# *An Electronic Dictionary of Persian Verbs*

**Bahareh Kakanaeeni**

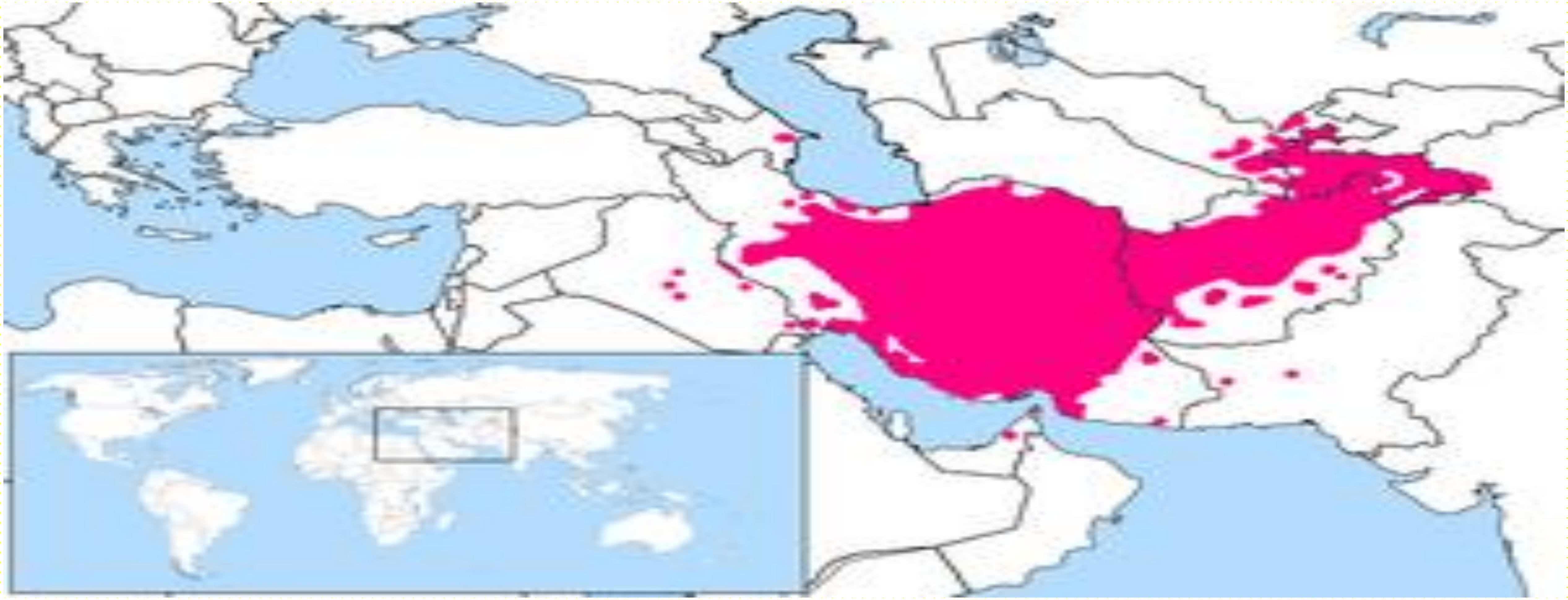
3rd UNITEX/GramLab Workshop

October 9-10, 2014

# Outline

- Motivation and problem
- Related work
- Objectives
- Methods
- Evaluation
- Conclusion
- future work
- References

# 1.Motivation and Problem



There are around 110 million Persian speakers worldwide. Persian is spoken in Iran, Afghanistan, Tajikistan and other countries that historically came under Persian influence.

Source: http://en.wikipedia.org/wiki/Persian_language

- In spite of its cultural and demographic importance, studies on Persian in view of Natural Language Processing (NLP) are scarce. One of the main problems is the fact that lexical resources for NLP are almost non-existent in Persian.

- Because of this, I intend to start to build a new module for the Unitex linguistic development platform (Paumier 2003, 2013) by building linguistic resources for Persian contribute to lay the basis for further studies of this language in a computational linguistics perspective. I start by describing the inflection of Persian verbs, which is the topic of this presentation.

# 2.Related work

1. Finite-State Morphological Analysis of Persian by Karine Megerdoomian from University of California, San Diego.

2. A Morphological Lexicon for the Persian Language by Beno ˆıt Sagot, G´eraldine Walther from Universit´e Paris.

3. Persian Morphology by John R. Perry from University of Chicago.

# 3.Objectives

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

1. Build a sufficiently large lexicon of Persian verbs 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 simple forms of Persian verbs, in order to produce a dictionary of inflected forms and build morphological FST to recognize compound verb forms in the text.

4. Build and apply BNFs to validate the formal correction of the simple words' dictionary and the compound forms matched in texts.

5. Apply the Persian dictionary to a sample of text in order to assess the lexical coverage and grammatical adequacy of the dictionary.

# 4. Methods

## 4.1. Synopsis of Persian verb morphology

| Verb forms | Prefix | Verb stem | Suffix |
|---|---|---|---|
| **Simple present (HS)** | — | rav (ro) “go” | i |
| **Simple past (GS)** | — | raft “gone” | i |
| **Past Continuous (GC)** | mi | raft “gone” | i |
| **Present Continuous (HC)** | mi | rav (ro) “go” | i |
| **Present Subjunctive (HU)** | be | rav (ro) “go” | i |
| **Imperative (I)** | bo | rav (ro) “go” | — |
| **Negative (n)** | na | rav (ro) “go” | — |
| **Present Perfect (HP)** | — | raft “gone” | eh i |
| **Past Perfect (GP)** | — | raft “gone” | eh bodeh i |

## 4.2. A frequency-based lexicon of Persian verbs

| 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 | V008 | 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 | V008 | 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** | ميشه | 121235 | 0.112834 | شدن | ش | sho | become | V008 | ش,V008 |

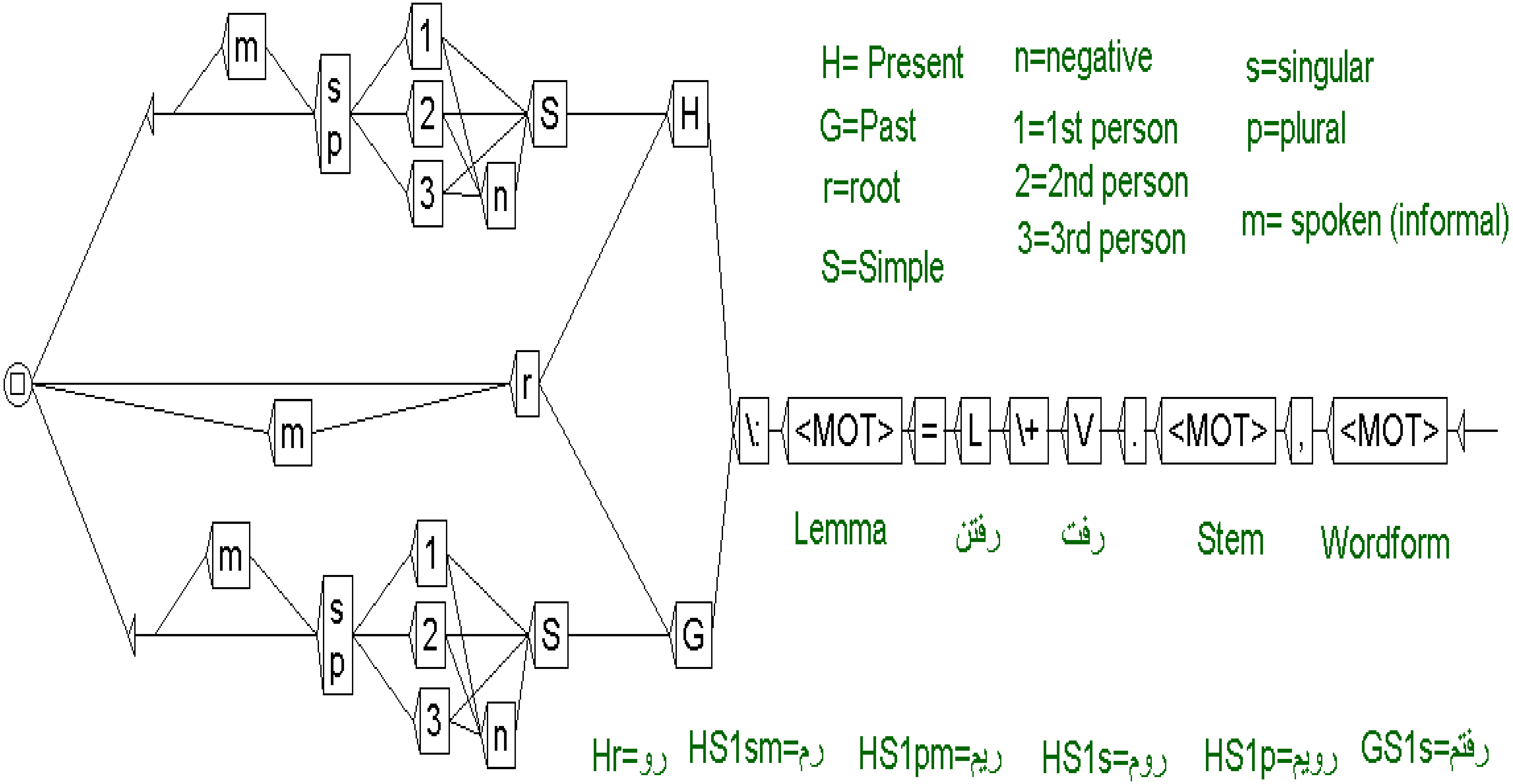
Source for token frequency values: TEP corpus (4,485,147 words)
[Ref:http://ece.ut.ac.ir/nlp/resources.htm]

## 4.3. A Persian verbs dictionary of inflected forms

### 4.3.1. BNF

BNFs in a form of FSA were built to provide formal guidelines for the description of the verb dictionary entries (simple and compound) and, I use these BNFs in order to validate automatically the inflected words' dictionary entries.

Source: http://en.wikipedia.org/wiki/Backus%E2%80%93Naur_Form



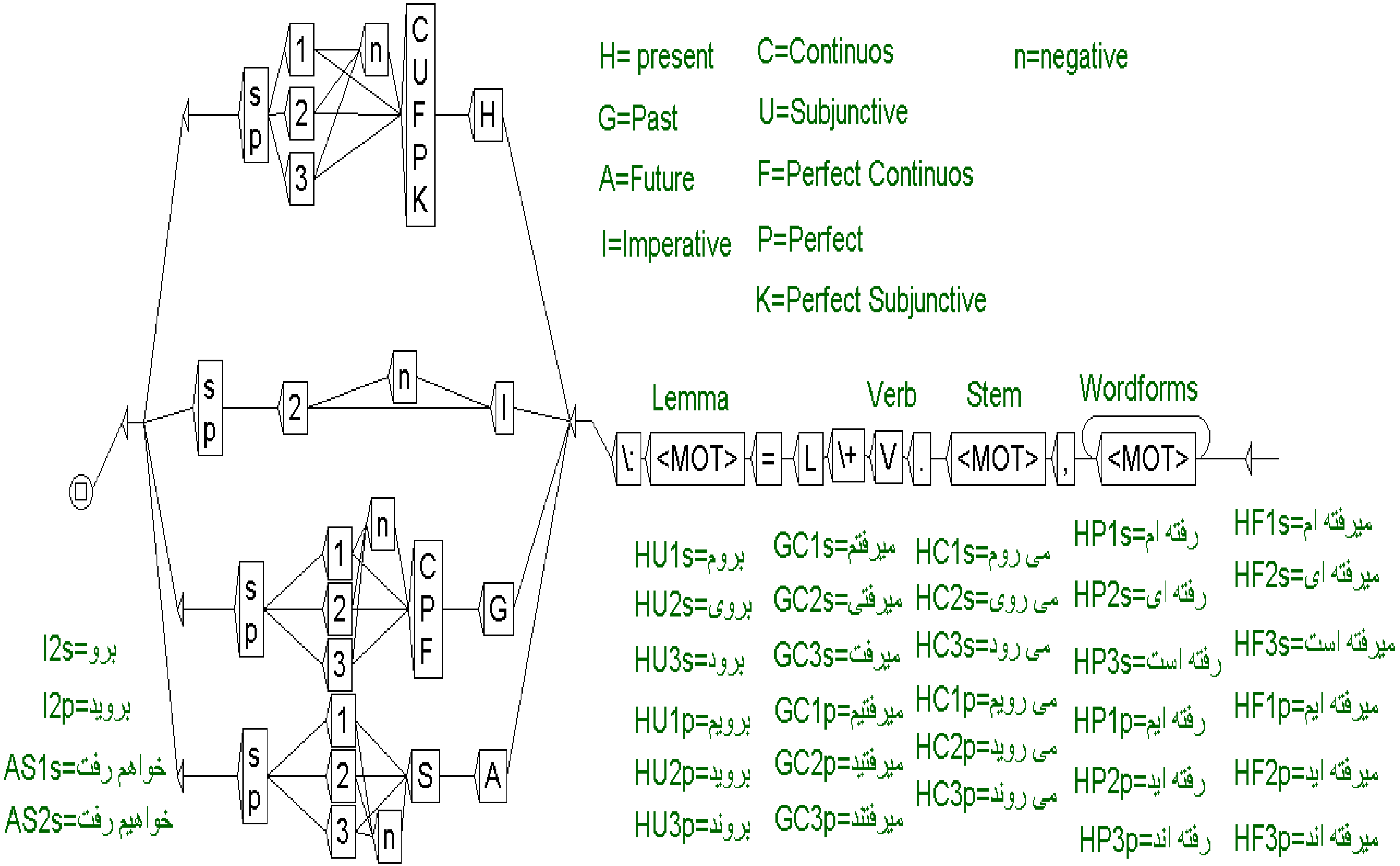
H= Present
G=Past
r=root
S=Simple

n=negative
1=1st person
2=2nd person
3=3rd person

s=singular
p=plural
m= spoken (informal)

| | | | | | | |
|---|---|---|---|---|---|---|
| Hr=رو | HS1sm=رم | HS1pm=ریم | HS1s=روم | HS1p=رویم | GS1s=رفتم | GS1p=رفتیم |
| Gr=رفت | HS3sm=ره | HS3pm=رون | HS3s=رود | HS3p=روند | GS2s=رفتی | GS2p=رفتید |
| Hrm=ر | HS2sm=ری | HS2pm=رید | HS2s=روی | HS2p=روید | GS3s=رفت | GS3p=رفتند |

Figure 1. BNF1 for Simple Tenses



H= present
G=Past
A=Future
I=Imperative

C=Continuos
U=Subjunctive
F=Perfect Continuos
P=Perfect
K=Perfect Subjunctive

n=negative

I2s=برو
I2p=بروید
AS1s=خواهم رفت
AS2s=خواهیم رفت

HU1s=بروم
HU2s=بروی
HU3s=برود
HU1p=برویم
HU2p=بروید
HU3p=بروند

GC1s=میرفتم
GC2s=میرفتی
GC3s=میرفت
GC1p=میرفتیم
GC2p=میرفتید
GC3p=میرفتند

HC1s=می روم
HC2s=می روی
HC3s=می رود
HC1p=می رویم
HC2p=می روید
HC3p=می روند

HP1s=رفته ام
HP2s=رفته ای
HP3s=رفته است
HP1p=رفته ایم
HP2p=رفته اید
HP3p=رفته اند

HF1s=میرفته ام
HF2s=میرفته ای
HF3s=میرفته است
HF1p=میرفته ایم
HF2p=میرفته اید
HF3p=میرفته اند

Figure 2. BNF2 for Compound Tenses

## 4.3.2. Inflection FSTs for written simple words

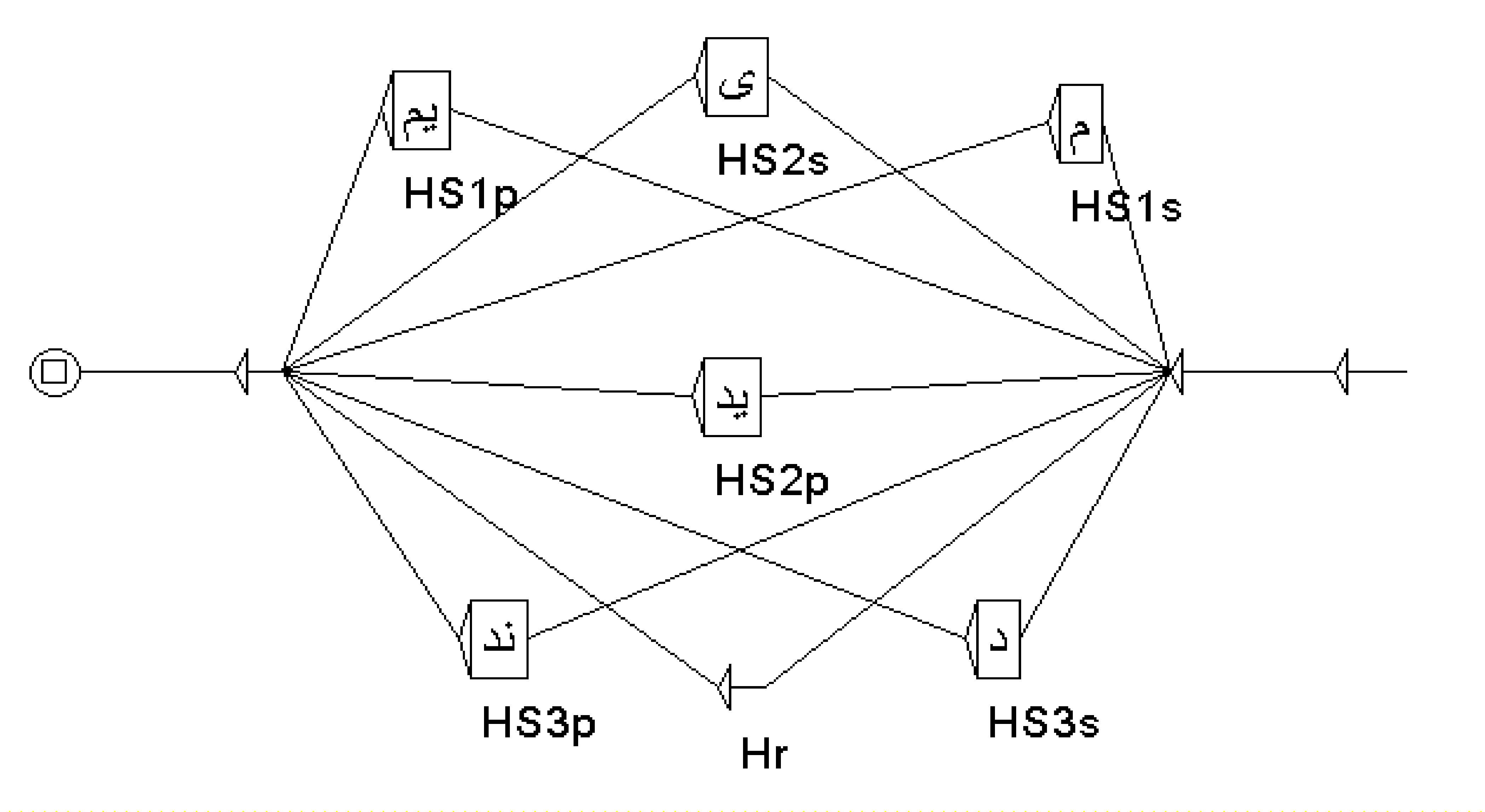


Figure 3. V001, Written Simple Present graph

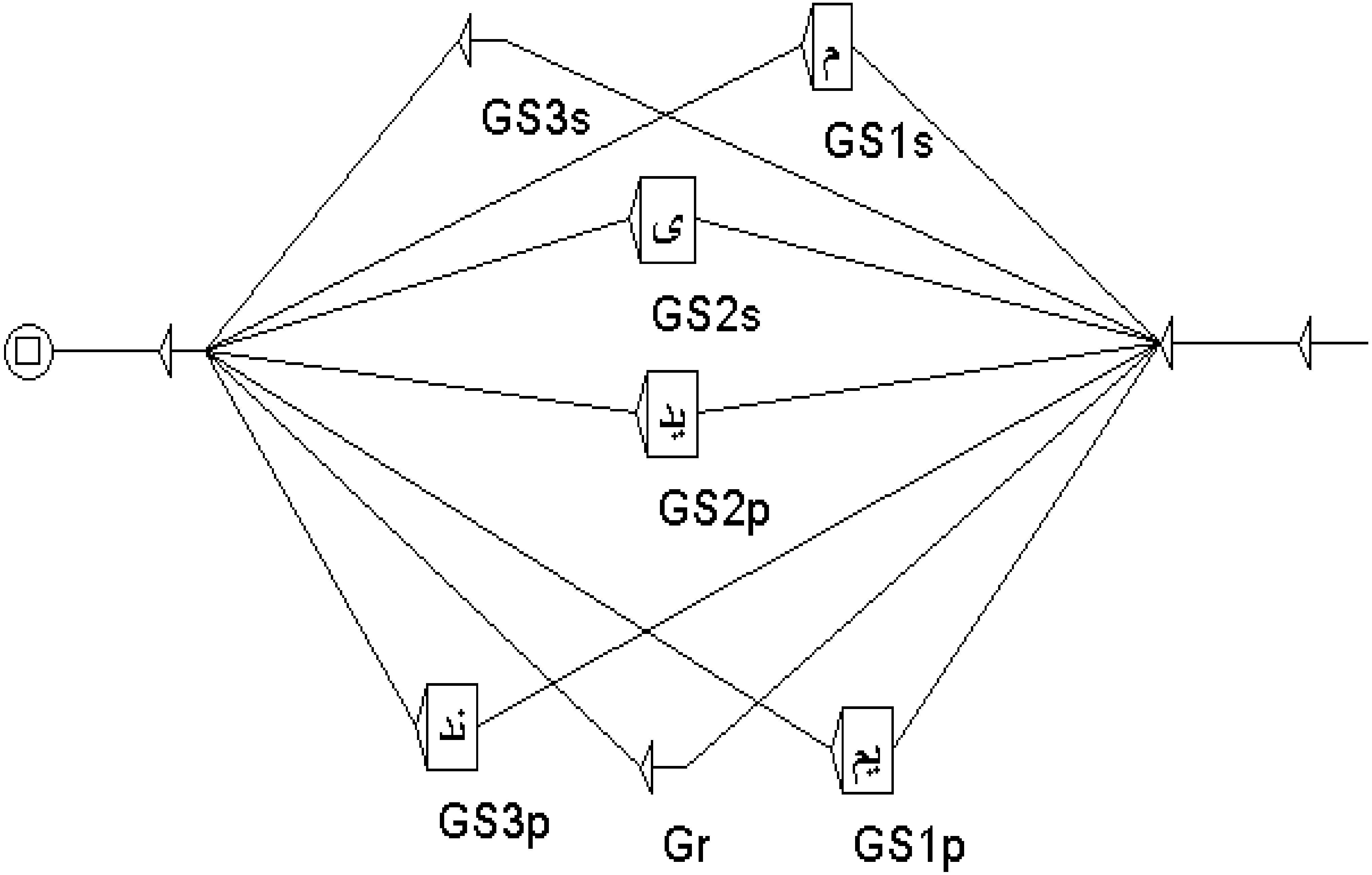


Figure 4. V002, Written Simple Past graph

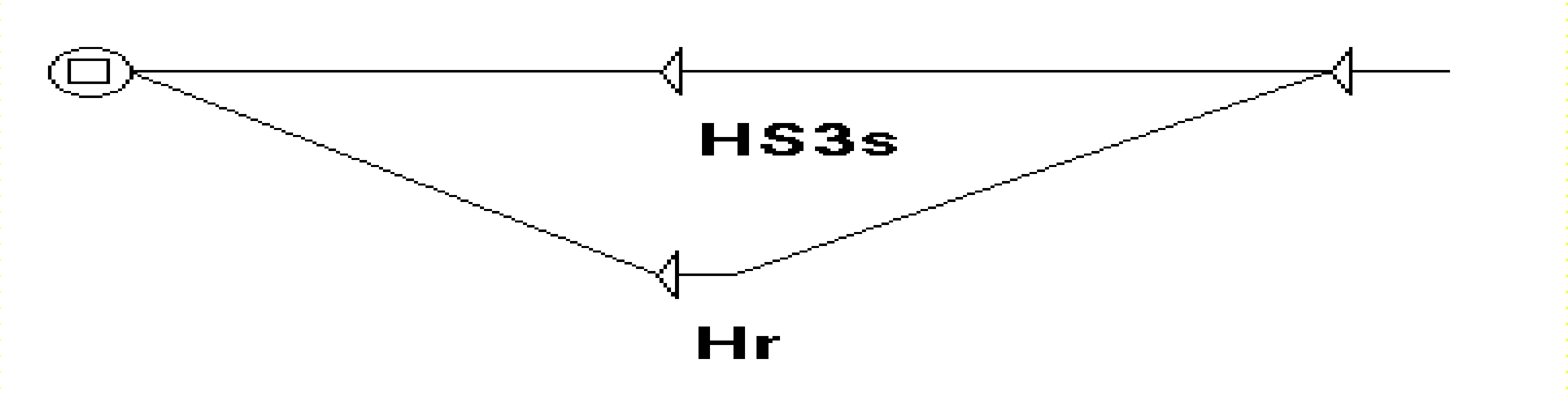


Figure 5. V003, است *ast* (be) graph

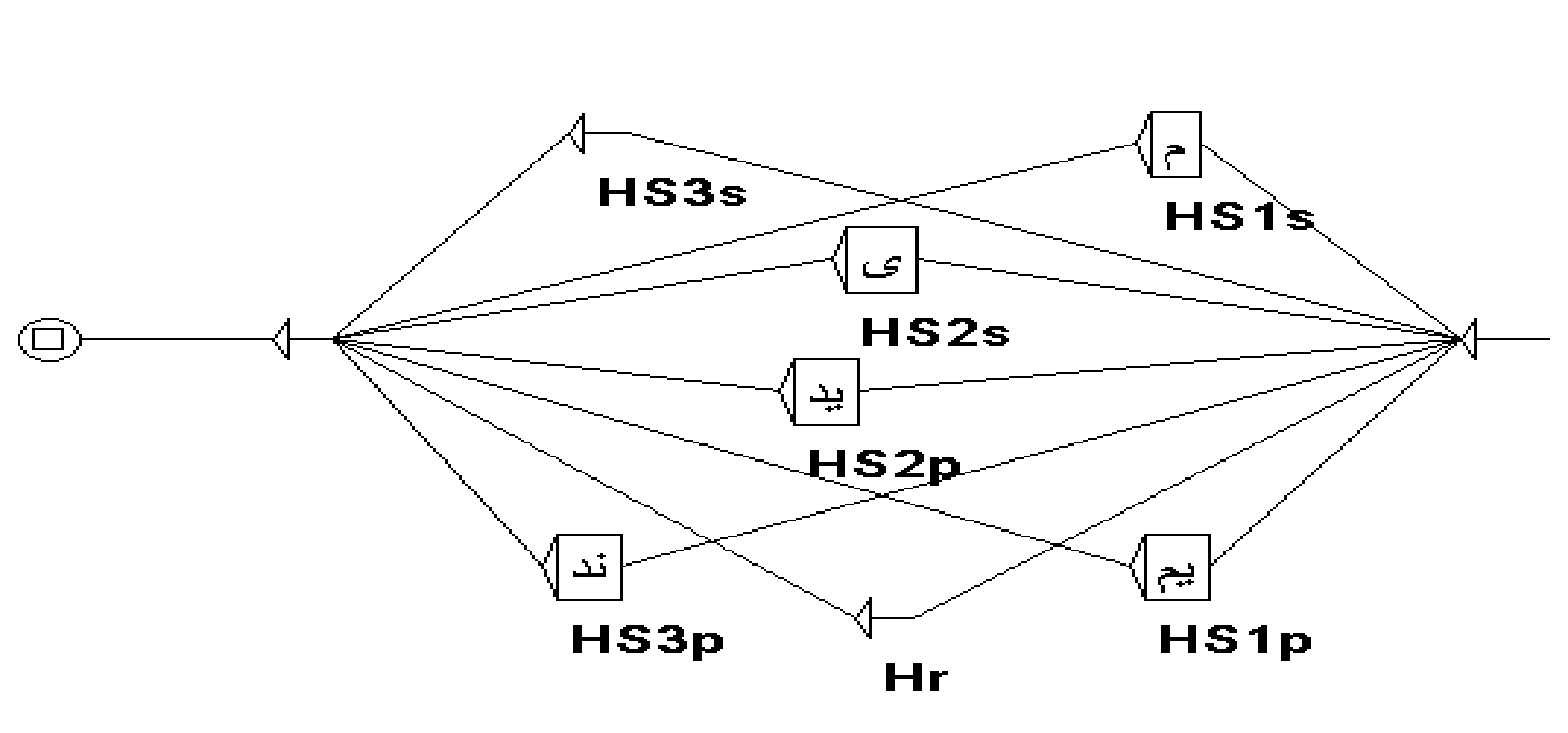


Figure 6. V004, هست *hast* (be) graph

### 4.3.3. Inflection FSTs for spoken simple words

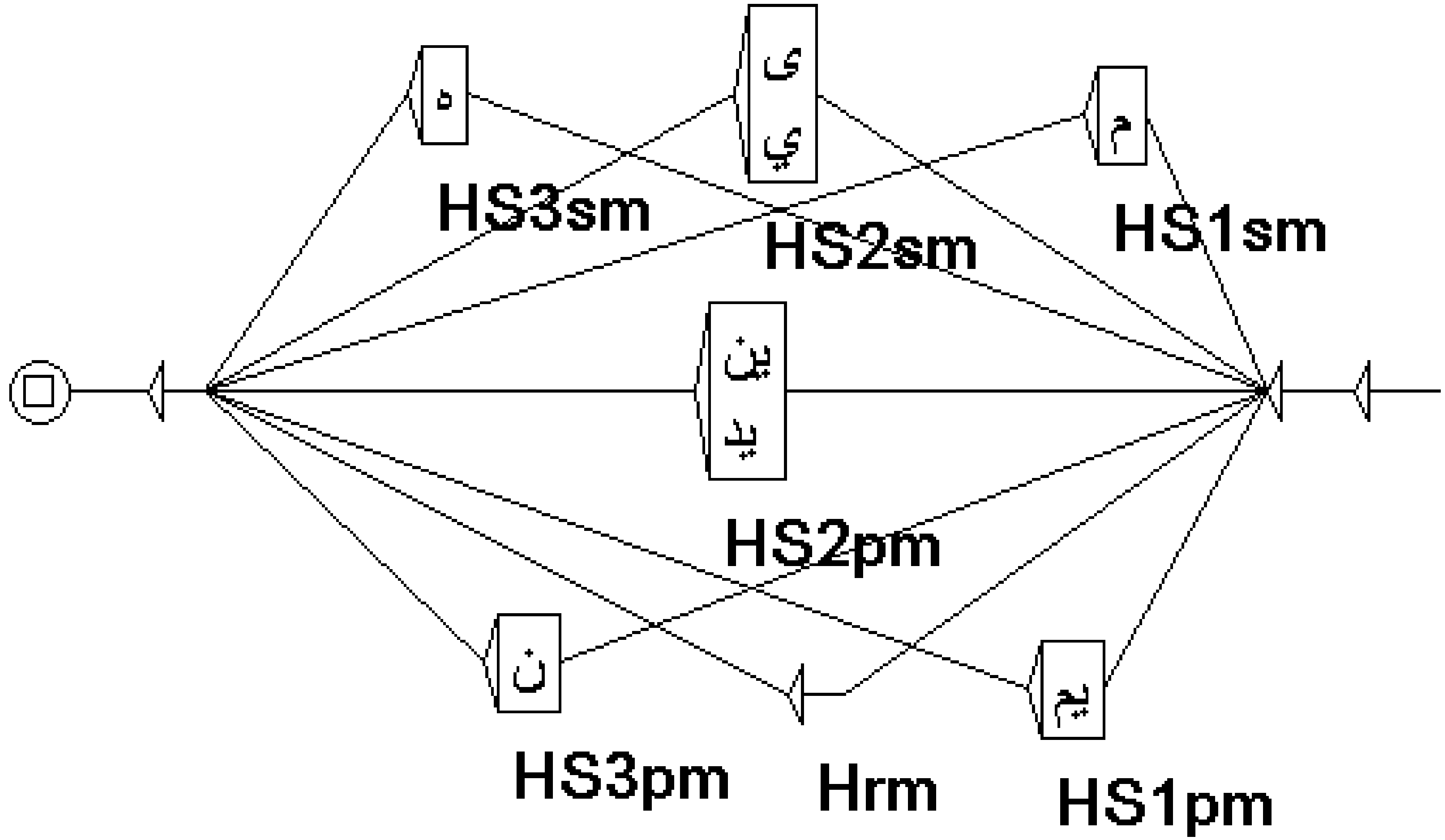


Figure 7. V005, Spoken Simple Present graph

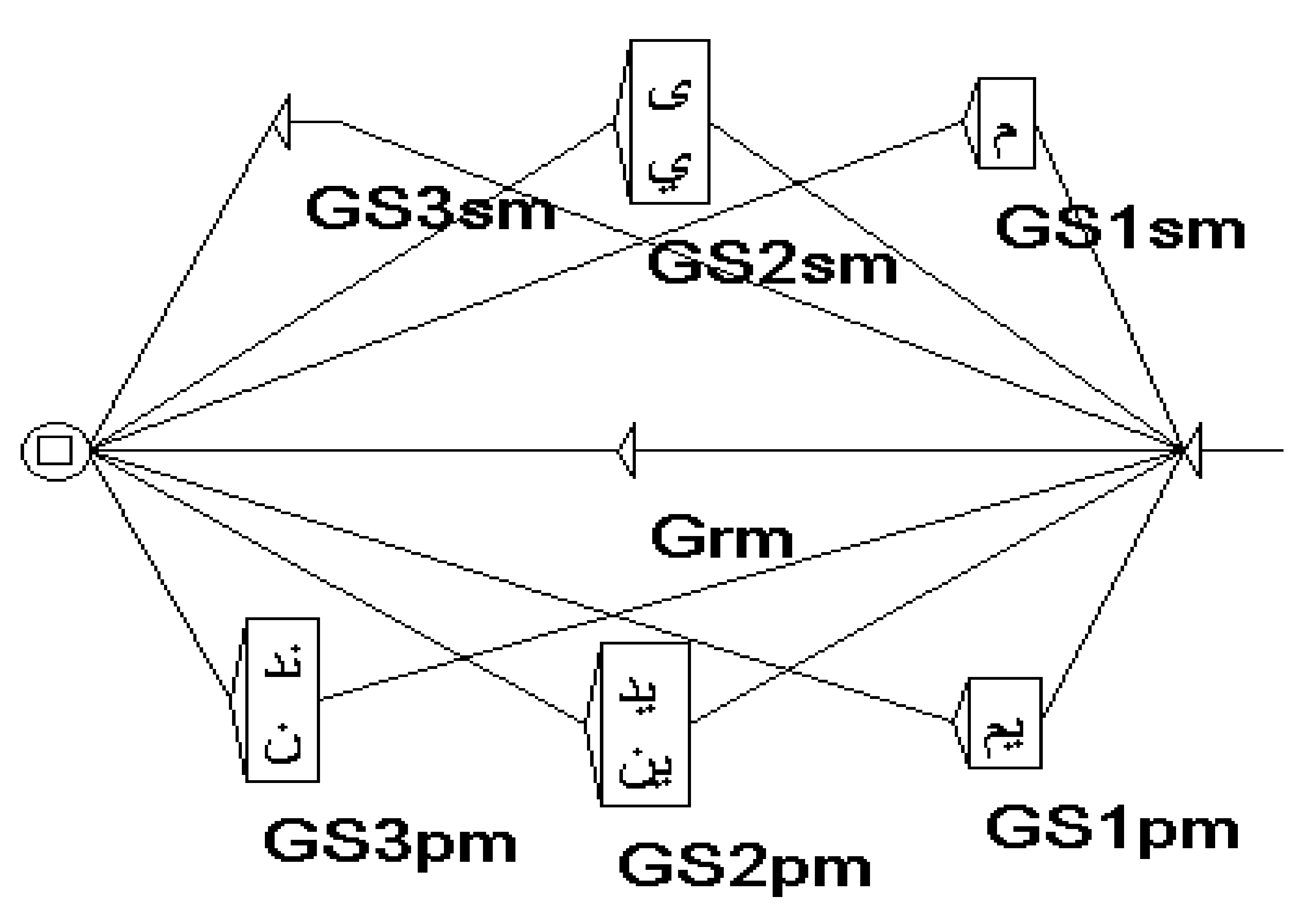


Figure 8. V009, Spoken Simple Past graph

## 4.4. Compound inflection FSTs

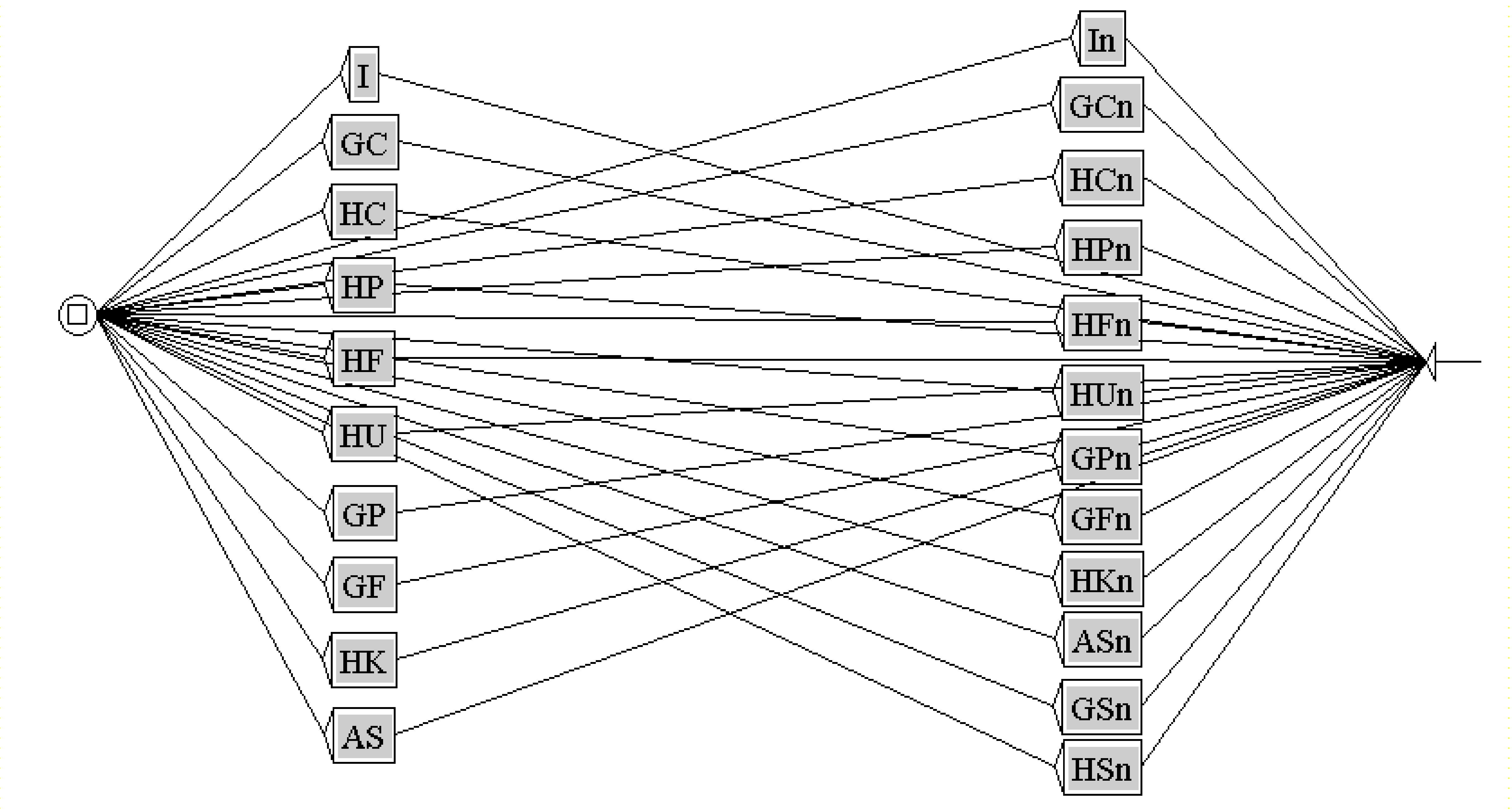
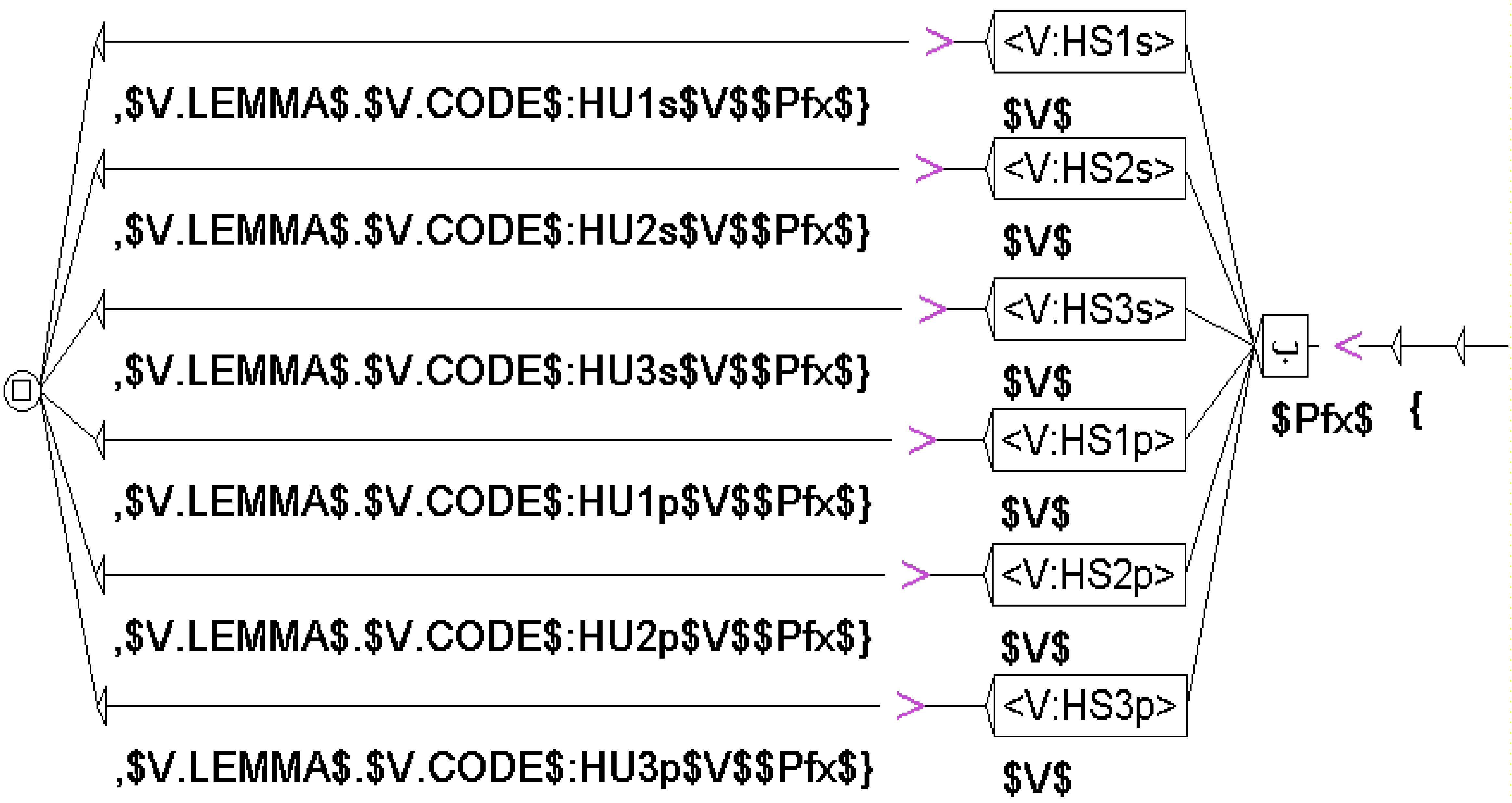


Figure 9. Compound-tenses general graph

This compound tense is formed by prefix بـِ *“be”* and the simple present forms of the verb.



<V:HS1s>
$V$
,$V.LEMMA$.$V.CODE$:HU1s$V$$Pfx$}

<V:HS2s>
$V$
,$V.LEMMA$.$V.CODE$:HU2s$V$$Pfx$}

<V:HS3s>
$V$
,$V.LEMMA$.$V.CODE$:HU3s$V$$Pfx$}

<V:HS1p>
$V$
,$V.LEMMA$.$V.CODE$:HU1p$V$$Pfx$}

<V:HS2p>
$V$
,$V.LEMMA$.$V.CODE$:HU2p$V$$Pfx$}

<V:HS3p>
$V$
,$V.LEMMA$.$V.CODE$:HU3p$V$$Pfx$}

بـ
$Pfx$ {

Figure 10. Morphological graph for Present Subjunctive (HU)

In this graph, prefix می “mi” can be a separate token or be joined with the base form of the verb.

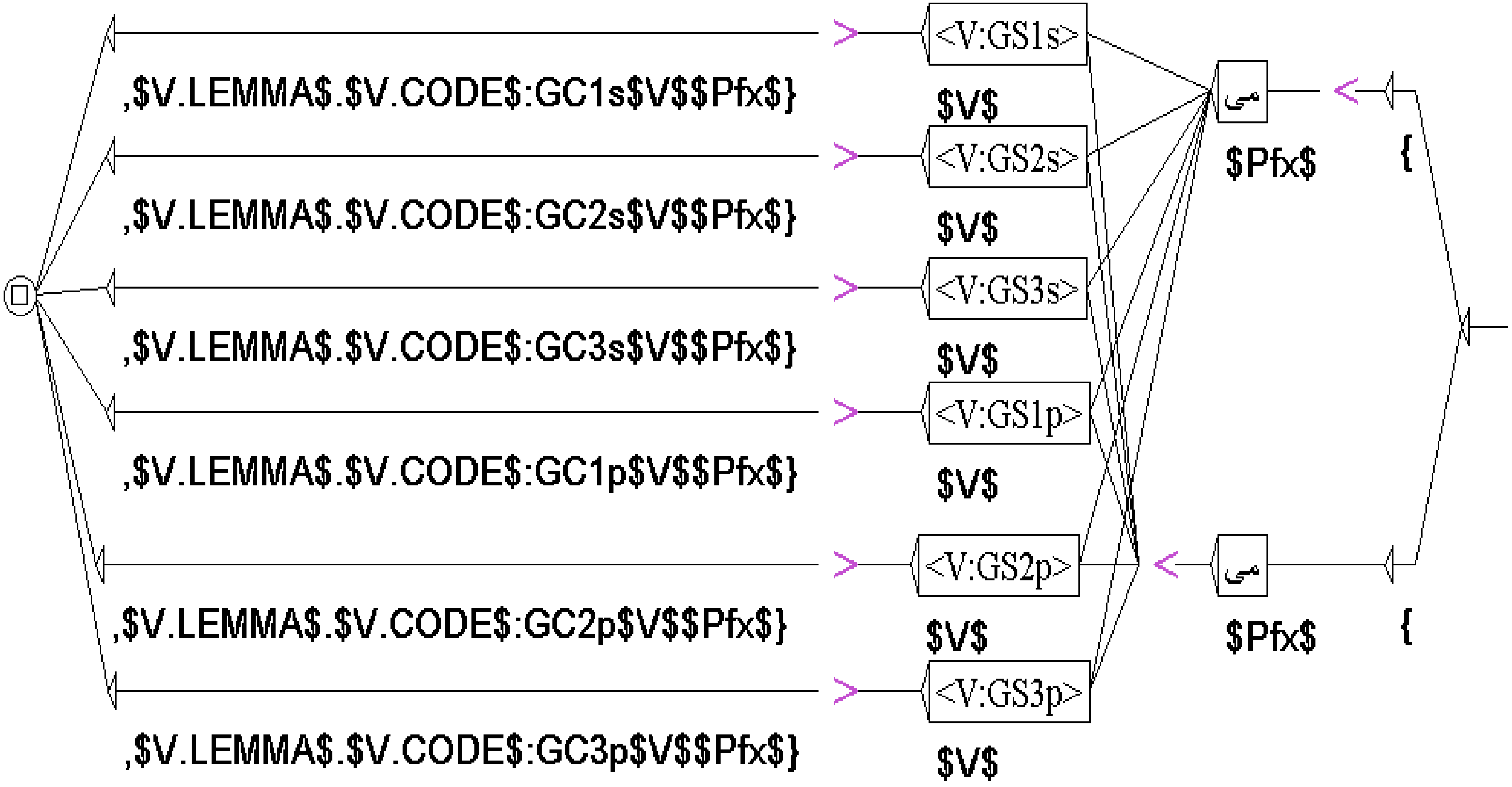

,$V.LEMMA$.$V.CODE$:GC1s$V$$Pfx$}
<V:GS1s>
$V$
,$V.LEMMA$.$V.CODE$:GC2s$V$$Pfx$}
<V:GS2s>
$V$
,$V.LEMMA$.$V.CODE$:GC3s$V$$Pfx$}
<V:GS3s>
$V$
,$V.LEMMA$.$V.CODE$:GC1p$V$$Pfx$}
<V:GS1p>
$V$
,$V.LEMMA$.$V.CODE$:GC2p$V$$Pfx$}
<V:GS2p>
$V$
,$V.LEMMA$.$V.CODE$:GC3p$V$$Pfx$}
<V:GS3p>
$V$
می
$Pfx$
{
می
$Pfx$
{

Figure 11. Morphological graph for Past Continuous (GC)

## 4.5.Generating the dictionary of inflected simple forms

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

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

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

## 4.6. Building 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.

Each entry in DELAF has the following structure:

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

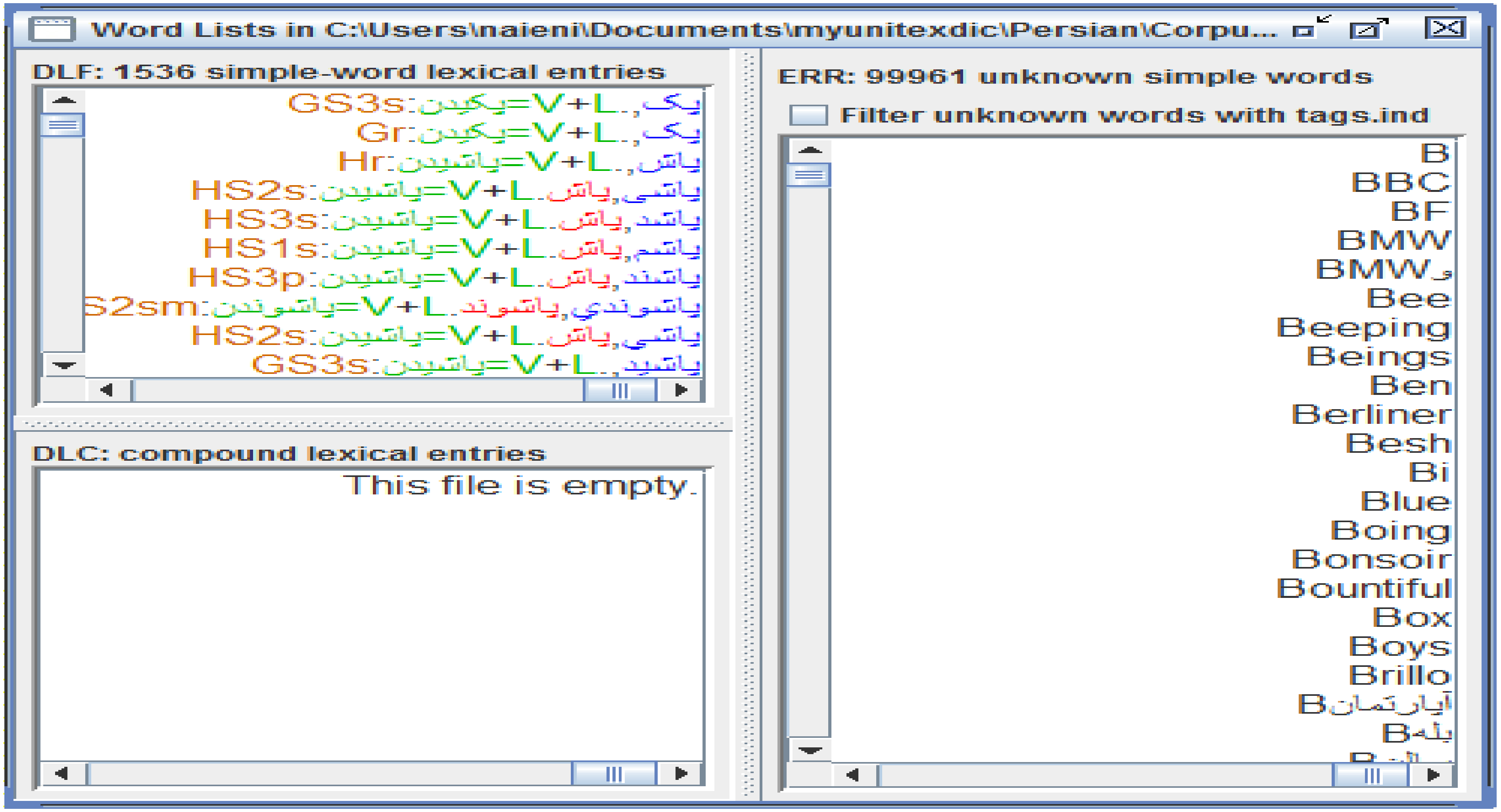
```
رفتم،رفت.V+L=رفتن:GS1s
رفتی،رفت.V+L=رفتن:GS2s
رفت،رفت.V+L=رفتن:GS3s
رفتیم،رفت.V+L=رفتن:GS1p
رفتید،رفت.V+L=رفتن:GS2p
رفتند،رفت.V+L=رفتن:GS3p
رفت،رفت.V+L=رفتن:Gr
```

Figure 12. Extract of Persian DELAF

- 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.

## 4.7. 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 and 1,370 different simple word forms (slightly 10% of corpus' simple words).



Word Lists in C:\Users\naieni\Documents\myunitexdic\Persian\Corpu...

DLF: 1536 simple-word lexical entries

```
GS3s:پکیدن=V+L.,پک
Gr:پکیدن=V+L.,پک
Hr:پاشیدن=V+L.,پاش
HS2s:پاشیدن=V+L.پاش,پاشی
HS3s:پاشیدن=V+L.پاش,پاشد
HS1s:پاشیدن=V+L.پاش,پاشم
HS3p:پاشیدن=V+L.پاش,پاشند
S2sm:پاشوندن=V+L.پاشوند,پاشوندي
HS2s:پاشیدن=V+L.پاش,پاشي
GS3s:پاشیدن=V+L.,پاشید
```

DLC: compound lexical entries

This file is empty.

ERR: 99961 unknown simple words

□ Filter unknown words with tags.ind

```
B
BBC
BF
BMW
BMWو
Bee
Beeping
Beings
Ben
Berliner
Besh
Bi
Blue
Boing
Bonsoir
Bountiful
Box
Boys
Brillo
Bآپارتمان
Bبله
```

Figure 13. Word list

## 4.8. Apply the compound tenses FST to the corpus

When I applied the compound tenses FST to the corpus, the graph found 82,605 compound tenses matches (1,145 different compound word forms).

| | |
|---|---|
| کرده بودید,کرد.V+L=کردن:GP2p | گذاشته ,گذاشت.V+L=گذاشتن:HP3s |
| کرده بودیم,کرد.V+L=کردن:GP1p | گذاشته است,گذاشت.V+L=گذاشتن:HP3s |
| کرده بودم,کرد.V+L=کردن:GP1s | گذاشته بودم,گذاشت.V+L=گذاشتن:GP1s |
| کرده بودند,کرد.V+L=کردن:GP3p | گذاشته بودند,گذاشت.V+L=گذاشتن:GP3p |
| کرده بوده ,کرد.V+L=کردن:GF3s | گذاشته بوده ,گذاشت.V+L=گذاشتن:GF3s |
| کرده بوده است,کرد.V+L=کردن:GF3s | گذشته ,گذشت.V+L=گذشتن:HP3s |
| کنده ,کند.V+L=کندن:HP3s | گذشته ای,گذشت.V+L=گذشتن:HP2s |
| کنده است,کند.V+L=کندن:HP3s | گذشته است,گذشت.V+L=گذشتن:HP3s |
| گذاشته بود,گذاشت.V+L=گذاشتن:GP3s | گذشته ام,گذشت.V+L=گذشتن:HP1s |

Figure 14. first entries of the compound tenses dictionary

# 5. Evaluation

My evaluation was two-fold:

- Evaluate *the formal correction* of the dictionary entries by using the BNFs for both simple and compound forms.

- Estimate *the lexical coverage* of the dictionaries by applying them to a sample text.

## 5.1. Evaluation of the formal correction of the dictionary

The inflected simple forms' dictionary contains 145 lemmas, 292 written verb stems and 127 spoken verb stems and 1,536 different entries. After applying the simple forms BNF (Figure 1) **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, I have 222,698 total entries matches and 3,953 different entries. After applying the Compound form BNF (Figure 2), **all entries were also matched, and no errors were found**.

## 5.2. Evaluation of the lexical coverage of the dictionary

A sample text with around 1,000 words was selected from the web and the text was POS-tagged by the system.

Results:

- For the simple form, we find the verb بـودن *bodan* “to be” that appears two times in the text that was not captured by the dictionary due to forgetting to add corresponding inflectional graph.

- In the compound forms’ evaluation, after manual verification, we find also the verb نـيـست *nist* “not to be” that appear one time in the text that was not taken because, we missing to add this compound graph for negation of the verb هست *hast “be”*.

# 6. Conclusion

- I have achieved the main objectives of this project. I built a sufficiently large lexicon of Persian verbs, until I attained a frequency threshold corresponding to a cumulative percentage of 90% of the total corpus word forms (cumulative sum in 3,349,920) corresponding to 1,466 different verb forms.

- I 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.

- Furthermore, I built 22 morphological graphs in order to produce a dictionary of compound verb inflected forms.

- I 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).

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

- In order to evaluate the dictionary lexical coverage, for the simple form, I found the verb بـودن *bodan* “to be” that appears two times in the text that was not captured by the dictionary due to forgetting to add corresponding inflectional graph and in the compound form evaluation, after manual verification, I found the verb نـيـسـت *nist* “not to be” that appear one time in the text that was not matched because, I was missing to add compound graph for negation of the verb هست *hast “to be”*.

# Future work

Naturally, much is still left to be done.

- First, I 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 the remaining parts of speech (POS), this will be the challenge for future work.

Thank You!