELAS (Dictionary of Simple Words Lemmas):

رفت،L+V002=رفتن
رو،L+V001=رفتن
آمد،L+V002=آمدن
آی،L+V001=آمدن
شد،L+V002=شدن
شو،L+V001=شدن

Figure 15. Extract of Persian DELAS

The full lexicon is provided in appendix A.

Table 2 shows the distribution of the verb stems by inflection graphs:

Graph	Total stems	Example (3rd person singular)	Transliteration	Gloss
V001	136	آید	ayad	He/she comes
V002	154	آمد	amad	He/she came
V003	1	است	ast	he/she/it is
V004	1	هست	hast	he/she/it is
V005	7	ده	deh	He/she gives
V006	120	آمده	amadeh	He/she came
V007	3	خواد	khad	He/she wants
V008	19	دونه	doneh	He/she knows
V009	21	اومده	omadeh	He/she came

Table 2. The distribution of the verb stems by inflection graphs

2.5. Generating the inflected forms

By intersecting the DELAS and the inflection FSTs, Unitex is able to generate all the inflected forms associated to each stem of the verbs represented in the dictionary. This constitutes the DELAF (Dictionary of Inflected Forms).

The output is shown in Figure 16:

رفتم،رفت.V+L=رفتن:GS1s
رفتی،رفت.V+L=رفتن:GS2s
رفت،رفت.V+L=رفتن:GS3s
رفتیم،رفت.V+L=رفتن:GS1p
رفتید،رفت.V+L=رفتن:GS2p
رفتند،رفت.V+L=رفتن:GS3p
رفت،رفت.V+L=رفتن:Gr

Figure 16. Extract of Persian DELAF

A sample of this DELAF is provided in Appendix BB.

Each entry in DELAF has the following structure:

<word-form>،<stem>.<POS>+L=<lemma>:<inflection>

Where the <word-form> is the inflected form of the verb, <stem> is the stem represented in the DELAS; <POS> is the grammatical category or part of speech (POS) of each word form and <L> is the value of the lemma that is associated with each form.

So far, the dictionary contains 145 lemmas, 292 written verb stems and 127 spoken verb stems that have been described in the DELAS. These allow for the generation of 1,536 inflected forms.[19]

[19] We evaluated the formal correctness of the DELAF entries by using the BNF (see 3.1, bellow)

2.6. Apply the DELAF to the corpus

The DELAF can now be applied to texts in order to tag the simple verbs. At this stage, 1,536 simple forms of the TEP corpus were tagged by our DELAF, which correspond to 368,831 occurrences (slightly 10% of corpus' simple words).

Figure 17 shows the word lists of the TEP corpus. On the left-upper window, one can see the words tagged by the DELAF.

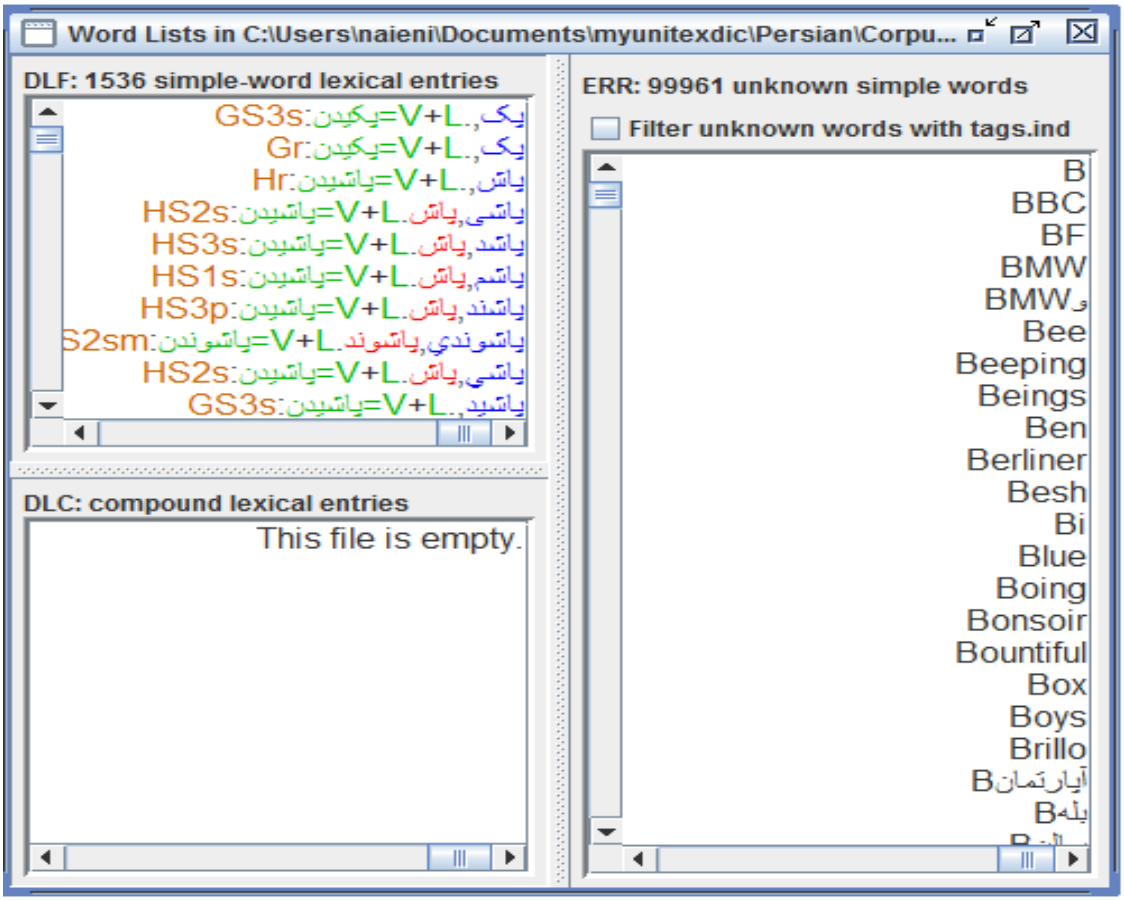


Figure 17. Word List

2.7. Apply the compound tenses FST to the text

When applied compound tenses FST to the corpus, the graph of compound tenses matches 222,698 instances of 3,953 different compound forms. In the merge mode, the output also shows the stem and the lemma of verbs. **Figure 18** illustrates the first entries of the compound tenses: one can see that the first two lines correspond to alternative wordings of present perfect third person singular with and without ست / *ast "is"*.

گذاشته ,گذاشت.V+L=گذاشتن:HP3s	کرده بودید,کرد.V+L=کردن:GP2p	کرده ,کرد.V+L=کردن:HP3s
گذاشته است,گذاشت.V+L=گذاشتن:HP3s	کرده بودیم,کرد.V+L=کردن:GP1p	کرده ای,کرد.V+L=کردن:HP2s
گذاشته بودم,گذاشت.V+L=گذاشتن:GP1s	کرده بودم,کرد.V+L=کردن:GP1s	کرده اید,کرد.V+L=کردن:HP2p
گذاشته بودند,گذاشت.V+L=گذاشتن:GP3p	کرده بودند,کرد.V+L=کردن:GP3p	کرده ایم,کرد.V+L=کردن:HP1p
گذاشته بوده ,گذاشت.V+L=گذاشتن:GF3s	کرده بوده ,کرد.V+L=کردن:GF3s	کرده است,کرد.V+L=کردن:HP3s
گذشته ,گذشت.V+L=گذشتن:HP3s	کرده بوده است,کرد.V+L=کردن:GF3s	کرده ام,کرد.V+L=کردن:HP1s
گذشته ای,گذشت.V+L=گذشتن:HP2s	کنده ,کند.V+L=کندن:HP3s	کرده اند,کرد.V+L=کردن:HP3p
گذشته است,گذشت.V+L=گذشتن:HP3s	کنده است,کند.V+L=کندن:HP3s	کرده باشی,کرد.V+L=کردن:HK2s
گذشته ام,گذشت.V+L=گذشتن:HP1s	گذاشته بود,گذاشت.V+L=گذاشتن:GP3s	کرده باشید,کرد.V+L=کردن:HK2p
گذشته اند,گذشت.V+L=گذشتن:HP3p	گذاشته بودی,گذاشت.V+L=گذاشتن:GP2s	کرده باشیم,کرد.V+L=کردن:HK1p
گذشته باشد,گذشت.V+L=گذشتن:HK3s	گذاشته ام,گذاشت.V+L=گذاشتن:HP1s	کرده باشد,کرد.V+L=کردن:HK3s
گذشته بود,گذشت.V+L=گذشتن:GP3s	گذاشته اند,گذاشت.V+L=گذاشتن:HP3p	کرده باشم,کرد.V+L=کردن:HK1s
گذشته بودم,گذشت.V+L=گذشتن:GP1s	گذاشته باشی,گذاشت.V+L=گذاشتن:HK2s	کرده باشند,کرد.V+L=کردن:HK3p
گذشته بوده ,گذشت.V+L=گذشتن:GF3s	گذاشته باشم,گذاشت.V+L=گذاشتن:HK1s	کرده بود,کرد.V+L=کردن:GP3s
گردونده ,گردوند.V+L=گرداندن:HP3s	گذاشته باشند,گذاشت.V+L=گذاشتن:HK3p	کرده بودی,کرد.V+L=کردن:GP2s

Figure 18. First entries of the compound tenses dictionary

3. Evaluation

Our evaluation was two-fold:

- Evaluate *the formal correction* of the dictionary entries by using the BNFs for both simple and compound forms.
- Evaluate (or rather, estimate) *the lexical coverage* of the dictionaries by applying them to a sample text.

3.1. Evaluation of the formal correction of the dictionary

The inflected simple forms' dictionary contains 1,536 different entries. After applying the simple forms BNF (**Figure 1**, page 8) no errors were found and all entries of the dictionary were matched by the BNF graph.

As for the compound forms retrieved from the corpus, we have 3,953 different entries. After applying the Compound form BNF (**Figure 2**), all entries were also matched, and no errors were found.

3.2. Evaluation of the lexical coverage of the dictionary

A sample text of around 1,000 words was selected from the web. This text is about the world's fastest car that is being built in a workshop on the outskirts of Bristol, England[20]. (the corresponding translation is also available at the same address). The text was POS-tagged by the system.

[20]http://persian.euronews.com/2014/05/01/bloodhound-ssc-the-1000-mph-car-doing-the-school-run, http://www.euronews.com/2014/05/01/bloodhound-ssc-the-1000-mph-car-doing-the-school-run/ Last access: 02/May/2014

3.2.1. Lexical coverage of simple forms' dictionary

The system found 43 simple verb entries from the simple word dictionary (DLF), corresponding to 107 instances in the text; after manual evaluation, no errors were found in the DLF, and 225 forms were unknown, that is, they are not in the dictionary because they correspond to other POS. These constitute 404 instances in the text (approx. 40%).

Naturally, only the verbs were matched, which means that if a word could belong to other POS, that POS could have been tagged as a verb. For example, the word یـا *ya "or"* is ambiguous with an inflected form of the verb آمـدن *amadan "come"* in the simple verb root of spoken form; so naturally, it was matched even though the POS was incorrectly assigned. 21 of 43 simple verb forms were in the same condition. These cases were ignored.

On the other hand, after manual verification, 1 verb from among the unknown words had not been captured by our dictionary. It is the verb بـودن *bodan "to be"* in third person plural of the spoken form that appears two times in the text. Our dictionary did not recognize it because we forgot to add this inflection of بـود *bod "be"* to the corresponding inflection graph, which was afterwards corrected.

To sum up[21], we obtained a Precision of 1, a Recall of 0.98 and F-measure of 0.99.

3.2.2. Lexical coverage of compound forms' dictionary

In the second part of this evaluation, we now deal with compound forms. In this text, we have 31 compound lexical entries and 20 different forms, corresponding to 46 instances in the text (DLC)[22].

[21] Precision (P)= Number of correct verb forms in the text recognized by the dictionary over the total number of verb forms recognized by the dictionary; Recall (R)= number of the correct verb forms in the text recognized by the dictionary over total number of verb forms in the text, and F-measure (F)= 2PR/P+R

[22] For producing a DLC, we apply our compound-tenses graph to the corpus and then we save the resulting concordance as a DLC after some post edition. This includes removing the morphological tags produced for the

In the end, for the existing verbs matched by the systems, no errors were found, which includes the adequate stem, the lemma and the inflection.

As for the evaluation of the compound verbs, we noticed the unknown form *nist* نیست *"not to be"*, and this form appears only one time and correspond to the contraction of the negation morpheme نـــ *ne- "not"* and the third person singular of simple present tense of هست *hast "to be"*.

Since this form was missing from the graph for the verb هست *hast "to be"*, it was therefore corrected.

To sum up, for the compound forms' dictionary, we obtained a Precision of 1, Recall of 0.96 and F-measure of 0.98.

Obviously, we are well aware of the size limitation of the sample text, and do not extrapolate these results any further.

In conclusion, though the dictionary is obviously incomplete, since there are many verbs that must still be included, this small experiment provided some evidence that its lexical coverage may be already quite satisfactory. Naturally, in future work, we will endeavor to extend the list of verbs encoded in the DELAS.

simple verb that is part of the compound form: {.کرد ,کرده اندV+L=کردن: GS3s:HP3p} > .کرد ,کرده اندV+L=کردن: HP3p (the verb کرد *kard "do"* in the compound present perfect tense کرده اند *karde and "have been done"*).

4. Conclusion

We have achieved the main objectives of this project. We built a sufficiently large lexicon of Persian verbs, until we attained a frequency threshold corresponding to a cumulative percentage of 90% of the total corpus word forms (3,349,920) corresponding to 1,466 different verb forms. We, therefore, expect to have achieved a reasonable lexical coverage.

We built 9 inflectional graphs to generate automatically all simple forms of Persian verbs; 4 graphs for the written forms and 5 graphs for "spoken" forms, in order to produce a dictionary of simple verb inflected forms, aiming at a highly granular and systematic description of all the relevant grammatical values pertaining to those forms.

Furthermore, we built 22 inflectional graphs to produce automatically all compound forms of Persian verbs, in order to produce a dictionary of compound verb inflected forms.

We produced a Persian verbs dictionary of lemmas (145), where each stem (399) is associated with its corresponding inflectional paradigms; each lemma is associated with 3 different stems (present, past and spoken stem); the distinction between lemmas and stems is overcome by the lemma being treated as a feature of the verb, so one can use it to capture all inflected forms associated with the same lemma with lexical mask in much the same way as one would do for other languages without the distinction between stems and lemmas.

In order to evaluate the correction of the both dictionaries, we applied the simple form BNF and compound form BNF to the main corpus, and no errors were found in both of them.

In order to evaluate the dictionary lexical coverage, we have applied the dictionaries to a sample of +1000 words text and, for the simple, form we find one verb that appears two times in the text

that was not captured by the dictionary due to forgetting to add this inflection to the DELAS and in the compound form evaluation, after manual verification, we find also one verb that appear one time in the text that was not taken because of the incomplete graph for the verb هست *hast "to be".*

Naturally, much is still left to be done: First, we want to complete the dictionary by adding many more verbs and, eventually, adding new inflectional paradigms and their corresponding inflection graphs.

Secondly, in order to continue to produce a full lexicon for Persian, attention must be given to another part of speech: this will be the challenge for future work.

References

Instructional Technology Services group. (2007). Persian online-grammar & resources. Retrieved March 10, 2013, from http://sites.la.utexas.edu/persian_online_resources/verbs/

Paumier. S. (2003). *De la reconnaissance de formes linguistiques a l'analyse syntaxique*. Univ. Marne-la-Vallée.

Paumier. S. (2013). *Unitex manual*. Univ. Marne-la-Vallée. Retrieved from http://www-igm.univ-mlv.fr/~unitex/

Persian in Texas group. (n.d.). Persian Simple Verbs. Retrieved December 13, 2013, from http://persian.nmelrc.org/pvc/simpleverbs.php

Persian verbs. (2014). Retrieved June 10, 2013, from http://en.wikipedia.org/wiki/Persian_verbs

Pilevar, M.T; H. Faili, and A. H. P. (n. d. . (n.d.). Natural Language Processing (NLP) Research lab, University of Tehran. Retrieved May 10, 2013, from http://ece.ut.ac.ir/NLP/resources.htm

Tabatabayi, A. (2010). *ساختمان واژه و مقوله دستوری : تشخیص مقوله دستوری واژه‌ها، بر اساس ملاک‌های صرفی,sakhtemane vaje va magholeye dastori: tashkhis magholeye dastoriye vajeha,bar asase melakhaye sarfi, vocabulary structure and grammatical categories: recognition of words grammmar* Tahran: Farhang, honar va ertebatat. Retrieved from http://dlib.ical.ir/site/catalogue/789489

فعل ساده, feele sade, Simple verb. (2014). Retrieved November 07, 2013, from http://fa.wikipedia.org/wiki/فعل_ساده

Appendices

A. DELAS (dictionary of Persian verb stems and associated lemmas)

ماند,L+V002=ماندن
ماند,L+V006=ماندن
مرد,L+V002=مردن
مرد,L+V006=مردن
مون,L+V008=ماندن
موند,L+V002=ماندن
موند,L+V006=ماندن
میر,L+V001=مردن
نگر,L+V001=نگریستن
نگریست,L+V002=نگریستن
نگریست,L+V006=نگریستن
نال,L+V001=نالیدن
نالید,L+V002=نالیدن
نالید,L+V006=نالیدن
نام,L+V001=نامیدن
نامید,L+V002=نامیدن
نامید,L+V006=نامیدن
نداز,L+V005=انداختن
نشان,L+V001=نشاندن
نشاند,L+V002=نشاندن
نشست,L+V002=نشستن
نشست,L+V006=نشستن
نشون,L+V008=نشاندن

خواب,L+V001=خوابیدن
خوابید,L+V002=خوابیدن
خوابید,L+V006=خوابیدن
خواست,L+V002=خواستن
خواست,L+V006=خواستن
خوان,L+V001=خواندن
خواند,L+V002=خواندن
خواند,L+V006=خواندن
خواه,L+V001=خواستن
خور,L+V001=خوردن
خور,L+V008=خوردن
خوران,L+V001=خوراندن
خوراند,L+V002=خوراندن
خورد,L+V002=خوردن
خورد,L+V006=خوردن
خورون,L+V008=خوراندن
خوروند,L+V009=خوراندن
د,L+V005=دادن
داد,L+V002=دادن
داد,L+V006=دادن
دار,L+V001=داشتن
داشت,L+V002=داشتن
داشت,L+V006=داشتن

گرفت,L+V002=گرفتن
گرفت,L+V006=گرفتن
گریخت,L+V002=گریختن
گریخت,L+V006=گریختن
گریز,L+V001=گریختن
گزار,L+V001=گزاردن
گزارد,L+V002=گزاردن
گزارد,L+V006=گزاردن
گشت,L+V002=گشتن
گشت,L+V006=گشتن
گفت,L+V002=گفتن
گفت,L+V006=گفتن
گند,L+V001=گندیدن
گندید,L+V002=گندیدن
گوی,L+V001=گفتن
گوز,L+V001=گوزیدن
گوزید,L+V002=گوزیدن
گوزید,L+V006=گوزیدن
گیر,L+V001=گرفتن
یا,L+V001=آمدن
یار,L+V008=آوردن
آ,L+V007=آمدن
آی,L+V001=آمدن

پک,L+V002=پکیدن
پکوند,L+V009=پکیدن
پکید,L+V002=پکیدن
پکید,L+V006=پکیدن
پایید,L+V002=پاییدن
پایید,L+V006=پاییدن
پاش,L+V001=پاشیدن
پاشوند,L+V009=پاشوندن
پاشید,L+V002=پاشیدن
پاشید,L+V006=پاشیدن
پخت,L+V002=پختن
پخت,L+V006=پختن
پذیر,L+V001=پذیرفتن
پذیرفت,L+V002=پذیرفتن
پذیرفت,L+V006=پذیرفتن
پر,L+V001=پریدن
پراکن,L+V001=پراکندن
پراکند,L+V002=پراکندن
پراکند,L+V006=پراکندن
پراند,L+V002=پراندن
پرداخت,L+V002=پرداختن
پرداخت,L+V006=پرداختن
پرداز,L+V001=پرداختن

نشوند,L+V009=نشاندن
نشین,L+V001=نشستن
نواخت,L+V002=نواختن
نواخت,L+V006=نواختن
نواز,L+V001=نواختن
نوش,L+V001=نوشیدن
نوشید,L+V002=نوشیدن
نوشید,L+V006=نوشیدن
نوشت,L+V002=نوشتن
نوشت,L+V006=نوشتن
نویس,L+V001=نوشتن
هست,L+V004=بودن
وز,L+V001=وزیدن
وزید,L+V002=وزیدن
یاب,L+V001=یافتن
یافت,L+V002=یافتن
یافت,L+V006=یافتن
فهم,L+v001=فهمیدن
فهمید,L+V002=فهمیدن
فهمید,L+V006=فهمیدن
کش,L+V001=کشتن
کش,L+V001=کشیدن
کشت,L+V002=کشتن
کشت,L+V006=کشتن
کشید,L+V002=کشیدن
کشید,L+V006=کشیدن

دان,L+V001=دانستن
دانست,L+V002=دانستن
دانست,L+V006=دانستن
در,L+V001=دریدن
درید,L+V002=دریدن
درید,L+V006=دریدن
دزد,L+V001=دزدیدن
دزدید,L+V002=دزدیدن
دزدید,L+V006=دزدیدن
ده,L+V001=دادن
دو,L+V001=دویدن
دوخت,L+V002=دوختن
دوخت,L+V006=دوختن
دوز,L+V001=دوختن
دوش,L+V001=دوشیدن
دوشید,L+V002=دوشیدن
ساز,L+V001=ساختن
سوخت,L+V002=سوختن
سوخت,L+V006=سوختن
سوز,L+V001=سوختن
سوزان,L+V001=سوزاندن
سوزاند,L+V002=سوزاندن
سوزون,L+V008=سوزاندن
سوزوند,L+V009=سوزاندن
ش,L+V005=شدن
شاش,L+V001=شاشیدن

آشامید,L+V002=آشامیدن
آفرید,L+V002=آفریدن
آفرید,L+V006=آفریدن
آفرین,L+V001=آفریدن
آمد,L+V002=آمدن
آمد,L+V006=آمدن
آموخت,L+V002=آموختن
آموخت,L+V006=آموختن
آموز,L+V001=آموختن
آمیخت,L+V002=آمیختن
آمیخت,L+V006=آمیختن
آمیز,L+V001=آمیختن
آور,L+V001=آوردن
آورد,L+V002=آوردن
آورد,L+V006=آوردن
آویخت,L+V002=آویختن
بخش,L+V002=بخشیدن
بخشید,L+V002=بخشیدن
بخشید,L+V006=بخشیدن
بر,L+V001=بردن
بر,L+V001=بریدن
برد,L+V002=بردن
برد,L+V006=بردن
برید,L+V002=بریدن
برید,L+V006=بریدن
بست,L+V002=بستن

پرس,L+V001=پرسیدن
پرسید,L+V002=پرسیدن
پرسید,L+V006=پرسیدن
پرست,L+V001=پرستیدن
پرستید,L+V002=پرستیدن
پرستید,L+V006=پرستیدن
پرور,L+V001=پروراندن
پروراند,L+V002=پروراندن
پروروند,L+V009=پروراندن
پروند,L+V009=پراندن
پرید,L+V002=پریدن
پرید,L+V006=پریدن
پز,L+V001=پختن
پسند,L+V001=پسندیدن
پسندید,L+V002=پسندیدن
پسندید,L+V006=پسندیدن
پندار,L+V001=پنداشتن
پنداشت,L+V002=پنداشتن
پوک,L+V008=پکیدن
پوس,L+V001=پوسیدن
پوسید,L+V002=پوسیدن
پوسید,L+V006=پوسیدن
پوش,L+V001=پوشیدن
پوشاند,L+V002=پوشاندن
پوشید,L+V002=پوشیدن
پوشید,L+V006=پوشیدن

کوب,L+V001=کوبیدن
کوبید,L+V002=کوبیدن
کوبید,L+V006=کوبیدن
کوش,L+V001=کوشیدن
کوشید,L+V002=کوشیدن
کوشید,L+V006=کوشیدن
لیس,L+V001=لیسیدن
لیسید,L+V002=لیسیدن
لیسید,L+V006=لیسیدن
مال,L+V001=مالیدن
مالید,L+V002=مالیدن
مالان,L+V001=مالاندن
مالاند,L+V002=مالاندن
مالون,L+V008=مالاندن
مالوند,L+V009=مالاندن
مان,L+V001=ماندن
جوشید,L+V002=جوشیدن
جوشید,L+V006=جوشیدن
جوي,L+V001=جویدن
جوید,L+V002=جویدن
جوید,L+V006=جویدن
خار,L+V001=خاریدن
خاران,L+V001=خاراندن
خاراند,L+V002=خاراندن
خارون,L+V008=خاراندن

شاشید,L+V002=شاشیدن
شتاب,L+V001=شتافتن
شتافت,L+V002=شتافتن
شتافت,L+V006=شتافتن
شد,L+V002=شدن
شد,L+V006=شدن
شست,L+V002=شستن
شست,L+V006=شستن
شکان,L+V001=شکاندن
شکاند,L+V002=شکاندن
شکست,L+V002=شکستن
شکست,L+V006=شکستن
شکن,L+V001=شکستن
شکون,L+V008=شکاندن
شکوند,L+V009=شکاندن
شمار,L+V001=شمردن
شمرد,L+V002=شمردن
شمرد,L+V006=شمردن
شناخت,L+V002=شناختن
شناخت,L+V006=شناختن
شناس,L+V001=شناختن
شناسان,L+V001=شناساندن
شناساند,L+V002=شناساندن
شناسون,L+V008=شناساندن
شناسوند,L+V009=شناساندن

بست,L+V006=بستن
بلع,L+V001=بلعیدن
بلعید,L+V002=بلعیدن
بلعید,L+V006=بلعیدن
بند,L+V001=بستن
بود،L+V002=بودن
بود،L+V006=بودن
بوس,L+V001=بوسیدن
بوسید,L+V002=بوسیدن
بوسید,L+V006=بوسیدن
بوي,L+V001=بوییدن
بویید,L+V002=بوییدن
بویید,L+V006=بوییدن
بین,L+V001=دیدن
تپ,L+V001=تپیدن
تپید,L+V002=تپیدن
تپید,L+V006=تپیدن
تراش,L+V001=تراشیدن
تراشید,L+V002=تراشیدن
تراشید,L+V006=تراشیدن
ترس,L+V001=ترسیدن
ترسید,L+V002=ترسیدن
ترسید,L+V006=ترسیدن
تکان,L+V001=تکاندن
تکاند,L+V002=تکاندن

پیچ,L+V001=پیچیدن
پیچاند,L+V002=پیچیدن
پیچوند,L+V009=پیچیدن
پیچید,L+V002=پیچیدن
پیچید,L+V006=پیچیدن
پیمای,L+V001=پیمودن
پیمود,L+V002=پیمودن
پیمود,L+V006=پیمودن
پیوست,L+V002=پیوستن
پیوست,L+V006=پیوستن
پیوند,L+V001=پیوستن
چپوند,L+V009=چپوندن
چین,L+V001=چیدن
چاپ,L+V001=چاپیدن
چاپید,L+V002=چاپیدن
چاپید,L+V006=چاپیدن
چراند,L+V002=چراندن
چرخ,L+V001=چرخیدن
چرخاند,L+V002=چرخاندن
چرخوند,L+V009=چرخاندن
چرخید,L+V002=چرخیدن
چرخید,L+V006=چرخیدن
چروند,L+V009=چراندن
چس,L+V001=چسیدن
چسب,L+V001=چسبیدن

خاروند,L+V009=خاراندن
خارید,L+V002=خاریدن
خارید,L+V006=خاریدن
خر,L+V001=خریدن
خراشید,L+V002=خراشیدن
خراشوند,L+V009=خراشاندن
خرید,L+V002=خریدن
خرید,L+V006=خریدن
خشك,L+V001=خشکیدن
خشکید,L+V002=خشکیدن
خشکید,L+V006=خشکیدن
خفت,L+V002=خفتن
خند,L+V001=خندیدن
خندید,L+V002=خندیدن
خندید,L+V006=خندیدن
خوا,L+V007=خواستن
گری,L+V001=گریستن
گریست,L+V002=گریستن
گرد,L+V001=گشتن
گردان,L+V001=گرداندن
گرداند,L+V002=گرداندن
گردون,L+V008=گرداندن
گردوند,L+V009=گرداندن

شنو,L+V001=شنیدن
شنید,L+V002=شنیدن
شنید,L+V006=شنیدن
شو,L+V001=شدن
شوي,L+V001=شستن
شین,L+V008=نشستن
صرف,L+V001=صرفیدن
صرفید,L+V002=صرفیدن
طلب,L+V001=طلبیدن
طلبید,L+V002=طلبیدن
فرست,L+V001=فرستادن
فرستاد,L+V002=فرستادن
فرستاد,L+V006=فرستادن
فروخت,L+V002=فروختن
فروخت,L+V006=فروختن
فروش,L+V001=فروختن
گذار,L+V001=گذاشتن
گذاشت,L+V002=گذاشتن
گذاشت,L+V006=گذاشتن
گذر,L+V001=گذشتن
گذشت,L+V002=گذشتن
گذشت,L+V006=گذشتن

تکاند,L+V006=تکاندن
توان,L+V001=توانستن
توانست,L+V002=توانستن
توانست,L+V006=توانستن
تون,L+V008=توانستن
تونست,L+V002=توانستن
تونست,L+V006=توانستن
جست,L+V002=جستن
جست,L+V006=جستن
جنگ,L+V001=جنگیدن
جنگید,L+V002=جنگیدن
جنگید,L+V006=جنگیدن
جه,L+V001=جهیدن
جهید,L+V002=جهیدن
جهید,L+V006=جهیدن
جوش,L+V001=جوشیدن
کرد,L+V002=کردن
کرد,L+V006=کردن
کن,L+V001=کردن
کند,L+V002=کندن
کند,L+V006=کندن
گ,L+V005=گفتن

چسباند,L+V002=چسباندن
چسبوند,L+V009=چسباندن
چسبید,L+V002=چسبیدن
چسبید,L+V006=چسبیدن
چسید,L+V002=چسیدن
چسید,L+V006=چسیدن
چش,L+V001=چشیدن
چشای,L+V001=چشاندن
چشید,L+V002=چشیدن
چشید,L+V006=چشیدن
چك,L+V001=چکیدن
چکاند,L+V002=چکاندن
چکوند,L+V009=چکاندن
چکید,L+V002=چکیدن
چکید,L+V006=چکیدن
چوس,L+V008=چسیدن
چوسید,L+V009=چسیدن
چید,L+V002=چیدن
چید,L+V006=چیدن
کار,L+V001=کاشتن
کاشت,L+V002=کاشتن
کاشت,L+V006=کاشتن

B. Sample of DELAF (dictionary of inflected simple forms)

پاشیدم,پاشید.L+V=پاشیدن:GS1s
پاشیدی,پاشید.L+V=پاشیدن:GS2s
پاشیدي,پاشید.L+V=پاشیدن:GS2s
پاشید,پاشید.L+V=پاشیدن:GS3s
پاشیدیم,پاشید.L+V=پاشیدن:GS1p
پاشیدید,پاشید.L+V=پاشیدن:GS2p
پاشیدند,پاشید.L+V=پاشیدن:GS3p
پاشید,پاشید.L+V=پاشیدن:Gr
پاشیدین,پاشید.L+V=پاشیدن:GS2pm
پاشیدن,پاشید.L+V=پاشیدن:GS3pm
پختم,پخت.L+V=پختن:GS1s
پختی,پخت.L+V=پختن:GS2s
پختي,پخت.L+V=پختن:GS2s
پخت,پخت.L+V=پختن:GS3s
پختیم,پخت.L+V=پختن:GS1p
پختید,پخت.L+V=پختن:GS2p
پختند,پخت.L+V=پختن:GS3p
پخت,پخت.L+V=پختن:Gr
پختین,پخت.L+V=پختن:GS2pm
پختن,پخت.L+V=پختن:GS3pm
پذیرم,پذیر.L+V=پذیرفتن:HS1s
پذیری,پذیر.L+V=پذیرفتن:HS2s
پذیري,پذیر.L+V=پذیرفتن:HS2s
پذیرد,پذیر.L+V=پذیرفتن:HS3s
پذیریم,پذیر.L+V=پذیرفتن:HS1p
پذیرید,پذیر.L+V=پذیرفتن:HS2p
پذیرند,پذیر.L+V=پذیرفتن:HS3p
پذیر,پذیر.L+V=پذیرفتن:Hr

پاییدم,پایید.L+V=پاییدن:GS1s
پاییدی,پایید.L+V=پاییدن:GS2s
پاییدي,پایید.L+V=پاییدن:GS2s
پایید,پایید.L+V=پاییدن:GS3s
پاییدیم,پایید.L+V=پاییدن:GS1p
پاییدید,پایید.L+V=پاییدن:GS2p
پاییدند,پایید.L+V=پاییدن:GS3p
پایید,پایید.L+V=پاییدن:Gr
پاییدین,پایید.L+V=پاییدن:GS2pm
پاییدن,پایید.L+V=پاییدن:GS3pm
پاشم,پاش.L+V=پاشیدن:HS1s
پاشی,پاش.L+V=پاشیدن:HS2s
پاشي,پاش.L+V=پاشیدن:HS2s
پاشد,پاش.L+V=پاشیدن:HS3s
پاشیم,پاش.L+V=پاشیدن:HS1p
پاشید,پاش.L+V=پاشیدن:HS2p
پاشند,پاش.L+V=پاشیدن:HS3p
پاش,پاش.L+V=پاشیدن:Hr
پاشوندم,پاشوند.L+V=پاشوندن:GS1sm
پاشوندی,پاشوند.L+V=پاشوندن:GS2sm
پاشوندي,پاشوند.L+V=پاشوندن:GS2sm
پاشوند,پاشوند.L+V=پاشوندن:GS3sm
پاشوندیم,پاشوند.L+V=پاشوندن:GS1pm
پاشوندید,پاشوند.L+V=پاشوندن:GS2pm
پاشوندین,پاشوند.L+V=پاشوندن:GS2pm
پاشوندند,پاشوند.L+V=پاشوندن:GS3pm
پاشوندن,پاشوند.L+V=پاشوندن:GS3pm
پاشوند,پاشوند.L+V=پاشوندن:Grm

پکم,پک.L+V=پکیدن:GS1s
پکی,پک.L+V=پکیدن:GS2s
پکي,پک.L+V=پکیدن:GS2s
پک,پک.L+V=پکیدن:GS3s
پکیم,پک.L+V=پکیدن:GS1p
پکید,پک.L+V=پکیدن:GS2p
پکند,پک.L+V=پکیدن:GS3p
پک,پک.L+V=پکیدن:Gr
پکوندم,پکوند.L+V=پکیدن:GS1sm
پکوندی,پکوند.L+V=پکیدن:GS2sm
پکوندي,پکوند.L+V=پکیدن:GS2sm
پکوند,پکوند.L+V=پکیدن:GS3sm
پکوندیم,پکوند.L+V=پکیدن:GS1pm
پکوندید,پکوند.L+V=پکیدن:GS2pm
پکوندین,پکوند.L+V=پکیدن:GS2pm
پکوندند,پکوند.L+V=پکیدن:GS3pm
پکوندن,پکوند.L+V=پکیدن:GS3pm
پکوند,پکوند.L+V=پکیدن:Grm
پکیدم,پکید.L+V=پکیدن:GS1s
پکیدی,پکید.L+V=پکیدن:GS2s
پکیدي,پکید.L+V=پکیدن:GS2s
پکید,پکید.L+V=پکیدن:GS3s
پکیدیم,پکید.L+V=پکیدن:GS1p
پکیدید,پکید.L+V=پکیدن:GS2p
پکیدند,پکید.L+V=پکیدن:GS3p
پکید,پکید.L+V=پکیدن:Gr
پکیدین,پکید.L+V=پکیدن:GS2pm
پکیدن,پکید.L+V=پکیدن:GS3pm

C. Sample of Compound forms (found in a text)

پیوسته اند,پیوست.L+V=پیوستن:HP3p

چپونده ,چپوند.L+V=چپوندن:HP3s

چپونده بودم,چپوند.L+V=چپوندن:GP1s

چاپیده ,چاپید.L+V=چاپیدن:HP3s

چرخانده ,چرخاند.L+V=چرخاندن:HP3s

چرخانده است,چرخاند.L+V=چرخاندن:HP3s

چرخیده ,چرخید.L+V=چرخیدن:HP3s

چرخیده اند,چرخید.L+V=چرخیدن:HP3p

چسباند ,چسباند.L+V=چسباندن:HP3s

چسبانده بود,چسباند.L+V=چسباندن:GP3s

چسبونده ,چسبوند.L+V=چسباندن:HP3s

چسبونده بود,چسبوند.L+V=چسباندن:GP3s

چسبیده ,چسبید.L+V=چسبیدن:HP3s

چسبیده اند,چسبید.L+V=چسبیدن:HP3p

چسبیده بود,چسبید.L+V=چسبیدن:GP3s

چسبیده بودم,چسبید.L+V=چسبیدن:GP1s

چشیده ,چشید.L+V=چشیدن:HP3s

چکیده ,چکید.L+V=چکیدن:HP3s

چیده ,چید.L+V=چیدن:HP3s

چیده باشند,چید.L+V=چیدن:HK3p

کاشته ,کاشت.L+V=کاشتن:HP3s

کاشته بود,کاشت.L+V=کاشتن:GP3s

کرده ,کرد.L+V=کردن:HP3s

کرده ای,کرد.L+V=کردن:HP2s

کرده اید,کرد.L+V=کردن:HP2p

پوسیده است,پوسید.L+V=پوسیدن:HP3s

پوسیده اند,پوسید.L+V=پوسیدن:HP3p

پوشانده ,پوشاند.L+V=پوشاندن:HP3s

پوشانده است,پوشاند.L+V=پوشاندن:HP3s

پوشانده اند,پوشاند.L+V=پوشاندن:HP3p

پوشانده بود,پوشاند.L+V=پوشاندن:GP3s

پوشانده بودم,پوشاند.L+V=پوشاندن:GP1s

پوشانده بوده ,پوشاند.L+V=پوشاندن:GF3s

پوشیده ,پوشید.L+V=پوشیدن:HP3s

پوشیده ام,پوشید.L+V=پوشیدن:HP1s

پوشیده اند,پوشید.L+V=پوشیدن:HP3p

پوشیده باشم,پوشید.L+V=پوشیدن:HK1s

پوشیده بود,پوشید.L+V=پوشیدن:GP3s

پوشیده بودم,پوشید.L+V=پوشیدن:GP1s

پوشیده بودند,پوشید.L+V=پوشیدن:GP3p

پوشیده بوده ,پوشید.L+V=پوشیدن:GF3s

پیچونده ,پیچوند.L+V=پیچیدن:HP3s

پیچونده باشم,پیچوند.L+V=پیچیدن:HK1s

پیچیده ,پیچید.L+V=پیچیدن:HP3s

پیچیده است,پیچید.L+V=پیچیدن:HP3s

پیچیده بود,پیچید.L+V=پیچیدن:GP3s

پیموده ,پیمود.L+V=پیمودن:HP3s

پیوسته ,پیوست.L+V=پیوستن:HP3s

پیوسته است,پیوست.L+V=پیوستن:HP3s

پیوسته ام,پیوست.L+V=پیوستن:HP1s

پخته ,پخت.L+V=پختن:HP3s

پخته اند,پخت.L+V=پختن:HP3p

پخته بود,پخت.L+V=پختن:GP3s

پذیرفته ,پذیرفت.L+V=پذیرفتن:HP3s

پذیرفته ام,پذیرفت.L+V=پذیرفتن:HP1s

پذیرفته اند,پذیرفت.L+V=پذیرفتن:HP3p

پذیرفته بودند,پذیرفت.L+V=پذیرفتن:GP3p

پراکنده ,پراکند.L+V=پراکندن:HP3s

پرداخته ,پرداخت.L+V=پرداختن:HP3s

پرسیده ,پرسید.L+V=پرسیدن:HP3s

پرسیده ایم,پرسید.L+V=پرسیدن:HP1p

پرستیده ,پرستید.L+V=پرستیدن:HP3s

پرورانده ,پروراند.L+V=پروراندن:HP3s

پرورانده بودند,پروراند.L+V=پروراندن:GP3p

پرونده ,پروند.L+V=پراندن:HP3s

پرونده است,پروند.L+V=پراندن:HP3s

پرونده ام,پروند.L+V=پراندن:HP1s

پرونده بود,پروند.L+V=پراندن:GP3s

پرونده بوده ,پروند.L+V=پراندن:GF3s

پریده ,پرید.L+V=پریدن:HP3s

پریده است,پرید.L+V=پریدن:HP3s

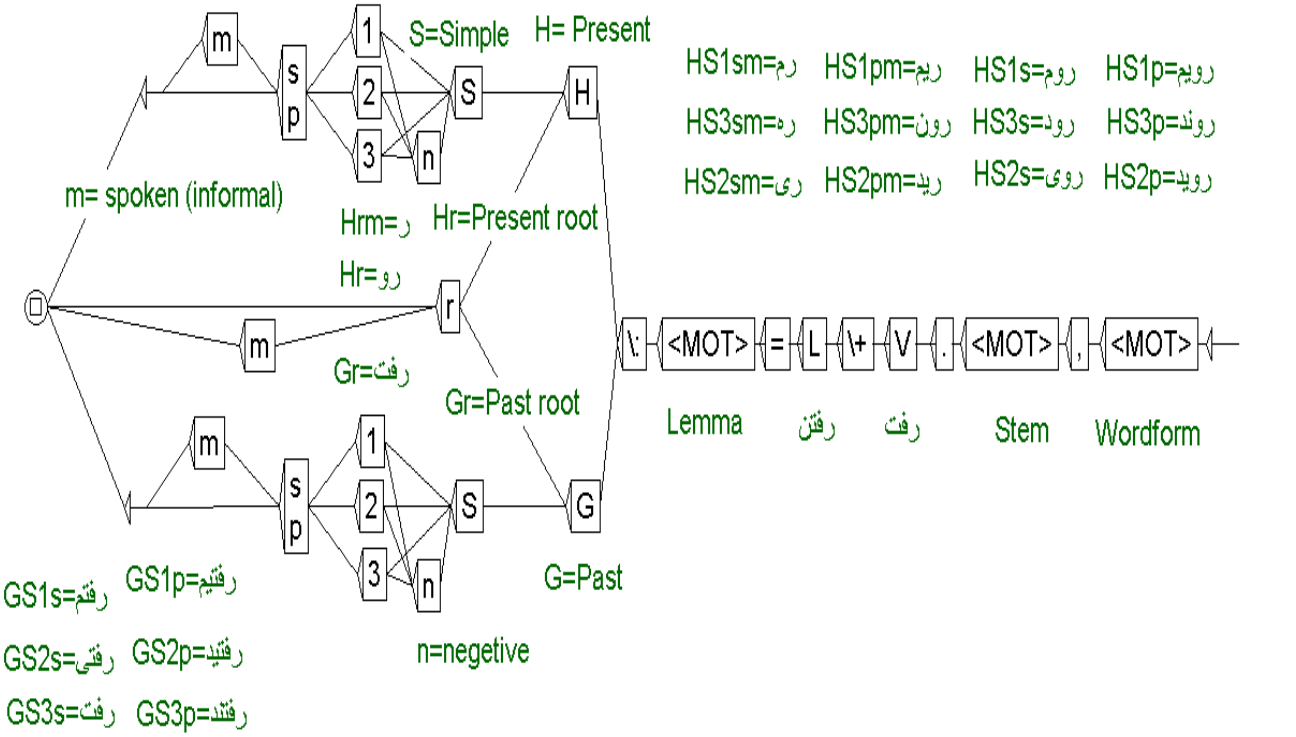
پریده بود,پرید.L+V=پریدن:GP3s

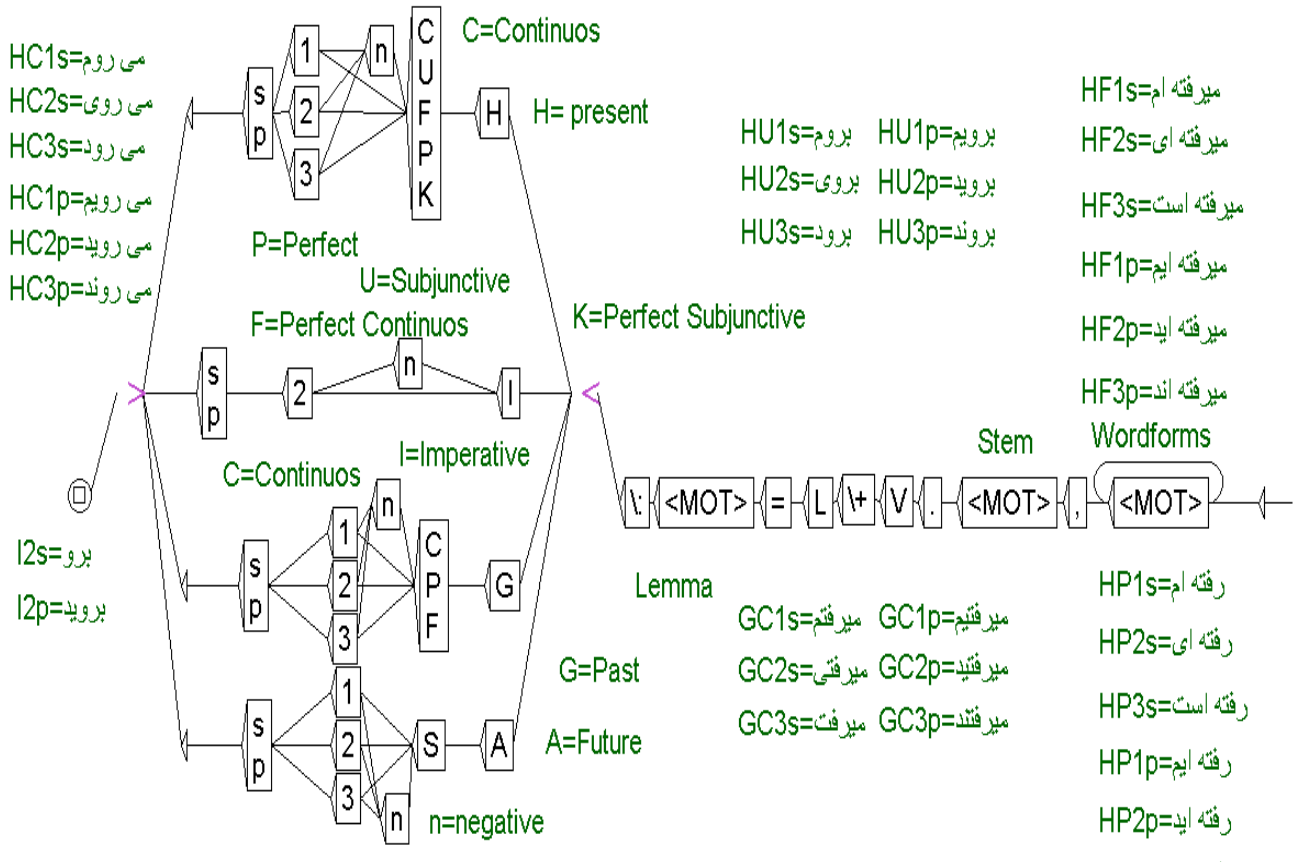
پسندیده ,پسندید.L+V=پسندیدن:HP3s

پنداشته ,پنداشت.L+V=پنداشتن:HP3s

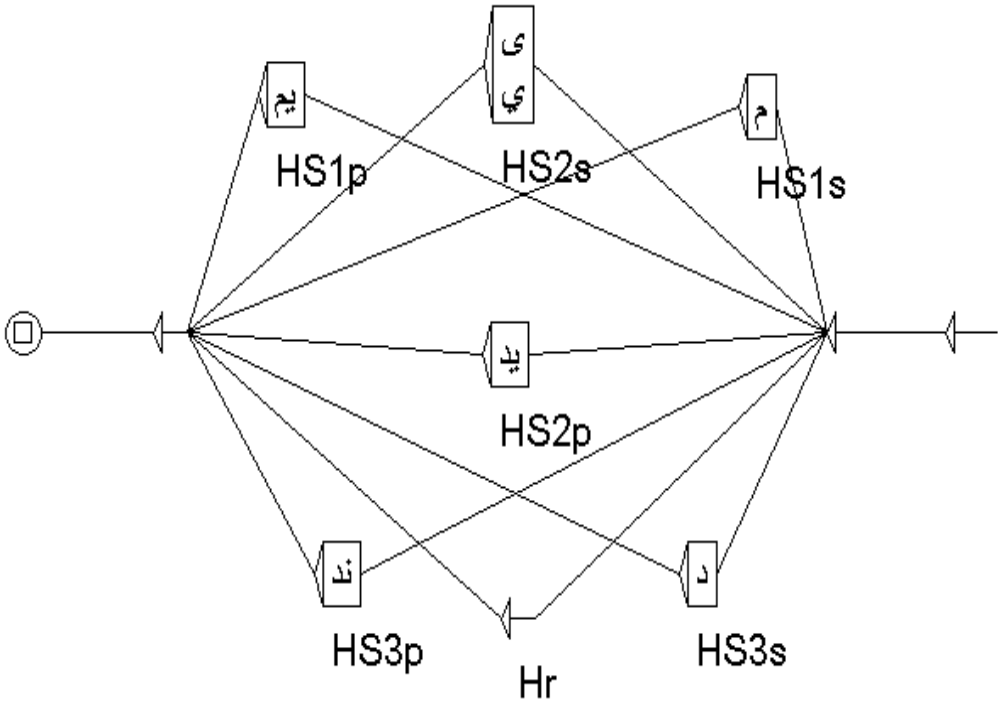
پوسیده ,پوسید.L+V=پوسیدن:HP3s

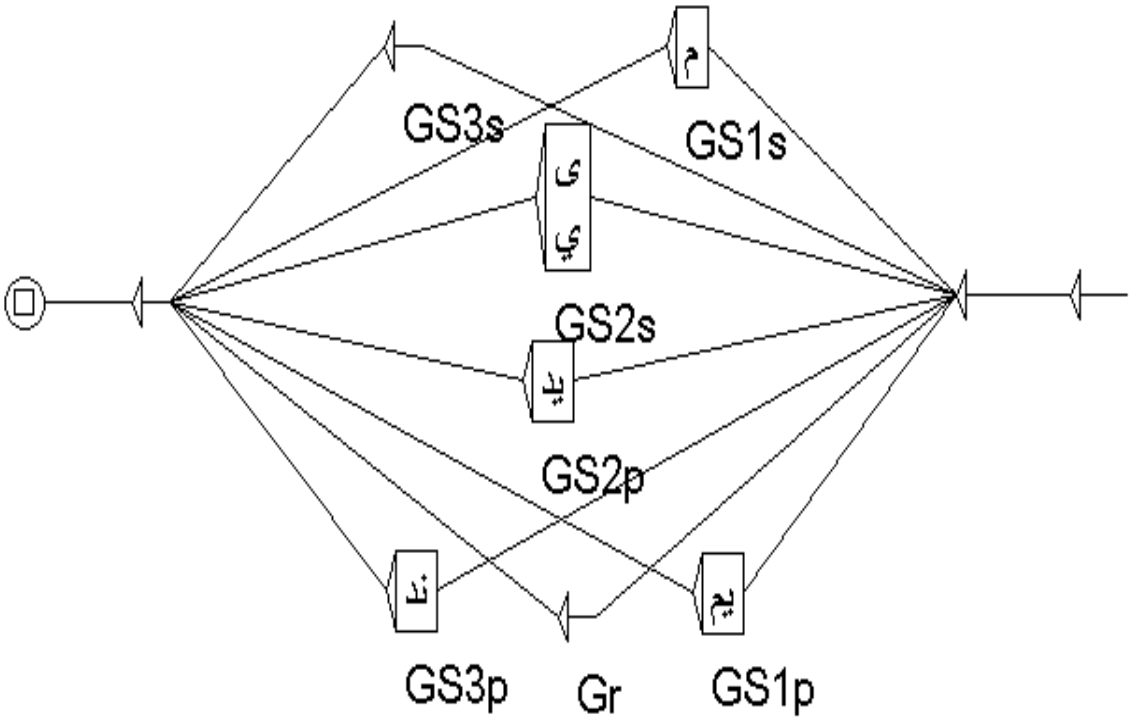
D. Simple BNF and Compound BNF



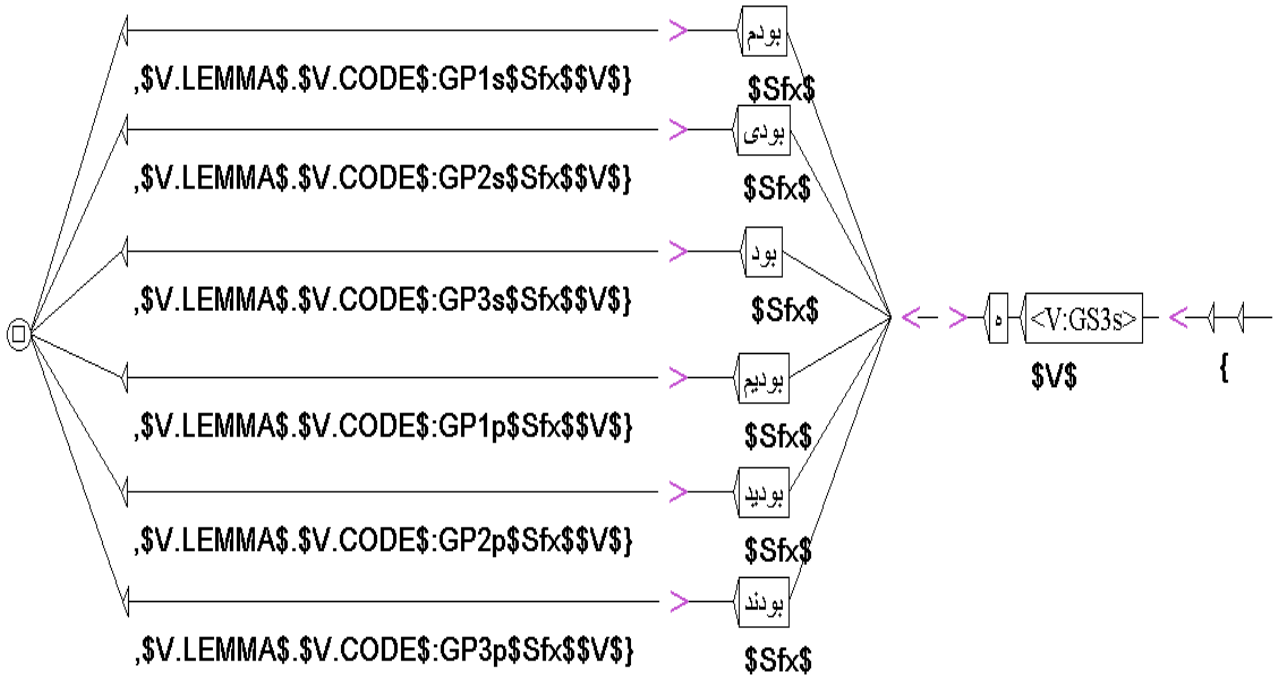


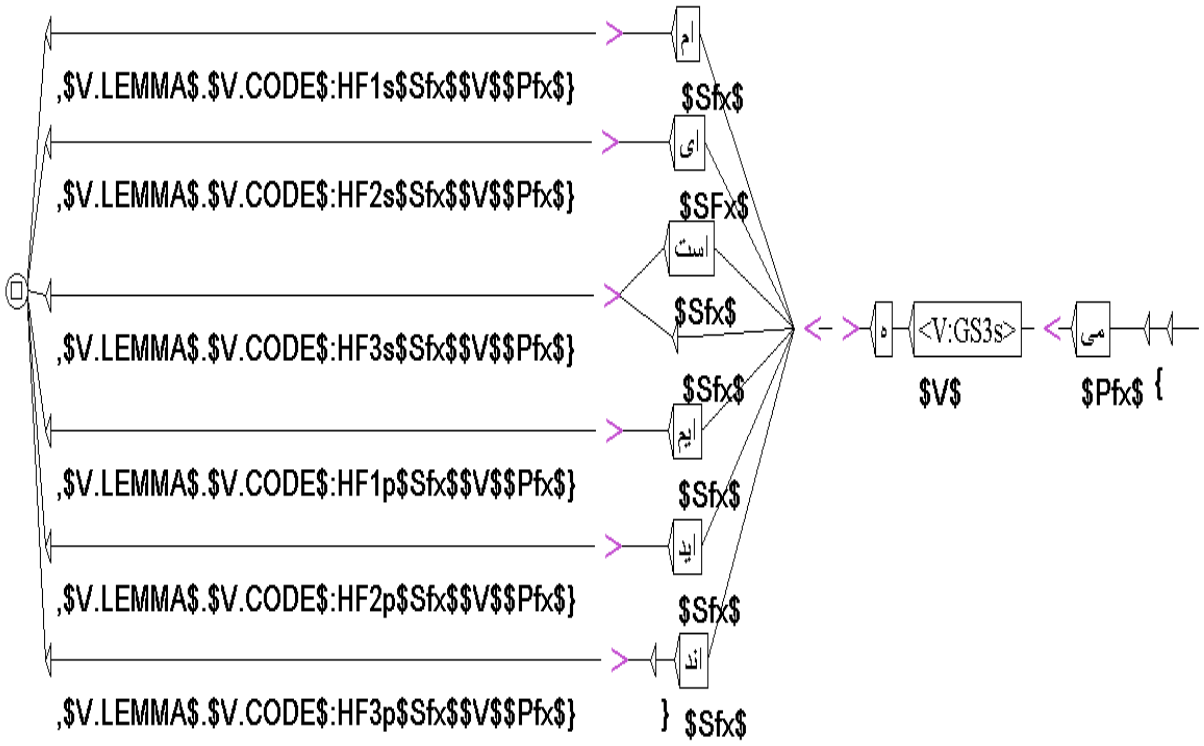
E. Sample of simple verbs FST





F. Sample of Compound verbs FST





G. Evaluation sample text

In Persian[23]:

رازهای غیر پنهان اتومبیلی با سرعت یکهزار مایل در ساعت

سریع ترین اتومبیل جهان قطعه قطعه در این کارگاه در بریتانیا ساخته می شود. هدف دست یابی به سرعت ۱۰۰۰ مایل یا ۱۶۰۹ کیلومتر در ساعت است.

این خودروی فراصوتی "بلدهوند اس.اس.سی" نام دارد و پشت فرمان نمونه آزمایشی آن اندی گرین، خلبان نظامی و صاحب رکورد جهانی سرعت نشسته است.

او برای نخستین بار دریچه خودرو را باز می کند تا صندلی راننده را به ما نشان دهد. اندی گرین می گوید: "این اتاقک خودروی فراصوتی بلدهوند اس.اس.سی، یعنی دفتر کار ۱۰۰۰ مایل در ساعت من است. دارای فرمانی با چاپ سه بعدی از تیتانیوم است که با دست های من جور است. همچنین دارای سه صفحه نمایش از کریستال مایع است. دو ابزار محاسبه آنرا رولکس ساخته، یعنی سرعت سنج و ساعت همراه با کرونو متر. دقت این ابزار لازمه کاری است که با خودرو انجام می دهیم"

این مدل، شبیه اتومبیلی است که تا یکسال دیگر تکمیل می شود. و این چیزی است که اکنون در حال ساخته شدن است.

سه موتور روی خودرو نصب است، یک موتور جت، یک موشک و یک موتور خودرو. فولادی بودن بدنه خود رو به خاطر محکم و قوی بودن آنست.

این خودرو یک رویای مهندسی است و رویای تیم سازنده این است که الهام بخش مهندسان جوان باشد. به همین دلیل هم هیچ چیز سری در اینجا وجود ندارد.

گزارشگر یورونیوز می گوید: "یکی از ویژگیهای این پروژه این است که تمام اطلاعات، و تمام طرحهای مربوط به ساخت خودرو و طرز رانندگی آن، برای تمام دنیا آنلاین و آزاد است."

در کنار تلاشی که برا ی ارائه باز منابع اطلاعات این خودرو می شود، تیم سازنده بلاد هوند هر سال از صدها مدرسه دیدن می کنند. این شیوه تشویق به یاد گیری، از ریچارد نوبل صاحب پیشین رکورد سرعت زمینی است.

ریچارد نوبل می گوید: "بریتانیایی ها در سال به صد هزار مهندس کارشناس نیاز دارند و این برای اماده کردن سی هزار نفر آن ها کافی ست. به این ترتیب وضعیت خیلی جدی است. مساله این است که باید راهی به دبستان و شگفت زده کردن کودکان یافت تا در سالهای بعد جوانان علاقمند شوند."

خودرویی که موتور آن ۱۳۵ هزار اسب قدرت دارد و می تواند سریع تر از جت جنگی حرکت کند، واقعا باید خیلی قابل توجه باشد. و هیچ کس بیشتر از مردی که پشت فرمان آن نشسته توجه برانگیز نیست.

اندی گرین، راننده نمونه آزمایشی این خودرو: "در حال حاضر من یاد می گیرم با تیمی که رکورد بیشترین سرعت زمینی را درتاریخ دارند، کار کنم، مهندسانی در سطح جهانی که مسائل سختی را حل کرده اند. در سال های اینده ما این ماشین را توسعه داده و تکمیل میکنیم و بعد به همراه تیم آن را می رانیم و تمرین می کنیم. همزمان

[23] http://persian.euronews.com/2014/05/01/bloodhound-ssc-the-1000-mph-car-doing-the-school-run
Date of access: 02/May/2014

ما جراجویی های خودمان را رسانه ای می کنیم تا از طریق مخاطبان جهانی الهام بخش نسل آینده باشیم."

در طول یکسال آینده بلد هوند اس اس سی آماده می شود تا به آفریقای جنوبی برود و هدف خود را که طی کردن یک هزار مایل در ساعت است، آغاز کند.

In English[24]:

<Inside Bloodhound SSC: the 1000 mph car>

Piece by piece, the world's fastest car is being built in a workshop on the outskirts of Bristol, England.

Known as Bloodhound SSC, it has the goal of hitting the heady target of 1,000 miles per hour, and inspiring millions of school children in the process.

Behind the wheel of the arrow-shaped machine is current world land speed record holder and military pilot, Andy Green.

For the first time, he opened the hatch to show us the driver's seat.

"This is the cockpit for the Bloodhound Super Sonic Car. This is my 1,000 mile per hour office," Green smiles.

"It has a 3D steering wheel printed in titanium to match my hands.

"We have three liquid crystal displays, and the two backup instruments that Rolex has produced, which is the backup speedometer, and the clock and stopwatch."

"The precision of these instruments is essential to what we're doing," he stresses.

At the other side of the workshop lies a full-scale mock-up of the car, which gives a sense of its scale. At 13.5 metres long and three metres high, it looks like a cross between a jet fighter and a 9-year-old's sketch of their dream dragster.

Bloodhound SSC chasis

Inside there will be more than a few aerospace ingredients in the Bloodhound recipe, including at its heart a Rolls Royce EJ220 jet engine from a Eurofighter Typhoon. This is paired with a Falcon hybrid rocket from Norwegian company Nammo, a supplier to the European space industry.

For the moment the real car is still being built. Chief Engineer Mark Chapman showed us around: "This is the front of the car, this is where the wheels attach to, the front suspension," he explains. "This is all carbon fibre, this is Andy's office," he says, gesturing to the cockpit. The lower assembly behind Green is made from steel. "The reason we've got a steel structure is it's very, very stiff and very very strong," Chapman says.

The rear part of the car where the rear wheels will be mounted is aluminum. "It's very complicated machining, but a beautiful part," he adds proudly.

The car is an engineer's dream, and the dream of the team is to inspire young engineers.

[24] http://www.euronews.com/2014/05/01/bloodhound-ssc-the-1000-mph-car-doing-the-school-run/
Date of access: 02/May/2014

For that reason, nothing here is secret. All of the data on how the car is built and how it will be tested and driven as it attempts to smash the current world land speed record will be shared with the world online, rights-free.

Alongside that push to offer open source information, the Bloodhound team visit hundreds of schools in Britain every year.

The vision for education came from former land speed record holder and Bloodhound SSC Project Director, Richard Noble.

“Britain needs 100,000 graduate engineers per annum, and the system is only delivering 30,000, so it’s a really serious situation. And the problem actually boils right the way down to primary school, theneed to get the kids really excited at a really interesting and young age,” he says.

The team have put enormous effort into creating eye-catching animations to bring the project to life. These include Bloodhound SSC racing and beating a fighter jet, and driving faster than a bullet from a pistol.

In Andy Green they have one of the world’s most convincing ambassadors for the thrill of driving faster than anyone has ever driven before in a car with the equivalent of 135,000 horsepower.

His eyes filled with excitement, he tells us: “right now I get to work with the best land speed record team in history, world class engineers solving problems nobody has solved before. Next year and the year after we actually get to test and develop this car, and I get to drive it and be the delivery part of that test team. And at the same time we get to transmit that message and share the adventure live to a global audience, we get a chance to inspire that next generation.”

The car should be ready for its first test run in spring next year – a humble 200 miles per hour (322 kph) on a Cornish air field. Then in summer next year, the team will decamp to Hakskeen Pan in South Africa, a dry lake that has been specially cleared of small stones to provide a perfect surface for the record attempt.

The aim is to break the current world land speed record of 763 miles per hour, or 1,227 kilometers an hour, in 2015.

Then, in 2016, the car will be readied for its ultimate challenge of breaking through the mythical barrier of 1,000 mph.