Frame approach to Persian verb generation for educational purposes

Artem Lukanin
Faculty of Linguistics
South Ural State University
pr. Lenina 76, Chelyabinsk
Russian Federation, 454080
artyom@popdict.com

Constance Bobroff
Middle Eastern Studies
University of Texas at Austin
1 Univ Stn, F9400
Austin, TX 78712
connieb@mail.utexas.edu

Abstract

This paper describes an on-line Persian Verb Conjugator using a frame based system. Created rules allow the generation of Persian verb forms in conjunction with a Past-Present stem lexicon to handle exceptions. This program conjugates prefixed and compound verbs as well as independent verbs and creates full paradigms of positive and negative verb forms. In addition, the program is used to generate interactive multiple-choice tests for self-practice.

1 Introduction

Along with dictionaries and grammars, many languages have a book of the most common one hundred or so verbs fully conjugated for all persons and tenses, including the exceptions, for students of that language. To date, the Persian language lacks such a reference work. To fill this void, the online Persian Verb Conjugator (PVC) was created as a free, open-access pedagogical tool for self-study or classroom use.

PVC is first and foremost a tool to teach the mechanical conjugation of modern Persian verbs, a most difficult task for beginning and intermediate students of Persian. The students should understand that PVC, having been supplied with the two stems of each verb, conjugates the verbs dynamically, "on the fly" according to linguistic rules with which it has been supplied and that they too, instead of rote memorization can also perform the same mechanical operations after learning the same rules driving PVC.

1.1 Input methods by the User

Most students will access PVC having been given the URL of the PVC index of the most common verbs of modern standard Persian (http://persianintexas.org/courses/PRS506/PVClist.html) listed in (Persian) alphabetical order along with the English translation. The student may

Picture 1. Infinitives may be manually typed in Perso-Arabic script or Latin transliteration from any page on PVC.

proceed to the conjugation of any verb by clicking either on the Perso-Arabic script version or the transliterated version in Latin characters.

Other users typically come upon PVC having typed or pasted some conjugated form of a verb into a search engine such as Google hoping to determine the infinitival form so that they may then look up the meaning of the verb in a dictionary.

The third way a user may call up the conjugation of a verb is by manually entering the infinitive from any page within PVC. The interface allows the user to toggle between Perso-Arabic script or Latin transliteration without having any need for Persian script or extended Latin capabilities on the user end.

1.2 Unicode Encoding Compliancy

Perso-Arabic: Unlike Arabic, Persian requires Unicode (UTF-8) encoding for proper display of all the letters of the alphabet. Because our target audience consists of students, part of whose training includes following scholarly and scientific international standards, the common practice of the substitution of Arabic (Microsoft Windows codepage 1256) letters, specifically Yeh (U+064A), Kaf (U+0643) and Space (U+0020) for respectively, Persian Yeh (U+06CC), Kaf (U+06A9) and ZWNJ (U+200C) in order that the program be viewable in non-Unicode compliant browsers and platforms was not an option.

Latin transliteration: The only character in the extended Latin subset not commonly found on US English keyboards is the "a with macron above" (U+0101) to represent the long /a/ sound. PVC allows the user to type "aa" (a common practice among Persian users in everyday situations such as email) which PVC then dynamically converts to "ā".

1.3 CSS Font-family and font-size

A few precautions were taken in consideration of the beginning level learners of Persian. While native speakers can read Persian fonts at very small font sizes, the beginners require at least an 18 pt font in order to not strain the eye. The well-hinted Tahoma font is very readable even at low resolution and is the font of choice for Windows. On the other hand, owing to the faulty nature of the Perso-Arabic subset of the Tahoma font supplied with Microsoft Office for Macs, Tahoma should be disallowed for such educational web pages for Mac users. Instead, the Geeza Pro font (unavailable on Windows) is more suitable for Macs. Thus, the PVC style sheet has been designed to display in Tahoma for the Windows users and Geeza Pro for Macs.

Past Perfect	Present Indicative
ایستاده بودم	می‌ایستم
ایستاده بودی	می‌ایستی
ایستاده بود	می‌ایستد
ایستاده بودیم	می‌ایستیم

Picture 2. Font display in Tahoma on a PC with Internet Explorer.

Past Perfect	Present Indicative
ایستاده بودم	می‌ایستم
ایستاده بودی	می‌ایستی
ایستاده بود	می‌ایستد
ایستاده بودیم	می‌ایستیم

Picture 3. Font display in Geeza Pro on a Mac with Safari.

1.4 Color-coding system

The verb forms in Persian differ mainly in stems. To help students to differentiate them a color-coding system was developed. The verb forms based on the Past stem are marked in yellow color, the verb forms based on the Present stem are marked in purple color. The stems are shown before the paradigm serving mostly the purpose of showing the legend of colors used in PVC. The Perfect tenses are based on the Past Participle, which in consequence is based on the Past stem. To group these tenses together the verb forms of the Perfect Subjective, Present Perfect and Past

Perfect are marked in green color. The same Past-Present stem color system is used in tests on conjugation. The background of tests on Present tenses (including the Imperative) is purple, while the background on Past tenses is yellow. Such a color-coding system provides users with an additional clue in memorization of tenses in Persian. (The self-test feature of PVC is taken up below in section 2.4.)

1.5 Pedagogical Considerations

Persian is a language known for its literature, especially poetry with ambiguities suggesting limitless interpretations. It is a language which rejects standardization at every level. Therefore, each school of Persian has its own methods and agreement among scholars and practitioners of the language is rare. Which verbs should be considered in a learner's basic verb list, tenses, the names of the tenses, the six default pronouns, the verb classification scheme, the transliteration system and the spelling conventions all required decision-making and compromise. In the end, those methods which favor the needs of the students, the target users and those which promote scholarly conventions were chosen.

2 System Description

The Persian Verb Conjugator is an on-line education service, which is written in PHP programming language and is located at http://www.laits.utexas.edu/persian/pvc/. It is based on the frame approach, wherein every verb form is a strictly defined frame. Each slot can remain empty or filled with a pseudo-affix, calculated differently for each verb. The pseudo-affix is a sequence of letters, which is not isomorphous to the affix in traditional grammar. For example, the Infinitive 'sukhtan' (سوختن) is divided into the primordial participle 'sukh' and the suffix 'tan' in (Fazel, 2006) whereas in PVC it is divided into the Past stem 'sukht' and the Infinitive pseudo-affix 'an'.

PVC generates all verb forms for the following tenses, including positive and negative forms:

1) the Simple Past;
2) the Imperfect;
3) the Perfect Subjunctive;
4) the Past Progressive;
5) the Present Perfect;
6) the Past Perfect;
7) the Present Indicative;
8) the Present Progressive;
9) the Present Subjunctive;
10) the Future;
11) the Imperative.

There are no negative forms in the Progressive tenses. For educational purposes it was decided to omit verb forms in 1st and 3rd persons for the Imperative in PVC. However, the missing persons are channeled to special usage tests on the Optative and Jussive grammatical functions. Thus PVC generates up to 112 verb forms. Nevertheless, the generation function used both in PVC and tests on conjugation can generate up to 240 verb forms (120 in Active voice and 120 in Passive voice). This means there are 120 frames for generation of these verb forms and 1 frame for generation of the Past Participle needed in formation of passive forms.

PVC takes infinitives in Perso-Arabic script or transliteration and generates verb forms in the corresponding script. In general PVC does not require any lexicons to generate verb forms but it uses an exceptions list (which we refer to as the lexicon) with correspondences between the Past stem and the Present stem for verbs not covered by the PVC rules.

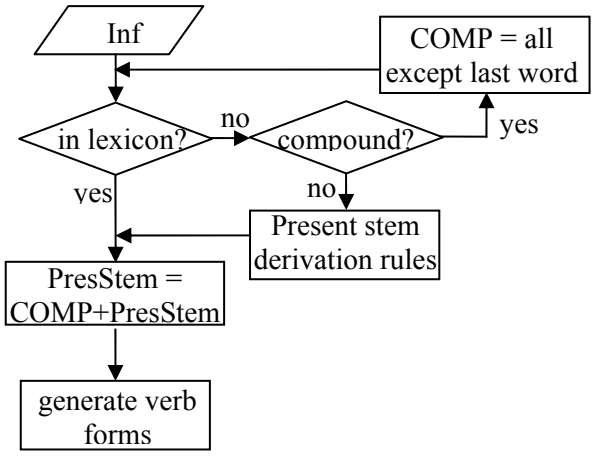
2.1 Present Stem Derivation

Every Persian verb form is built upon one of two different stems for surface forms. As the Past stem is known before verbal paradigm generation, being derived from the Infinitive, we have to derive the Present stem from the Infinitive as well to generate 3 Present tenses and the Imperative. Though as Megerdomian, (2004, p. 5) states, "the two stems … cannot be derived from each other", we developed several rules for Present stem derivation. These rules can be used to conjugate verbs not present in lexicons, e.g. neologisms.

Due to the large number of exceptions, so far no one has come up with a fool-proof scheme to organize Persian verbs into patterns, however Boyle (1966, pp. 30-36) has come up with a manageable system which defines 10 patterns for Present stem derivation. Boyle's system defines a regular pattern for each of the 10 classes and then lists exceptions to each class. After analyzing these

rules we developed 10 rules for deriving Present stems from the Infinitives written in transliteration and 9 rules for verbs in Perso-Arabic script. The difference is that we use only one rule for the infinitives ending in -ndan (e.g. afkandan) and -rdan (e.g. āvardan) instead of two. Thus, to get the Present stem the final 3 letters must be deleted so that the Present stems for these verbs will be 'afkan' and 'āvar'. The verbs ending in these pseudo-affixes, but requiring additional transformations (e.g. 'bordan-bar') are placed into the lexicon of exceptions. The first step of the algorithm is to search for the Infinitive in the lexicon of exceptions. If the verb is not found, the derivation rules are applied to get the Present stem (see Picture 4). So, if the verb 'bordan' is not placed into the lexicon of exceptions, we get a wrong Present stem *bor.

Also we use two different rules for verbs ending in -estan (e.g. bāyestan) and -stan (e.g. jastan), while Boyle defines one rule for such verbs. To derive the Present stems for these verbs the first rule removes the pseudo-affix -estan, while the second rule removes the pseudo-affix -stan and adds -h, so that we get the correct stems 'bāy' and 'jah'.



Picture 4. PVC algorithm.

Thus the Present stem derivation algorithm is defined so that the fourth letter from the end of the Infinitive is compared with the list of letters. Depending on every letter the definite number of letters is subtracted from the end and a number of certain letters are added to the subtracted form. In some rules there is an additional condition whereby we must look at the fifth letter from the end to select a rule for deriving the Present stem.

Due to the fact that the short vowels are omitted in words written in the Persian script, parsing of infinitives in the Persian script is necessarily more complex. But as we take the fourth letter from the end of infinitives as the base for comparison, only one short vowel can be in this position in transliteration — letter 'a' (e.g. verb 'zadan'). Thus we have the same 10 rules for the Persian script, but the '-adan' rule is executed last. Here the base of comparison is not the fourth, but the third letter from the end as the second letter — the short vowel 'a' — is omitted.

Difficulty still exists when we choose between two rules — '-stan' and '-estan'. Short vowel 'e' is omitted and cannot be the base for comparison. As we have only four examples for the '-stan' rule we have made a rule depending on their structure: two of them consist only of four letters, i.e. there is one letter before the pseudo-affix –stan (e.g. **جستن**) while the fourth letter from the end of two others is Alef (e.g. **خواستن**). Therefore the infinitives not satisfying these conditions are supposed to be verbs parsed by the '-estan' rule.

Compounding is very productive in Persian, even more productive than creation of new verbs using –idan suffixing (about 90% of all Persian verbs are compound). The most frequent verbs used for the creation of compound verbs are kardan (**کـردن**), shodan (**شـدن**), dādan (**دادن**), zadan (**زدن**), budan (**بـودن**), dāshtan (**داشـــتن**), sākhtan (**ســـاختن**), etc. As the Present stems of most of them cannot be derived using our rules (except for dāshtan and sākhtan), they are listed in the lexicon of exceptions. If such verbs are listed in the lexicon, PVC can correctly produce the paradigms of the compounds, based on them (see Picture 4). For example, if verb 'kardan' is listed in the lexicon, the Present stem of verb 'tabdil kardan' (**تبدیل کردن**) will be derived correctly (tabdil kon, **تبدیل کن**). The Present stem for verb 'bāz dāshtan' (**باز داشتن**), for example, will be derived correctly (bāz dār, **باز دار**) using only the –shtan rule (remove –shtan and add –r) without the lexicon of exceptions.

Presently there are 95 exceptions in the lexicon. As most of them are used for compounding, PVC is estimated to correctly generate paradigms of 90-95% of all Persian verbs (including possible neologisms).

2.2 Verb Form Frames

Every verb form in PVC is a frame, i.e. a pattern, where each slot is filled in with a pseudo-affix or a non-conjugated verb of a compound verb. Each slot can be empty, e.g. every tense frame has a COMPOUND slot, which is filled in with the verb attached to the main verb to form a compound verb. If the given verb is compound, the COMPOUND slot is filled in, otherwise it is left empty (see verb 'bar gozidan' in Table 1).

Every pseudo-affix is calculated for every verb before filling in the slots, e.g. if the final letter of the Present stem is Alef, (a or ā), pseudo-affix VyV is set to Persian Yeh, otherwise it is left empty (see verb 'afzudan' in Table 1).

Additional rules can deny some frames. In the current version of PVC there are only two verbs denying frames: 'budan' does not have forms in the Present Progressive, the Past Progressive, the Perfect Subjunctive and the Past Perfect, denying the corresponding frames; 'dāshtan' denies the frames for the Present Progressive, the Past Progressive and the Perfect Subjunctive.

The Passive voice in Persian is formed with the Past Participle and the conjugated form of shodan (**شدن**, to become). Because this is a syntactical construction and therefore does not involve any morphological or phonological changes it was decided not to generate passive forms in PVC itself, but as with all matters relating to usage rather than mechanical manipulation, to make an additional self-test on Active/Passive voice formation. It should be noted, that the base function, used in both PVC and conjugation tests, takes an additional parameter to return the required frame in active or passive voice. When the function is called with the passive voice parameter, the Past Participle of the required verb is added to the contents of the COMPOUND slot and the PastStem and PresentStem variables are rewritten with the Past and Present Stems of the verb shodan (shod and shaw respectively). For example, if it is required to generate the passive verb form for the Past Progressive 3rd singular of 'bar chidan' (**بر چیدن**), the contents of the COMPOUND slot ('bar ', **بر**) become 'bar chide ' (**بر چیده**) and then all the required contents of the rest of the slots of the Past Progressive 3rd singular frame are calculated based on the verb 'shodan'. The contents of the slots can be seen in Table 2. So after concatenating slots the resulting form will be 'dāsht bar chide mishod' (**داشت بر چیده می‌شد**).

2.3 Control Panel

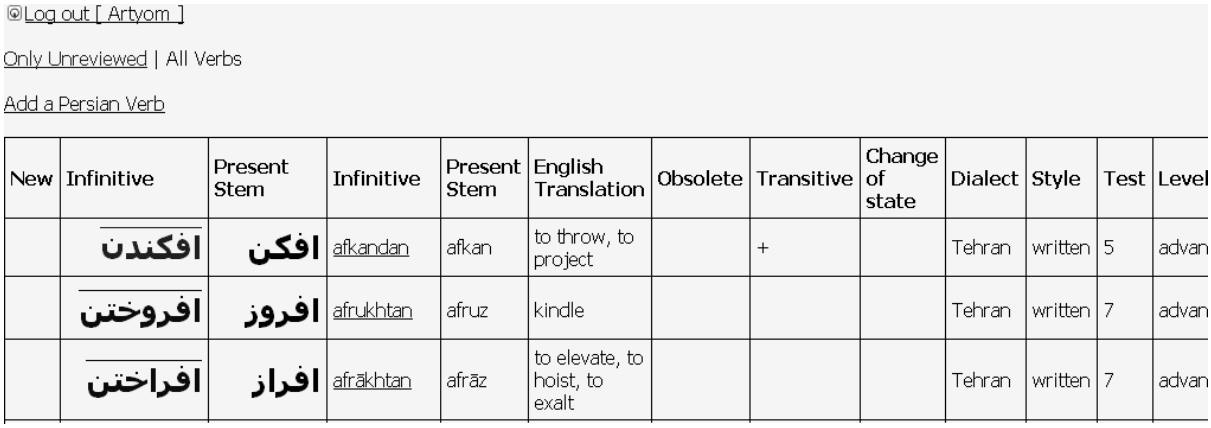
PVC was originally designed to be an open system. That is, in order to encourage participation among scholars and teachers of Persian, a facility was added to allow users to submit new or modified

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5	Slot 6	Slot 7	Slot 8
DAAR	**SG1**	**SPACE**	**COMPOUND**	**PREF_MI**	**PresStem**	**VyV**	**SG1**
dār	*am*	*SPACE*	*bar+SPACE*	*mi*	*gozin*		*am*
دار	م	*SPACE*	بر+*SPACE*	می+*ZWNJ*	گزین		م
dār	*am*	*SPACE*		*mi*	*afzā*	*y*	*am*
دار	م	*SPACE*		می+*ZWNJ*	افزا	ی	م

Table 1. Present Progressive 1st singular frame with 2 examples: 'bar gozidan' (**بر گزیدن**) and 'afzudan' (**افزودن**).

Slot 1	Slot 2	Slot 3	Slot 4	Slot 5
DAASHT	**SPACE**	**COMPOUND**	**PREF_MI**	**PastStem**
dāsht	*SPACE*	*bar+SPACE+chide+SPACE*	*mi*	*shod*
داشت	*SPACE*	بر+*SPACE*+چیده +*SPACE*	می+*ZWNJ*	شود

Table 2. Past Progressive 3rd singular frame with the Passive Voice parameter for verb 'bar chidan' (**بر چیدن**).



Log out [Artyom]

Only Unreviewed | All Verbs

Add a Persian Verb

New	Infinitive	Present Stem	Infinitive	Present Stem	English Translation	Obsolete	Transitive	Change of state	Dialect	Style	Test	Level
	افکندن	افکن	afkandan	afkan	to throw, to project		+		Tehran	written	5	advan
	افروختن	افروز	afrukhtan	afruz	kindle				Tehran	written	7	advan
	افراختن	افراز	afrākhtan	afrāz	to elevate, to hoist, to exalt				Tehran	written	7	advan

Picture 5. Control panel.

verbs or suggestions. However, after several months, the only individuals taking advantage of this facility appeared to be mischief-makers and so this facility was disabled.

In the interests of future uses of PVC, it was decided to use the lexicon of exceptions not only for exceptions, but for providing additional information about every verb:

- English translation;
- dialect;
- style;
- usage (if the verb is obsolete or not);
- transitiveness;
- change of state;
- Boyle pattern;
- level (basic or advanced);
- person for the passives.

The administrators has a special control panel, where they can add new verbs, edit or tag existing ones and delete incorrect verbs.

The same lexicon is also used as the base for conjugation tests due to the fact that some information is required only for test generation purposes. For example, Boyle pattern number is used to select verbs from the list belonging only to one of ten patterns or to all of them to make a comprehensive test on Boyle patterns. The level field is used to generate conjugation tests for students of different skill levels. The transitiveness field is used to select only transitive verbs for the test on Voice. Both 'tabdil kardan' (**تبدیل کردن**) and 'tabdil shodan' (**تبدیل شدن**) verbs, for example, are not selected for this test, because the former has 'only active' and the latter has 'only passive' values in this field. This means that the program selects only those verbs, which can be used in both active and passive constructions.

Not all verbs can be used in all persons in Passive voice. For example, verb 'bar chidan' (**بر چیدن**) can be used only in 3rd person singular and plural. That is why only these two verb forms are used in tests on Passive voice for verbs with this tag.

2.4 Conjugation Tests

However much PVC may be a useful tool in performing mechanical operations on verb stems to correctly conjugate verbs, from the new learner's perspective, the number of forms can be overwhelming. The student may not be able to digest and assimilate the data given in the form of charts or paradigms such as presented in PVC. Furthermore, PVC makes no claims to assist with matters of usage. Therefore, a need was felt for an interface between student and PVC whereby the student could practice one kind of manipulation at a time taking actual usage into consideration.

This interface has taken the form of interactive, multiple-choice tests (http://laits.utexas.edu/persian/tests/index.php) which are fed from PVC using the PVC algorithm. The advantage of such tests in comparison to other grammar tests is that they are dynamically created every time meaning the student may retake them multiple times and always have a fresh test preventing monotony and fatigue. The randomly chosen verbs are automatically conjugated into the required verb forms using the verb form generation

function of PVC. The order of questions and answers is randomized each time.

The tests take the verbs from the PVC lexicon and thus require only the generation function. The Present Stem Derivation algorithm is not used here, because additional information is required for generation of different tests (see section 2.3). Presently the exceptions list consists of 182 verbs, the true exceptions of which are 95. The rest of the tagged verbs are listed there for the purposes of generating the self-tests. Exception here refers to any verb, the Present stem of which cannot be derived from its Infinitive using our derivation rules.

Some verbs are marked to be used in special tests. Thus the verbs from Boyle patterns are marked with the number of the pattern, e.g., Boyle Pattern 1 test selects the verbs only from this pattern. Other conjugation tests select verbs randomly from the whole lexicon.

The generation function requires 6 arguments:

- the Past stem;
- the Present stem;
- the tense number;
- the form number (there are 6 forms: 3 persons singular and plural);
- the script (transliteration or the Persian script);
- the voice (passive by default).

Every test is presented both in transliteration and the Persian script allowing self-learning for students of different levels. Many tests are presented in two versions: basic vs. advanced levels, the former being more suited to beginning level students.

Different tests have different algorithms for selecting questions and distractors (wrong answer choices). For example, the test on Negative Imperfect selects random questions of eight types and wrong answers of four types. It is offered to either translate a verb from English to Persian or to change the verb form from the Present Indicative to the Imperfect. Any Persian verb in Imperfect can be translated by seven different sentences. For example, the verb form 'midādand' can be translated into English as 'They were in the habit of regularly giving' or ‘They used to give’ or ‘They would give’. Each Persian infinitive has one or more English translations in the lexicon. A random translation is taken from the lexicon to fill in the required slot. The terminal slot can be a function for generation of required the English verb form. Three English-language functions were created to generate Simple Present 3rd singular (e.g. carry→carries), Participle I (e.g. tie→tying) and Past forms (e.g. take→took, take→taken or change→changed). The functions use special phonological rules and a lexicon of irregular verbs for this purpose. The wrong answers can be:

- the verb in different tense;
- the verb in different person;
- incorrectly spelled verb;
- a different verb in correct tense and person.

It was noted that students of Persian often mistakenly write the negative prefix separately from the verb. While such practice is allowed in transliteration, it is a true mistake in the Persian script. For this reason, the wrong spelling **نه خورده باشیم** can be used in this test as a distractor among the multiple-choice answers along with the correct spelling **نخورده باشیم** but not for the counterpart test in the transliterated script.

2.5 Classroom experiences and feedback

PVC was integrated into the syllabus of the first-year Persian class at the University of Texas at Austin and was used by the students with great success for the academic year 2006-2007. Since the students generally have access to the internet, either on their own laptops or through campus computer labs, this online tool is always at hand. However, the most important use of PVC in the life of the students is in the self-testing opportunities provided by the tests which not only serve the student well in the initial teaching phase but in review and maintenance as well. The improved performance on class written tests and speaking practice testifies to the effectiveness of this tool. The only recommendation from the students so far seems to be that they want more of the same and more variety of tests.

2.6 Future Plans

So far, PVC has been primarily used for modern, standard Persian. However, by tagging each verb, there is scope for adding obsolete and dialectal variants. There is also a need to add usage notes for each verb and deal with suppletive forms and

verbs requiring cross referencing of various kinds. All verbs should ideally be available in the colloquial forms as well as the current written styles. The developers have kept all these considerations in mind and the verbs are being tagged accordingly with these future uses in mind. It is hoped that one day, an audio component will also be available for each form.

Acknowledgements

The online version of PVC received inspiration and guidance from an earlier downloadable version based on different technology and different methodology and currently still available from its creator, Ali Jahanshiri at

http://alijsh.googlepages.com/pvc.htm

The developers of the online PVC gratefully acknowledge much help and guidance — sometimes on a daily basis — from Dr. Gernot Windfuhr, University of Michigan. We wish to also thank University of Texas Liberal Arts Instructional Technology Services for generously providing server space including the facility for long-distance collaboration involved in this project and in particular we would like to mention Justin Davis and Nick Lauland who keep the servers up and running and always respond promptly to all manner of emergencies.

References

John A. Boyle. *Grammar of modern Persian*. Wiesbaden, Harrassowitz, 1966.

Navid Fazel. 2006. *Academic Grammar of New Persian.* http://www.fazel.de/dastur/EN/index.html

Karine Megerdomian. *Finite-state morphological analysis of Persian*. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Coling 2004, University of Geneva. August 28, 2004.