

UNCLASSIFIED

MTR080206
MITRE TECHNICAL REPORT

MITRE

Analysis of Farsi Weblogs

**Karine Megerdooian
August 2008**

UNCLASSIFIED

MITRE

Analysis of Farsi Weblogs

**Karine Megerdooian
August 2008**

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

Approved for public release;
Distribution unlimited.

©2008 The MITRE Corporation.
All Rights Reserved.

UNCLASSIFIED

UNCLASSIFIED

Executive Summary

Description. This technical report is a compilation of the findings in the project entitled "Analysis of Farsi Blogs" (the project ran from April through October 2007). It includes an overview of the project, a survey of the literature on Persian (Farsi) language blogs, and a detailed description of the linguistic features encountered in Persian blogs created in Iran or in the Iranian expatriate community. It also contains the results of the statistical correlations between bloggers' gender, blog topic, and language style, as well as the effects of the conversational dialect on the results of automatic morphological analysis.

Problem. Persian websites have traditionally been written in the literary or formal language, which differs significantly from the standard conversational language. Within the last few years, however, the conversational variant of the language has been used in writing emails, chats and weblogs. With the exponential growth of Persian weblogs, in particular, the conversational language has become the main variant especially among the youth for writing journals, expressing personal opinions, and providing social criticism on the web. Nevertheless, existing computational systems are focused on the analysis of the literary variant and there is a great need for a systematic investigation of the linguistic characteristics and for computational systems able to process Persian blogs.

Insights. The main goal of the project was to improve the state of knowledge of conversational Persian text and provide tools that can help in the analysis of Iranian weblogs. The first six phases of the planned 14 phases have been accomplished providing several interesting results and deliverables: The study carried out in the project provides a better understanding of the language used in Persian blogs, which differs significantly from the literary text found in news sources, and shows that the level of ambiguity and variance is higher than previously anticipated. Findings show that there is a direct correlation between the number of conversational forms in blog text and reduced performance for traditional morphological analyzers built on literary text, suggesting ways to quickly improve the morphological system. In addition, statistical studies provide interesting evidence for correlations of gender and language, pointing to the topic of the blog post as the main factor in determining language style. A number of resources were also created within this project such as a corpus of blog material; two sets of annotated files, one tagged for Part-of-Speech and the other for individual blogger differences, topic, and language variant. A tagset for Persian and interface tools for annotating blog posts were also developed.

Prospects. Section 11 of this report discusses areas for future research and development. The remaining phases of the original project proposal (approximately 5 staff months of effort) were to include an analysis of boundary recognition issues and partial parsing of Persian blogs, relevant for event detection or entity extraction. In addition, the project proposes to implement the findings to improve state-of-the-art morphological analyzers and parsers for higher level applications such as Machine Translation or Information Extraction. Continued exploration of the sociological characteristics of bloggers, in analyzing a target audience, and specifically how language is used to unify subgroups could be used in applications of opinion analysis or for detecting emergent threats or ideas.

Table of Contents

1	Project Overview	1
1.1	Accomplishments	1
1.1.1	Deliverables	1
1.1.2	Resources created	2
1.1.3	Key insights	3
1.2	Prospects	4
2	Survey of the Literature	7
2.1	Summary	7
2.2	Introduction	8
2.3	Weblogstan: The Persian Blogosphere	10
2.3.1	Characterizing Weblogs	10
2.3.2	State of Persian Language Weblogs	11
2.4	Blogs, Politics and Society in Iran	13
2.4.1	Blogs and Censorship	14
2.4.2	The Evolution of a Political Consciousness	15
2.4.3	A New Public Space	15
2.4.4	Islamist Bloggers	17
2.5	Quantitative Research on Iranian Blogs	17
2.5.1	Demographic Profile	17
2.5.2	Content Analysis	18
2.5.3	Social Network Analysis	20
2.6	Computational Linguistic Analysis of Persian Blogs	25
2.6.1	Linguistic Features	25
2.6.2	The Vulgarly Debate	26
2.6.3	Computational Systems	28
2.7	Future Directions	29
2.7.1	Blogspeak: An Emerging Linguistic Genre	31
2.7.1.1	Linguistic Features of Netspeak	31
2.7.1.2	Social Networks and the Language of Blogs	33
2.7.2	Blogs and Individual Differences	34
2.7.2.1	Determining Personality Traits	34

2.7.3	Age, Gender and Language	35
2.7.4	The Intermediary Factor: Blog Topic	35
2.7.5	Applications for Persian Blogs	37
3	Individual Differences in Persian Blogs	39
3.1	Introduction	39
3.2	Data Selection and Preparation	39
3.3	General Statistics about Data	40
3.4	Analysis Results	41
3.5	Evaluation	42
3.6	Discussion of Results	45
4	Morphological Analysis of Conversational Text	46
4.1	Introduction	47
4.2	Background	47
4.3	Data Set and Annotation	47
4.4	Morphological Analyzer	49
4.5	Data Analysis	50
4.6	Evaluation of Representative Sample	54
4.7	Conclusion and Discussion	55
5	The Language of Persian Weblogs	57
5.1	Summary	57
5.2	Persian Blogspeak	58
5.3	Persian Writing System	59
5.4	Lexicon	61
5.4.1	Phonological Alternations	61
5.4.2	Lexical Items	68
5.4.3	Neologisms and Loans	69
5.5	Morphology	71
5.5.1	Introduction	71
5.5.2	Nominal Inflection	72
5.5.2.1	Plurals	72
5.5.2.2	Ezafe	75
5.5.2.3	Indefinite article & relativizing particle	77

5.5.2.4	Definite marker	77
5.5.2.5	Object marker	78
5.5.3	Complex Tokens	79
5.5.3.1	Pronominal clitics	79
5.5.3.2	Functions of pronominal clitics	83
5.5.3.3	Copula verb	85
5.5.3.4	Other complex categories	88
5.5.4	Verbal Inflection	90
5.5.4.1	Stems in conversational Persian	90
5.5.5	Other verbs:	93
5.5.5.1	Personal inflections	94
5.5.5.2	Simple tenses on the present stem	96
5.5.5.3	Simple tenses on the past stem	97
5.5.5.4	compound tenses on the past stem	97
5.5.5.5	Progressive	102
5.5.5.6	Passive voice	103
5.5.5.7	Negation	104
5.5.5.8	Progressive prefix	104
5.5.5.9	Subjunctive prefix	105
5.5.6	Morphotactics	105
5.6	Syntax	106
5.6.1	The verb 'to be'	106
5.6.2	The indefinite 'ye'	106
5.6.3	Topicalization	106
5.6.4	The uses of 'ke'	107
5.6.5	Dropped prepositions	108
5.6.6	Free word order	108
5.6.7	Other constructions	109
5.7	Other Orthographic Issues	109
5.8	Conclusion	110
Appendix A	Project Schedule	117
Appendix B	Table of Persian Letters and Transliteration	118

Appendix C	Persian Parts of Speech for Morphological Annotation Tasks	121
Appendix D	Additional References	126
D.1	Blog sites used in this document	126
D.2	A sampling of recent and upcoming conferences on weblogs	126

List of Figures

Figure 1 – Correlation of conversational entries and accuracy of POS tag per correctly aligned entry	4
Figure 2 – Correlation of conversational entries and ambiguity per system analysis	4
Figure 3 – Map of weblogs that include GeoURL data. <i>Source:</i> NITLE Blog Census (map courtesy of Paul Hastings)	9
Figure 4 – The Global Blogosphere: Posts by Language. <i>Source:</i> The Technorati State of the Live Web, April 2007	12
Figure 5 – The Global Blogosphere: Hourly Posts by Language. <i>Source:</i> The Technorati State of the Live Web, April 2007.	12
Figure 6 – Content analysis of Iranian weblogs. <i>Source:</i> Kelly and Etling (2008)	19
Figure 7 – Gender organized by subgroups in Iranian weblogs. <i>Source:</i> Kelly and Etling (2008)	20
Figure 8 – Number of comments on each day of the week. <i>Source:</i> Qazvinian et al (2007a)	21
Figure 9 – Time graph shows number of comments on each day with tagged outliers. <i>Source:</i> Qazvinian et al (2007a)	22
Figure 10 – Venn diagram for different links distribution. <i>Source:</i> Qazvinian et al (2007b)	22
Figure 11 – Social network map of the Iranian blogosphere. <i>Source:</i> Kelly and Etling (2008)	23
Figure 12 – Split of English-language Blogs based on Gender.	36
Figure 13 – Correlation of Gender and Blog Subjects in English-language Blogs.	36
Figure 14 – Distribution of the topics in the data	40
Figure 15 – Distribution of the gender in the data	40
Figure 16 – Distribution of the formality in the data	41
Figure 17 – Sample annotation output in XML from Callisto	48
Figure 18 – Correlation of conversational entries and accuracy of morphological analysis	52
Figure 19 – Correlation of conversational entries and accuracy of POS tag per correctly aligned entry	53
Figure 20 – Correlation of conversational entries and ambiguity per system analysis	53

List of Tables

Table 1 – Most discriminating word n-grams for detecting some moods; <i>Source: Mishne (2005)</i>	30
Table 2 – Spoken language criteria applied to BlogSpeak (adapted from Crystal 2001); <i>Source: Nilsson (2003)</i>	33
Table 3 – Written language criteria applied to BlogSpeak (adapted from Crystal 2001); <i>Source: Nilsson (2003)</i>	33
Table 4 – A-Priori output for the training data set	42
Table 5 – Confusion matrix for rule #1.....	43
Table 6 – Confusion matrix for rule #2.....	43
Table 7 – Confusion matrix for rule #3.....	43
Table 8 – Confusion matrix for rule #4.....	44
Table 9 – Confusion matrix for rule #5.....	44
Table 10 – Confusion matrix for rule #6.....	44
Table 11 – Chi Square significance test results	44
Table 12 – Groundtruth files ordered by percent of conversational forms in posts	49
Table 13– Preliminary results of morphological analysis test	50
Table 14 – Evaluation scores (after minor adjustment to alignment scores)	51
Table 15 – Breakdown of system analysis per category	54
Table 16 – System tags analyzed per POS category	55

1 Project Overview

The main goal of the "Analysis of Farsi Blogs" project was to improve the state of knowledge of conversational Persian (Farsi) text and provide tools that can help in the analysis of Persian-language weblogs created in Iran or in the Iranian expatriate community. Most existing work in computational analysis of Persian has been focused on processing of formal or literary writing, but there is a shortage of tools that can be used in the analysis of Persian conversational text found in weblogs, leaving this resource untapped. Furthermore, most textbooks of Persian focus on the literary variant of the language and a complete description of the linguistic features of the conversational language is not currently available. The aim of the project was to investigate the language used in Persian blogs and to provide a comprehensive description of the linguistic differences of conversational Persian text. Additional goals included the building of a corpus of blog material that could be used in further investigations and for training new systems, and the investigation of the effects of conversational Persian on the performance of a morphological analyzer developed for literary language.

The project was originally planned in 14 phases (see project schedule in Appendix A). To this day, 6 of the planned phases and one additional task (correlations of gender and formality) have been completed, and will be described in the following sections. The remaining phases were to develop a morphological analysis tool that would be able to process conversational wordforms, provide a study of partial parsing and entity extraction issues on conversational blog text, and develop preliminary rules for partial parsing of Persian weblogs. The MITRE team members were Dr. Tim Allison, Dr. Karine Megerdooian and Dr. Zohreh Nazeri. All deliverables as well as tools and resources developed are posted at the project Sharepoint site at <http://communityshare.mitre.org/sites/persianblogs/> and could be made available.

1.1 Accomplishments

1.1.1 Deliverables

The project had several deliverables:

- **Analysis of Persian Weblogs: A Survey of the Literature**

The survey of the literature on Persian blogs presents the state of the Iranian Blogosphere and provides a review of the research and publications on the topic: Research on Persian blogs has mainly centered around a socio-political study of this new medium, and several quantitative investigations have provided preliminary studies on sociological characteristics and content analysis in weblogs. However, rigorous research and investigation of the linguistic aspects of Persian blogs and computational analysis of these online resources are lacking. This survey also presents a summary of the existing literature on the language of weblogs in English and discusses how the results may be relevant for a computational study of Persian Blogospeak.

- **The Language of Persian Blogs**

The report provides a first account of the language of Persian blogs that contain instances of both literary language and of the very distinct conversational Persian, sometimes mixed within the same blog post. It presents a comprehensive description of the orthographic, lexical, phonological, morphological, and syntactic properties of Persian Blogospeak. The

examples and patterns discussed show that there is a large amount of linguistic and orthographic variance encountered in Persian blogs. The diverse spellings also give rise to more ambiguity and provide new challenges for analyzing and processing online Persian documents.

- **Evaluation of Morphological Analysis on Informal Blog Text**

This report presents the results obtained from the application of a morphological analyzer originally developed for literary Persian text on blog posts. Accuracy and ambiguity measures of alignments and Part-of-Speech (POS) tags suggest that the presence of conversational forms in text do play a significant role in morphological analysis for tools developed primarily based on literary Persian, pointing to a direct correlation between the number of conversational forms and reduced system performance.

1.1.2 Resources created

- **Corpus of Persian Blogs**

For the purposes of the project, we downloaded blog posts from one of the major Iranian weblog servers, *blogfa*. The corpus after stripping off of html markers has a size of about 400 MB, consisting of about 88,000 files. It should be noted that the corpus has not been edited to remove junk material, advertisements, or poetry sites.

- **Annotated Corpus of Persian Blogs**

a) A small corpus of files annotated with *Part-of-Speech information* was prepared for the project to evaluate the performance of a morphological analyzer on conversational blog text. We downloaded 11 blog posts from three sites. Each post was manually annotated for Part-of-Speech information using the Callisto annotation tool (<http://callisto.mitre.org/>) and saved in the Callisto XML format. All posts were manually annotated by a Persian native speaker in a first round and checked and edited by a second annotator in the second round. The final ground truth contained 11 files of varying length, totaling 3193 words (some of which were compounds). Statistics on the number of conversational forms in each document was also computed.

b) Files labeled for *individual differences* such as bloggers' gender, topic, and language variant were created and used in the study of correlations between these three features. To ensure that the data used in the experiment represent the whole population, 22,000 posts were downloaded from 521 websites containing a variety of Persian blogs; 10% of these data were randomly selected for analysis and irrelevant posts such as poems, non-Persian text, and posts that did not contain enough text were eliminated. The final filtered data consisted of 1,012 records which were then reviewed manually and labeled for *gender* (female, male, unknown, or group), *morphology* (formal/literary vs. informal/conversational), and *category* (e.g., journal, art/culture, news, socio-political, technical).

- **Persian Tagset**

For the purposes of the morphological annotation, a tagset was developed specifically for Persian and integrated within Callisto. The tagset includes 37 annotation tags for Nouns, Adjectives, Adverbs, Verbs and closed class items such as Prepositions, Quantifiers, etc. Attempts were made to keep the number of tags low, hence the tags do not represent all possible morphology on a particular part-of-speech unless the information is important for higher-level tasks such as POS disambiguation or parsing. A complete description of the

tagset and relevant annotation guidelines can be found in Appendix C. The corresponding Callisto properties file is available on the project Sharepoint.

- **LabelIt Interface Tool**

LabelIt is a Graphical User Interface for browsing and labeling blogs. The LabelIt tool was developed using Microsoft Office ACCESS 2003. It provides an easy to use interface to browse Persian blogs and to label them based on their topic, formality of the language used, and gender of the author. The tool can be used for blogs in other languages and the labeling criteria can be modified by the user. The LabelIt tool and a user's guide are available on the project Sharepoint.

- **Various Scripts**

Several scripts were created for this project to

- (i) strip off HTML encoding on Persian blogs from the Blogfa server
- (ii) extract alignment and annotation information from the Callisto XML files serving as reference and provide automatic scoring of the performance of a morphological analyzer

1.1.3 Key insights

The survey of the literature clearly showed that not much work has been done in describing the linguistic differences of conversational text in Persian, and existing computational systems have not been developed for blog material. In addition, as part of the main goal of the project, a detailed investigation of the language used in Persian weblogs was carried out and provided as a deliverable. The study showed that blog text is more ambiguous than originally thought and that there exists a large amount of individual variation in terms of the language variant used in the post and the orthographic patterns encountered.

The preliminary results of the evaluation of a morphological analyzer developed for literary Persian text on blog posts point to a direct correlation between the number of conversational word forms found in the blog text and reduced performance of the system in terms of tag accuracy (Figure 1) and ambiguity of analyses (Figure 2). In addition, a closer examination of a representative sample shows that the majority of mistagged elements are of conversational form. A system that provides guesses based on the word forms can provide better results although the ambiguity is increased as more (unknown) conversational forms are encountered in text. Furthermore, the results show that the conversational forms missed most often are verbs and frequent elements such as the object marker *ro*. This suggests that the addition of verbal conjugation rules and including frequent wordforms for conversational language can significantly improve the results of analysis on blog text.

A statistical analysis of Persian blog posts suggests strong associations between the three features of gender, formality and topic. Past work on individual differences of bloggers in other languages has argued that women write in informal language while men prefer formal language. Our results suggest that the decisive factor on the language variant used is not gender but rather the topic of the document.

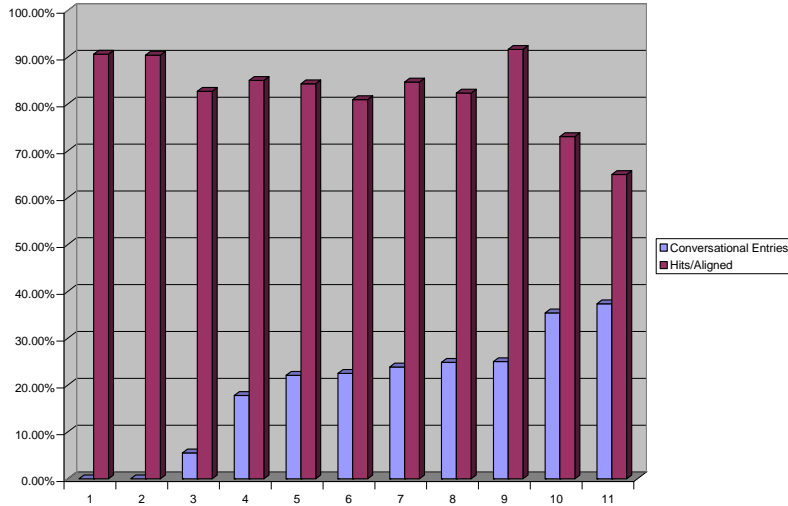


Figure 1 – Correlation of conversational entries and accuracy of POS tag per correctly aligned entry

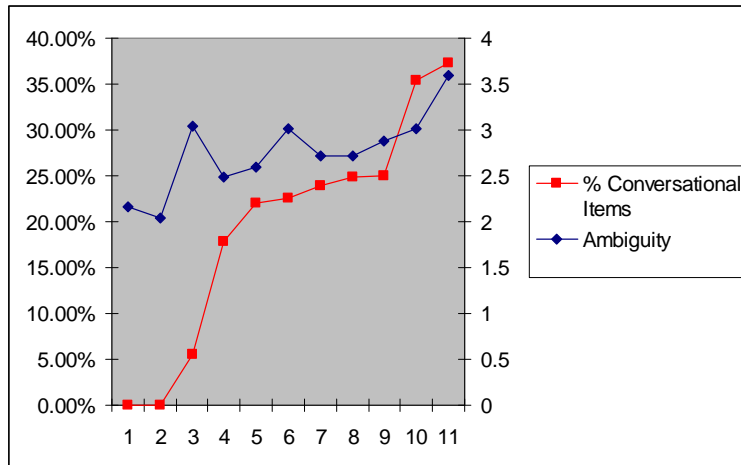


Figure 2 – Correlation of conversational entries and ambiguity per system analysis

1.2 Prospects

Although some foundational work has been carried out on online social networks of Persian blogs, and on their socio-political ramifications, there is much to be done in the field of Natural Language Processing (NLP). The many differences identified in blogs compared to newsprint and traditional online sources suggest that existing computational systems (such as machine translation, information retrieval or summarization applications) and their components (such as morphological analyzers or syntactic parsers) require extensions to cover Persian blog language. These extensions will ensure that the wealth of information contained in Persian blogs can be correctly and efficiently processed. The following phases of the original project plan, requiring 5 staff months, respond to some of these needs.

- **Extending an existing morphological analyzer to cover Persian blog text**

The fundamental distinction between literary and conversational text in Persian involves the morphology of the language. A morphological analysis tool that can process conversational text found in blogs and offer the literary Persian equivalent will be able to associate blog vocabulary with dictionary forms. This tool could help junior language analysts bridge the gap between their training in formal Persian and the requirements of their jobs, and increase their efficiency in analyzing Persian blog material. The extended analyzer could also be a first step towards automatic applications such as entity extraction or machine translation. Based on the results of the evaluation performed, we believe that the extension of the morphological analyzer could be performed in stages as some modifications (e.g., rules for the verbal paradigm) seem to deliver more value than others. Each stage can be followed with an informal evaluation and cost/benefit analysis to show the effect of the modifications and provide strategies for the improvement or adaptation of other existing systems.

- **Analysis of entity extraction issues on conversational blog text**

One of the major issues in automatic processing of Persian text is the difficulty in detecting word boundaries due to the ambiguities of the Perso-Arabic script. New orthographic variants used in conversational text and the high level of ambiguity at the level of morphological analysis of blogs suggest that a higher level of analysis such as parsing of the text would be crucial in order to disambiguate the output. We propose to investigate the issues for Persian entity extraction with respect to noun, preposition and verb phrases encountered in blog text, and to propose the most appropriate strategy for partial parsing of conversational Persian.

- **Development of partial parsing rules**

Based on the analysis performed in the last step, develop rules for parsing of noun phrases, preposition phrases, and basic verb phrases encountered in conversational Persian blogs. The partial parsing output will be evaluated with detailed error analysis, including the impact of the modifications made. The results could be used in extending existing systems to cover Persian blogs.

- **Automatic classifiers**

The blog posts that have been labelled for the level of formality can be used to train an Automatic Classifier to separate Persian documents based on the style of language used. This would allow for a categorization of documents that is relevant to the knowledge of language analysts (e.g., heritage speakers would be better at analyzing conversational language than high-level, formal writing). In addition, classifiers can be trained on language features to successfully categorize blogs by topic or ideology.

- **Uncovering social group dynamics in the blogosphere**

Further exploring the research on the individual characteristics of Iranian bloggers, we can develop automatic tools to successfully wade through the information flood of blogs by detecting leading issues, credible posts, and influential people. By combining deeper linguistic features with social network analysis of the Iranian blogosphere, we can enhance the study of online social groups and the behavior of their members. This research and resulting tools will allow analysts to focus on the most influential blogs, to gain a deeper and

broader contextual understanding of social groups represented in blogs and to better harness the information to identify developing threats or impending activity.

- **Dissemination of resources and findings**

Effort will be made to put the findings from the project and resources created in forms that can easily be disseminated to the NLP and intelligence community working on Persian. The deliverable describing the linguistic properties of blog text and features of conversational Persian can be provided as a handbook that could also be used for training of language analysts or for high-level language training in classrooms; the corpus of blog material and annotated files could be used in further investigations and for training of new systems; and the computational systems for morphological analysis and partial parsing of conversational text could be used either directly as a tool for language analysts or as a crucial component within a larger automatic application, such as for entity extraction, event analysis, machine translation, or sentiment detection.

2 Survey of the Literature

2.1 Summary

This survey presents a comprehensive overview of the literature studying Persian-language weblogs created in Iran and within the Iranian expatriate community. As the number of Internet users in Iran has increased since 2001, the number of Persian language blogs (websites where entries are made and displayed in a reverse chronological order) has also undergone a dramatic growth making Persian one of the top ten languages of the global Blogosphere. The exponential growth of Persian blogs has attracted much attention in the international media which has described how marginalized groups in Iran, such as the youth and women, use blogs to evade the strict regulations imposed upon them by the state, express their thoughts and opinions on the political and social situation, coordinate or influence political activities, or record their daily lives. Most research on the Persian-language blogs has focused on a socio-political study, yet rigorous and systematic investigation of the linguistic characteristics and the development of computational systems able to process Persian blogs have been quite rare. The survey presents the existing literature and suggests areas for further exploration of Persian-language blogs.

Socio-political analyses aim to capture the developing political consciousness of the young generation. Demographic analysis of bloggers suggests that most are computer-literate, university-aged youth, and live in urban settings. The growth of Iranian weblogs has been attributed to state censorship over traditional forms of media: Blogs provide a relatively free space – despite government attempts at filtering – for reformists, journalists, authors, women and young Iranians to discuss taboo issues. The very strict rules of conduct imposed by the Islamic government and societal norms have created a strong divide between the outside world or the *public sphere* and the home or the *private sphere*. The separate behavior in the two realms has given rise to an intensive sense of duality in Iranian society, but the advent of blogs and the possibility of blogging anonymously have blurred this boundary by providing a new public space where bloggers are able to discuss issues traditionally confined to the private realm, and to develop new friendships and support networks. In turn, this has allowed the youth and women in particular to create a new sense of self and identity. Interestingly, some of the attacks on young bloggers have come from intellectuals and journalists who have criticized their conversational style of writing, the disregard for orthographic and grammatical standards, and the fact that they "place issues ahead of analysis", paralleling similar reactions in the Western world. Researchers have therefore noted a clash between the intellectual elite tied to traditional institutions of print and the young "nonintellectuals" calling into question their linguistic and cultural authority. This debate has been analyzed as the struggle of an emerging new elite from the professional and young middle class, distinct from both the state authority and the traditional class of language authority.

The intense interconnection at the core of the blogging community through the generous use of hyperlinks and comments has given rise to research on blogs from the perspective of social network analysis (SNA). Studies of Persian language blogs suggest that the linking patterns of bloggers are not homogeneous; there exist instead several smaller communities sharing common interests. This, combined with content analysis of blog posts, reveals that the Persian-language blogosphere is dominated by four major poles, each with its own interesting structural and social characteristics: 1) *Secular/Reformist*,

2) *Conservative/Religious*, 3) *Persian Poetry and Literature*, and 4) *Mixed Networks* (i.e., no particular issue or ideology and varied interests). In addition, focusing on the distribution of comments over time, researchers have detected temporal bursts of activity in the blogosphere during major political or cultural events. SNA also helps identify the A-list – popular and authoritative bloggers – and the effect they have on the blogging community.

Preliminary exploration of the language of Persian blogs shows parallels with English Bloggspeak. The main characteristic of blog language is the use of a conversational style in writing. Non-standard spelling that reflects the colloquial pronunciation of words is often used. Blog entries are usually written in short sentences and include a large number of hyperlinks. Deviant spelling is common and standard orthography is often ignored, opting instead for a more intimate style. Emotions are expressed with emoticons, repetition of letters and punctuation marks, the use of ellipsis, and special symbols and capitals (if available in the language) for emphasis. Jargons and neologisms abound in Bloggspeak, especially based on technical or computer-related terms. In addition to all these features, Persian Bloggspeak in particular has some properties corresponding to the conversational language such as shortened verbal stems, frequent use of attached pronoun forms, and affixes that are not part of the standard formal grammar. There are more instances of free word order, idiomatic expressions, loan words, and an inordinate amount of orthographic variance partly due to the flexibility and ambiguity of the Perso-Arabic script. Investigation of Bloggspeak points to the emergence of a new variety of language which combines elements from both spoken and written forms of communication, as well as possessing novel characteristics related to its technological context.

Although some foundational work has been carried out on Persian Bloggspeak, on online social networks of Persian blogs, and on their socio-political ramifications, there is much to be done in the field of natural language processing. The many differences identified in blogs compared to newsprint and traditional online sources suggest that existing computational systems (such as machine translation, information retrieval or summarization applications) and their components (such as morphological analyzers or syntactic parsers) require extensions to cover Persian blog language. These extensions will ensure that the wealth of information contained in Persian blogs can be correctly and efficiently processed. The review of the literature on English-language blogs suggests several areas for future research and development including the need for a comprehensive analysis of Persian Bloggspeak in conjunction with empirical investigations on blog corpora, continued exploration of SNA and the sociological characteristics of bloggers and specifically how language is used to strengthen and unify subgroups within the Iranian blogging community, which may even result in the possible identification of future opinion leaders.

2.2 Introduction

In September 2001, a young Canadian-Iranian journalist named Hossein Derakhshan established one of the very first weblogs in Persian. Derakhshan also created a how-to guide for setting up Persian weblogs which, combined with the appearance of blog hosts dedicated to Persian language and a somewhat less strict atmosphere created by the then-ruling reformist government in Iran, gave rise to the explosion of the Iranian blog community. Persian is now among the top ten languages in weblogs in the world. The exponential growth of the Persian blog community has been attributed to the strict state censorship of print media in Iran: Blogging provides a safe space for Iranians, in particular the youth, to write about a wide variety of topics. Bloggers can express their thoughts and

opinions on the political and social situation in Iran relatively freely, coordinate or influence political activities, or record their daily lives in these online journals. In addition, expatriate Iranians often use blogs to communicate with people in Iran.

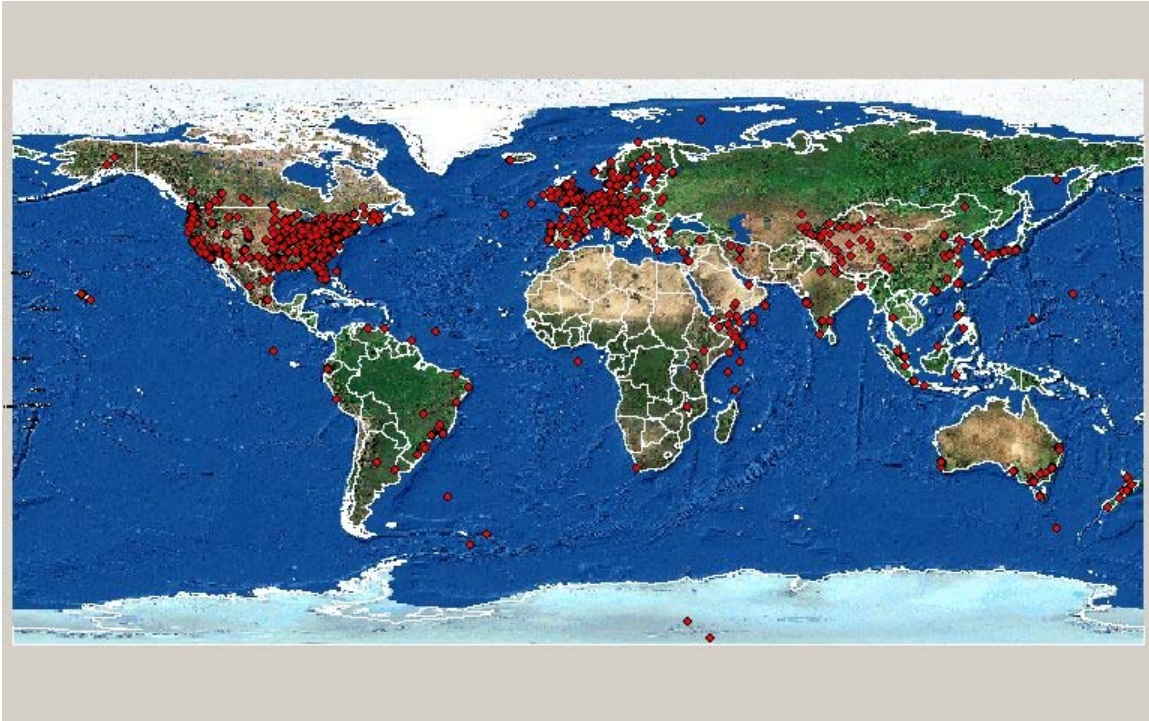


Figure 3 – Map of weblogs that include GeoURL data¹.
Source: NITLE Blog Census (map courtesy of Paul Hastings)

The volume and power of blogging has not gone unnoticed by the Iranian regime. In April 2003, Iran became the first government to imprison a blogger for views expressed online when Sina Motallebi, a reformist journalist blogger was detained by the authorities. Since then, an estimated 28 bloggers have been imprisoned on various charges². As a result, an increasing number of blogs are now written anonymously. The Iranian government has also been targeting weblogs through sophisticated filtering systems. Despite these obstacles, the Internet has become an important information medium in Iran. The country has experienced dramatic growth in Internet usage from one million users in 2001 to 18 million in 2006 – from a total population of nearly 66 million – placing it above Australia, Sweden, or Switzerland (CIA World Factbook 2008a, 2008b), and polls indicate that people trust the Internet more than any other media outlet (OpenNet 2006). This growth is partly due to the fact that two-third of the population is under 30 (the median age is 26) and many are

¹ Since registering one's blog location with GeoURL is on a voluntary basis, this map is not representative of all existing and active blogs in the world.

² This number is based on an estimate by Persian Impediment, an initiative of Article 19 (www.article19.org). The 2008 report by World Informational Access puts the number at only 8 with the following caveat: "The real number of arrested bloggers is probably much higher, since many arrests in China, Zimbabwe, and Iran go unreported in the international media."

technologically savvy, while the country's literacy rate is 77% – and 83.5% for males (CIA World Factbook 2008b). The Persian blog community now provides a powerful source of information on Iranian society both within Iran and to the outside world. Reporters Without Borders have noted that "the Internet has grown faster in Iran than any other Middle Eastern country since 2000 and has become an important medium, providing fairly independent news and an arena for vigorous political discussion" (RSF 2004). According to Nasrin Alavi, "through the anonymity that blogs can provide, those who once lacked voices are at last speaking up and discussing issues that have never been aired in any other media in the Islamic world." (Alavi 2005; p.6)

Not surprisingly, a large number of articles have been published discussing the unprecedented popularity of Persian language blogs and their impact on society and politics in Iran. The informal style used by authors of weblogs has also given rise to a number of critical commentaries, especially from intellectuals and professional journalists. Research on Persian blogs has mainly centered around a socio-political study of this new medium, and several quantitative investigations have provided preliminary analyses on sociological characteristics and content analysis in weblogs. However, rigorous research and investigation of the linguistic aspects of Persian blogs and computational analysis of these online resources are lacking. This survey provides an overview of the literature on Persian-language blogs. It also presents a summary of the existing literature on the language of weblogs in English and discusses how the results may be relevant for a computational study of the language of Persian blogs. The focus of this survey is on weblogs written in Persian either within Iran or in the expatriate Iranian community and it does not address Afghani or Tajiki sites.

2.3 Weblogestan: The Persian Blogosphere

2.3.1 Characterizing Weblogs

A weblog or blog is a website where entries are made and displayed in a reverse chronological order, with the most recent post featured most prominently. Typically, weblogs are published by individuals, their style is personal and informal, and are frequently updated. Attempts have been made to define weblogs based on features not regularly found on traditional websites such as comments or trackbacks, or based on the genre or topic discussed in blogs. However, there is a great variety in the features used or the content of the blog, resulting in very distinct weblog types depending on the author. In this section, I will review some of these optional attributes.

Weblogs have a number of optional features which can be turned on or off at the discretion of the authors. They may combine text and images, and often use a generous amount of links to other blogs and websites, allowing readers to track conversations between blogs. Trackbacks (which allow authors to obtain notifications when another site links to one of their documents), permalinks (URLs that point to a specific entry even after the entry has passed from the front page into the blog archives), and blogrolls (a collection of links to other blogs, often found on the front page sidebar) are generally common in blogs but may be omitted. In addition, weblogs often provide the ability for readers to leave comments in an interactive format. Although weblogs were traditionally maintained by a single author, there are now a number of group or multi-user blogs available. The frequency of updates is another variable: while some blogs may be updated daily, others may remain untouched for a long time.

The contents of weblogs vary widely. Some weblogs provide commentary or news on a particular subject such as politics, literature, travel, fashion, technology, or legal issues. Other blogs function more as personal online diaries, focusing on personal experiences or reflection. A single blog may also touch on various issues, not allowing the reader to classify it under one category. The media type represented in blogs can also vary — although most weblogs are textual, they may focus on displaying photography (photoblogs), sketches (sketchblogs), videos (vlogs), music (MP3 blogs), audio podcasts, or even spam or fake articles (splogs). It has also been very difficult to describe weblogs as a specific genre since they may display diverse writing styles, and present very differing perspectives.

Ó Baoill (2004) therefore proposes three distinct features for a classification of weblogs that together can provide a better identification of a particular blog site: (1) format of the blogs; (2) the audience or content of the blog; and (3) organization of the weblog as a hobby, an income-generating operation, or professional endeavor.

According to Nabavi (2004), one main characterizing trait of the blog, in particular of Iranian blogs, is the fact that authors of weblogs may remain anonymous. This, combined with the fact that Iranian blogs are not a professional medium and thus are not subject to the same governmental controls, strongly shape the structure and content of Persian-language blogs. Another important feature that differentiates blogs from other online publishing tools such as traditional websites is the "intrinsic ability to allow individuals to write and publish multimedia content on the web without needing to know anything about HTML or about any other technical issue" (Good 2005). These traits have strongly contributed to what could be considered the most significant phenomenon in end-user content creation on the web.

2.3.2 State of Persian Language Weblogs

There is some discrepancy as to the exact ranking of Persian in terms of blog languages, mostly due to the different metrics used and the blog services included. In 2004, based on data from BlogCensus, the Times Online³ and The Guardian UK⁴ mentioned that Persian is the fourth largest blog language in the world. Alavi (2005) claims that Persian is the third most used language in weblogs but no source is quoted. In 2006, the blog search engine Technorati reported that "Farsi has pushed its way into the top 10 languages in use in the blogosphere, bumping Dutch, which had held the number 10 spot over the last couple of quarters, into the number 11 spot" (Sifry 2006). According to the 2007 report by Technorati, Farsi has maintained its 10th place in terms of blog languages as shown in Figure 4 (Sifry 2007). The latest data from NITLE Blog Census, however, place Persian in the 9th rank (NITLE Census 2007). Technorati is currently tracking more than 71 million blogs and NITLE has about 3 million indexed blogs. It should be noted that these sites use automatic language identification software, the metrics are based on the number of posts or frequency of updates, and they may not include some of the larger blog services for a particular language, all of which may affect the results of the rankings.⁵

³ <http://www.timesonline.co.uk/tol/comment/article390484.ece>

⁴ <http://technology.guardian.co.uk/online/weblogs/story/0,14024,1377538,00.html>

⁵ To find out more about the methodology used by Technorati and NITLE Blog Census, the reader is referred to <http://www.sifry.com/alerts/archives/000433.html> and <http://www.hirank.com/semantic-indexing-project/census/methodology.html>, respectively.

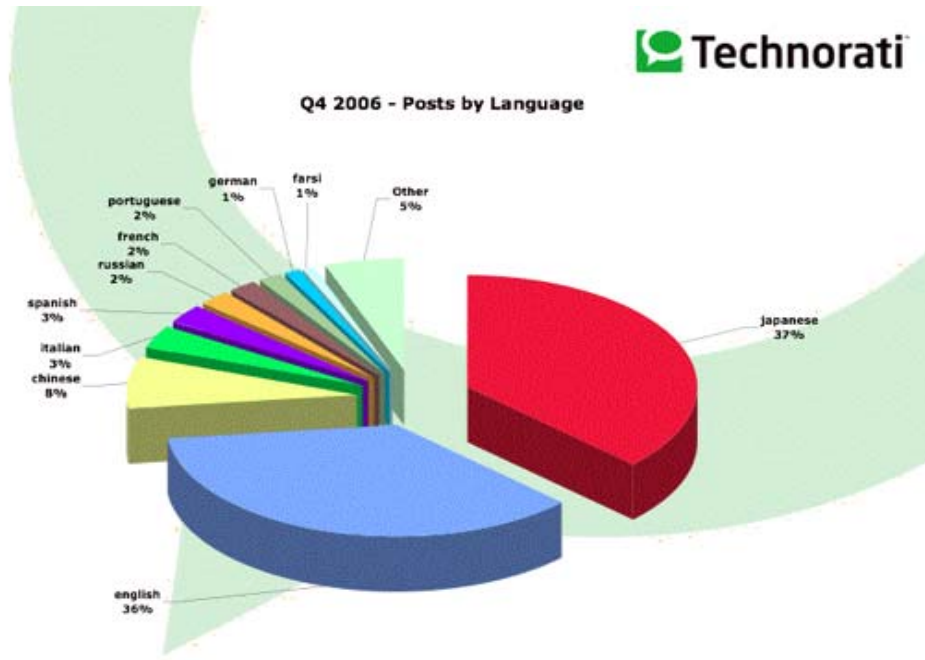


Figure 4 – The Global Blogosphere: Posts by Language. *Source: The Technorati State of the Live Web, April 2007*

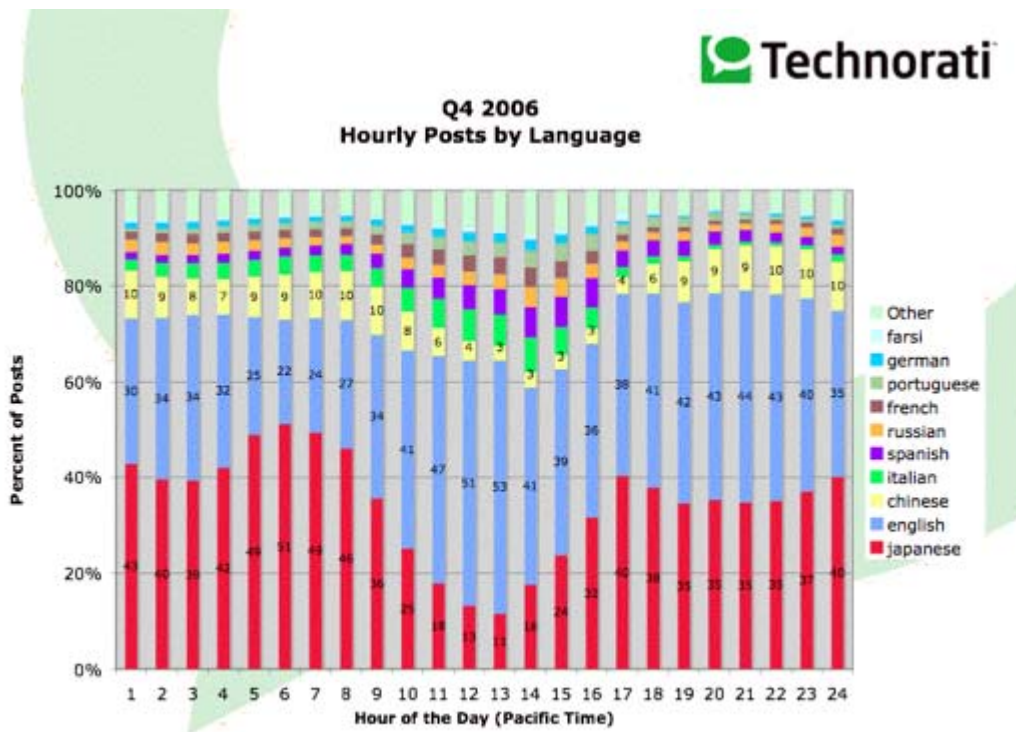


Figure 5 – The Global Blogosphere: Hourly Posts by Language. *Source: The Technorati State of the Live Web, April 2007.*

Originally, some of the features of Persian language, such as encoding, font, and right-to-left direction, hindered the creation of weblogs. Shortly after Hossein Derakhshan posted instructions on creating Persian language blogs (Derakhshan 2001) and the launching of Persianblog.com, the first weblog host in Iran, Persian blogs began appearing both in Iran and in the expatriate community. Currently, there are several hosts for Persian language blogs like Blogfa, Mihanblog, Parsiablog, Persianblog, Blogsky, and more recently Muslimblog.ir. Persianblog is the largest and oldest blog host in Iran and includes more than 50% of Persian blogs (Sheykh Esmaili et al 2006).

2.4 Blogs, Politics and Society in Iran

Although there has been a lot of discussion in international media on how weblogs are being used by the reformist journalists, by the youth and by women in Iran, the pro-government and conservative blogs have been rather ignored. Within the last couple of years, new weblogs have emerged that are either pro-establishment, fundamentalist, or conservative. Several mainstream politicians have established their own Persian language weblogs, such as President Ahmadinejad⁶ or Mohammad Ali Abtahi⁷. Recently, the new on-line community of the Iranian Muslim Bloggers' Association⁸ was announced. The pro-active response of the establishment and of conservative groups and individuals in Iran to counter the reformist views expressed on Iranian weblogs is still a new phenomenon and has not been discussed in detail. Derakhshan described the multifaceted nature of Persian blogs in an interview with the Columbia Journalism Review Daily:

From trendy art students in the north of Tehran to young clerics in the holy city of Qom, it's very mainstream. There are secular, anti-religious, anti-regime people and there are also some fundamentalist-supporting revolutionary guards. (Beckerman 2006)

Despite the pervasiveness and variety of Persian-language blogs, the majority of published articles have studied the social and political aspects of the weblogs of reformists, of marginalized groups such as women and the youth, as well as blogs by Iranians in exile.⁹ At the center of the discussion is the *duality* of Iranian society. Due to very strict rules of conduct imposed by the Islamic state and societal norms, Iranians learn from childhood to separate the home or the *private sphere* from the outside world or the *public sphere*. Children learn at a young age to lie about their parents' praying habits, their sibling's political activities or even their favorite book, lest they be judged or imprisoned. Along with this schism comes a "dual language" (Behrouzan 2005) that affects all aspects of a person's daily

1 October, 2003

As a nation we all have dual personalities ... At home we are as free as can be ... we have fun, drink, have parties ... and pay no attention to the religious dictates of the Supreme Leader. But in public we are forced to act devout and show support for the regime...This has destroyed our culture and has turned us into the worst kind of hypocrites ... and as a society we are rotting from the inside.

Website: fozool.blogspot.com

(Extracted from Alavi 2005)

⁶ <http://www.ahmadinejad.ir/>

⁷ Abtahi was the Vice-President and advisor to former President Mohammad Khatami. His weblog can be found at <http://www.webneveshteha.com/>.

⁸ <http://www.muslimbloggers.ir/>

⁹ It should be noted that despite all the studies on the blogs of marginalized groups in Iran, publications on the weblogs of religious and ethnic minority groups are tangibly absent.

life; a subtle language that does not directly express one's true feelings or thoughts, where one needs to "develop an ear for the whisperings of irony and an eye for the traces of paradox" (Behrouzan 2004). The advent of weblogs has managed to blur the divisive line between the public domain and the private realm for young Iranians who take advantage of the anonymity offered by this new medium to discuss taboo subjects and to express what they feel or think with relatively little fear of being judged or punished. The interactive nature of blogs allows bloggers to create an online network where participants develop friendships and often provide support for each other. This new public sphere allows bloggers to develop a sense of self and identity that they cannot display in the real public domain in Iran. The publications surveyed in this section discuss these very issues and use Iranian blogs to present a perspective on the changing consciousness of the youth in Iran.

2.4.1 Blogs and Censorship

Blogs as the new medium. Nabavi (2004)¹⁰ attributes the growth of Iranian weblogs to the crisis in mass media in Iran, arguing that bloggers have turned to this new medium as a reaction to the government monopoly, judicial control, and censorship over more traditional forms of communication. In addition, blogs have slowly taken the place of societal structures that are absent in Iran. Nabavi points out that most Iranian bloggers do not overtly discuss politics and rather focus on personal, literary or cultural issues. In clear contrast, bloggers writing anonymously freely discuss politics and other taboo issues and "what has never been put on paper in any Persian writing now finds its way into written literature". Nabavi feels, however, that the main influence of Persian language weblogs will be in the development of identity and individuality for the bloggers and the youth before it can become a force in bringing about social change in Iran. But more importantly, the main purpose of Iranian blogs is to claim a freedom of expression which is expressed in Derakhshan's blog title: "Editor, myself".

Effect of censorship on blogs. Jensen (2004) presents a case study on English language weblogs by Iranians to study the effects of state censorship on this new medium. The study tracks 20 English or bilingual blogs with a focus on socio-political issues (10 of which are maintained by Iranians in Iran and the others are authored by Iranians living in Western countries), between April and September 2004. Jensen finds that there were many references by bloggers to increased Internet censorship in Iran during that timeframe, which also dramatically affected the readership of outside blogs. However, the attempts to censor websites are not always successful in Iran since the network is not centralized and bloggers use various technical methods such as proxies, mirror sites and RSS readers, to surpass filters. Not surprisingly, Jensen's study suggests that bloggers inside Iran tend to be more anonymous and reveal less about their identities than outside bloggers. The study shows that there is a significant cross-linking between bloggers inside and outside of Iran, building a social network despite the geographic boundaries.

It should be noted that this case study includes only a small sample of English-language weblogs by Iranians, which is itself a minority in the Iranian weblog community. Hence, although this thesis points to certain trends in Iranian blogs, it is not possible to extrapolate the conclusions drawn from such a small percentage of the whole structure. Furthermore,

¹⁰ Seyyed Ebrahim Nabavi is a sociologist, prolific satirist, writer, and journalist. His articles were published in various reformist newspapers. He was imprisoned twice for his political satire. He currently lives in Belgium.

as Jensen notes, the censorship situation for English language blogs is arguably very different from that of Persian language sites. Jensen's main conclusion is that the Internet "is not a magic vessel and it does not exist in a political and cultural vacuum." State censorship through the control of the access to the Internet and the constant crackdown, although not 100% effective, does result in the intimidation and discouragement of bloggers. According to Jensen, the Iranian Internet and blogs in particular are playing a slow-paced yet significant role in facilitating democratization.

2.4.2 The Evolution of a Political Consciousness

Iranian society through weblogs. Alavi (2005)¹¹ is the first book dedicated to a discussion of Persian blogs. The main premise of the book is that weblogs offer a unique perspective on the changing consciousness of the Iranian youth and present an image of Iranian society that is rarely portrayed in the mainstream Western media. Nasrin Alavi uses posts from actual blogs in English translation to paint the issues that the younger generation discusses

November 17, 2004

I keep a blog so that I can breathe in this suffocating air. In a society where one is taken to history's abattoir for the mere crime of thinking, I write so as not to be lost in my despair, so that I feel that I am someplace where my calls for justice can be uttered. I write a blog so that I can shout, cry and laugh, and do the things that they have taken away from me in Iran today.

Website: lolivashaneh.blogspot.com

(Extracted from Alavi 2005)

and struggles with, such as the conflict between globalization and maintaining old traditions, the conflict between adopting a Western lifestyle and the strict Islamic code that needs to be followed, and social issues such as drug addiction and unemployment. Alavi provides the social and political backdrop for the discussions in the weblogs by presenting an overview of the history of modern Iran. Other issues that are discussed heavily in the

selected blogs are the problems of censorship in Iran and the arrest of dissidents, especially bloggers; women's rights; struggle for reform and human rights; and arts and culture. The book manages to present a portrait of the Iranian youth and the challenges they face by using their own words as the primary source.

Another socio-political study is Golkar (2005) which states that 91% of the Persian language blogs are written by youth between 13 and 30 years old and argues that, given the lack of freedom in expressing one's opinions in an overt interview, weblogs provide a useful medium for gaining insight into the political consciousness of the young generation in Iran. The article presents a very qualitative summary of some of the political discourse on the blogs.

2.4.3 A New Public Space

Blogs and the public sphere. Alexanian (2006) investigates the blogs of several Iranian immigrants living in Orange County, California, and their motivations for blogging, in order to study how they contribute to the formation of Iranian communities and identities. There is a very poignant separation between the *public* (outer) and *private* (inner) spheres for the youth in Iran given all the state-imposed restrictions such as the ban on drinking, dancing,

¹¹ Nasrin Alavi spent her formative years in Iran, attended university in Britain and worked in London, and then returned to her birthplace to work for an NGO for a number of years. She has taught engineering in Britain and the USA, and now lives in the UK.

or dating in public spaces. The blogs, however, represent a space where the bloggers can be themselves. Alexanian argues that blogs blur the boundary between the *public* and *private* spheres and reconfigure the meanings associated with each of these realms.

Urban sociologist and geographer Masserat Amir-Ebrahimi's (2004)¹² article studies the socio-cultural ramifications of blogs in Iran through an investigation of Persian language weblogs, focus groups and interviews with bloggers.¹³ In every society, there exist certain social masks in public spaces. Since the establishment of the Islamic Republic, however, the behavior of Iranians has become strictly regulated in the public space with laws and societal pressures governing appearance, body language, and speech. The urban women and youth, in particular, have since formed a new appearance and demeanor as a reaction to the strict codes imposed upon them. Amir-Ebrahimi notes that "two decades of continuously playing contradictory roles in different spaces [...] has led to a kind of identity crisis, especially among youth whose only lived experience has been under the Islamic Republic."

For many bloggers, the weblog becomes a mirror into their souls; a place where they represent their true selves and define themselves according to their liking, without the social and cultural constraints that impede them in real spaces. For women, who are constantly playing roles in a moralistic society, this takes on added significance. The internet and weblogs become a mirror in which youth and women can see their "hidden selves" and/or "repressed selves." (Amir-Ebrahimi 2004)

The virtual space of weblogs provides a space for the youth to redefine their "self" and to shape their repressed identities through writing as distinct from the one prescribed by society. In democratic societies, this hidden self tends to be one that the individual does not reveal due to psychological or social impediments. In the case of Iran, this identity is one that the bloggers have been forced to repress due to legal and cultural limitations. Interestingly, Amir-Ebrahimi notes that women bloggers tend to use a pseudonym more often than men and present a more guarded persona in weblogs. In her analysis, this is due to the fact that men are less obliged to play a predetermined role in the physical public space whereas women's behavior and speech is heavily guarded by tradition and judged by society.

Although Anderson (1999) mainly analyzes media in the Arabic world, it does contain

20 July 2003

Has everyone noticed the spooky absence of graffiti in our public toilets since the arrival of weblogs? Remember the toilets at university we used to call our 'Freedom columns'?

Website: python.persianblog.com

(Extracted from Alavi 2005)

certain discussions that are directly related to the state of the blogosphere in Iran. Anderson argues that Internet writing does not represent a new genre as has been suggested in the literature (see Section 2.6 for a discussion). In his view, the Internet is a continuum of existing forms: what

¹² Masserat Amir Ebrahimi is an associate researcher at 'Le Monde Iranien' CNRS in Paris. She holds a PhD in Human, Economic and Regional Geography from the Université de Paris X-Nanterre. She was the executive coordinator of the [Atlas of Tehran Metropolis](#) (Published in 2005 in Tehran by TGIC) and a 2006-2007 Nikki Keddie-Balzan Fellow in the Department of Sociology and Geography at UCLA. She has worked extensively on Tehran, particularly on the Southern parts of the city, and is currently conducting research on gender and public spaces in Iran.

¹³ This article is part of a larger research project, "Authority and Public Spaces in Iran", assisted by an International Collaborative Research Grant from the Social Science Research Council's Program on the Middle East and North Africa. The results of the study are still to be published.

previously circulated in smaller, face-to-face interpersonal settings such as coffee houses, university dormitories, or dissident cells, has now moved into the virtual public space. This new form of information is less like the traditional centralized communication where the information was presented in a top-down, asymmetric model, and more like the decentralized form of face-to-face communication structured rather horizontally where there are nearly as many senders of information as receivers. According to Anderson, the medium of "blurred genres" dramatically lowers the barrier between the public and private realms and its effects are not unlike the effects of various media introduced in the past such as printing, desktop publishing, or home-produced tapes.

2.4.4 Islamist Bloggers

Research on weblogs where Islamic or political Islamist values and beliefs play an important role has been strikingly absent. In an article, Tehrani (2007) discusses the creation and rise of the Muslim Bloggers Association (MBA) in Iran, primarily aimed at countering secular and feminist bloggers and promoting conservative Shiite Islamic values. According to the author, MBA bloggers want to "export revolution" and are anti-American and anti-Zionist. Tehrani argues that "Iranian Islamist blogs probably provide one of the best places to learn information and news about power and state-related issues in the Islamic Republic, because some of their writers have close ties with Iranian leaders and some of them even are leading figures in the regime."

22 July 2008

These days, whoever makes people disillusioned and distrusts himself, the officials, and the future ends up helping the enemy. Today, whoever promotes dissent helps the enemy of Iran. Those who have a pen, who have a voice, who have a rostrum or a position should be careful; they should not allow the enemy to take advantage of them. Psychological warfare is the most important aspect of the enemy's struggle against the people of Iran.

Website: hamedtalebi.blogfa.com
(Muslim Reporter)

Tehrani also points out that the Islamist blogs are not monolithic and represent a diversity of opinion. This is confirmed by a case study carried out by Harvard University's Birkman Center for Internet and Society that finds that "criticism of government policies and leaders is routine, even among conservatives, though among the latter there is a clear conceptual separation between the *government* and the Islamic Republic/Supreme Leader" (Kelly and Etling 2008). The study describes several subgroups within the conservative pole: while one cluster tends to focus on religion and Shiite Islam, another group discusses socio-political issues like Iran's poor economic condition, gas prices or the nuclear issue. Crucially, these sites include both criticism as well as support of Ahmadinejad's leadership.

2.5 Quantitative Research on Iranian Blogs

2.5.1 Demographic Profile

Halavi (2006) investigates the pervasiveness of the Iranian blogging phenomenon using an online survey with the goal of providing quantitative data on the demographics, habits, and experiences of Iranian blog-readers. He conducted an online survey of the readers of a number of weblogs written by Iranians living inside as well as outside of Iran, consisting of 36 questions on the demographic profile and blog-reading habits of the respondents. The survey lasted four months (from 1 November 2005 through 4 March 2006) and resulted in 325 responses. The survey collected data on a number of attributes such as age, gender,

ethnicity, language skills, religion, education and socio-economic status, blog-reading habits, level of Internet access and censorship experiences.

The results show that most bloggers are members of the lower to upper middle class, have completed at least some post-secondary education while a large number have attended university. The overwhelming majority is between the ages of 20 to 32. The Iranian blog-readers are fairly computer-literate, university-aged youth, and live in urban settings. There were, however, certain issues with the way this survey was conducted. Since the online survey was voluntary and anonymous, control factors are weak and the validity of responses could not always be verified. The sample collected is quite small and not controlled for the demographics, thus is not truly representative of the vast Iranian blogging community. Nevertheless, Halavi's work is a first step towards a collection of quantitative demographic data on the readership of the Iranian blogosphere.

2.5.2 Content Analysis

Nabavi (2004) studies the content of forty popular Persian language weblogs on a single day and lists the following topics:

1. Personal issues and reflection (26%): daily events, travelogs, personal thoughts.
2. Discussion of issues related to other bloggers, the Internet and computers (19%): computer issues comprise 11.5% of this category, protests of bloggers' arrests consist of 5.8% of the writings, and the remaining posts discuss other bloggers' issues.
3. Political issues (32.3%): international and middle eastern events are usually discussed in these posts. On the particular day the study was run, the main issues were Arafat's death and problems in Palestine (12.7%) and US foreign policy in Iraq and Afghanistan (7%). Foreign political news comprised 25.4% of the discussions while internal politics consisted of only 7% (most of it dedicated to arrests of bloggers).
4. Religious topics (9%): the study was conducted during the month of Ramadan and various discussions on this topic (with diverse perspectives) were present in the blogs.
5. Literature and the arts (7%)
6. Other (6.7%)

Based on this brief case study, Nabavi argues that the topics discussed are limited to the world of blogs. Hence, the main role of blogs is not to provide news reports. The news items discussed by the most popular blogs are generally repeated under various forms by the others. The issues tend to stay away from internal politics unless bloggers are directly involved.

Kelly and Etling (2008) describe a more systematic methodology that (i) used human researchers to identify key topics and manually examine blog posts, (ii) studied the relative frequencies of a number of terms found in blogs, and (iii) investigated outlinks from blog sites to other blogs, websites and news sources. The results showed that the topics discussed on Persian language weblogs are quite diverse, including a number of subjects beyond the socio-political discussions mentioned earlier:

Religion is a major topic for bloggers, and not predominantly in its overtly political aspects, but more often in its historical, theological, and deeply personal ones. Persian culture and history, including music, visual arts and performance, but most especially poetry, are very big topics.

Sports are popular too, as are movies. And as in the American blogosphere, a great many bloggers write simply about their day-to-day lives, seemingly with mnemonic rather than polemical purposes in mind. (Kelly and Etling 2008)

According to this case study, discussion of socio-political issues is encountered in the secular (secPol), reformist (refPol) and conservative (ConPol) subgroups of the Iranian blogosphere, while religious issues are rarely discussed by secular and reformist bloggers (cf. Figure 6). In addition, personal diaries were found most often among the *secular/expatriate* bloggers and *poetry* sites.

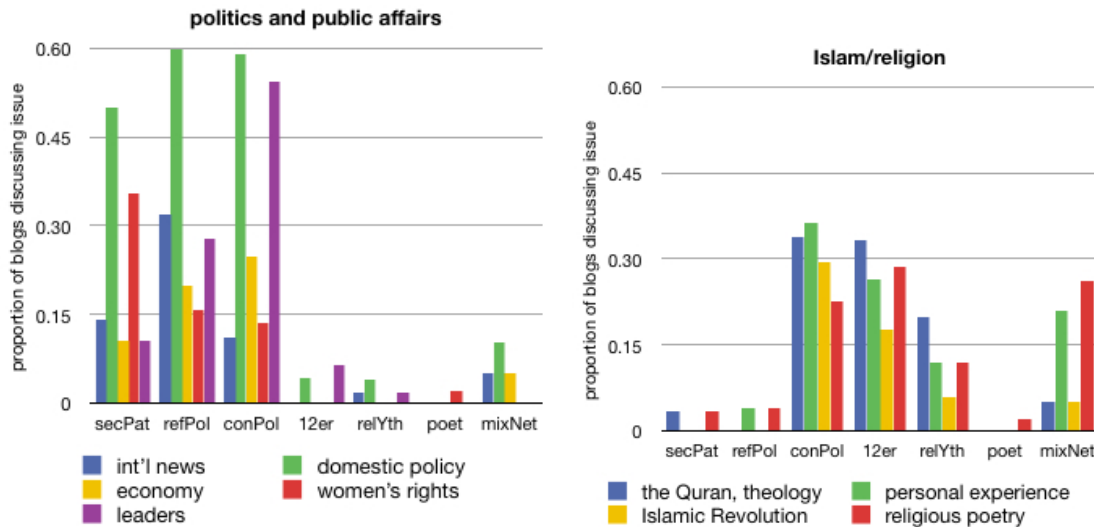


Figure 6 – Content analysis of Iranian weblogs.
Source: Kelly and Etling (2008)

Focusing on the individual characteristics of bloggers, the study finds that the majority of bloggers are men, although there is a significant number of women bloggers among the *secular/expatriate* and *poetry* clusters. The bloggers in the *Twelver* Shi'a cluster are mostly men (see Figure 7). With the exception of the *secular/expatriate* subgroup, the vast majority of bloggers live inside Iran. The case study also notes that, despite government censorship and filtering of websites, most of the weblogs investigated are visible inside Iran, although the most frequently blocked blogs are clearly those in the secular and reformist groups.

The patterns detected in this study show that there are two main ideological formations, one more progressive and the other conservative, each with its own sub-clusters. In addition, there is a large group of bloggers concerned mainly with poetry, and finally a *mixNet* group that could not be easily classified. There is cultural affinity in the domain of sports and popular entertainment among the *reformists* and the broader, unpoliticized members of the *mixNet*. Although the two ideological poles of *secular/reformist* and *conservative/religious* show common interest in certain issues, they also diverge in others. For instance, the *secular/reformist* clusters are concerned about political prisons, women's rights, and arrested bloggers, while the conservative groups focus primarily on domestic issues and religion.

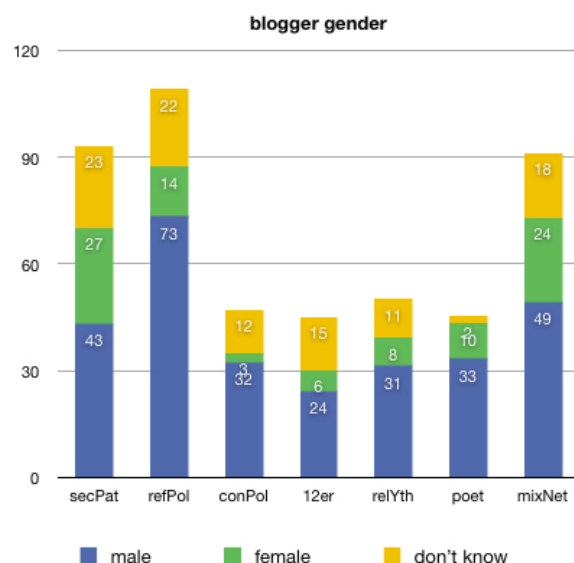


Figure 7 – Gender organized by subgroups in Iranian weblogs.
Source: Kelly and Etling (2008)

2.5.3 Social Network Analysis

A significant trait of weblogs is that they form an informal local community of bloggers. Each blog tends to have a regular readership that leaves comments and makes links to its entries. In addition, bloggers interact by listing each other's blogs in a blogroll linking to content on other community members' blogs. The community structure of blogs has a strong temporal feature since higher rate of activity and responses takes place in bursts as new topics arise. Furthermore, blog entries are associated with a *time stamp* which provides data for a temporal analysis of the blogosphere.

The intense interconnection at the core of the blogging community has triggered research on blogs from the perspective of social network analysis. Social network analysis (SNA) deals with mapping and measuring relationships and associations among people, groups, organizations and every other entity that can process information and knowledge. Nodes in the network represent people and groups while edges show ties or relationships among them. In the blogosphere, SNA can take on several forms. Linking patterns of the blogging community provide implicit information on the authority and content of blogs: a hyperlink from one blog to another suggests that the content of the blog is recommended by the author making the link and that the two blogs are probably related in topic. The authority of a site grows if it contains a large number of inlinks from blogs with a high authority. Through the analysis of link structure, one may also be able to detect subgroups sharing common interests and opinions, or to determine the level of trust associated with each site.

Sheykh Esmaili et al (2006) describe the preliminary social network analysis obtained by applying the results of several ranking algorithms to a collection of Persian weblogs. The paper describes the collection of a corpus of approximately 106,699 blogs with 215,765 hyperlinks from the host Persianblog which includes more than 50% of Persian blogs. The authors run several ranking algorithms on the dataset:

1. Ranking based on inlink count (links directing to the site)
2. Ranking based on outlink count (links directing away from the blog to other sites)

3. PageRank ranking based on a link graph (Brin and Page 1998)
4. Hypertext Induced Topic Search (HITS), based on hubs (strong central points with high numbers of outbound lists) and authorities (highly-referenced pages), used for exploring web communities related to a specific topic (Kleinberg 1999)

Results suggest that the linking patterns of bloggers are not homogeneous and seem to point to the existence of several smaller communities sharing common interests. In addition, the authors contrast results obtained from the various ranking algorithms. The research sets the ground for future research on Persian blogs by providing essential tools for researchers such as a blog dataset¹⁴, list of links between nodes in the graph, and a list of all connected components. Sheykh Esmaili et al (2005) further discusses the application of the HITS algorithm to the collected Persian blog set.

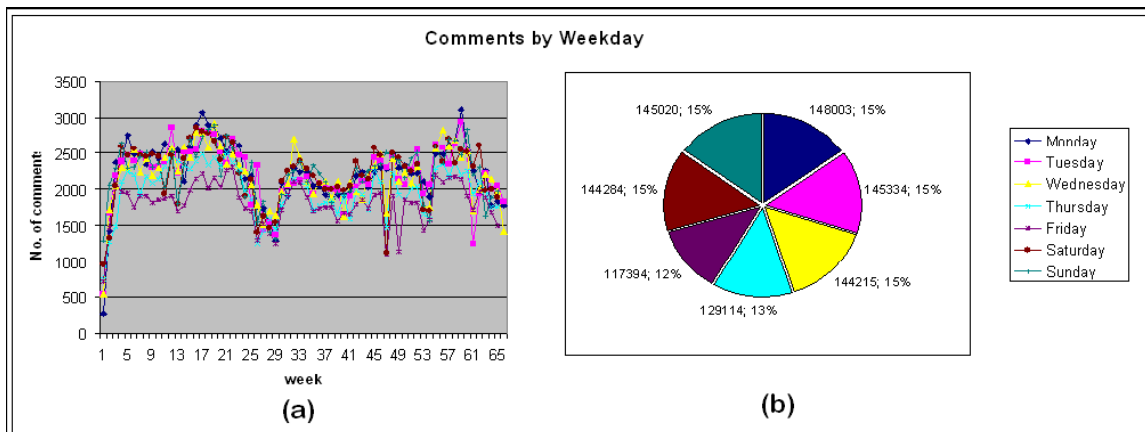


Figure 8 – Number of comments on each day of the week.
Source: Qazvinian et al (2007a)

Qazvinian et al (2007a) investigate the commenting behavior of Persian bloggers, and present a model on the distribution of comments following a posting. The dataset consists of the archives of over 22,000 blogs from Persianblog comprising about 347,800 posts and 1,258,000 comments¹⁵. The authors state that the number of comments left on weblogs are at their highest on Thursdays and Fridays, which correspond to the weekend in Iran (Figure 8). Time graph analysis shows that, although long holidays and the beginning of the school year correspond to low comment activity periods, certain events such as the presidential election trigger a large number of comments (Figure 9).

A crucial aspect of the cyber-community is the importance of the *A-list*, expert bloggers who update their sites often, are well-known and often quoted, and maintain the most popular blogs with the most number of visitors and links. The effect of these powerful bloggers cannot be underestimated in a SNA as the authors discovered when a "fall" occurred as several of these bloggers terminated their accounts with Persianblog and moved to other hosts.

¹⁴ The authors indicate the dataset containing the link-structure of the blogs is available at <http://ce.sharif.edu/~shesmail/persianweblogs/>.

¹⁵ The dataset is available at <http://www.blogscience.org/data.html> for research purposes.

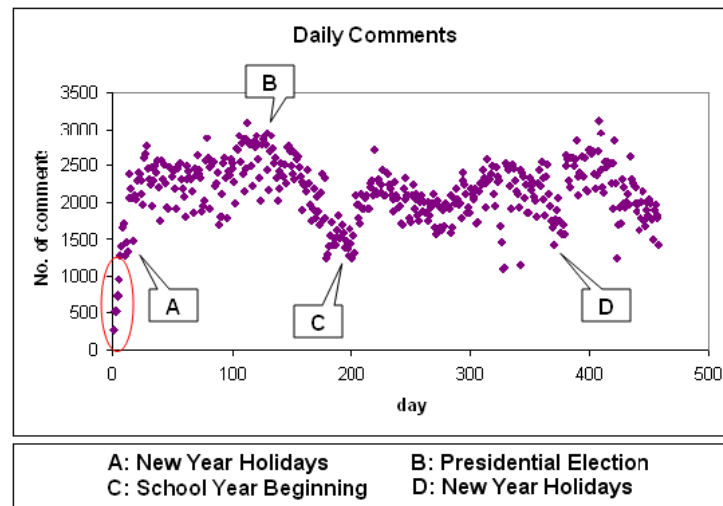


Figure 9 – Time graph shows number of comments on each day with tagged outliers.
 Source: Qazvinian et al (2007a)

Qazvinian et al (2007b) argue for the importance of taking into account comments in SNA. As Figure 10 demonstrates, comment inlinks at about 64% of all links in their Persian-language corpus make a noticeable contribution to the formation of the blogspace graph. The authors focus on *failure* in blogs, a term they use to refer to situations where a blogger quits writing in the blog (commitment-failure) or when a blogger ceases to receive comments from readers (connection-failure). Both instances affect the results of a social network analysis since the identity of a blog is generally defined by its interactions or posts.

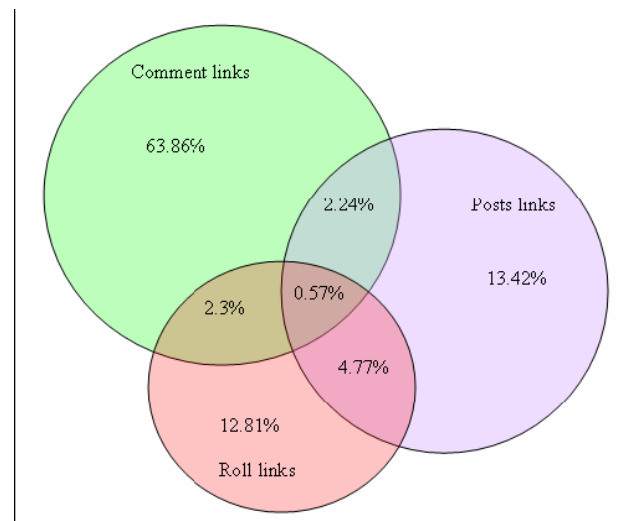


Figure 10 – Venn diagram for different links distribution.
 Source: Qazvinian et al (2007b)

Kelly and Etling (2008) describes a large-scale social network analysis using links captured from the 6018 most prominent Iranian weblogs over a period of seven months (from July 2007 to March 2008). The network structure was then mapped by (i) identifying large scale

groupings of densely linked blogs and (ii) clustering methods used to detect outlink patterns (i.e., links from these blogs to all other Internet resources), defining *attentive clusters* of bloggers who link to similar things. The resulting network map was drawn with a Fruchterman-Rheingold 'physics model' algorithm and can be seen in Figure 11, where each dot represents a blog and its size depicts its popularity measured by the number of blogs that link to it. The relative position of each dot represents the function of its links with its neighbors – blogs come closer to each other either by direct links or by the links among their shared neighbors, allowing large groups of blogs to cluster up into densely interlinked network neighborhoods. The color of each dot on the map indicates the assignment of a blog to a particular *attentive cluster*, which is a group of blogs that link to similar online resources.

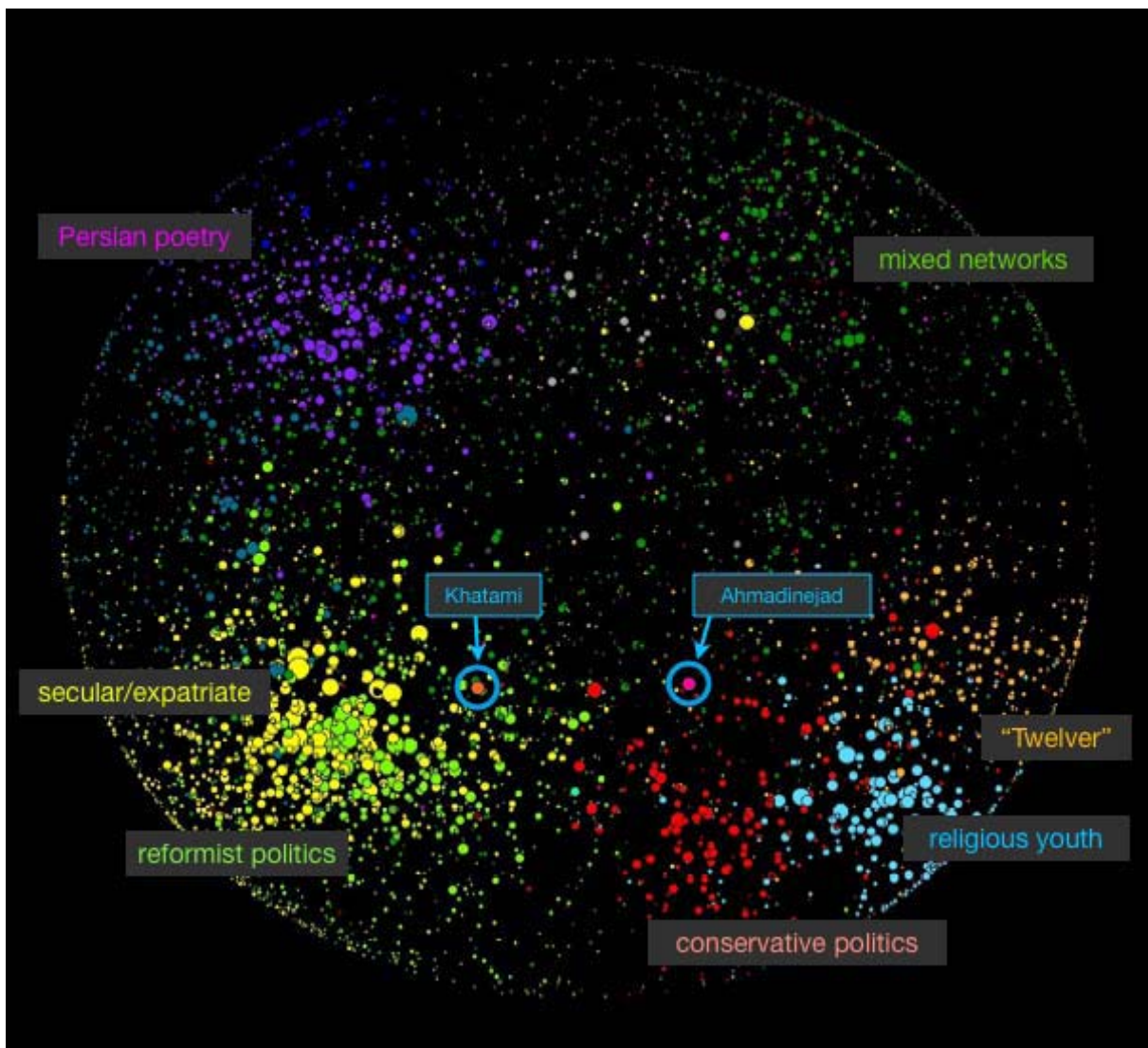


Figure 11 – Social network map of the Iranian blogosphere.
Source: Kelly and Etling (2008)

As the map shows, by leveraging content study of Iranian weblogs against a large-scale social network analysis, Kelly and Etling (2008) delineate four major groups in the Iranian blogosphere, two of which (the *secular/reformist* and the *conservative/religious*) form their own subclusters. These are described below:

1. Secular/Reformist

(i) Secular/Expatriate (secPat)

This cluster includes a large proportion of women and expatriates, including prominent dissidents and journalists who have left Iran recently. Common issues include women's rights, political prisoners, and cultural issues.

(ii) Reformist Politics (refPol)

This cluster is more focused on news and politics, including issues like drug abuse and environmental degradation in Iran. Bloggers in this subgroup are overwhelmingly male, and live inside Iran.

2. Conservative/Religious

(i) Conservative Politics (conPol)

This cluster is focused on power politics, tracking news and current public affairs, with emphasis placed primarily on domestic issues although some attention is given to foreign policy as well. This subgroup frequently quotes speeches of politicians. These bloggers are generally supportive of the Iranian Revolution and Islamist political philosophy but criticisms of government policies and leaders are common.

(ii) Twelver Discourse (12er)

Twelvers are the dominant Shi'a sect in Iran and this cluster focuses first and foremost on religious matters.

(iii) Religious Youth (relyth)

This cluster includes a lot of students and younger bloggers whose main concern is religion.

3. Persian Poetry and Literature (poet)

This group is devoted mainly to poetry which is a very important part of Iranian cultural expression.

4. Mixed Networks (mixNet)

This group displays a less centralized structure and does not focus on any particular issue or ideology. It is a loosely interconnected network of many smaller communities of interest such as sports, celebrity, minority cultures, and popular media.

Crucially, the authors of the study emphasize that the "online dissent" narrative centered around the large group dominated by expatriates, reformists, and secular bloggers and seen in most of the current literature on Iranian weblogs only describes one part of a very diverse and complex online public communications network:

Iranian bloggers include members of Hezbollah, teenagers in Tehran, retirees in Los Angeles, religious students in Qom, dissident journalists who left Iran a few years ago, exiles who left thirty years ago, current members of the Majlis (parliament), reformist politicians, a multitude of poets, and quite famously the President of Iran, among many others. (Kelly and Etling 2008)

2.6 Computational Linguistic Analysis of Persian Blogs

Youth, women and intellectuals in Iran—often members of an educated, middle class excluded from the physical public space—use weblogs to make their voices heard. While journalists and intellectuals often use blogs to bypass strict state censorship of the print media, youth and women generally express their thoughts on daily issues that cannot be spoken in the real public space. The relative freedom of the virtual world allows those who participate to meet new friends, form new relations and communities, and rediscover or shape their own identities. Clearly, blogs have become a form of struggle between the youth and the traditional and state authorities over the limits placed on public discourse, and the language used by young bloggers has become a crucial tool in this struggle. Interestingly, the main individuals that have strongly criticized the language of young bloggers have been the traditional intellectuals and journalists. The debate over the language of blogging known as the *vulgarity debate* reached its peak in 2003 and has continued since in various forms. Literature on the two sides of this issue abound in weblogs and online articles, yet academic and computational investigations of the language of Persian weblogs are rare.

2.6.1 Linguistic Features

According to Nabavi (2004), the language of Persian blogs is of utmost importance as it has become a medium for political engagement and social change. He characterizes this language as follows:

- Avoidance of the literary and formal written language;
- Brief and telegraphic in style, and lacking lengthy descriptions;
- Generous use of images, audio and links;
- Straightforward and direct style, avoiding unnecessary literary and social formulations. In certain instances, purposeful mistakes and misspellings can be found.
- Creation of new vocabulary, which finds its source in what has been called the "hidden" or "street language" of the youth. This clearly shows the dynamic and evolving nature of this language, which has also met strong resistance from previous generations and in particular authors living in the expatriate community (who are still using the more traditional, static Persian language).
- Self-reflexive and self-assertive in nature. This language is defined by the use of the first person singular, and is void of the modesty typical of Iranian literature (especially of the religious literature). Mixed with the slang of the young generation, the language of Persian blogs includes humor as well as individuality and self-assertiveness.

The main characteristic of blog language in Iran is therefore the use of colloquial forms in writing. The colloquial or conversational Persian language differs considerably from its formal counterpart in several domains including lexicon, morphology, and syntactic structure – this difference is much larger than the one found in English where the standard written and spoken forms are relatively similar. Traditionally, the colloquial language has never been used in writing, even in familial text such as letters written to family members. This trend was broken at times by modern authors who used colloquial forms in their writings; however, current journalistic and literary prose both in print media and online is confined to the formal language.

Megerdooomian (2006) describes some of the linguistic features of informal text in Persian blogs:

- Informal text makes more frequent use of clitic pronouns that attach to the verb instead of appearing as separate pronouns.
ex. the informal *gereftætæm* گرفته‌ام vs. the formal *mæra gerefte æst* مرا گرفته‌است 'has caught me'
- Has shortened verbal stems (especially in the case of the present stem) and inflectional endings.
ex. informal *migæ* میگن vs. formal *miguyænd* می‌گویند 'they say'
- Includes morphemes that do not exist in the formal language such as the definite article, ex. *forushændehe* فروشنده 'the salesperson'
- The spelling in these texts reflects the pronunciation in colloquial speech, such as the plural *ha* or *haye* being pronounced as 'â' or the pronoun clitic '*tan*' as '*tun*' as shown in the example.
ex. informal *næzæratun* نظراتون vs. formal *næzærhayetan* نظرهایتان 'your views'
- Informal text contains more instances of scrambling, idiomatic expressions, loan words, jargons and other non-dictionary words, as well as cultural inferences.

Blogs written in informal text often do not follow a standard set of orthographic rules and may write morphemes attached or detached as the author sees fit. Blogs contain ellipses, emoticons and hyperlinks, which require special tokenization. Blogs also contain a larger number of interjections and English words (especially from the technical domain). In addition, spelling errors are much more common than one may encounter in non-blog websites. It should also be noted that even blogs written in formal text display a varying range of orthographic patterns when it comes to the printing of derivational and inflectional affixes in Persian, which differ significantly from the standard rules followed by newsprint media online. For instance, while a number of bloggers (even those using a formal writing style) always write the clitic copula in detached form as in *mvafq~ænd*¹⁶ موافق‌اند 'they agree', the traditional websites use the attached form *mvafqnd* موافقند. Bloggers writing in formal text also use new spellings such as *hæta* 'even' written with an "alef" (حتا) instead of the original Arabic writing with an 'alef maqsura' (حتى) which is usually treated as an exception in Persian.

Thus, the representation of the conversational form of Persian language, online scripting devices such as emoticons and hyperlinks, the preponderance of loan words and non-dictionary items, and the large variation in orthographic forms and standards are the main linguistic features of Persian blogs.

2.6.2 The Vulgarity Debate

The so-called *bæhs-e ebtezal* or Vulgarity Debate was triggered in 2003 by a post by Hossein Derakhshan on his blog on the incompatibility of Islam with human rights. This entry provoked a large number of responses, notably one by prominent blogger journalist and literary critic Seyyed Reza Shokrollahi:

¹⁶ The *tilde* (~) in this example represents the shortspace or zero-width nonjoiner (Unicode \uz00c) which is used to force a final (i.e., unattached) form on the previous character.

Blogging, after laying waste to the Persian script and language, has been able to drag every serious and intellectual topic into the scum of the disease of vulgarity, grow like a cancerous tumor, and trash the writer, the reader and everyone else.¹⁷ (Shokrollahi 2003)

Shokrollahi's complaints about the sloppy language, orthographic inconsistencies, and low argumentative standards of Persian-language weblogs launched a very heated debate, which lasted for several weeks and continues to this day in many impassioned writings and interviews. The discussions have centered around three distinct issues that have often been muddled in the debate: (i) The lack of professional, thoughtful and researched opinions in the blogosphere; (ii) the disregard for the traditional spelling and orthographic principles of the Persian language; and (iii) the choice by bloggers to write in colloquial rather than in formal Persian. These three issues are clearly interrelated as they all represent the professional and aesthetic standards of traditional publications and have all been used to challenge the language used in Persian blogs.

In a paper that appeared in the *American Anthropologist* in 2004, Alireza Doostdar¹⁸ presents an incisive ethnographic study of the Vulgarity Debate and conceptualizes the controversy as "a clash between two classes of people with unequal access to cultural capital." The identification of blogs as having a "vulgar spirit", he argues, is an interpretation made by the dominant intellectual class of journalists, writers, and literary critics. Basing his analysis on Bourdieu (1984), Doostdar suggests that the Vulgarity Debate should be viewed as part of a battle for hegemony and authority online.

Traditionally, the *roshānfekr* (intellectual) class has held a highly respected place in Iranian society and has historically come to represent a liberal humanist individual who publicly critiques the values and policies of the authority. However, intellectuals generally consider themselves an authority in matters of language and culture. The opponents of these critics, on the other hand, are not intellectuals by social function or merit and have thus far been excluded from this class. Hence, the Vulgarity Debate can be seen as a dominant class of intellectuals imposing their taste of a "legitimate" and "correct" mode of writing in blogs. In response, the bloggers have used language as the main tool of engagement in challenging the linguistic authority and cultural hegemony of the intellectual class. For most bloggers, the conversational form of writing is crucial in the expression of their views or thoughts since it represents an intimate and authentic form free from the burden of standards. Other bloggers have even made deliberate orthographic or grammatical "mistakes" as a form of resistance against the authority of the intellectuals.

December 8, 2003

I'm disgusted by intellectualist pretense and everything else like it in weblog [...] Sit down and give your opinions in the language of your grandfathers and brag about being an intellectual. Keep mistaking this place as a literary conference when others consider it to be an informal and safe place for chatting [...] Come sit down wearing a suit and tie and mock those who are wearing jeans.

Website: khorshidkhanoom.com

(Extracted from Doostdar 2004)

¹⁷ As noted in Doostdar (2004), Shokrollahi adopts an aesthetic definition of *vulgarity* as facileness, disharmony between content and form, etc. Others, however, have taken the word to refer to the more "commonplace" moralistic sense.

¹⁸ Alireza Doostdar is a Ph.D. candidate in Anthropology and Middle Eastern Studies at Harvard University.

The generic clash can also be seen as one dimension of a struggle for the creation of hegemonies and counterhegemonies: An intellectual class sees its own linguistic and cultural authority threatened by the "vulgar" practices of bloggers and a disparate class of nonintellectuals deliberately undermines this authority by neglecting or flouting grammatical and orthographic standards and calling into question the linguistic and cultural authority of the intellectuals. (Doostdar 2004)

Anderson (1997) states that parallel clashes are taking place in the rest of the Middle East as well as in Western countries. He presents a very similar analysis stating that the Internet and blogs in particular are creating a large gap between "elite, super-literate, authoritative discourse" and the "nonintellectual", popular discourse which "places issues ahead of analysis". The main critiques are once again liberal humanists, particularly academics and journalists, who are tied to traditional institutions of print. According to Anderson, Internet communities harbor potential new elites, an emerging class of authority distinct from both the existing modes of state authority and the traditional class of language authority tied to previous informational-communicational regimes. Anderson states that this third force is currently loosely identified as civil society, generally associated with notions of a professionalized middle class, suggesting that an analysis and investigation of blogs may potentially help identify future leaders in a society.

2.6.3 Computational Systems

Within the frame of the vulgarity debate, much energy has been spent pushing for orthographic standards for online publishing¹⁹, yet very little work has been done on the development of computational systems to analyze and process Persian-language blogs.

Shokrollahi (2006) discusses the lack of Persian Machine Translation (MT) systems in e.g., Google online translation and based on a blog article by Nima Akbarpour²⁰, he argues that in order to facilitate work in Natural Language Processing (NLP), the Persian writing system should be standardized. Shokrollahi points to the ambiguities of encoding (e.g., the use of the Persian letters *ye* or *kaf* as well as the Arabic versions of these letters in online publications) and the fact that detached morphemes such as the plural *ha* (ها) and the superlative *taerin* (ترین) can be written with either a whitespace separating them from the word or with a short space, and claims that a standardized online orthography would undoubtedly facilitate work on Persian MT. It should be noted, however, that such encoding issues are relatively easy to handle in computational systems and the main issues hindering automatic Persian translation into English are related to linguistic factors such as word order or the absence of certain vowels in the writing system.

Imran (2006) investigates the features of colloquial Persian in venues such as instant messaging systems and bulletin boards, which can also be extended to the informal language used in chatrooms, blogs, and emails. However, the focus of this work is on Romanized Persian (i.e., Persian written in the Latin alphabet) and does not look at the effects of the Perso-Arabic script. This work describes the common patterns of linguistic behavior such as code-switching (the use of English and Romanized Persian terms in the

¹⁹ Interestingly, the orthographic standards proposed by intellectual bloggers such as Shokrollahi (<http://www.khabgard.com/?id=-965599463>) are not consistent with the more traditional standards such as the regulations put forth by the Persian Language Academy, *Farhangestan*. (<http://www.persianacademy.ir/fa/das.aspx>).

²⁰ http://www.osyan.net/2006/04/post_412.php.

same sentence) and acronyms. It also looks at certain computational linguistic issues associated with the language used in these domains, such as those caused by the colloquial Persian variants based on the speaker's cultural background and educational level.

Megerdoomian (2006) presents an overview of the issues that the characteristic traits of blog language raise for computational morphological analysis of Persian. Departing from an existing morphological analyzer for standard Persian of online newsprint, she proposes extensions that will allow the system to successfully analyze informal text. The paper describes a morphological analyzer for Persian based on the Xerox Finite State Technology or XFST (Beesley and Karttunen 2003). A large part of the paper is dedicated to a discussion of issues in Persian text in general, such as complex tokens consisting of more than one lexical category or part of speech, detached inflectional morphemes, phonological alternations at affix boundaries, and long-distance dependencies. Solutions to each of these within the XFST system are described. In addition, the paper presents further challenges placed on the morphological analyzer by the many orthographic variations in Persian language blogs, informal language morphology, and the conversational forms of affixes²¹, and provides potential solutions to be integrated in the existing morphological system. The author points out, however, that the rules for informal text have not been incorporated in the system described. Megerdoomian (2006) also provides several categories of words that need to be added in the lexicon in order to be able to process informal blog text; these include

- Conversational forms or pronunciations of words:
colloquial *khune* (خونه) 'house' vs. formal *khane* (خانه); *hæmdige* (همدیگه) 'each other' vs. *hæmdigar* (همدیگر); *bæram* (برام) 'for me' vs. *bærayæm* (برایم)
- Colloquial counterparts to words:
colloquial *vase* (واسه) 'for' vs. formal *bæraye* (برای); *tu* (تو) 'in' vs. *dær* (در)
- Loan words appearing in Romanized form or in the original language:
anlayn (آن لاین) 'online'; *chætrum* (چتروم) 'chatroom'; *danlod* (دانلود) 'download'
- Neologisms or new words created by bloggers – these words often follow Persian word-formation rules:
linkduni (لینکدونی) 'blogroll'; *kament-gozar* (کامنت گذار) 'commenter'; *lagidæn* (لاگیدن) 'to blog'
- Interjections:
aaakh! (آآخ!) 'ouch'; *ouuuh!* (اووووه!); *vay* (وای); *vala* (والا) 'well'

2.7 Future Directions

Recent years have seen a number of workshops and conferences dedicated solely to the study of weblogs²². Blogs represent a number of challenges for computational analyses since they are generally unedited, represent a fragmented topic structure, contain inconsistent grammar, and are vulnerable to spam. However, the blogosphere, by virtue of being a highly dynamic subset of the World Wide Web that forms an online social network, uses a more informal writing style, and evolves and responds to real world events, opens up several new interesting research areas for NLP. As we have already seen, not much work has been performed on computational approaches to Persian-language blogs, with the

²¹ See Section 4 for a more detailed description of these features of informal text.

²² See Appendix D for examples of recent and upcoming computational conferences, symposia and workshops related to blogs.

possible exception of social network analysis. The work on English language blogs can provide the foundation for developing research projects for Persian blogs, which may also present interesting and contrasting results.

With the launch of the Blog Track at TREC 2006, there was a need to create a test collection of English-language blog data to be shared among the participants; the criteria for the development of this corpus are discussed in Macdonald and Ounis (2006)²³. The opinionated nature of most blogs has given rise to a large number of works on mood analysis, opinion mining, and sentiment detection, generally using a statistical classification algorithm (Ounis et al 2006, Mullen and Malouf 2006, Glance et al 2005, Mishne 2005), and Mishne and de Rijke (2006) find that the type of queries relevant for information retrieval of blogs differs from those used in conventional web search engines. One major area of research is on weblog-based social network analysis, investigating link patterns of blog communities to determine the network structure of the online community (Adamic 1999, Gibson et al 1998). Kumar et al (2003) model temporally-concentrated bursts of connectivity within blog communities over time. SNA has also been used to identify *authoritative* bloggers (Marlow 2004, Nakajima et al 2005), dynamics of information propagation (Gruhl et al 2004, Adar et al 2004), and content similarity analysis (Kurland and Lee 2005, Kritikopoulos et al 2006). Many researchers have tried to identify the characteristic traits of *Blogspeak* (Nilsson 2003a, Tavosanis 2006) and to determine whether it is representative of a new genre, often through content analysis of blogs (Herring et al 2005, Herring et al 2006b). In addition, research has been performed on the identification of bloggers' demographic information, such as gender or geographic region (Schler et al 2005, Yasuda et al 2006, Nowson 2006).

Table 1 - Most discriminating word n-grams for detecting some moods; Source: Mishne (2005)

Mood	Top words	Top bigrams	Top trigrams
hungry	hungry eat bread sauce	am hungry hungry and some food to eat	I am hungry is finally happened I am starving ask my mother
frustrated	n't frustrated frustrating do	am done can not problem is to fix	I am done am tired of stab stab stab I do not
love	love me valentine her	I love love you love is valentines day	I love you my god oh I love him love you so

In this section, I will address research that is relevant for a computational linguistic investigation of weblogs by reviewing publications describing the language of blogs as well as the research on the use of language features to determine a blogger's profile such as age,

²³ The BLOGo6 collection is available for research purposes. The dataset and relevant statistics can be found at http://ir.dcs.gla.ac.uk/test_collections/blogo6info.html.

gender and personality traits. Although most of these works target English weblogs, the issues raised can provide a starting point for future research on Persian-language blogs.

2.7.1 Blogspeak: An Emerging Linguistic Genre

In an entertaining piece, Nunberg (2004) contrasts traditional journalistic writing to the style used in weblogs.

[Most journalists] do all the things you should do in a newspaper feature. They fashion engaging ledes, they develop their arguments methodically, they give context and background, and tack helpful ID's onto the names they introduce -- "New York Senator Charles E. Schumer (D)."

In contrast, Nunberg describes the language used in blogs as a kind of "anti-journalese" – "it's informal, impertinent, and digressive, casting links in all directions." Print journalists tend to address their readers as anonymous citizens, while bloggers see their audience as "co-conspirators who are in on the joke". Echoing the analysis by Anderson (1999), Nunberg states that the "blogging world sounds a lot less like a public meeting than the lunchtime chatter in a high-school cafeteria, complete with snarky comments about the kids at the tables across the room." However, Nunberg believes that, despite sharing certain elements with past styles of public discourse, Blogspeak is indeed a new genre. He goes on to point out the paradox in this form of communication: it is a democratic form of expression allowing those previously excluded from the world of print to participate, yet it is not a neutral language (as was the formal style of newspaper op-eds) since it represents the conversational language of the urban middle class and privileges the speech of a particular class of society.

2.7.1.1 Linguistic Features of Netspeak

Crystal (2001) is the only book currently published, as far as this author is aware, that investigates the Internet solely from a linguistic perspective. David Crystal studies the various forms of computer-mediated communication in English and provides a comprehensive description of language use in different Internet situations. He identifies five such situations which are sufficiently different (email, synchronous chatroom, asynchronous chatroom, virtual world or role-playing sites, and the World Wide Web) and demonstrates that the language used in each instance is significantly distinctive. He contrasts *Netspeak* to the characteristics of speech and writing and argues that it is different enough to be a genuine "third medium". He concludes that, although not homogeneous across all language-using situations, there is a clear "emergence of a distinctive variety of language, with characteristics closely related to the properties of its technological context as well as to the intentions, activities and (to some extent) personalities of the users." Although the book does not include a discussion of weblogs, many linguistic features of Netspeak can also be found in Blogspeak.

While *speech* is typically time-bound, spontaneous, face-to-face, socially interactive, loosely structured, immediately revisable, and prosodically rich, *writing* is typically space-bound, contrived, visually decontextualized, factually communicative, elaborately structured, repeatedly revisable, and graphically rich. Crystal argues that Netspeak is interesting as a form of communication in that it relies on characteristics belonging to both sides of the speech/writing divide. It differs from face-to-face or phone-based conversational speech (and parallels writing) if one considers the lack of real-time feedback or turn-taking, and absence of prosody or paralanguage (expressed through vocal variations in pitch, stress, rhythm, speed, tone of voice). Netspeak resembles speech, however, in its short

constructions, phrasal repetition, and looser sentence construction. Netspeak usually lacks the informality and intimacy of face-to-face conversation. However, this can be improved through the use of colloquial grammar and vocabulary. In addition, internet users employ several methods to express emotions in writing such as repeated letters and punctuation marks (*aaaaahhhh!!!!, whohe????*), capitals and special symbols for emphasis (*the *real* answer*), and emoticons and smileys. Another interesting feature of Netspeak is its creative nature. Neologisms (*to mouse over, clickthrough rate, webzine, webster, cyberian, Netspeak, geekification, a screenful, to clipboard*), non-standard spelling which reflects pronunciation (*boyz, yup*), deviant spellings (*fone, kool, phreak*), and the use of acronyms are all common features in Netspeak. Although punctuation tends to be minimalist in most situations, ellipsis is used more often to indicate pause in the speech. In fact, Crystal states:

A strong personal, creative spirit imbues Netspeak as an emerging variety [...] The rate at which [internet users] have been coining new terms and introducing playful variations into established ones has no parallel in contemporary language use.

Hence, Netspeak seems to be a hybrid resource, combining elements from both spoken and written forms. Yet it also does things that neither of these other mediums do and must accordingly be seen as a "new species of communication". According to Crystal, its electronic nature gives it fluidity, simultaneity (being available on an indefinite number of machines) and permeable boundaries through the use of hyperlinks. It also allows documents to be disseminated beyond the traditional limits of published text and demonstrates an unprecedented amount of anonymity by Internet users. Crystal concludes that Netspeak is neither 'spoken writing' nor 'written speech', making computer-mediated language a new medium of linguistic communication. "And as a new linguistic medium, it will grow in its sociolinguistic and stylistic complexity to be comparable to that already known in traditional speech and writing."²⁴

As with language change in general, most features that distinguish Netspeak from previous genres are currently found chiefly in graphology and the lexicon – the levels of language where it is relatively easy to introduce innovation and deviations. In fact, syntactic or grammatical variation is less frequent or widespread. Crystal also discusses the potential growing linguistic variation on the Internet as different Internet communities begin developing their own, distinct 'dialects', and as the content of a site (e.g., information, education, diary) influences the general character of the language being used.

Several linguistic studies are also worthy of mention in this section: The Internet is a global phenomenon as different languages come into contact in computer-mediated communication. The interface between different languages on the Internet has been discussed for Chinese (Gao 2006 discusses the impact of English on Mandarin Chinese) and Jamaican Creole (Hinrichs 2006 studies code-switching in emails). Finally, an interesting

²⁴ A lot of energy has been spent on developing prescriptive style manuals and list of standards for Internet writing, which have often been ignored. The *Writing Style* (Hale and Scanlon 1999), however, recommends that users embrace colloquial language, creativity and wordplay, and jargon use. The manual proposes to simplify spellings and remove hyphens from compound words, to "violate journalism's cardinal rules and toy with conventions" yet maintain typographical conventions. The manual states: "Welcome inconsistency, especially in the interest of voice and cadence. Treat the institutions and players in your world with a dose of irreverence. Play with grammar and syntax. Appreciate unruliness." In other words, the manual *prescribes* a new and rebellious form of Netspeak!

study by Tsujimura (2007) analyzes certain innovative patterns of intransitivization constructions by young Japanese Internet users, pointing to evidence for language change.

2.7.1.2 Social Networks and the Language of Blogs

According to linguistic social networking theory which studies the variations of a language in relation to the social network in which it is used, an *in-group* language is often a crucial feature of a closed and dense network, which is used to strengthen network ties and unify the members of the group. The in-group language is generally represented by personal pronouns such as *we*, *us*, *our*, while out-group language consists of pronouns referring to the other such as *they* and *them*. Nilsson (2003a, 2003b) find a dichotomy in the nature of blogs since graphically they represent a closed and dense social network, yet do not use an in-group language. Nilsson finds instead that blogs are written in the first person, out-group language keywords are much more common, and hyperlinks rather than in-group language are used to signify solidarity within the network. These results hold regardless of the length of the post or the status of a blogger within the social network.

Nilsson applies the criteria used by Crystal (2001) to BlogSpeak and concludes that it is indeed a distinct genre, "a new variety of language that has evolved from spoken and written communication and has adapted itself to flourish in the virtual environment." (see Table 2 and Table 3). Yet she argues that BlogSpeak is somewhat different from the language used in the five situations that Crystal investigated, since it combines writing and speech in a unique environment which allows both written internal monologue and threads of conversation.

Table 2 – Spoken language criteria applied to BlogSpeak (adapted from Crystal 2001); Source: Nilsson (2003)

Spoken Language Criteria	The Web (in general)	Blogs (in general)
1 Time-bound	No	Yes
2 Spontaneous	No	No
3 Face-to-face	No	No
4 Loosely structured	Variable	Variable
5 Socially interactive	No, with increasing options	Variable
6 Immediately revisable	Variable; depends on tools	Yes
7 Prosodically rich	Variable	Variable, but limited

Table 3 – Written language criteria applied to BlogSpeak (adapted from Crystal 2001); Source: Nilsson (2003)

Written Language Criteria	The Web (in general)	Blogs (in general)
1 Space-bound	Yes, with extra options	Yes
2 Contrived	Yes	Yes
3 Visually decontextualized	Yes, but with considerable adaptation	Variable
4 Elaborately structured	Yes	Yes
5 Factually communicative	Yes	Yes
6 Repeatedly revisable	Yes	Yes
7 Graphically rich	Yes, but in different ways	Variable, but limited

Blogs combine both the monologue and the dialogue in a space-bound, electronic environment. They are simultaneously self-reflective thoughts presented publicly and continuous conversations. Blogs utilise both the attributes of on-line, informal spoken language with those of the conventional written monologue.

Nilsson notes some other features of blogs: Posts are usually written in short, paratactic sentences, in fairly informal and non-standard language; use of slang and blogging and professional jargon are common; conversation is carried out through comment features and trackbacks. More importantly, because of the closeness of the blogging networks, varieties of language can be standardized, strengthening further the sense of group identity.

2.7.2 Blogs and Individual Differences

Most work on text classification has traditionally been focused on identifying the topic of the text, rather than detecting stylistic features. The availability of weblogs with their subjective characteristics has given rise to a number of publications on "stylometric" research, which can be applied to sentiment and mood detection, gender classification, and identification of other types of individual differences such as age and personality traits. Most approaches use a statistical association level between words in the text and a set of keywords.

2.7.2.1 Determining Personality Traits

Nowson et al (2005) and Nowson and Oberlander (2006) attempt to delineate certain English language features that can be used in automatically identifying individual differences such as personality traits. The study is performed on 71 bloggers (24 males and 47 females) who have completed a sociobiographic questionnaire and an IPIP Five Factor Personality Inventory. The corpus is a set of personal weblogs written by the subjects amounting to over 410,000 words. The personality traits being studied are characterized as follows:

- *Neuroticism*: anxiety, angry hostility, depression, self-consciousness, impulsiveness, and vulnerability
- *Extraversion*: warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotion
- *Openness to experience*: fantasy, aesthetics, feelings, actions, ideas, and values (characterized by culture, intellect and originality)
- *Agreeableness*: trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness
- *Conscientiousness*: competence, order, dutifulness, achievement striving, self-discipline, and deliberation

The authors use several language features for content analysis:

1. Frequency of word categories from the *Linguistic Inquiry and Word Count* (LIWC) collection. Examples of this set are 'talk', 'us', 'friend' from the Social Processes category; and 'because', 'hence', 'effect' from the Causation group.
2. Frequency of relevant word collocations in forms of bigrams and trigrams.
3. Frequency of parts-of-speech (POS), based on the F-measure by Heylighen and Dewaele (2002). This measure is based on the frequency of two categories of POS:
 - a) *category 1*. Pronouns, Verbs, Adverbs, Interjections
 - b) *category 2*. Nouns, Adjectives, Prepositions

$$F=0.5 * [(nounfrq + adjfrq + prepfrq + artfrq) - (pronfrq + verbfrq + advfrq + intfrq) + 100]$$

Heylighen and Dewaele claim that a lower score on a text implies higher contextuality (i.e., informality) characteristics. Application of the F-measure shows that spoken language is more contextual than written language and fiction is more contextual than newspapers. Nowson and Oberlander (2006) apply the F-measure to their blog corpus and determine that blogs are more contextual than essays or non-academic social science, but that they are more formal than scripted speeches, e-mail, personal letters, fiction prose, and sermons.

The results reported in Nowson et al (2005) indicate that Neuroticism and Extraversion scores correlate positively with the frequencies of contextual parts of speech, and negatively with those POSs considered formal. However, they admit that the correlations, though in the expected direction, are small and do not reach significance. The opposite correlation holds for Agreeableness and Openness, while the Conscientiousness correlation is negligible. The results therefore do not support the claim in the literature that category 1 POSs should correlate with Extravert tendencies and category 2 POSs with Introvert characteristics. The application of the F-measure indicates that females score lower, suggesting that they prefer a more contextual style, while men prefer a more formal style. The authors conclude, "within the blog genre, there is variability in contextuality/formality due to individual differences. But the differences that make the most difference are not Extraversion or Neuroticism, but Openness and – especially – Agreeableness and gender."

2.7.3 Age, Gender and Language

Schler et al (2005) study a corpus of about 37,478 blogs from blogger.com (comprising 1,405,209 blog entries and 295,526,889 words). They select three style-related features (selected parts-of-speech, function words, blog words such as *lol*, *haha*, *ur*, and hyperlinks) and content-related features (simple content words and special classes of words from LIWC), and measure the frequency with which the features appear in the corpus per gender per age bracket. The authors note a pattern of more "personal" writing by female bloggers while male writing contains more references to politics and technology. They also note that the frequency of "male" words (e.g., *data*, *software*, *democracy*, *linux*) increase monotonically with age whereas usage of "female" words (e.g., *shopping*, *pink*, *husband*, *skirt*) seems to decrease with age.

The authors provide the following generalizations: "For each age bracket, female bloggers use more pronouns and assent/negation words while male bloggers use more articles and prepositions. Also, female bloggers use blog words far more than do male bloggers, while male bloggers use more hyperlinks than do female bloggers. [...] Prepositions and articles, which are used more frequently by male bloggers, are used with increasing frequency by all bloggers as they get older. Conversely, pronouns, assent/negation words and blog words, which are used more frequently by female bloggers, are used with decreasing frequency as bloggers get older."

2.7.4 The Intermediary Factor: Blog Topic

An interesting study performed in 2003 by Blog Census (NITLE Census 2003) set out to analyze the split along gender lines in the English blogosphere. The study involved manual analysis of a random sample of 776 blogs out of a total of 490,000 English-language weblogs. The researchers tagged the blogs for male or female when unambiguous evidence was detected, such as photos or gendered pronouns in reported speech, and marked the gender as unknown in the lack of such evidence. The results, shown in Figure 12 indicate

that female and male bloggers are split almost equally, with 39.8% of men bloggers vs. 36.3% of women bloggers (the results falling well between the margin of error of +/- 3.5%).

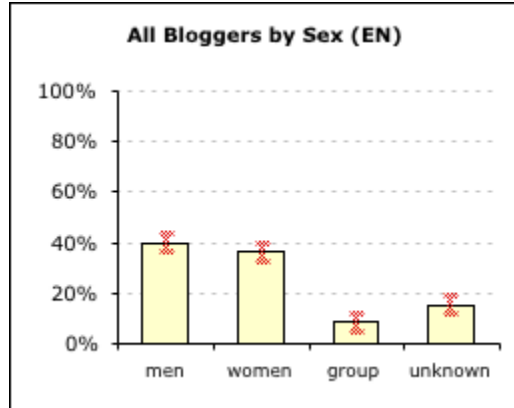


Figure 12 – Split of English-language Blogs based on Gender.

A closer content analysis of the blogs, however, showed that nearly half the blogs in the sample (368, or 47%) fell within the category of 'personal diary', which is dedicated entirely to recording the events of the blogger's life. Among the members of this group, women outnumbered men by about two to one (56% to 28%, with a margin of error of +/-4.8%). On the other hand, in the category of 'politics' which comprises about 6.2% of the total sites, only 4% were written by women (margin of error was 14.5%). These blogs focus primarily on politics, current events, foreign policy, and various ongoing wars (Figure 13).

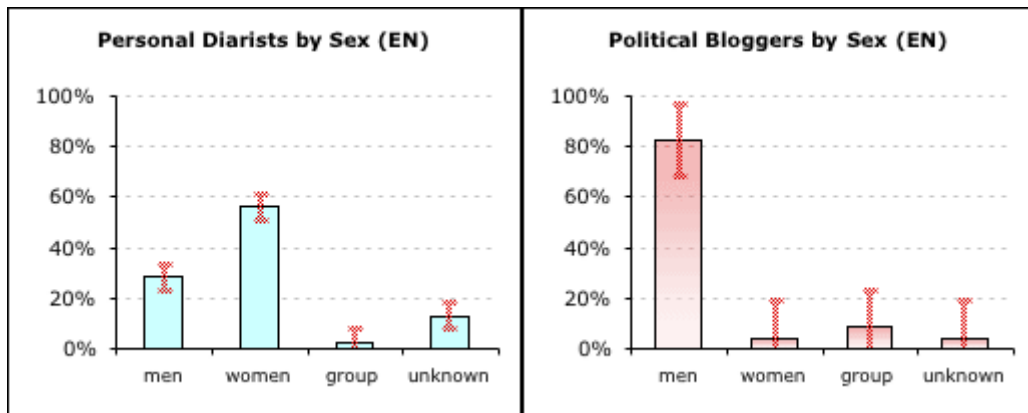


Figure 13 – Correlation of Gender and Blog Subjects in English-language Blogs.

Although the sample in this study was small, it does seem to point at a significant difference in what men and women tend to write about on weblogs. A similar study on Persian weblogs would be of interest, but more importantly, it could perhaps shed light on the claims that women use more informal language than men in blogs. The data here suggest that the linguistic distinction between the two genders could be directly related to the topic

of discussion since political issues are often written in formal style, while personal events are portrayed in colloquial language.²⁵

2.7.5 Applications for Persian Blogs

This section presented several research topics on the language of weblogs that could be used for a computational study of the Persian blogspace. An empirical investigation of the linguistic characteristics of Persian-language blogs would provide a tool for the comparative exploration of the global aspect of the blogosphere. In addition, it would set the stage for the development of computational models for analysis and processing of Persian language blogs. Furthermore, the application of text classification and stylometric methods can be of interest for sociolinguistic purposes as well as helping to pave the way for detection of sentiment and identification of individual differences in Persian-language blogs.

²⁵ One should not make the mistake of categorizing personal diaries in Iran as devoid of political import. Since overt discussion of much of daily life and issues are strictly regulated and even considered taboo in Iran, all these subjects take on socio-political value. In fact, they probably portray more of the true state of the society than more impersonal, formal political writing may communicate.

3 Individual Differences in Persian Blogs

3.1 Introduction

This section describes a preliminary analysis performed on text data in Persian blogs in order to delineate correlations between the gender of the author, the topic of the post, and the choice of language variant used. The steps taken to prepare the data set, perform data analysis, as well as the results and evaluation are discussed below.

3.2 Data Selection and Preparation

To ensure that the data used in the experiment represents the whole population, 22,000 posts were downloaded from 521 websites containing a variety of Persian blogs. A total of 22,000 posts were downloaded and 10% of the data was randomly selected for analysis. The random sample was then checked for irrelevant posts such as poems, non-Persian text, and posts that didn't contain enough text for the analysis. After removing the irrelevant posts from the randomly sampled data, 1,012 records remained for analysis.

The filtered data were reviewed manually and each record (post) was checked for three labels:

- **Gender**

If the author's name has been provided in the post, the gender is determined based on the name as either *male* or *female*. If the name is such that gender cannot be determined, such as *Farrokh* which refers to either male or female, or *yek bikar* 'a bum; unemployed' which is a nickname without revealing the gender, then gender is labeled as *unknown* for that record. Also if no name is explicitly provided, gender is labeled as *unknown*. In some cases, a group or organization name is given in place of an individual's name. In these cases, instead of *unknown*, the gender is labeled as *group/org* which gives a little more information than the *unknown*.

- **Morphology**

This parameter is determined based on the language variant used, namely *formal/literary* vs. *informal/conversational*. Note that this distinction is more significant in Persian than one may find in English text, affecting in particular the morphology in the word. If the record uses formal wordforms throughout the post, such as در هوای ابری آسمان را نگاه کنید (*dær hævaye æbri aseman ra negah konid* = 'Look at the sky in cloudy weather'), the record is labeled as *formal* and if conversational wordforms are used, such as تو هوای ابری آسمون رو نگاه کنین (*tu hævaye æbri asemoon ro negah konin*), the record is labeled as *informal*. In cases where formal forms are used mixed with conversational ones, such as *dær hævaye æbri aseman **ro** negah konid* (where the conversational form is shown in bold), then the record is labeled as *mixed*.

- **Category**

Depending on the topic discussed in the post, the record is labeled as one of the following categories:

- Art/Culture – discussion of topics related to art or culture (such as movies or certain cultural ceremonies)

- Journal – report of everyday events and diaries
- Literature – discussion of books, authors, or narrating a story/novel
- News – news about different topics without discussion of personal opinion
- Religion – discussion of topics related to religion (such as how to perform a certain prayer, stories about leaders of a faith, etc.)
- Social/Political – discussion of topics related to social or political events
- Technical – providing information about technical topics (such as computer software, hardware, network, etc.)

3.3 General Statistics about Data

Each record in the data consists of a unique ID in English, a narrative text in Persian, and three fields corresponding to the three labels of Gender, Morphology, and Category. After reviewing the narratives and labeling them, the IDs and narratives are removed from the records and the rest of the data is analyzed as described in the next section. The following Figures show general statistics about the data.

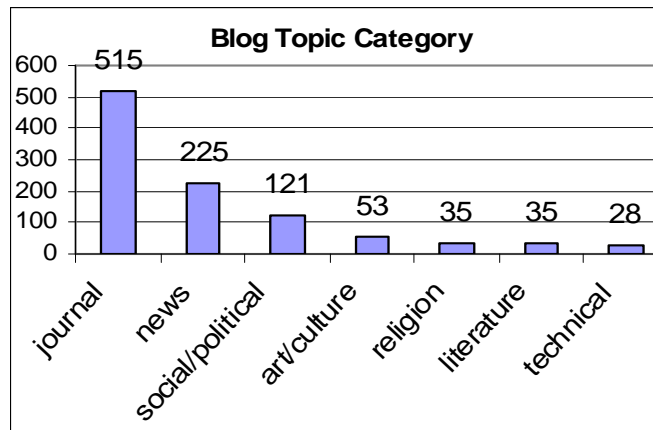


Figure 14 – Distribution of the topics in the data

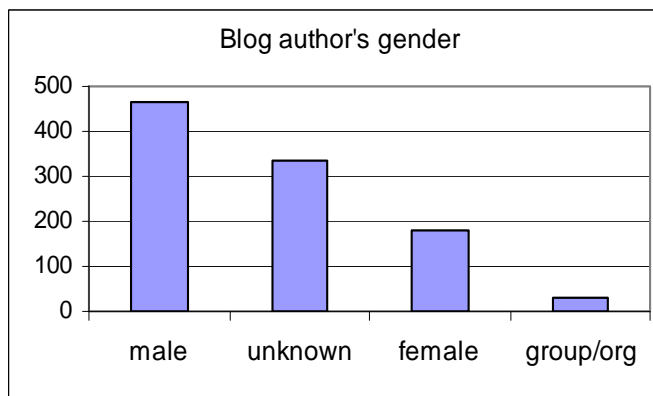


Figure 15 – Distribution of the gender in the data

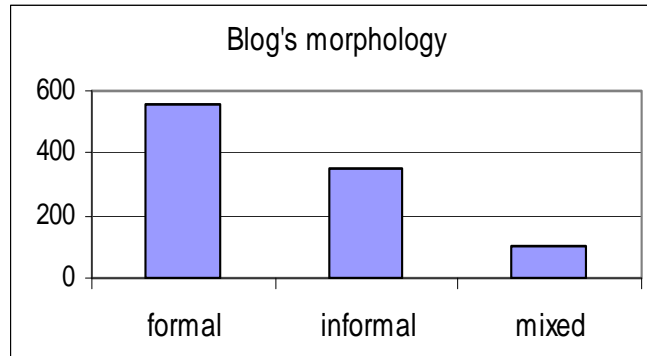


Figure 16 – Distribution of the formality in the data

Data Analysis. The following analysis was performed to identify strong patterns among the three label attributes of topic category, gender, and morphology. The SPSS Clementine implementation of the A-Priori algorithm is used to identify *strong* associations among the data attributes. The strength of the associations is determined by two user-defined thresholds called *support* and *confidence*. To understand these parameters, consider the following example:

(category = news) → (morphology = formal) supp = 21.7%, conf. = 90.4%

This association rule indicates that 90.4% of all news postings in the data have used formal morphology. So we can conclude that news postings use formal morphology, with 90.4% confidence. It also indicates that 21.7% of the data consists of news postings with formal morphology. In other words, 21.7% of the data support the rule.

To evaluate the findings of each technique, 80% of the data is randomly selected as the training set to be used by the algorithm. The remaining 20% of the data is used to test the generated rules and is referred to as the test data set. The rules obtained based on the training set are applied to the test data set to predict their attribute associations. Accuracy and precision of the rules are then measured based on comparison of actual and predicted values.

3.4 Analysis Results

Running the A-Priori algorithm on the training data set generated a number of "association rules". To limit the output to the top-most significant rules, the following parameters were used when running the algorithm:

support = 5%
 confidence = 75%
 optimize=speed
 expert system= rule confidence

Results of the analysis are shown in Table 4. Rules that contained 'unknown' category of authors are removed from the results.

Table 4 – A-Priori output for the training data set

	Support	Confidence	Consequent	Antecedent 1	Antecedent 2
1	21.7	90.4	morph = formal	cat = news	
2	34.6	81.6	cat = journal	morph = informal	
3	5.4	75.0	morph = formal	cat = art/culture	
4	11.3	94.6	morph = formal	gender = male	cat = news
5	11.8	82.3	cat = journal	gender = male	morph = informal
6	6.1	78.0	cat = journal	gender = female	morph = informal

The findings shown in Table 4 are interpreted as follows:

rule #1: (news) → formal

90.4% of the news postings in the data were formal

21.7% of the data contained this rule (i.e., news postings with formal morphology)

rule #2: (informal) → journal

81.6% of the informal postings were journal type

34.6% of the data contained this rule (i.e., informal postings with journal topic)

rule #3: (art/culture) → formal

75% of the formal postings in the data were art/culture topics

5.4% of the data contained this rule (art/culture postings with formal morphology)

rule #4: (male, news) → formal

94.6% of the formal posts were news written by male authors

11.3% of the data contained this rule (i.e., formal postings on news topics written by male authors)

rule #5: (male, informal) → journal

82.3% of the informal postings by male authors were journal type

11.8% of the data contained this rule (i.e., informal journal postings by male authors)

rule #6: (female, informal) → journal

78.0% of the informal postings by female authors were journal type

6.1% of the data contained this rule (i.e., informal journal postings by female authors)

3.5 Evaluation

To test the accuracy of these rules, they were applied to the test data set.

Table 5 through Table 10 indicate statistics pertaining to correct and incorrect predictions by the rules compared to the actual data points in the test data set. For example,

Table 5 indicates there were 40 postings with news topics written in formal morphology in the data and all 40 were correctly predicted by rule #2. There were also 8 postings with

news topics that did not have a formal morphology, the rule has correctly predicted 5 of them, 3 were predicted incorrectly.

The information shown in Tables 5 through 10 is referred to as a *confusion matrix*²⁶ and is used to calculate *accuracy* and *precision* measures for the rules. The *Accuracy* of a rule is calculated as the total number of its predictions (positive and negative) that were correct, divided by the total number of data points. *Precision* is calculated as the number of positive cases that were predicted correctly, divided by the total number of cases predicted as positive. Accuracy and precision calculated for each of the rules are indicated below, following their corresponding tables.

Table 5 - Confusion matrix for rule #1

(news)→ formal	predicted +	predicted -	total
actual +	40	0	40
actual -	3	5	8
total	43	5	48

Rule #1 accuracy: $45/48 = 93.75\%$

Rule #1 precision: $40/43 = 93.02\%$

Table 6 - Confusion matrix for rule #2

(informal)→ journal	predicted +	predicted -	total
actual +	55	0	55
actual -	26	19	45
total	81	19	100

Rule #2 accuracy: $74/100 = 74\%$

Rule #2 precision: $55/81 = 67.9\%$

Table 7 - Confusion matrix for rule #3

(art/culture)→ formal	predicted +	predicted -	total
actual +	7	0	7
actual -	1	2	3
total	8	2	10

Rule #3 accuracy: $9/10 = 90\%$

Rule #3 precision: $7/8 = 87.5\%$

²⁶ For more information see:

http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

Table 8 – Confusion matrix for rule #4

(male, news)→formal	predicted +	predicted -	total
actual +	29	0	29
actual -	0	2	2
total	29	2	31

Rule #4 accuracy: $31/31 = 100\%$

Rule #4 precision: $29/29 = 100\%$

Table 9 – Confusion matrix for rule #5

(male, informal)→journal	predicted +	predicted -	total
actual +	8	0	8
actual -	2	2	4
total	10	2	12

Rule #5 accuracy: $10/12 = 83.3\%$

Rule #5 precision: $8/10 = 80\%$

Table 10 – Confusion matrix for rule #6

(female, informal)→journal	predicted +	predicted -	total
actual +	13	1	14
actual -	0	4	4
total	13	5	18

Rule #6 accuracy: $17/18 = 94.4\%$

Rule #6 precision: $13/13 = 100\%$

To examine statistical significance of the above results, Chi-Square significance test was conducted. The result of this test is determined by the p-value associated with the Chi Square. Commonly a p-value of .05 or 5% is used as the threshold. Tests with a p-value of less than 0.5 are considered significant. As seen in Table 11, all relationships identified by the rules were found to be statistically significant.

Table 11 – Chi Square significance test results

rule #	Chi Square	p value	statistically significant
1	27.907	<0.0001	yes
2	28.669	<0.0001	yes
3	5.833	0.157	yes
4	31.00	<0.0001	yes
5	4.80	0.0285	yes
6	13.371	0.003	yes

3.6 Discussion of Results

Preliminary analysis points to interesting correlations between gender, topic, and language variant used in Persian language blogs. The results show strong patterns between the topic of blog posts and the language choice: As expected, reports and discussions of news items tend to be written predominantly in formal language, whereas the majority of posts written in conversational or informal language are journals. Taking into consideration the gender of the author indicates that although a male author writing a news post will use formal language by a significant margin, both male and female authors use the conversational language variant in journal writing.

The results seem to suggest that the choice of language variant is not determined solely based on the gender of the author as previously claimed (see Section 2.7), since the topic category plays an important determining role. Hence, journals – as opposed to news – are often written in conversational text and both males and females tend to use this variant in this particular blog genre. However, a larger set of annotated data is needed to be able to draw stronger conclusions on the influence of individual author differences in affecting language use.

Morphological Analysis of Conversational Text

3.7 Introduction

This report presents the findings of an informal evaluation of a Persian morphological analyzer on blog text. The goal is to study the output of a morphological analyzer that has been developed for literary or formal Persian on text written in conversational or informal language. The testfiles are posts obtained from three Persian blog sites. The report describes the preparation of the dataset, data analysis and the results of the evaluation.

3.8 Background

Persian websites have traditionally been written in the literary or formal language, which differs significantly from the everyday conversational language of Iran. The literary language is used mainly for writing and is the variant taught at school. Within the last few years, however, the conversational variant of the language has been used in writing emails, chats and weblogs. With the exponential growth of Persian weblogs, in particular, the conversational (or colloquial) language has become the main variant especially among the youth for writing journals, expressing personal opinions, and providing social criticism on the web.

The fundamental distinction between modern conversational (or informal) Persian and literary (or formal) Persian lies in the choice of lexical items and the inflection of word forms. As most existing computational systems of Persian have been developed for formal writing usually found in news reports, their application on conversational text has not really been studied. The goal of this evaluation, therefore, was to examine the output of an existing Persian morphological analyzer on blog text containing conversational language. A morphological analysis tool that can process conversational text found in blogs and offer the literary Persian equivalent will be able to associate blog vocabulary with dictionary forms which could ultimately benefit applications from Part-of-Speech (POS) tagging to machine translation and entity extraction.

For this informal evaluation, we selected the Shooka morphological analyzer, a unification-based Persian analyzer which is an extension of the system developed in the Shiraz project at CRL (Computing Research Laboratory, New Mexico). The tool was made available to MITRE for free for research purposes and no licensing was required.

3.9 Data Set and Annotation

For the purposes of the evaluation, we used 9 posts from two popular Persian blog sites, written by women bloggers who often use conversational Persian and discuss social, political, and personal issues. These two sites are www.khorshidkhanoom.com and www.z8un.blogfa.com. The posts varied in topic from personal journals to discussions of societal issues, to technical and computer-related subjects. Two additional posts were downloaded from www.4shanbe.blogfa.com, a male blogger, who discusses news items and only writes in the literary variant. Each post was manually annotated for Part of Speech information using the Callisto annotation tool (<http://callisto.mitre.org/>) in order to develop a groundtruth of a little over 3,000 entries.

TAGSET. For the purposes of the annotation, a tagset was developed specifically for Persian and integrated within Callisto. The tagset includes 37 annotation tags for Nouns, Adjectives, Adverbs, Verbs and closed class items such as Prepositions, Quantifiers, etc. Attempts were made to keep the number of tags low, hence the tags do not represent all possible

morphology on a particular part-of-speech. For instance, a singular noun is tagged as Nn (e.g., *ketab* 'book'), a plural noun is tagged as NnPl (e.g., *ketabha* 'books'), but a plural noun with an attached pronoun and an object marker would still be tagged as NnPl (e.g., *ketabhayeshuno* 'their books [obj]'). However, if a noun or adjective appears with the copula verb (i.e., the verb 'to be'), a special tag is used since the copula represents a verbal element and thus can be important for higher-level tasks such as POS disambiguation or parsing. The complete description of the tagset and relevant annotation guidelines can be found in Appendix C. All posts were manually annotated by a Persian native speaker in a first round and checked and edited by a second annotator in the second round.

```

- <Analysis id="Ana7" type="ptbpos_lex-tagset" role="ptbpos_lex-tagset">
- <AnnotationSet containedType="ptbpos_lex">
- <Annotation id="Ann790" type="ptbpos_lex">
  <RegionRef xlink:href="#Reg790" role="text-extent" xlink:type="simple" />
- <Content type="ptbpos_lex-content">
  <Parameter type="string" unit="NULL_UNIT" role="type">Det</Parameter>
  </Content>
  </Annotation>
- <Annotation id="Ann791" type="ptbpos_lex">
  <RegionRef xlink:href="#Reg791" role="text-extent" xlink:type="simple" />
- <Content type="ptbpos_lex-content">
  <Parameter type="string" unit="NULL_UNIT" role="type">Prop</Parameter>
  </Content>
  </Annotation>
- <Annotation id="Ann792" type="ptbpos_lex">
  <RegionRef xlink:href="#Reg792" role="text-extent" xlink:type="simple" />
- <Content type="ptbpos_lex-content">
  <Parameter type="string" unit="NULL_UNIT" role="type">NnPl</Parameter>
  </Content>
  </Annotation>
- <Annotation id="Ann793" type="ptbpos_lex">
  <RegionRef xlink:href="#Reg793" role="text-extent" xlink:type="simple" />
- <Content type="ptbpos_lex-content">
  <Parameter type="string" unit="NULL_UNIT" role="type">Nn</Parameter>
  </Content>
  </Annotation>

```

Figure 17 – Sample annotation output in XML from Callisto

GROUNDTRUTH. The final annotated files were saved in the Callisto XML format, which tags the document encoding, word spans, and the annotation tag for each span. A sample of the annotated file is shown in Figure 17.

The final groundtruth contained 11 files of varying length, totaling 3193 words (some of which were compounds). The number of conversational forms in each document was computed and can be seen in Table 12. The conversational forms were classified as follows:

- *Verb*: conversational forms of inflectional morphology on the verb (conversational *darǣn* 'they have' vs. literary *darǣnd*), and of the verbal stem form (*nǣgin* 'don't say' vs. *nǣguyid*).

- *Nonverbal*: conversational forms of inflectional morphology on nouns, adjectives, adverbs, pronouns (*dastanetun* 'your story' vs. *dastanetan*) and of the word stem (*un* 'he/she' vs. *an*).
- *Closed Class*: conversational forms of inflection on closed class items (*behes* 'to him/her' vs. *be u*) and of the word stem or word choice (*vase* 'for' vs. *bæraye*).
- *Loans*: words of foreign origin that are transcribed into Persian. Examples are *homofob* 'homophobe' or *padkæst* 'podcast'.
- *Foreign Words*: words of foreign origin written in Latin as in "https".
- *Interjection*: words such as 'uh', 'oh' used in text (*ey vay* 'oh god').

Table 12 – Groundtruth files ordered by percent of conversational forms in posts

File	Topic	Total Entries	Conversational Entries in Text						Total Conv.	% of Conv. In Text
			Verb	Non-Verbal	Closed Class	Loans	Foreign Words	Interj		
1	News	271	0	0	0	0	0	0	0	0.00%
2	News	141	0	0	0	0	0	0	0	0.00%
3	Politics	219	8	0	3	1	0	0	12	5.5%
4	Politics/Book	467	45	13	20	3	0	2	83	17.8%
5	Politics	340	31	19	20	3	0	2	75	22.1%
6	Journal	391	33	16	34	0	0	5	88	22.5%
7	Journal	725	55	51	49	13	1	4	173	23.9%
8	Tech/ Book	326	25	14	15	20	5	2	81	24.9%
9	Journal	16	0	0	4	0	0	0	4	25.0%
10	Tech/ Blogs	147	20	5	18	8	1	0	52	35.4%
11	Tech/ Blogs	150	14	8	20	11	3	0	56	37.3%
TOTAL NUMBERS:		3193	231	126	183	59	10	15	624	19.5%

As can be seen from Table 12, the two news-related posts did not contain any conversational forms at all. Some conversational morphology was found in the files related to politics (files 3 through 5), but the posts categorized as journals have a higher rate of conversational morphology and lexical items. Discussions of technical issues (often related to filtering and ways to avoid them) contain a large number of loans and foreign words.²⁷

3.10 Morphological Analyzer

The Shooka morphological analyzer used in the evaluation is a unification-based system that provides full morphological analysis and lexicon lookup. The results of the system are provided in a feature structure format which were mapped to the POS tagset developed for

²⁷ It should be noted that technology related documents received a high score in conversational forms due to the large amount of foreign and loan words they contain, while journals receive a high score mainly because of the conversational morphology on nominal, verbal, and closed class elements.

the project. The analyzer is completely knowledge-base and does not use statistical or machine learning methodology in disambiguation of the final output, hence the results may be ambiguous. In addition, when unable to analyze a word, the system "guesses" the part of speech based on the possible morphological rules. The morphological grammar was originally developed by the author for the Shiraz project²⁸ and reimplemented and extended for Shooka. The grammar was developed to cover only literary Persian. The system came with a very basic tokenizer, however, which gave rise to some issues which will be further discussed in the section entitled "Evaluation of Results".

3.11 Data Analysis

Each blog post was run through the Shooka morphological analyzer and the results were mapped to the Persian tagset developed for this project (see the Appendix). The results were judged based on accuracy and ambiguity level: For each post, the system results were compared to the groundtruth files by taking into account the number of correct alignments of the word entries, the number of correct POS tags, and the ambiguity the system generated in each instance. These numbers are shown in Table 13.

The "aligned" category refers to entries whose spans were correctly aligned in the groundtruth or reference file and the output of the Shooka system. For example, if a word with 2 characters was assigned the span 109 to 111 in the groundtruth, it is expected to receive the same span on the Shooka output file. There were several issues with the alignments that will be discussed below. Ambiguity is defined as the number of POS tags in the Shooka system output, divided by the total number of entries in the output.

Table 13- Preliminary results of morphological analysis test

File	Topic	Hits	Aligned	Number Entries in Reference	Number Entries in System	Number Tags in System	Ambiguity
1	News	0	0	271	303	654	2.158416
2	News	1	1	141	171	348	2.035088
3	Politics	148	179	219	246	749	3.044715
4	Politics/Book	291	336	467	550	1370	2.490909
5	Politics	187	216	340	394	1021	2.591371
6	Journal	230	275	391	401	1209	3.014963
7	Journal	340	396	725	615	1675	2.723577
8	Tech/ Book	173	208	326	366	992	2.710383
9	Journal	11	12	16	17	49	2.882353
10	Tech/ Blogs	81	105	147	175	530	3.028571
11	Tech/ Blogs	68	104	150	167	602	3.60479

There were a number of issues with the matching of alignments, arising from both the Shooka system and the Callisto-tagged annotation files, that affected the final results:

²⁸ <http://crl.nmsu.edu/Research/Projects/shiraz/>

- **Unknowns: Proper Names**

Certain proper names consisting of more than one token were not recognized by the Shooka system and were not tokenized as compounds, whereas they were tagged as a single entry in the annotation files. These were valid unknown entries and were expected in the evaluation.

- **Tokenization: compounds and separated morphemes**

The Shooka system tokenizer segments the words based on whitespace only, hence it does not recognize compounds or words with separated morphemes (affixes that are written detached from the stem). The recognition of such elements is performed at the pre-parsing stage in the Shooka system, which we did not have access to. Hence, all compounds and separated morphemes were analyzed as separate entries and were not put together as a single entry, giving rise to misalignments with respect to the groundtruth files.

- **Callisto annotation spans**

One major issue, however, was the spans in the annotation tags of the reference or groundtruth files. The annotation through Callisto sometimes includes whitespace along with the entry when the word is manually highlighted. The spans for each entry often included the whitespace elements. This gave rise to issues with the alignment since the spans ended up being larger than the actual entry itself and did not match the alignment determined by the Shooka system.

Table 14 - Evaluation scores (after minor adjustment to alignment scores)

File	Topic	Total Entries	Percent of Conv. in Text	Hits/ Entries	Hits/ Aligned	Ambiguity
1	News	271	0.00%	74.91%	90.62%	2.16
2	News	141	0.00%	73.76%	90.43%	2.04
3	Politics	219	5.48%	72.15%	82.72%	3.04
4	Politics/Book	467	17.77%	68.09%	85.02%	2.49
5	Politics	340	22.06%	61.76%	84.33%	2.59
6	Journal	391	22.51%	64.19%	80.96%	3.01
7	Journal	725	23.86%	49.66%	84.70%	2.72
8	Tech/Book	326	24.85%	55.83%	82.35%	2.71
9	Journal	16	25.00%	68.75%	91.66%	2.88
10	Tech/Blogs	147	35.37%	57.14%	73.04%	3.02
11	Tech/Blogs	150	37.33%	50.67%	64.95%	3.60
TOTAL NUMBERS:						
		624	19.54%	63.36%	82.80%	2.75

In order to correct the third alignment issue, the original annotation files had to be manually edited, but due to time constraints, we implemented a simple adjustment for the

purposes of this informal evaluation. The new scorer allowed for the referent to have one spurious space character at the end of the entry before matching the alignments. This took care of some of the alignment issues noted here but the others remained. The adjusted results for the alignment are shown in Table 14 in terms of percentages. This table lists "Hits/Entries" which consists of the number of correct POS tags over the total number of entries in the post; "Hits/Aligned" which represents the number of correct POS tags over the correctly aligned entries in the post; and "Ambiguity" measured as the number of tags in the system output over the number of entries in the system output.

The issues described above still play a large role in reducing the analysis scores overall. The issue with the alignment matching lowers the *hits/entries* results, while the inability of the system to treat compounds and separated morphemes within the analyzer module hurts the overall analysis results in both *hits/entries* and in *hits/aligned*. These two issues would need to be corrected and another evaluation performed in order to obtain more stable results. Nevertheless, the relative results on the various blog posts seem to indicate an interesting pattern, as illustrated in Figure 18, which shows the correlation between the percent of conversational entries in a blog post vs. the hits per total entries in the text and hits per correctly aligned entries in the document.

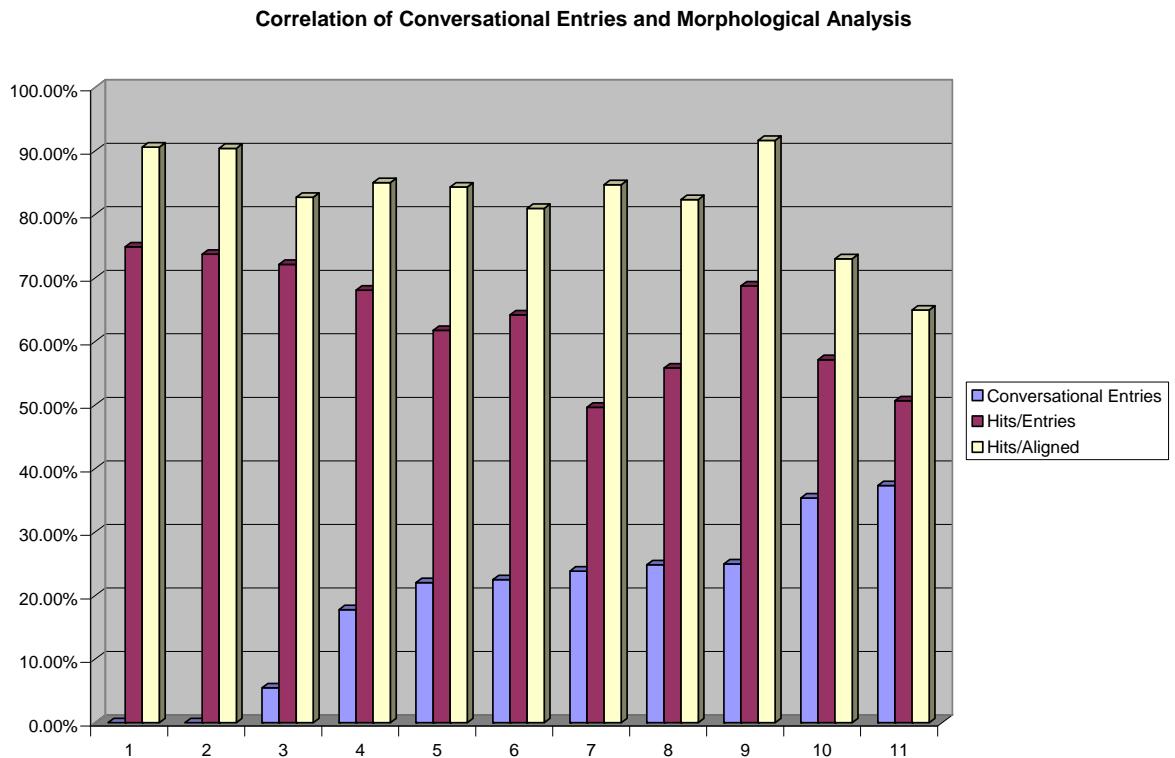


Figure 18 – Correlation of conversational entries and accuracy of morphological analysis

This correlation is perhaps better illustrated in Figure 19 which shows only the conversational entries (shown in blue) vs. hits per correctly aligned entries (shown in red) for each file. The morphological analyzer tends to get better scores for the news items that have very low or null conversational entries. And the lowest accuracy scores were obtained for the last two files that have the highest number of conversational forms. The only obvious

exception to this pattern is file #9, which had a number of conversational forms but had high accuracy scores; this file was a very short blog post consisting of only two sentences (16 entries).

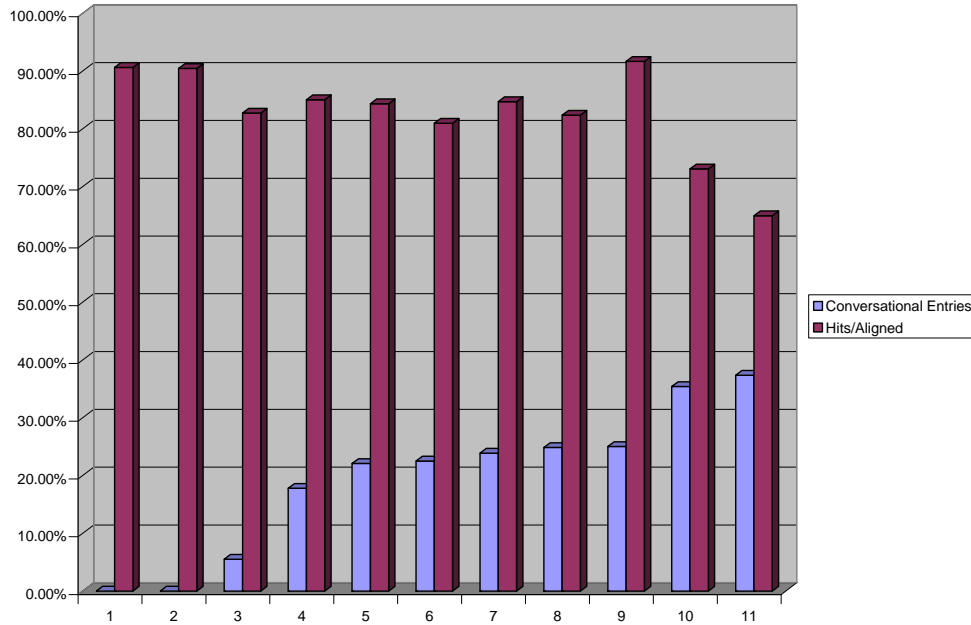


Figure 19 – Correlation of conversational entries and accuracy of POS tag per correctly aligned entry

Figure 20 also shows an interesting correlation emerging from the data with respect to ambiguity scores. As the level of conversational forms in the text increases (File 1 having the least amount of conversational morphology and File 11 having the most, shown on the x-axis) the ambiguity also seems to increase. This indicates that the morphological system encounters more unknowns in the text and thus generates more guesses as to their POS tag.

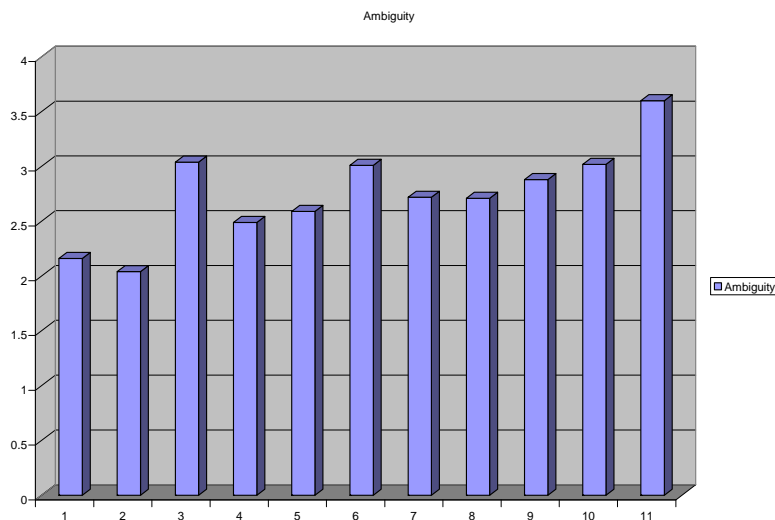


Figure 20 – Correlation of conversational entries and ambiguity per system analysis

The correlations noted in the dataset and the results suggest a direct relation between the number of conversational forms in a text and the difficulty of analysis for the system. There were, as already mentioned, certain issues with the alignment measurements which may have hindered the results. We therefore selected a representative sample of the results and manually examined them for effects of conversational forms on the system output, discussed in the following section.

3.12 Evaluation of Representative Sample

A representative sample was selected from the testfiles and analyzed manually. The total number of entries in the sample was 458, of which 398 were literary Persian forms and 60 consisted of conversational forms.

A close examination of the results of the Shooka system showed that 338 entries (about 74% of the total entries) were tagged correctly while 120 entries (about 26% of total entries) were mistagged. As previously mentioned, however, the Shooka system had a problem with segmenting and recognizing compound forms and separated morphemes; these entries formed about 12% of the whole document (57 entries). Note that this is close to half of the mistagged entries, hence a large chunk of the mistagged results were due to issues of tokenization.²⁹ The Table 15 represents the breakdown of the results per category where *Alignment* refers to the mis-segmented entries. *Literary* and *Conversational* refer to the variant of the forms found in the text. *Guessed Conversational* denotes elements of conversational form that were unknown but guessed correctly by the system. These numbers clearly demonstrate that most of the mistagged elements are of conversational form while wordforms from the literary variant are generally correctly tagged by the system.

Table 15 - Breakdown of system analysis per category

Tag	Number of Entries	Percent of Total	Category	Number of Entries	Percent of Total
Mismatch	120	26.20%	Alignment	57	12.44%
			Literary	14	3.06%
			Conversational	49	10.70%
Match	338	73.80%	Literary	327	71.40%
			Guessed Conversational	11	2.40%

Table 16 provides a breakdown of the literary and conversational system tags for this representative sample based on Part-of-Speech category, for informational purposes. Note that these represent the number of token entries tagged with a particular POS in the reference or groundtruth file; hence we have counted each entry whether it was previously

²⁹ If we eliminate the misaligned tokens and treat them as segmentation issues, then the number of correctly tagged entries becomes 84% of the total entries while the mistagged entries are about 16% of the total. Of the mistagged entries, about 22% are literary forms and 78% are conversational forms. On the other hand, of the correctly tagged entries, about 97% are literary while only about 3% are conversational forms.

encountered or not (e.g., the postposition mistagged 7 times was the same one, namely the object marker *ro* in its conversational form).

Table 16 – System tags analyzed per POS category

POS	Literary Matched	Literary Mismatched	Conversational Guessed	Conversational Mismatched
Noun	100	6	5	1
Verb	35	0	1	26
Adjective	22	1	0	3
Adverb	27	1	1	1
Proper Name	9	5	0	0
Pronoun	9	0	0	3
Preposition	42	0	1	3
Postposition	0	0	0	7
Conjunction	43	1	0	0
Determiner	11	0	0	1
Numeral	7	0	0	1
Quantifier	5	0	0	1
Relativizer	13	0	0	0
Question Word	4	0	0	0
Interjection	0	0	0	2
Loan Word (Noun)	0	0	3	0

It is interesting to note that the POS category missed most often is the verb in conversational form. These results suggest that adding morphological rules for the conversational variant of the verbal conjugation and the addition of the postposition *ro* could significantly improve the analysis results for blog text containing colloquial language.

3.13 Conclusion and Discussion

Despite issues with segmentation, the results obtained based on a scoring of the alignments and POS tags combined with a closer examination of a representative sample suggest that the presence of conversational forms in text do play a significant role in morphological analysis for tools developed primarily based on literary Persian. The statistical results suggest a direct correlation between the number of conversational forms and reduced performance. In addition, a closer examination shows that the majority of mistagged elements are of conversational form. A system that provides guesses based on the word forms can provide better results although the ambiguity is increased as more (unknown) conversational forms are encountered in text. Furthermore, the preliminary results of this informal evaluation suggest that the addition of verbal conjugation rules for conversational language can significantly improve the results of analysis on blog text.

4 The Language of Persian Weblogs

4.1 Summary

Since its beginnings in 2001, the Persian blogosphere has undergone a dramatic growth making Persian one of the top ten languages of the global blog community in 2007. The Persian blogosphere has opened the door to journalists, intellectuals, and University students who use blogs to evade government censorship or social and political restrictions. This new medium has also provided a forum for bloggers to express their opinions and thoughts in their everyday speech rather than the traditional literary language. This has given rise to a variance in vocabulary, word forms, and sentence structures on Persian weblogs. This report provides the first overview of the main features of Persian blog language and provides a detailed analysis of its word-level properties.

Standard conversational Persian³⁰ – which more directly reflects the way people speak – is very distinct from the more prestigious, literary variant of the language mainly used for writing and in news. With the popularity of blogs, however, conversational Persian has become pervasive in the written domain giving rise to a variation of linguistic forms that did not exist before in text. In addition, this diversity is accentuated by variants of orthographic forms found online as each social group has defined its own standards or writing approaches: (i) the traditional orthography taught at school and recommended by the Persian Language Academy has strict rules of spelling and spacing; (ii) journalist and intellectual bloggers have recently proposed their own guidelines for Persian orthography that differ from the traditional rules; and (iii) bloggers using conversational language, mostly the youth, write the words as they are pronounced in spoken Persian and do not have any set standards.

This section begins with an introduction to the Perso-Arabic writing system and its intricacies. The lexicon of conversational language is then studied by presenting a detailed description of the modifications in the pronunciation of Persian words and how they influence the spelling seen on blogs (e.g., *hizhdæ* 'eighteen' vs. literary *hejdæh*). A large number of new words have been borrowed from other languages, in particular English, and are often used by young bloggers as in *anlayn* 'online' or *es em es* 'SMS, text message'. Furthermore, bloggers continually create new words to fit their needs, especially in the domain of the Internet (e.g., *filtershekæn* 'anti-filter software'). A comprehensive description of nominal and verbal inflection in both the literary and the conversational variants are provided, discussing the modifications and permutations seen in Persian word structures (e.g., how literary *khanemanra* 'our house' becomes *khunæmuno*). The verbal paradigm is the one affected most in conversational language as it has shortened verbal stems and inflectional endings (e.g., *migæn* 'they say' vs. literary *miguyænd*) and makes more frequent use of attached pronouns giving rise to forms absent from the literary dialect (e.g., *gereftætæm* 'has caught me' vs. literary *mæra gerefte æst*). At the sentential level, the conversational language possesses certain features that do not exist in the literary form such as the definite article and subject pronouns that attach to verbs, and makes more

³⁰ Standard conversational Persian reflects mainly the dialect spoken in Tehran, the capital of Iran. There are many other conversational dialects in Iran but these are not represented as prominently in blogs.

frequent use of existing constructions such as freer word order permutations and topicalization of the main theme by moving it to the beginning of the sentence. Conversational text also contains more instances of idiomatic expressions, jargons and other non-dictionary words, as well as cultural inferences. In addition, blogs often contain deviant spelling, spelling errors, as well as non-standard spacing and punctuation, giving rise to bigger variation for a computational processing of online text.

The main focus of this section is to provide a description of Persian for computational linguistic implementation and thus emphasis is placed on textual and orthographic issues encountered on blogs. The format and content can easily be adapted to serve different audiences. For instance, the format and level of detail can be changed to serve as a descriptive aid for language analysts of Persian not intimately familiar with the conversational variant.

4.2 Persian BlogSpeak

Since its beginnings in 2001, the Persian blogosphere has undergone a dramatic growth making Persian one of the top ten languages of the global blog community in 2007 (Sifry 2007, NITLE Census 2007). The exponential growth of Persian blogs – websites where entries are made and displayed in a reverse chronological order – has attracted much attention in the international media. Many reports in the United States have described how marginalized groups in Iran, such as the youth and women, use blogs to evade the strict regulations imposed upon them by the state, express their thoughts and opinions on the political and social situation, coordinate or influence political activities, or record their daily lives. Research on Persian blogs has mainly centered around a socio-political study of this new medium, and several quantitative investigations have provided preliminary analyses on sociological characteristics and content analysis in weblogs. However, rigorous research and investigation of the linguistic aspects of Persian blogs and computational analysis of these online resources are lacking.

The *diglossic* situation of Persian, whereby two distinct varieties of the language coexist in the society, is also reflected in the language found in the Iranian blog community. Traditionally, Persian literature and news media have been written in the literary dialect which holds a higher prestige over the conversational form of the language. Although the latter has been used in some works of modern literature, its usage is generally limited to the informal, conversational domains and is rarely seen in written form. With the advent of blogs, the restrictions against the use of the conversational dialect in writing have been challenged and, despite strong criticisms from intellectuals and professional journalists, bloggers often use the conversational Persian variant in their posts. This creates a new challenge for the analysis of Persian language websites as current grammars and academic textbooks of Persian focus mainly on the literary dialect and existing text-based computational systems often fail to analyze or process conversational Persian.

Preliminary exploration of the language of Persian blogs shows parallels with English BlogSpeak. As noted by Crystal (2001) for English, the content of a site (e.g., information, education, diary) strongly influences the general character of the language being used leading to linguistic variation on the Internet. This observation holds for the Persian language websites as well. Hence in both English and Persian, the language of blogs that address personal thoughts, opinions, and issues has been characterized as a conversational style in writing. As previously mentioned in Section 2.7.1, non-standard spelling that reflects the colloquial pronunciation of words is often used. Blog entries are usually written

in short sentences and include a large number of hyperlinks. Deviant spelling is common and standard orthography is often ignored, opting instead for a more intimate style. Emotions are expressed with emoticons, ellipsis, repetition of letters and punctuation marks, and emphasis is shown with capitals (if available in the writing system) and special symbols. Jargons and neologisms abound in Blogspeak, especially based on technical or computer-related terms.

Persian Blogspeak differs from that of English, however, due to its strong diglossic situation. Crystal notes that in English, as with language change in general, most features that distinguish Netspeak from previous genres are currently found chiefly in graphology and the lexicon – the levels of language where it is relatively easy to introduce innovation and deviations – while syntactic or grammatical variation is less frequent. Yet, the distinction between the literary and conversational language is especially poignant in Persian, affecting morphology and syntax as well. Persian Blogspeak often includes properties corresponding to the conversational language such as shortened verbal stems, frequent use of attached pronoun forms, and affixes that are not part of the standard formal grammar. There are more instances of free word order, idiomatic expressions, loan words, and an inordinate amount of orthographic variance partly due to the flexibility and ambiguity of the Perso-Arabic script.

This section provides a detailed descriptive analysis of Persian Blogspeak from a computational perspective, presenting both the traditional, literary variant of the language as well as the characteristics of the conversational variant. In addition, orthographic variances are explored in each instance. As in the rest of this report, the focus of this section is on weblogs written in Persian either within Iran or in the expatriate Iranian community and it does not specifically address Afghani or Tajiki sites.

4.3 Persian Writing System

The Persian language spoken in Iran and Afghanistan uses an extended version of the Arabic alphabet; it includes four additional letters that do not exist in standard Arabic: *pe* (پ), *che* (چ), *zhe* (ژ) and *gaf* (گ). Tajiki Persian, however, is written in an extended version of the Cyrillic alphabet. The Perso-Arabic script possesses a number of the features of the Arabic alphabet, including its more ambiguous properties such as the lack of certain vowels in the script.³¹ Texts are written from right to left. The vowels /æ/, /e/ and /o/ are usually not written; the vowels /i/, /u/ and /a/ are represented in the text. This of course creates certain ambiguities: Since the short vowels are not inscribed, the word کرم [krm], for instance, can be pronounced with different vowel combinations resulting in five possible lexical elements. A reader uses the context to determine the word in the sentence.

kerm 'worm', *karam* 'generosity', *kerem* 'cream', *korom* 'chrome', *karm* 'vine'

Persian has adopted the Arabic alphabet, but since the languages do not have the same sounds, the characters are pronounced quite differently in the two languages. Hence in Persian the four letters *ze* (ز), *zal* (ذ), *zat* (ض) and *za* (ظ) are all pronounced /z/ and *sin*, *sat*, *se* are all pronounced /s/. Since diacritics are not marked in text, the Persian letter *alef* (ا) in the beginning of a word can be pronounced as /æ/, /e/, or /o/. If *alef* appears with the

³¹ See Appendix B for the Persian alphabet as well as the transliteration and transcription schemes used in this report. Throughout this document, transliteration of words is given in square brackets while the actual pronunciation is shown in italics.

madd (also known as 'the hat') (ّ) it is pronounced as /a/. The letters *vav* (و), *he* (ه) and *ye* (ی) can be either a consonant or a vowel given the context. As consonants, they are pronounced /v/, /h/ and /y/, respectively. As vowels, they double as /u/ (as in 'food') or /ow/ (as in 'door'), /e/ and /i/.

In the Perso-Arabic writing system, letters in a word are often connected to each other. Most characters have a different form depending on their position within the word. The initial form indicates that no element is attached to the element from the right (i.e., there is no "attaching" character before it, but there is one following the character). Note that an initial form does not mean that the character is in the beginning of a word, it only indicates that the character is not at the end of the word. Characters are in medial form if they have an attaching character both before and after them. The final form denotes that the character is at the end of a word. Final forms can therefore be used to mark word boundaries. Certain characters (*alef* (ا), *dal* (د), *zal* (ذ), *re* (ر), *ze* (ز), *zhe* (ژ), *vav* (و)) have only one form regardless of their position within the word.

In traditional written text, words are usually separated by a space. Compounds and detachable morphemes (i.e., morphemes following a word ending in final form character), however, are written without a space separating them. In these situations, the two parts of a compound appear next to each other but the first element in the compound will usually end in a final form character³²; hence it would be possible to recognize the two parts of the compound. This format, however, is not very consistent: words may appear without a space between them in online documents. If the first word ends in a character that has a final form, then we can easily distinguish the word boundary. But if the first word ends in one of the characters that have only one form, the end of the word is not clear. Although this latter case is usually avoided in traditional written text, it is not rare and it is even more common in blogs. Furthermore, a space is sometimes inserted between a word and the detached morpheme.

One of the intricacies of the Persian writing system is the 'silent *he*' character. This character is used to represent a word-final, stressed /e/ sound. Hence, if the *he* character follows a vowel it will be pronounced as a /h/, as in *shah* 'king' written شاه [šah] or *kuh* 'mountain' written کوه [kvh]. Note that the vowel may be unwritten as in *meh* 'fog' written as مه [mh] since the vowel /e/ is not generally transcribed. If it follows a consonant, however, *he* is pronounced as /e/ as in the Persian-origin word *name* 'letter' written as نامه [namh] and the Arabic loanword *kælæme* 'word' written as کلمه [klmh]. What is important is that the words where *he* is pronounced as /h/ can appear with attached suffixes, whereas the words in which *he* is pronounced as the vowel /e/ do not allow any suffixes to attach to them. This is illustrated in the table below, with the plural suffix *ha*, for *kuh* 'mountain' and *name* 'letter'.

³² This is achieved by placing a "half-space" or "short space" in Windows (by holding together the Control key and the minus sign in Windows XP or by holding down the Control, Shift and 2 simultaneously: Ctrl+Shift+2). In Unicode this is achieved by the ZWNJ (zero-width nonjoiner) control character, which has the code \u200c. For transliteration purposes in this report the tilda (~) has been used to denote this control character.

Isolated Form (intervening full space)	Detached Form (intervening half-space)	Attached Form
کوه ها [kvh ha]	کوه‌ها [kvh~ha]	کوهها [kvhha]
نامه ها [namh ha]	نامه‌ها [namh~ha]	N/A

As we will see in the following sections, there is large orthographic variance in the material on Persian blogs, raising new ambiguities and providing many challenges for an analysis of Persian Blogspeak. Any analysis needs to take into account the features encountered in both literary and conversational writing, as well as the enormous amount of variation existent within these textual forms.

4.4 Lexicon

When a language undergoes change, the domain affected first and foremost is the lexicon. The words used in conversational Persian are very distinct from their literary equivalents as they have experienced a number of phonological changes that are reflected in conversational text. In addition, conversational language uses new loanwords as well as newly coined terms, jargon and colloquial expressions, and idiomatic expressions that are rarely used in literary text. Finally, a number of orthographic tendencies have modified the spelling of lexical elements. The following describe some of the issues to consider:

PHONOLOGICAL ALTERNATIONS. Words are often written as they are pronounced in modern conversational Persian resulting in new wordforms. These will be discussed at length in Section 4.4.1.

COLLOQUIAL FORMS. Certain literary words have been replaced by colloquial counterparts that are used frequently in blogs, but never appear in traditional literary text. Some of these are introduced in Section 4.4.2.

LOAN WORDS. A large number of new loan words (in the original language or transcribed into Persian) can be found in blogs, in particular words relating to technology and computers. Some examples are provided in Section 4.4.3.

NEOLOGISMS. Bloggers often create new words, generally following the word-formation rules of the language. These words often contain a loanword as a subpart. In addition, one of the responsibilities of the Persian Language Academy is to coin new words to replace newly entered loanwords. See Section 4.4.3 for some examples.

INTERJECTIONS. Conversational blogs use many forms of interjections and emoticons, such as آآآخ! [AAAx!] (*aaakh!*) 'ouch', اووه! [avvvh!] (*oooh!*) 'oh', وای [vay] (*vay*) 'ah', والا [vala] (*vala*) 'well'.

4.4.1 Phonological Alternations

This section discusses several phonological changes in the Standard Persian of Iran that has affected the pronunciation of words in the conversational language. These patterns have to be reflected in the lexicon or applied to the lexical elements in order to be able to provide a computational analysis of the conversational variant of Persian.

AN TO UN ALTERNATION

In most phonetic environments an /a/ vowel changes to /u/ when followed by the alveolar nasal /n/as shown in the following examples.³³

<i>zæban-eman</i>	→	<i>zæbun-emun</i>	زبونمون
language-our		language-our	
'our language'			

Certain inflectional affixes such as the clitic pronouns discussed in Section 4.5.3 also undergo this alternation. The example above shows the /an/ to /un/ alternation both in the stem of the word *zæban* 'language, tongue' and in the clitic pronoun *-eman* 'our'. Further examples are shown below:

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>khane</i>	<i>khune</i>	خانه [xvnh]	'house'
<i>irani</i>	<i>iruni</i>	ایرونی [ayrvny]	'Iranian'
<i>anja</i>	<i>unja</i>	اونجا [avnja]	'there'
<i>nan</i>	<i>nun</i>	نون [nvn]	'bread'
<i>pushandæn</i>	<i>pushundæn</i>	پوشوندن [pvšvndn]	'to cover'
<i>zendani</i>	<i>zenduni</i>	زندونی [zndvny]	'prisoner'
<i>jævane</i>	<i>jævune</i>	جوونه [jvvnh]	'sprout'
<i>mian</i>	<i>miun</i>	میون [myvn]	'between'
<i>baran</i>	<i>barun</i>	بارون [barvn]	'rain'

This phonological alternation does not apply in all conditions where the word contains /an/, however:³⁴

- Proper person names usually do not change, such as *mowlana* (name of Rumi the poet) which is never pronounced as *mowluna*. Similarly, if *ræfsænjani* is used as an adjective and refers to a person from the city of Rafsanjan, it can be modified to *ræfsænjuni*; but if it refers to the former president of Iran, Ayatollah Rafsanjani, it cannot have the form *ræfsænjuni*. The female name *iran* can not be modified to *irun*, although the name of the country is often pronounced as *irun*. Another similar example is *tærane* which can become *tærune* if it means 'song' but cannot be changed if it is referring to a female name.

- If /an/ is the plural morpheme, it can never appear as /un/ (ex. *mærdan* → **mærdun* 'men', but *mærdane* → *mærdune* 'of or pertaining to men'). On the other hand, the present participle affix /an/ can undergo the alternation (*khændan* → *khændun* 'laughing') as can the causative affix in *tærsandæn* 'to frighten' which can become *tærsundæn*.

- The /ane/ suffix that forms an adverb or adjective is generally modified to /une/ (*zænane* → *zænune* 'of or pertaining to women', *asheghane* → *asheghune* 'loving, lovingly'). These

³³ This feature is also shared by many Iranian dialects.

³⁴ In the following examples in this section, an asterisk (*) indicates an ungrammatical or disallowed construction or word.

forms can be used as both adjective or adverb based on the context. It seems, however, that this alternation is much more limited in cases where the /ane/ suffix is used to form an adverb only that cannot be used as an adjective. Hence, *motæsefane* → *motæsefune* 'unfortunately' is very rare compared to *jævanmærdane* → *jævanmærdune* 'gentlemanly' which can be used as an adjective. In these instances, the acceptability tends to vary based on the word, however. For example, for most speakers the adverbs *khoshbækhtane* 'fortunately' and *bædbækhtane* 'unfortunately' do not allow the alternation at all: **khoshbækhtune*, **bædbækhtune*.³⁵

- Note that words that originally include a glottal sound before the /an/ do not undergo this alternation. For instance, the words *ghoran* 'Koran' and *ælan* 'now' are both conversational variants of the original *ghor'an* (written as قرآن [qrAn]) and *æl'an* (written as الآن [alAn]), respectively, where the apostrophe represents the glottal stop. These words then do not have a conversational form *ghorun* or *ælung*.

- Loanwords (words that were borrowed relatively recently from another language) do not show the alternation: *maman* → **mamun* 'mom', *bank* → **bunk* 'bank', *tank* → **tunk* 'tank', *roman* → **romun* 'novel'.

- Examples of some words that are not modified: *dastan* → **dastun* 'story', *rayane* → **rayune* 'computer', *ostovane* → **ostovune* 'cylinder', *khyanæt* → **khyunæt* 'betraying'. Also, *dæbestan* → **dæbestun* 'elementary school' but compare to *tabestan* → *tabestun* 'summer' which has a similar word structure yet allows the /an/ to /un/ alternation. Note also that certain Iranian cities undergo the alternation while others don't: *tehran* → *te:run* 'Tehran', *esfæhan* → *esfu:n* 'Isfahan' vs. *abadan* → **abadun* 'Abadan', *hæmædan* → **hæmædun*. One possibility is that the /an/ to /un/ alternation depends on the frequency of the word so that less frequent words in conversational speech may not be subject to this phonological alternation. Although frequency may play an important role, it fails to explain some of the instances observed. Another hypothesis proposed in Esmaili (1998) suggests that this alternation has stopped being a productive part of the language grammar, and therefore new words such as the newly coined *rayane* 'computer' do not change. Similarly, *dæbestan* 'elementary school' and *dæbirestan* 'high school' are relatively modern concepts in Iran (since the end of the 19th century) and thus may have been coined after the /an/ to /un/ phonological alternation process had stopped being productive in Persian.

AM TO UM ALTERNATION

In a limited set of words the /a/ vowel changes to /u/ when followed by the bilabial nasal /m/.

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>hæmam</i>	<i>hæmum</i>	حمام [Hmvm]	'bath'
<i>tæmam</i>	<i>tæmum</i>	تموم [tmvm]	'finished; all (of)'
<i>amædæn</i>	<i>umædæn</i>	اومدن [avmdn]	'to come'

³⁵ A search on Google provided the following comparative results: *jævanmærdune* (11 instances used as adverb), *motæ:sefune* (1 instance), *khoshbækhtune* (3 instances), *bædbæxtune* (5 instances).

aram *arum* آرام [Arvm] 'calm'

NB TO MB ALTERNATION

A very consistent pattern in Persian is to pronounce the alveolar /n/ preceding the bilabial /b/ as a bilabial nasal /m/. In other words, any /n/ sound followed by /b/ is pronounced as /mb/. Note that this alternation does not only happen in the conversational variant but also in the literary variant. However the original, literary writing of these words is maintained as /nb/ while conversational spelling often reflects the actual pronunciation of the words with /mb/.

Pronunciation	Script/Transliteration (lit.)	Script/Transliteration (conv.)	Translation
<i>shæmbe</i>	شنبه [ʃnbh]	شمبه [ʃmbh]	'saturday'
<i>zæmbur</i>	زنبور [znbvr]	زمبور [zmbvr]	'bee'
<i>dombal</i>	دنبال [dnbal]	دمبال [dmbal]	'following, after'
<i>æmbar</i>	انبار [anbar]	امبار [ambar]	'storage'
<i>tæmbæl</i>	تنبل [tnbl]	تمبل [tmb]	'lazy'
<i>pæmbe</i>	پنبه [pnbh]	پمبه [pmbh]	'cotton'
<i>jomb-o-jush</i>	جنبوجوش [jnb~vjvʃ]	جمبوجوش [jmb~vjvʃ]	'motion'

VOICE ASSIMILATION 1

Some voiced phonemes may become voiceless prior to a voiceless consonant, and vice versa³⁶. For example, the word written as [vqt] is actually not pronounced *væght* but rather *vækht*. This is because the 'gh' sound appears before a 't' which is a voiceless sound; it therefore assimilates in its voice properties and modifies to its voiceless equivalent, namely 'kh'. Similarly, the word written as [hfdh] is actually pronounced *hivdæ* rather than *hefdæ*. In this instance, the 'f' changes to its voiced equivalent 'v' because the following letter 'd' is voiced. The following are some examples of devoicing (voiced going to voiceless) which are often represented in the spelling of conversational Persian:

³⁶ A voiced sound is one in which the vocal cords vibrate, and a voiceless sound is one in which they do not. Examples of voiced and voiceless pairs in Persian (with a pronunciation guide) are given below:

Voiceless sounds	Voiced equivalent
<i>p</i> (pin)	<i>b</i> (bin)
<i>t</i> (ten)	<i>d</i> (den)
<i>k</i> (con)	<i>g</i> (gone)
<i>ch</i> (chin)	<i>j</i> (gin)
<i>f</i> (fan)	<i>v</i> (van)
<i>s</i> (sip)	<i>z</i> (zip)
<i>sh</i> (pressure)	<i>zh</i> (pleasure)
<i>kh</i> (German Bach)	<i>gh</i> (French 'r')

Script/Transliteration (lit.)	Pronunciation	Translation
وقتیکه [vq̄tykh]	væ <i>kh</i> tike	'when'
نقشه [nq̄ʃh]	næ <i>kh</i> she	'map, plan'
رقصیدن [r̄q̄Sydn]	ræ <i>kh</i> sidæn	'to dance'
اسب [asb]	æ <i>s</i> p	'horse'
ضبط کردن [ZbT kr̄dn]	zæ <i>pt</i> kær̄dæn	'to record'

VOICE ASSIMILATION 2

Another case of voice assimilation occurs in verbs in conversational Persian, where the /t/ sound is pronounced closer to /d/ before a vowel. The pronunciation is sometimes reflected in writing.

داشتم می‌رفتم dash <i>t</i> æm miræf <i>t</i> æm have-1SG going-1SG	→	داشدم می‌رفدم dash <i>d</i> æm miræf <i>d</i> æm 'I was (in the process of) going'
---	---	--

ST AND ZD CLUSTERS

In the /st/ and /zd/ clusters the /t/ and /d/ are sometimes not pronounced (they are assimilated into the preceding sound). Before a consonant, the /st/ and /zd/ clusters become simply /s/ and /z/, respectively. Before a vowel, they are pronounced as /ss/ and /zz/. Hence, the word *dæst* 'hand' is pronounced as *dæs* in conversational form. But if it appears with the possessive pronoun clitic as in *dæstæm* 'my hand' then it is pronounced as *dæssæm*. Note however that this alternation is optional.

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>bæstæni</i>	<i>bæssæni</i>	بسنی [bsny]	'ice cream'
<i>nuzdæh</i>	<i>nuzzæ</i>	نوزه [nvzh]	'nineteen'
<i>kojast</i>	<i>kojas</i>	کجاس [kojas]	'where is he/she/it?'
<i>nist</i>	<i>nis</i>	نیس [nis]	'isn't'
<i>hæstim</i>	<i>hæssim</i>	هسیم [hsym]	'we are (here)'
<i>peste</i>	<i>pesse</i>	پسه [psh]	'pistachio'
<i>shekæste</i>	<i>shekæsse</i>	شکسه [ʃksh]	'broken'
<i>pæræstar</i>	<i>pæræsar</i>	پرسار [prsar]	'nurse'
<i>ezevaj</i>	<i>ezevaj</i>	ازواج [azvaj]	'marriage'

WORD-FINAL CLUSTERS

Similar to the clusters discussed above, the final /t/ and /d/ are optionally not pronounced in word-final consonantal clusters as shown below.

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>ræft</i>	<i>ræf</i>	رف [rf]	'he/she/it left'
<i>anvækht</i>	<i>unvækh</i>	اونوخ [avnvx]	'then [lit. that time]'
<i>bolænd</i>	<i>bolæn</i>	بلن [bln]	'tall, long, high (sound)'
<i>fekr mikonæm</i>	<i>fek mikonæm</i>	فک میکنم [fk myknm]	'I think'

J TO ZH ALTERNATION

The sound /j/ changes to /zh/ before /d/:

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>hejdæh</i>	<i>hizhdæ</i>	هیژده [hyždh]	'eighteen'
<i>æjɖad</i>	<i>æzhdad</i>	اژداد [aždad]	'ancestors'
<i>mæjbur</i>	<i>mæzhbur</i>	مژبور [mžbvr]	'forced'

CH TO SH ALTERNATION

Although the 'ch' to 'sh' alternation is more common in other dialects, such as in Hamadani Persian, it can still be noticed in certain contexts in Standard Persian as in *hichkæs* 'nobody' becoming *hishki* or *hichvæght* 'never' being pronounced as *hishvækh* in conversational language.

GLOTTALS

The sound /h/ (represented by ه or ح) and the glottal sound represented by 'eyn' (ع) and 'hamze' (ء) are normally dropped in the middle or end of the word in conversational Persian. Instead, if there is a vowel preceding the glottal sound, it is lengthened (lengthened vowels are indicated with a colon (:)) below and glottal sounds are marked with the apostrophe).

Literary	Script/Transliteration (lit.)	Conversational	Translation
<i>tehran</i>	تهران [thran]	<i>te:run</i>	'Tehran'
<i>dæh</i>	ده [dh]	<i>dæ:</i>	'ten'
<i>ketabha</i>	کتابها [ktabha]	<i>ketaba:</i>	'books'
<i>chahar</i>	چهار [čhar]	<i>cha:r</i>	'four' ³⁷
<i>'æza</i>	اعضا [æZa]	<i>æ:za</i>	'members'
<i>næ'na</i>	نعنا [nena]	<i>næ:na</i>	'pistachio'
<i>ghor'an</i>	قرآن [qrAn]	<i>ghora:n</i>	'Koran'

³⁷ The word [čhar] is also sometimes pronounced as *chahar* and written as چاهار [čahar] in the conversational variant.

motæ'sefane متأسفانه [mta'sfanh] *motæ:sefane* 'unfortunately'

The orthography of these words is not always modified in the written form of the conversational variant in order to reflect the pronunciation. Instead, apart from the orthography of the plural morpheme (as in the third example) where the 'h' is dropped, the original spelling is often maintained.

/E/ TO /I/ ALTERNATION

The mid front vowel /e/ is sometimes changed to the high front vowel /i/ in the conversational variant. This change is optional.³⁸

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>kelid</i>	<i>kilid</i>	کيليد [kylyd]	'key'
<i>belit</i>	<i>bilit</i>	بيليط [bylyT]	'ticket'
<i>hejdæh</i>	<i>hizhdæ</i>	هيزده [hyždh]	'eighteen'
<i>hefdæh</i>	<i>hivdæ</i>	هيوده [hyvdh]	'seventeen'
<i>kuchek</i>	<i>kuchik</i>	کوچیک [kvčyk]	'small'
<i>shesh</i>	<i>shish</i>	شیش [šyš]	'six'
<i>englis</i>	<i>inglis</i>	اينگليس [aynglys]	'England'
<i>negah kon</i>	<i>niga kon</i>	نيگا کن [nyga kn]	'look!'
<i>chekar</i>	<i>chikar</i>	چيکار [čykar]	'what' (as in 'what are you doing?')

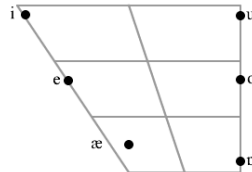
VOWEL HARMONY

In certain bisyllabic words, the vowels tend to match each other. This change is optional.

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>sahebkhane</i>	<i>sahabkhane</i>	صاحبخونه [SaHab~xvnh]	'landlord'
<i>sholugh</i>	<i>shulugh</i>	شولوغ [švlvQ]	'busy, packed'

/Y/ INSERTION

³⁸ The vowel chart for Persian below shows that /i/, /e/ and /æ/ are the front vowels (courtesy of Wikipedia: Persian Phonology).



The /y/ insertion occurs mainly at morpheme boundaries, i.e., at points where two affixes are put together. When two vowels are placed following each other, a /y/ is inserted to facilitate the pronunciation. This is very productive in the literary variant of Persian. In the conversational variant, the /y/ insertion occurs mainly when a vowel is followed by /i/. Some examples are:

<i>pa + æm</i>	'foot + Poss.1sg'	→	<i>payæm</i>	'my foot'	[literary]
<i>daneshju + an</i>	'univ. student + PL'	→	<i>daneshjuyan</i>	'university students'	[literary]
<i>to + i</i>	'you + are'	→	<i>toyi</i>	'it's you'	[literary, conversational]
<i>seda + i</i>	'sound/voice + a'	→	<i>sedayi</i>	'a sound/voice'	[literary, conversational]

4.4.2 Lexical Items

This section provides some of the words that have changed form in the conversational language and would need to be included as part of the lexicon of Persian. The list introduced is incomplete but provides several categories of new lexical items to look into.

FORM CHANGES

- The indefinite *yek* (یک [yk]) which means 'a' or 'one' becomes *ye* (یه [yh]) in the conversational variant.³⁹
- In certain words, the word-final *-gær* ending becomes *-ge* (written with a 'silent he').

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>ægær</i>	<i>æge</i>	اگه [agh]	'if'
<i>digær</i>	<i>dige</i>	دیگه [digh]	'other, no more'
<i>hæmdigær</i>	<i>hæmdige</i>	همدیگه [hmdigh]	'together'
<i>mægær</i>	<i>mæge</i>	مگه [mgh]	('isn't it the case that?') ⁴⁰

DISTINCT LEXICAL ITEMS

The preposition *dær* (در [dr]) is rarely used in conversational Persian, except for lexicalized compounds such as *dær ja mord* (in place died) meaning 'he/she/it died on the spot'. In the

³⁹ Numbers in conversational Persian reflect a number of the changes discussed in the previous section, some of which were already mentioned: *yek* → *ye* 'one', *chahar* → *cha:r* 'four', *shesh* → *shish* 'six', *hæft* → *hæf* 'seven', *hæst* → *hæsh* 'eight', *dæh* → *dæ:* 'ten', all those between 11 to 19 such as *sizdæh* → *sizzæ:* 'thirteen', *panzdæh* → *punzdæ:/punzæ:* 'fifteen', *hefdæh* → *hivdæ:* 'seventeen', *hejdæh* → *hizhdæ:* 'eighteen', *bist* → *bis* 'twenty', *chehel* → *che:l* 'forty', *pænjah* → *pænja:* 'fifty', *shæst* → *shæs* 'sixty'.

⁴⁰ There is no equivalent lexical item in English.

conversational variant, the preposition *tu* (تو [tv]) is used instead, which may also take the "ezafe" (see Section 4.5.2) and appear as *tuye* (توی [tuy]).

The preposition *bæraye* is often replaced by *vase* (واسه [vash]) or *vaseye* when it appears with the "ezafe" morpheme.

The preverbal preposition *bær* 'on, above, upon' is often pronounced as *vær* in conversational Persian. Hence, the verb *bær dashtæn* 'to pick up' becomes *vær dashtæn*.⁴¹

The word for nothing in the literary variant is *hich chiz*, but it becomes *hich-chi* or simply *hichi* in the conversational form.

ARABIC LOANS

Despite claims for "purism" and for eliminating the Arabic words from the language (especially among the Iranian diaspora), the modern conversational Persian has a preference for the Arabic loanwords that are commonly used in the language as opposed to their more literary equivalents of Persian origin. For instance, to say 'thank you' in Persian, the most common expression is *mersi* which is of French origin, or one can say *mæmnun* or *motshækkeræm* which are both of Arabic origin. The Persian word *sepasgozaræm* sounds very marked and literary and is almost never used in conversational language.

In fact, most heritage speakers of Persian (who have not been trained in Persian at school but have learned the language from conversation in the family) would have a very difficult time understanding the sentence (1) below formed with words of Persian origin, as opposed to (2) which conveys the same exact meaning but contains mainly Arabic-based loanwords. This suggests that the expressions of Persian origin are being eliminated from the language and the Arabic-origin loanwords are the common lexical items in modern Persian.

- (1) بکوشیم که واژگان تازی را به کار نبریم
bekushim ke vazhegan-e tazi ra be kar næbærim
 try.subj-1pl that words-of Arabic obj to work not-take.1pl
 'Let us try not to use any Arabic words.'
- (2) سعی کنیم که از کلمات عربی استفاده نکنیم
sæy konim ke æz kælæmat-e æræbi estefade nækonim
 try do.subj-1pl that from words-of Arabic use not-do.1pl
 'Let us try not to use any Arabic words.'

4.4.3 Neologisms and Loans

Blogs contain a large number of loanwords, especially from English which currently exerts a big influence on Persian because of computers and technology. Scientific and technological terms are widely used on blogs. This is in particular intensified when the Iranian government tries to crack down on the blogs, and bloggers begin posting ways to break the filtering technologies and provide technical support to each other online. These words of course follow the morphological rules of Persian and can take affixes as in *filteringeshun* (فیلترینگشون) = *filtering* 'filtering' + *eshun* 'their' ('their attempts at/act of filtering').

⁴¹ Note that this does not affect the preverbal *bær* that has the meaning 're-' as in *bær gæshæn* 'to return'.

Examples of technical terms are: *anlayn* (آنلاین [An~layn]), *pabliš* (پابلیش [pabliš]), *chætrum* (چتروم [čt~rvm]), *imeyl* (ایمیل [ay~myl]), *monitowr* (مونیتور [mvnytvr]), *es-em-es* (اساماس [as~am~as]), *filtering* (فیلترینگ [fyltryng]), *si-di* (سی‌دی [sy~dy]), *di-vi-di* (دی‌وی‌دی [dy~vy~dy]), *vindož* (ویندوز [vyndvz]), *afis* (آفیس [Afys]), *fotoshap* (فتوشاپ [ftvšap]), *kibord* (کیبورد [kybvrd]).

Other English and older French loans are also quite common: *pazel* (پازل [pazl]), *partner* (پارتنر [partnr]), *holokast* (هولوکاست [hvlvkast]), *nostalzhi* (نوستالژی [nvstalži]), *seksualite* (سکسوالیته [sksvalyth]).

The Persian Language Academy creates new words to replace loanwords such as Persian *rayane* (رایانه) to replace the word *kampyuter* (کامپیوتر) 'computer', or *balgærd* (بالگرد) to replace *helikopter* (هلیکوپتر) 'helicopter'. These attempts have not been very successful in recent years and bloggers do not use these newly coined words very often and instead use the foreign equivalents. On the other hand, bloggers themselves create new words based on Persian word-formation rules that are often based on a loanword. These neologisms are quite frequent in blogs. They include:

Script/Transliteration	Pronunciation	Consists of...	Translation
لینکدونی [lynkdvny]	<i>linkduni</i>	<i>link</i> 'link' + <i>duni</i> 'storage place'	'blogroll'
کامنت‌گذار [kamnt~gDar]	<i>kamentgozar</i>	<i>kament</i> 'comment' + <i>gozar</i> 'putter'	'commenter, one who leaves comments'
تابوسازی [tabvsazy]	<i>tabusazi</i> :	<i>tabu</i> 'taboo' + <i>sazi</i> 'creating'	'make taboo'
فیلترشکن [filtrškn]	<i>filtershekæn</i>	<i>filter</i> 'filter' + <i>shekæn</i> 'breaker'	'anti-filter software'

A number of new verbs have been formed by combining a loanword with a light verb such as *kærdæn* 'to do' or *zædæn* 'to hit'. Light verb constructions are very pervasive in Persian and consist of a preverbal element (noun, adjective or preposition) followed by a verb that is somewhat bleached in meaning called a "light verb". Examples of these new constructions are:

Script/Transliteration	Pronunciation	Consists of...	Translation
کلیک کردن [klyk krɔn]	<i>klik kærdæn</i>	<i>klik</i> 'click' + <i>kærdæn</i> 'do'	'to click (on mouse)'
چت کردن [čt krɔn]	<i>chæt kærdæn</i>	<i>chæt</i> 'chat' + <i>kærdæn</i> 'do'	'to chat (online)'
ایمیل زدن [ay~myl zɔn]	<i>imeyl zædæn</i>	<i>imeyl</i> 'email' + <i>zædæn</i> 'hit'	'to email'
دان کردن [dan krɔn]	<i>dan kærdæn</i>	<i>dan</i> 'down' + <i>kærdæn</i> 'do'	'to download'

More recently, verbs are formed on the simple verb construction instead of the compound forms above by adding the *-idæn* infinitival morpheme. So now we also have the verbs listed below:

Script/Transliteration	Pronunciation	Consists of...	Translation
کلیکیدن [klykyɔn]	<i>klikidæn</i>	<i>klik</i> 'click' + <i>idæn</i>	'to click (on mouse)'
چتیدن [čtyɔn]	<i>chætidæn</i>	<i>chæt</i> 'chat' + <i>idæn</i>	'to chat (online)'
دانلودیدن [danlvdyɔn]	<i>danlodidæn</i>	<i>danlod</i> 'download' + <i>idæn</i>	'to email'

لاگیدن	[lagydn]	<i>lagidæn</i>	<i>lag</i> 'blog' + <i>idæn</i>	'to download'
میلیدن	[mylydn]	<i>meylidæn</i>	<i>meyl</i> 'wish' + <i>idæn</i>	'to wish'

These verbs are conjugated regularly as can be seen in the following examples:

برای دانلودیدن به روی ادامه مطلب بکلیکید

[bray **danlvdydn** bh rvy adamh mTlb **bklykyd**]
bæraye danlodidæn be ruye edameye mætlæb bekelikid
 for download.Inf to on continuation subject click.2pl

'To download click on continuation of subject' (i.e., to download, click on Continue)

اگه اینترنت ایران اجازه بده از اونجا هم خواهیم لاگید اگر نه که هیچ

[agh ayntnrnt ayran ajazh bdh az avnja hm xvahm **lagyd** agrm nh kh hyč]
æge internete iran ejaze bede æz unja khahæm lagid ægæræm næ ke hich
 if Internet-of Iran allowance gives from there will.1sg blog if-also no that nothing
 'If the Iranian Internet allows it I will blog from there, otherwise not'

اگر میلید لطفا لینک من را در سایتتون قرار بدهید

[agr **mylyd** lTfa lynk mn ra dr sayttvn qrar bdhyd]
ægær meylid lotfæn linke mæn ra dær saytetun ghærar bedæhid
 if wish.2pl please link my OBJ in site-your placement give.2pl
 'If you wish please put my link on your site'

Misspellings are also very common in weblogs, but sometimes they are done on purpose as for the many forms of spelling the word *seks* 'sex' in order to be able to discuss a very taboo subject in Iranian society among the youth without being subject to filtering by the government. Bloggers write this word with the various characters representing the /s/ sound in Persian: صکس [SkS], نکس [ckc], نکس [ckS]. In addition, the spelling of some words that have been borrowed from Arabic and have maintained their original spelling despite the fact the pronunciation does not represent Persian orthographic rules are being modified by bloggers. Examples are the words that end in a "tanvin" as in لطفًا [lTfa] 'please' which is sometimes spelled as لطفن [lTfn]; this word is pronounced *lotfæn*. Similarly, موسى [mvsy] 'Moses' and حتی [Hty] 'even', pronounced *musa* and *hæta*, respectively, are now often spelled as موسا [mvsa] and حتا [Hta].

4.5 Morphology

4.5.1 Introduction

Persian morphology (word forms) is affixal consisting mainly of suffixes and prefixes, which generally follow a regular morphotactic order. Ambiguities arise in a computational analysis due to the use of the Arabic script since certain vowels are not marked in written text and spacing between words and morphemes is sometimes inconsistent. Furthermore, some affixes can represent different morphemes. For instance, the suffix *-ی* 'ye' can be an indefinite article, a relativizing particle, the second person singular form of the copula verb, or a derivational form creating adjectives out of nouns, as illustrated in example (3) below⁴².

In this case, the lack of the short vowel can add to the ambiguity since the word can also be analyzed as a verb as shown in (4).

⁴² There is less ambiguity in speech since the stress pattern can distinguish some of the constructions. For instance, the 'ye' suffix in (3a-c) are all inflectional and do not take the word stress which always remains on *mærd*, but in (3d) the stress falls on the derivational suffix: *mærd-i*.

- (3)
- | | | | |
|-------------------------------|------------------|---|-----------------------|
| مردی | | | |
| a. mærd-i | [man-indef] | → | a man |
| b. mærd-i (ra ke be ma didim) | [man-rel] | → | the man (that we saw) |
| c. mærd-i | [man-copula.2sg] | → | you are a man |
| d. mærd-i | [man-adj.affix] | → | manhood, manliness |
- (4)
- | | | | |
|--------|------------|---|----------|
| مردی | | | |
| mord-i | [died-2sg] | → | you died |

Conversational forms of morphemes give rise to further ambiguity. For instance, *zænha* 'women' (5a) would be pronounced as *zænna* in the conversational variant. Since the 'æ' vowel is not usually written in Persian script, it would have the form [zna] in a text, as shown in (5b). Without the overt vowel in the first syllable, however, the word is now ambiguous between *zænna* 'women' and *zēna* 'adultery'.

(5)	Persian script	Transliteration	Translation
a.	زنها	[znha]	women
b.	زنا	[zna]	women; adultery

Another instance of ambiguity arising from the conversational orthographic form can be found with words that end in the sound /e/ which is written as the 'silent *he*'. As will be discussed in more detail in this section, the word-final /e/ assimilates (or merges) with the following /æ/ sound of some affixes. For example, the word *khune* 'house' when used with the pronominal suffix *-æm* becomes *khunæm* in the spoken language, meaning 'my house'. It is often represented as in (6a) in the written form. However, this form can be ambiguous with the word *khunæm* meaning 'my blood' shown in (6b). There is no ambiguity in the spoken language since the stress pattern of the words distinguish them: *khunÆm* 'my house' vs. *khUnæm* 'my blood'. Yet the orthographic forms remain ambiguous.

(6)		<u>Persian word</u>		<u>word+pronoun (in conversational variant)</u>
a.	خونه	[xvnh]	'house'	→ خونم [xvnm] 'my house'
b.	خون	[xvn]	'blood'	→ خونم [xvnm] 'my blood'

As these examples show, the variant forms introduced by writing the conversational form of Persian add to the complexity and ambiguity of the morphological analysis. This section presents a description of the inflectional morphemes found in Persian blog text. Since blog language can be in either the literary or the conversational variant of the language or can include both within the same text, both forms will be described. Each morpheme is first presented in its traditional, literary form. If the morpheme has a different realization in the conversational variant, the distinctions are presented as well. Finally, the orthographic issues are discussed for each morpheme.

4.5.2 Nominal Inflection

4.5.2.1 Plurals

There are several plural markers in Persian, some of which were borrowed from Arabic:

- ها - 'ha'

This is the most productive plural morpheme. It can appear on any noun and it may also attach to an adjectival element if the latter is used as a noun. Infinitivals and certain adverbials may also take this plural morpheme.

Sing.	Plural	Persian script/transliteration	Translation
<i>zæban</i>	→ <i>zæbanha</i>	زبانها [zbanha]	'languages'
<i>jomle</i>	→ <i>jomleha</i>	جمله‌ها [jmlh~ha]	'sentences'

- ان - 'an'

Generally attaches to nouns that refer to an animate being, but it can also be used on temporal nouns or nouns representing parts of the human body.

Sing.	Plural	Persian script/transliteration	Translation
<i>vblagnevis</i>	→ <i>veblagnevisan</i>	وبلاگنویسان [vblagnvysan]	'bloggers'
<i>irani</i>	→ <i>iranian</i>	ایرانیان [ayranyan]	'Iranians'

variants:

یان 'yan': If the suffix attaches to a word ending in a written vowel (i.e., *alef* or *vav* indicating the /a/ and /u/ sounds, respectively), a *ye* is inserted between the vowel and the suffix.

Sing.	Plural	Persian script/transliteration	Translation
<i>daneshju</i>	→ <i>daneshjuyan</i>	دانشجویان [danšjuyan]	'university students'

گان 'gan': If the suffix attaches to a word ending in the /e/ sound (which is written as a 'silent *he*' word-finally), the *he* /h/ is replaced by *gaf* /g/.

Sing.	Plural	Persian script/transliteration	Translation
<i>nevisænde</i>	→ <i>nevisændegan</i>	نویسندگان [nvysndgan]	'writers'

- ات - 'at'

Appears on nouns that are of Arabic origin; these nouns have feminine gender in Arabic but Persian does not have any gender distinctions.

The suffix can attach to a word ending in a consonant. It can also attach to words ending in /æ/ or in the /e/ sound (written as 'silent *he*'). In these two cases, the word-final -æt (written [t]) and -e (written [h])– bolded in the examples below– are eliminated.

Sing.	Plural	Persian script/transliteration	Translation
<i>entekhab</i>	→ <i>entekhabat</i>	انتخابات [antxabat]	'elections'
<i>loghæt</i>	→ <i>loghat</i>	لغات [lQat]	'words'
<i>tæjrobe</i>	→ <i>tæjrobat</i>	تجربات [tjrbat]	'experiences'

- جات - 'jat'

This suffix denotes a collective meaning and generally attaches to words of Persian origin. If the suffix attaches to a word ending in the /e/ sound (written as a 'silent *he*' word-finally), the [h] is deleted.

Sing.	Plural	Persian script/transliteration	Translation
-------	--------	--------------------------------	-------------

<i>karkhane</i>	→	<i>karkhanejat</i>	کارخانه	[karxanh]	→	کارخانجات	[karxanjat]	'factories'
<i>mive</i>	→	<i>mivejat</i>	میوه	[myvh]	→	میوجات	[myvjat]	'fruits'
<i>sæbzi</i>	→	<i>sæbzijāt</i>	سبزی	[sbzy]	→	سبزیجات	[sbzyjat]	'vegetables'

- **ین 'in'**

Usually attaches to nouns that are originally arabic participles and end in a consonant.

Sing.	Plural	Persian script/transliteration	Translation
<i>mosafer</i>	→ <i>mosaferin</i>	مسافریں	[msafryn] 'travelers, passengers'

- **ون 'un'**

Attaches to certain nouns of Arabic origin ending in *ye*. This plural suffix is very rare and the plural nouns formed using this method are usually considered lexicalized.

Sing.	Plural	Persian script/transliteration	Translation
<i>rowhani</i>	→ <i>rowhaniun</i>	روحانیون	[rvHanyvn] '(the) clergy'

Conversational Variant

Only the '*ha*' morpheme is affected in the conversational variant of Persian, where word-medial *he* is generally deleted resulting in the form '*a*' for the plural marker.

Literary Plural form	Conversational Plural form	Translation
<i>ketabha</i>	<i>ketaba</i>	'books'
<i>daneshjua</i>	<i>daneshjua</i>	'University students'

Orthographic Variance

The '*ha*' morpheme can be written either attached or detached. Persian writing system allows certain morphemes to appear either as bound to the host or as free affixes – free affixes could be separated by a half-space (the control character \u200C in Unicode, also known as the zero-width non-joiner) or an intervening full space. The three possible cases are illustrated below for the plural suffix in *felestiniha* 'Palestinians'. As shown, the affixes may be attached to the stem, they may be separated with the final form control marker, or they can appear with intervening whitespace. All of these surface forms are attested in online text.

Isolated Form (intervening full space)	Detached Form (intervening half-space)	Attached Form
فلسطینی ها [flsTyny ha]	فلسطینی~ها [flsTyny~ha]	فلسطینیها [flsTynyha]

There is less variance, however, with nouns that end in the /e/ sound written as 'silent *he*', since this character can never appear attached to the following letter. Hence, only the

detached or isolated forms are available as shown below for the plural of *khane* [xanh] 'house'.

Isolated Form (intervening full space)	Detached Form (intervening half-space)	Attached Form
خانه ها [xanh ha]	خانه‌ها [xanh~ha]	N/A

In the conversational variant, the 'a' morpheme attaches to the nominal element if the latter ends in a vowel or consonant. However, if the word ends in the /e/ (written as 'silent *he*'), the suffix is not attached to the noun stem as in the example with *khune* 'house' shown in the table below.

Sing.	Plural	Persian script/transliteration	Translation
<i>ketab</i>	→ <i>ketaba</i>	کتابا [ktaba]	'(the) books'
<i>miz</i>	→ <i>miza</i>	میزا [miza]	'(the) tables'
<i>sændæli</i>	→ <i>sændælia</i>	صندلیا [Sndlya]	'(the) chairs'
<i>meshki</i>	→ <i>meshkia</i>	مشکیا [mškya]	'the black ones'
<i>daneshju</i>	→ <i>daneshjua</i>	دانشجوا [danšjva]	'university students'

Isolated Form (intervening full space)	Detached Form (intervening half-space)	Attached Form
خونه آ / خونه ا [xunh A/xunh a]	خونه‌آ / خونه‌ا [xunh~A/xunh~a]	N/A

4.5.2.2 Ezafe

The elements within a noun phrase and preposition phrase are linked using the *ezafe* morpheme in Persian. The *ezafe* is used to relate a head noun to its modifiers and possessors, and a preposition to its nominal complement. This morpheme is pronounced as /e/ after consonants and /ye/ after vowels, shown in bold in the transcription of the example below.

فروش جت های جنگنده بریتانیایی
forush-e jetha-ye jængænde-ye beritanyayi
 sale-EZ jets-EZ fighter-EZ British
 'the sale of British jet fighters.'

Conversational Variant

The *ezafe* morpheme is maintained in the conversational variant. It is sometimes dropped, however, following prepositions as in *bala shæhr* 'uptown'. This is particularly true of prepositions that end in /u/ such as *tu* 'in', *ru* 'on', *pæhlu* 'next to' as in *ru mize* 'it's on the table' vs. *ruye mize*.

Orthographic Variance

Following a consonant, the *ezafe* is an unstressed /e/ sound which is generally not written in the Persian writing system. Following vowels, it is pronounced as 'ye' and can take the following forms:

- ی - 'ye'

Occurs following written vowels /u/ or /a/. It is obligatory in this context.

Transcription	Persian script/transliteration	Translation
<i>jethaye</i>	جت های [jthay]	'jets (of)'
<i>buye</i>	بوی [bvy]	'the smell (of)'

- ی or ء 'ye'

Occurs following the sound /e/ word-finally (written as silent *he*). The *ezafe* is detached (isolated form or with intervening half-space) as shown in the two possible forms below. It is optional in this context and could be left unwritten.⁴³

Transcription	Persian script/transliteration	Translation
<i>jængænde^{ye}</i>	جنگنده ی [jngndh~y]	'fighter (of)'
	or جنگنده ء [jngndh~y]	

Hence in our example above ('the sale of British jetfighters'), the first *ezafe* on *forush* 'sale' is not written because diacritics are generally unmarked in the Persian writing system. The second *ezafe* on *jethaye* 'jets' appears after the vowel /a/ and is obligatorily written as 'ye'. Finally, in the third case of the *ezafe* on *jængænde* 'fighter', the suffix is not written because it follows the 'silent *he*' and is therefore optional.

As mentioned above, the prepositions may appear without the *ezafe*, especially in the conversational variant, which is reflected in the orthography as well. Hence either of the two forms below, with the *ezafe* (shown in red) and without it are attested⁴⁴:

باید امروز جلوی جنگ را گرفت.
bayæd emruz jelo^{ye} jæng ra gereft
 must today front-EZ war OBJ caught
 'One must stop the war today.'

می توانیم جلو جنگ را بگیریم.
mitævanim jelo jæng ra begirim
 we-can front war OBJ catch
 'We can stop the war'

In the conversational variant, the unwritten *ezafe* vowel is sometimes overtly written, not as a diacritic but as a 'silent *he*' which represents the word-final /e/ in Persian.⁴⁵ Although rare, the following forms may appear in blogs:

Transcription	Literary form	Conversational form	Translation
<i>daneshj^{ye}</i>	دانشجوی [danšjvy]	دانشجویه [danšjvyh]	'student-EZ'
<i>forushga^{he}</i>	فروشگاه [frvšgah]	فروشگاهه [frvšgahh]	'the store-EZ'

⁴³ Another form is جنگنده [jngndY]. The last character transliterated as [Y], is a *he* with a *hamze* above; it is a Persian letter and is pronounced /eye/. It represents the *ezafe* on a silent *he* character, but it does not occur very often in online documents.

⁴⁴ Based on an informal search on Google the form *jelo^{ye} jæng* 'front-Ez war' was found 410 times while the phrase *jelo jæng* 'front war' without the *ezafe* morpheme was located 96 times on the Persian Internet.

⁴⁵ The silent *he* is used in traditional Persian orthography to mark word-final stressed /e/ but in this instance, it is being used to represent the unstressed /e/ sound of the *ezafe* morpheme. This form is considered an "error" by traditionalists.

as in:

دانشجویه خوش‌شانس
daneshjuye khoshshans
 university student-EZ lucky
 'a lucky university student'

فروشگاهه دیزل
forushgah e dizel
 store-EZ Diesel
 'a Diesel store'

4.5.2.3 Indefinite article & relativizing particle

Although these two morphemes have different functions, the indefinite article and the relativizing particle both attach to the last element in the noun phrase. They are pronounced *i* after consonants and *yi* following vowels; it may also be pronounced *i* after the /e/ vowel (written as 'silent *he*').

The indefinite article marks indefiniteness for the noun phrase. The relativizing particle attaches to the head of a relative clause, or to the last element of the head noun phrase to be exact, as in the following example:

فروشگاه بزرگی که او برای کتاب تاسیس کرد ...
forushgah-e bozorg-i ke u bəraye ketab tæsis kærd...
 store-EZ big-REL that he/she for book establishment did ...
 'the big store that he/she established for books...'

Conversational Variant

There is no change in the conversational variant.

Orthographic Variance

- ی - 'i'

Occurs after consonants.

Transcription	Persian script/transliteration	Translation
<i>kudækani</i>	کودکانی [kvdkany]	'children-INDEF/REL'

- یی / ئی - 'yi'

Occurs after vowels. The hamze version is less frequent.

Transcription	Persian script/transliteration	Translation
<i>cheshmhayi</i>	چشمهایی [čšmhayi]	'eyes -INDEF/REL'

- ای - 'i'

Occurs after silent *he*. The suffix has to be written in unattached form (detached or isolated).

Transcription	Persian script/transliteration	Translation
<i>bæstei</i>	بسته‌ای [bsth~ay]	'package -INDEF/REL'

4.5.2.4 Definite marker

The definite marker exists only in the conversational variant of Persian and can never be found in the literary text. The definite marker is pronounced as a stressed /e/. If the definite marker follows a word ending in /e/, the sound /h/ is inserted between the word and the definite morpheme as in *forushænde + e* → *forushændehe* 'the salesperson'.

Orthographic Variance

Since the definite marker is a word-final /e/ it is written as a 'silent *he*'.

- ه - 'e'

Occurs after consonants and word-final /e/ (silent *he*).

Transcription		Persian script/transliteration	Translation
<i>ketab</i>	→ <i>ketabe</i>	کتابه [ktab → ktabh]	'the book'
<i>forushænde</i>	→ <i>forushændehe</i>	فروشندهه [frvšndh → frvšndhh]	'the salesperson'

It can also follow the vowels /a/ and /u/. Note that this goes against the traditional orthographic rule where the 'silent *he*' only appears after consonants while the [h] written after vowels is pronounced *h*. In the examples below, the word-final 'h' appears after vowels yet it is pronounced as *e*.

Transcription		Persian script/transliteration	Translation
<i>baba</i>	→ <i>babae</i>	باباه [baba → babah]	'the dad'
<i>daneshju</i>	→ <i>daneshjue</i>	دانشجوه [danšjv → danšjvh]	'the university student'

The word-final /e/ sound becomes /æ/ when followed by the *object marker* (see next section). The orthography does not change, however, since word-final /e/ and word-final /æ/ are both written with the 'silent *he*' character.

ketabe +ro → *ketabæro*
forushændehe+ro → *forushændeheæro*

4.5.2.5 Object marker

There is no case system in modern Persian but the *ra* morpheme is used to mark specific direct objects as illustrated in the example below from BBC. In addition, *ra* has been argued to mark topicalized elements as well.

ایران هنوز این خبر را تأیید نکرده است
iran hænuz in khæbær ra tæ'id nækærde æst
 Iran yet/still this news **OB** confirmation not-done is
 'Iran has not yet confirmed the news'

Conversational Variant

The object marker is pronounced *ro* (after both vowels and consonants) or *o* (after consonants or the vowel /i/) in spoken Persian. Whether the *ro* or *o* form is used following a consonant depends on the speaker.

Orthographic Variance

The object marker has only one form in the literary variant of the language: را [ra]. It is in general separated from the word, but it may also appear attached to consonants:

کتاب را پیدا نکردم
ketabra peyda nækærdæm
 book OBJ found not-did.1sg
 'I didn't find the book'

In this conversational form, the orthography of the object marker once again reflects its pronunciation in spoken Persian. The object marker may appear either as *ro* (رو [rv]) or simply as *o* (و [v]). It can appear detached or attached to consonants.

Transcription	Persian script/transliteration	Translation
<i>ketabro</i>	کتابرو [ktabr ^{rv}]	'the book-OBJ'
<i>ketabo</i>	کتابو [ktab ^v]	'the book-OBJ'

4.5.3 Complex Tokens

Complex tokens refer to word-like affixes that appear attached to the stem of a different part of speech category. For instance, affixal prepositions, conjunctions, pronoun clitics or verbal copulas attached to nouns or adjectives would be considered complex tokens in Persian.

4.5.3.1 Pronominal clitics⁴⁶

Persian has a series of suffixal personal pronouns representing different functions depending on the part of speech or syntactic context in which they occur. A clitic pronoun can be a possessive pronoun, an object pronoun, or a subject pronoun, or it could be interpreted as a partitive or impersonal clitic. Despite their function, they all have the same pronunciation or surface realization.

- **Pronunciation of post-consonantal forms**

Literary Form	Conversational Form	Features	Conversational Example
<i>æm</i>	<i>æm</i>	First person, singular	<i>ketabæm</i> 'my book'
<i>æt</i>	<i>et</i>	Second person, singular	<i>ketabet</i> 'your book'
<i>æsh</i>	<i>esh</i>	Third person, singular	<i>ketabesh</i> 'his/her book'
<i>eman</i>	<i>emun</i>	First person, plural	<i>ketabemun</i> 'our book'
<i>etan</i>	<i>etun</i>	Second person, plural	<i>ketabetun</i> 'your (pl.) book'
<i>eshan</i>	<i>eshun</i>	Third person, plural	<i>ketabeshun</i> 'their book'

- **Pronunciation of post-vocalic forms**

Literary	Conversational	Features	Conversational Example
----------	----------------	----------	------------------------

⁴⁶ A clitic is a linguistic term that refers to elements that display a combination of word-like and affix-like behavior. A clitic attaches to a host in a way similar to affixes, yet maintains properties of an independent word (e.g., does not take word-level stress). One characteristic shared by many clitics is a lack of prosodic independence; these clitics need to attach to a word that carries a high stress, which is the case of the pronominal clitics in Persian.

Form	Form		
<i>yæm</i>	<i>m</i>	First person, singular	<i>bæradæram</i> 'my brothers'
<i>yæt</i>	<i>t</i>	Second person, singular	<i>bæradærat</i> 'your brothers'
<i>yæsh</i>	<i>sh</i>	Third person, singular	<i>bæradærash</i> 'his/her brothers'
<i>yeman</i>	<i>mun</i>	First person, plural	<i>bæradæramun</i> 'our brothers'
<i>yetan</i>	<i>tun</i>	Second person, plural	<i>bæradæratun</i> 'your brothers'
<i>yeshan</i>	<i>shun</i>	Third person, plural	<i>bæradærashun</i> 'their brothers'

Orthographic Variance

In the literary variant, the pronominal clitics have different forms depending on whether they follow a consonant, the vowels /a/ and /u/, or the word-final /e/ written as a 'silent *he*'. This is shown in the table below. As expected, the clitic appearing following the word-final /e/ may not be written in the attached form.

After a consonant	After vowels /a/ and /u/	After word-final /e/ (silent <i>he</i>) [detached or isolated]
م [m]	یم [ym]	ام [~am]
ت [t]	یت [yt]	ات [~at]
ش [š]	یش [yš]	اش [~aš]
مان [man]	یمان [yman]	مان [~man]
تان [tan]	یتان [ytan]	تان [~tan]
شان [šan]	یشان [yšan]	شان [~šan]

In the conversational variant, the pronominal clitics that follow a consonant or the vowels /a/ and /u/ are written the same way.

After a consonant or After vowels /a/ and /u/
م [m]
ت [t]
ش [š]
مون [mvn]
تون [tvn]
شون [švn]

There is more variance in the orthography of the clitic following a word-final /e/, reflecting the pronunciation in the conversational variant. In spoken Persian, the final /e/ sound is assimilated into the initial /æ/ sound of the pronominal clitic. So for instance, the traditional (literary) pronunciation of *jælæsé-æm* 'my meeting' becomes *jælæsåm* where

the accent represents the main word-level stress. Similarly, the plural form *jælæsé-tan* 'your meeting' is usually pronounced as *jælæsætun* although *jælæsétun* as well as *jælæsé-etun* are also possible adding to the orthographic variants found for the plural forms.

For instance, in order to say 'my opinion', one can add the basic conversational endings shown in the above table to the word *æghide* 'opinion' (عقیده) as in column 1 below. In this case, the morpheme appears in detached or isolated form, and the word is pronounced *æghidæm* 'my opinion'. However, to represent the conversational pronunciation more directly (i.e., to represent the assimilation of the word-final /e/ sound), oftentimes the word-final 'he' character is removed in orthography as seen in column 2. Finally, in the plural forms, an additional 'alef' may be used as shown in column 3; this is done to represent e.g., the pronunciation *æghide-emun* vs. the more usual *æghidæmun*.⁴⁷

After word-final /e/ (silent <i>he</i>)		
Column 1: detached or isolated form	Column 2: eliminate final 'he' attached form	Column 3: detached form add 'alef' for plural forms
عقیده م [eqydh m]	عقیدم [eqydm]	
عقیده ت [eqydh t]	عقیدت [eqydt]	
عقیده ش [eqydh š]	عقیدش [eqydš]	
عقیده مون [eqydh mvn]	عقیدمون [eqydmvn]	عقیده امون [eqydh amvn]
عقیده تون [eqydh tvn]	عقیدتون [eqydtvn]	عقیده اتون [eqydh atvn]
عقیده شون [eqydh švn]	عقیدشون [eqydšvn]	عقیده اشون [eqydh ašvn]

The pronominal clitic can also attach to verbs, especially in the conversational language. In these cases, it often acts as the object of the action as in:

میشناسیش؟ میشناسیت؟
mishnasish? mishnasætet?
 'do you know him? does he know you?'

zædæm (I hit) + *-esh* (3rd person sing. clitic) → *zædæmesh* 'I hit him/her/it'

The various functions of the pronominal clitic will be discussed in more detail in the

⁴⁷ To illustrate the relative frequency of each orthographic form for 'my opinion', a Google search of *عقیده م* [eqydh m] gave 30 hits, while the version with the final 'he' eliminated (*عقیدم* [eqydm]) gave 162 hits. Similarly, we obtained the following for the variant forms of 'our opinion':

عقیده مون [eqydh mvn]	→	10 hits
عقیدمون [eqydmvn]	→	16 hits
عقیده امون [eqydh amvn]	→	1 hit

In addition, more variant forms are possible, since the '*man*' to '*mun*' alternation in the pronominal form may not be followed:

عقیدمان [eqydman]	→	2 hits
عقیده امان [eqydh aman]	→	2 hits

For comparison, the traditional literary spelling *عقیده مان* gives 53 hits.

following section. What is of interest in this section is the particular form the clitic takes when it follows the verbs of the third person singular that end in the vowel /e/ (written as 'silent *he*'). For instance, in the Present tense, the third person singular agreement ends in the /e/ vowel as shown below for the conversational pronunciation of the verb *zædæn* 'to hit' (*mi* is the progressive prefix):

Pronunciation			Persian Script	Transliteration
Prefix	Verb Stem	Person Ending		
<i>mi</i>	<i>zæn</i>	<i>æm</i>	میزنم	[myznm]
<i>mi</i>	<i>zæn</i>	<i>i</i>	میزنی	[myzny]
<i>mi</i>	<i>zæn</i>	<i>e</i>	میزنه	[myznh]
<i>mi</i>	<i>zæn</i>	<i>im</i>	میزنیم	[myznym]
<i>mi</i>	<i>zæn</i>	<i>in</i>	میزنین	[myznyn]
<i>mi</i>	<i>zæn</i>	<i>æn</i>	میزنن	[myznn]

For all but the 3rd person singular, the pronominal clitic simply attaches to the end of the person ending. Hence, to say 'I am hitting you' we can add *-et* 'you' to the verb *mizænæm* 'I hit; I am hitting' to obtain *mizænæmet*. However, since the 3rd person singular form ends in a vowel, an additional /t/ is added between the verb and the pronoun clitic. To be more precise, the /e/ vowel ([h] in orthography) is replaced by /æ/ before adding the pronoun. For example, to say 'He is hitting you' we get the following:

mizæne + et → *mizæn+æt+et* → *mizænætet*
 he hits + you 'he is hitting you'

The actual orthography for this example is as follows:

[myznh] + [t] → [myzn]+[t]+[t] → [myzntt]
 [میزنه] + [ت] [میزن]+[ت]+[ت] [میزنتت]

Similarly, the conversational form of the third person singular in the Present Perfect tense ends in the /e/ sound (written as 'silent *he*') and follows the same pattern as shown:

zæde + et → *zæd +æt+et* → *zædætet*
 he has hit + you 'he has hit you'

There is some variance in orthography here as some writers drop the 'silent *he*' character of [zdh] while others maintain it as shown below:

[zdh] + [t] → [zd]+[t]+[t] → [zdt]
 [زده] + [ت] [زد]+[ت]+[ت] [زدت]

[zdh] + [t] → [zdh]+[t]+[t] → [zdh~tt]

[ت] + [زده]

[ت] + [ت] + [زده]

[زدهنت]

The following are some more examples from Iranian weblogs. In each example, the extra *æt* is shown in blue and the object pronoun is in red. In the pairs in (7) and (8) we can see the two distinct orthographic variants: the examples on the right drop the 'silent *he*' character for *gereftætæm* 'it has caught/got me' in (7) and *mikhorætæsh* 'eats him/her/it' for (8), but the examples on the left maintain the [h] in the writing.

(7)

تازه استرس امتحان گرفته تم.
[tazh astrs amtHan grfth tm]
taze estres-e emtehan gereftætæm
just stress-EZ exam has caught-1SG

'The stress of the exam has just got me.'
(i.e., I just started stressing about the exam)

من از حالا استرس انتخاب واحد گرفتم.
[mn az Hala astrs antxab vaHd grfttm]
*mæn æz hala estres-e entekhab-e vahed
gereftætæm*
I from now stress-EZ selection-EZ unit has
caught-1SG

'The stress of selecting a unit has already got me' (i.e., I am already stressing about choosing a unit)

(8)

میدونم حرفهای من میخوره تش.
[mydvnM Hrfhay mn myxvrh tš]
midunæm hærfha-ye mæn mikhorætæsh
I-know words-EZ my eats-3SG
'I know that my words eat him/her up.'

گرگه میخور تش.
[grgh myxvrtš]
gorg-e mikhorætæsh
wolf-DEF eats-3SG
'the wolf eats him/her'

(9)

استاده هم دیگه میشناسنمون.
[astadh hm dygh myšnastmvn]
ostad-e hæm dige mishnasætæmun
prof-DEF also anymore knows-1PL
'The professor knows us by now.'

آدم واقعا فکر میکنه هزار ساله میشناسنت.
[Adm vaqea fkr myknh hzar salh myšnastt]
adæm vaqæn fekr mikone hezar sal-e mishnasætæt
human really thought does thousand year-is knows-2SG
'One really thinks that he knows you for a thousand years.' (i.e., you'd really think we've known each other for a thousand years')

(10)

دوست ندارم هیچ کس به خاطر این که آنلاین دیدم بهم سلام کنه.
[dvst ndarm hyč ks bh xaTr ayn kh Anlayn dydtm bhm slam knh]
dust nædaræm hich kæs be khater-e in ke anlayn didætæm behem sælam kone
like not-have any one to reason-EZ this that online saw-1SG to-me hello does
'I don't like it when people say hello just because they have seen me online.'

4.5.3.2 Functions of pronominal clitics

The pronominal elements discussed in the previous section can fulfill many roles; an example of each function is provided below. All examples are given in the spoken Persian forms.

Possessive: The possessive clitics are equivalent to possessive pronouns in English. The possessive clitic pronoun attaches to the last element in the noun phrase.

Examples:

<i>ktab + et</i>	'book + Poss.2sg'	→	<i>ketabet</i>	' <u>your</u> book'	کتابت
<i>dæva + mun</i>	'medication + Poss.1pl'	→	<i>dævamun</i>	' <u>our</u> medication'	دوامون
<i>khune + shun</i>	'house + Poss.3pl'	→	<i>khunæshun</i>	' <u>their</u> house'	خونشون

Object: The object clitics are the accusative form of the personal pronoun when they appear on transitive verbs. They can also occupy the position of the complement of the preposition. The object clitics are more frequent in the conversational variant of Persian.

Examples:

<i>zædæn + et</i>	'they hit + Obj.2sg'	→	<i>zædænet</i>	'they hit <u>you</u> '	زدنت
<i>didæm + eshun</i>	'I saw + Obj.3pl'	→	<i>didæmeshun</i>	'I saw <u>them</u> '	دیدمشون
<i>jelo + mun</i>	'in front + Obj.1pl'	→	<i>jelomun</i>	'in front of <u>us</u> '	جلومون
<i>bala + t</i>	'above + Obj.2sg'	→	<i>balat</i>	'above <u>you</u> '	بالات

In compound verbs the pronominal clitics can be on the preverbal or the light verb. For instance, with the compound verb *tæmbih kærdæn* [punishment do] meaning 'to punish', the pronoun can be attached to either the *tæmbih* or the *kærdæn* part of the verb.

Example:

<i>tæmbih kærdæm + et</i>	'I punished+ Obj.2sg'	→	<i>tæmbih-et</i>	<i>kærdæm</i>	تنبیهت کردم
		→	<i>tæmbih</i>	<i>kærdæm-et</i>	تنبیه کردم
			'I punished <u>you</u> '		

In both the literary and conversational variants, all prepositions that can take the *ezafe* (e.g., *bæra(ye)* 'for', *ru(ye)* 'on', *pæhlu(ye)* 'beside', *posht(e)* 'behind', *birun(e)* 'outside', *næzdik(e)* 'near') can appear with an attached pronominal clitic. Prepositions that don't take the *ezafe* morpheme (e.g., *ta* 'until, up to', *dær* 'in', *joz* 'except') do not usually take an attached pronominal clitic. However, a few of the prepositions that will not appear with a clitic in the literary language can appear with the attached clitic in the conversational variant; these prepositions are *æz* 'from', *be* 'to', and *ba* 'with'. Note in the examples below that a /h/ is inserted between two vowels in these cases, which is also represented in the orthography.

Examples:

<i>æz + æm</i>	'from + Obj.1sg'	→	<i>æzæm</i>	'from <u>me</u> '	ازم
<i>be + et</i>	'to + Obj.2sg'	→	<i>behet</i>	'to <u>you</u> '	بهت
<i>ba + shun</i>	'with + Obj.3pl'	→	<i>bahashun</i>	'with <u>them</u> '	باهاشون
<i>bæra(ye) + mun</i>	'for + Obj.1pl'	→	<i>bæramun</i>	'for <u>us</u> '	برامون
<i>hæmra + sh</i> ⁴⁸	'along with+Obj.3sg'	→	<i>hæmrash</i>	'along with <u>him/her</u> '	همراش

Subject: The subject clitics appear on some intransitive verbs (in a limited number of

⁴⁸ Note that the preposition همراه [hmrah] loses its final /h/ sound in conversational Persian (see Glottals in Section 4.4.1) and becomes *hæmra*. It then behaves as a word ending in a vowel in terms of the clitic forms it appears with

tenses) and only in the third person singular. It is almost never found in the literary variant of the language; it is a very conversational trait of Persian and especially of the Tehrani dialect of Persian.⁴⁹

Examples:

<i>oftad + esh</i>	'fell + Subj.3sg'	→	<i>oftadesh</i>	' <u>he/she/it</u> fell'	افتادش
<i>mund + esh</i>	'stayed + Subj.3sg'	→	<i>mundesh</i>	' <u>he/she/it</u> stayed'	موندش
<i>goft + esh</i>	'said + Subj.3sg'	→	<i>goftesh</i>	' <u>he/she/it</u> said'	گفتش
<i>hæst + esh</i>	'exists + Subj.3sg'	→	<i>hæstesh</i>	' <u>he/she/it</u> is'	هستش

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

Examples:

<i>væsæt + emun</i>	'middle + Part.1pl'	→	<i>væsætemun</i>	'in the middle <u>of us</u> '	وسطمون
<i>hæme + tun</i>	'all + Part.2pl'	→	<i>hæmætun</i>	'all <u>of you</u> '	همتون
<i>chahar ta + shun</i>	'four CLAS ⁵⁰ + Obj.3pl'	→	<i>chahar tashun</i>	'the four <u>of them</u> '	چهار تاشون
<i>kodum + esh</i>	'which + Part.3sg'	→	<i>kodumesh</i>	'which one (<u>of this set</u>)?'	کدومش

Impersonal: The clitic is also used in what are called "impersonal" verbal constructions. In these instances, the pronoun always attaches to the preverbal element. This construction is rare in newsprint but is common in the conversational variant.

<i>khosh + shun umæd</i>	'pleasure + Imp.3pl came'	→	<i>khosheshun umæd</i>	' <u>they</u> liked it'	خوششون اومد
<i>khāb + æm bord</i>	'sleep + Imp.1sg took'	→	<i>khābæm bord</i>	'I fell asleep'	خوابم برد

4.5.3.3 Copula verb

The present indicative of the verb *budæn* 'to be' has a series of clitic forms used in the function of verbal copula when attached to nominal and adjectival elements.

⁴⁹ An interesting word from the conversational language is *kushesh* (کوشش) which is comprised of the question element *ku* 'where' and *esh* (subject clitic, 3sg). An additional *sh* is inserted inbetween in this case. Note that this form does not exist in the literary variant of the language, but in conversational text gives rise to ambiguity with the noun *kushesh* (کوشش) meaning 'struggle'. Once again, the stress pattern in the spoken language can disambiguate these two tokens (*kUshesh?* 'where is he/she/it?' vs. *kushEsh* 'struggle') but they remain ambiguous on text.

⁵⁰ CLAS refers to a 'classifier' which is used in Persian to indicate the class of the noun in numerical expressions. Different classifiers can be used for countable nouns vs. mass nouns (e.g., *ta* for count nouns). Also, there are specific classifiers to indicate cattle (*ræ:s*), aircraft (*færvænd*), people (*næfær*), book (*jeld*), etc.

- **Pronunciation of post-consonantal forms and following the word-final vowels /i/ and /e/(silent *he*)**

Literary Form	Conversational Form	Features	Conversational Example
<i>æm</i>	<i>æm</i>	First person, singular	<i>khubæm</i> 'I am good/fine' <i>zendæm</i> 'I am alive'
<i>i</i>	<i>i</i>	Second person, singular	<i>khubi</i> 'you are good/fine' <i>zendei</i> 'you are alive'
<i>æst</i>	<i>e / æs*</i>	Third person, singular	<i>khube</i> 'he/she is good/fine' <i>zendæs</i> 'he/she is alive'
<i>im</i>	<i>im</i>	First person, plural	<i>khubim</i> 'we are good/fine' <i>zendeim</i> 'we are alive'
<i>id</i>	<i>in</i>	Second person, plural	<i>khubin</i> 'you are good/fine' <i>zendein</i> 'you are alive'
<i>ænd</i>	<i>æn</i>	Third person, plural	<i>khubæn</i> 'they are good/fine' <i>zendæn</i> 'they are alive'

* *e* after consonants and /i/, and *æs* after the word-final /e/

- **The pronunciation of post-vocalic forms (after vowels /a/ and /u/)**

Literary Form	Conversational Form	Features	Conversational Example
<i>yæm</i>	<i>m / æm**</i>	First person, singular	<i>injām</i> 'I am wise' <i>tærsuæm</i> 'I am fearful'
<i>yi</i>	<i>yi</i>	Second person, singular	<i>injāyi</i> 'you are wise' <i>tærsuyi</i> 'you are fearful'
<i>st</i>	<i>s</i>	Third person, singular	<i>injas</i> 'he/she is wise' <i>tærsus</i> 'he/she is fearful'
<i>yim</i>	<i>yim</i>	First person, plural	<i>injāyim</i> 'we are wise' <i>tærsuyim</i> 'we are fearful'
<i>yid</i>	<i>yin</i>	Second person, plural	<i>injāyin</i> 'you are wise' <i>tærsuyin</i> 'you are fearful'
<i>yænd</i>	<i>n / æn**</i>	Third person, plural	<i>injan</i> 'they are wise' <i>tærsun</i> 'they are fearful'

** *m* and *n* after /a/ and *æm* and *æn* after /u/

Orthographic Variance

Similar to the orthography of pronominal clitics, the verbal clitic has different forms in the literary variant depending on the last character of the word it attaches to. When the copula

verb follows a consonant, it attaches to it (with the exception of the *æst* 'is' form). A glide (the source /y/) is inserted after the vowels /a/ and /u/. And finally, the letter *alef* [a] is inserted following the word-final /e/ (written as [h]). As expected, the clitic appearing following the word-final /e/ may not be written in the attached form. This is shown in the table below where the [i] in the transliteration represents the "ye with hamze" (ئى) sometimes used in writing.

After a consonant		After vowels /a/ and /u/		After word-final /e/ (silent <i>he</i>) [detached or isolated]	
م	[m]	يم	[ym]	ام	[~am]
ى	[y]	يى / ئى	[yy / iy]	اى	[~ay]
است*	[~ast]	ست	[st]	است	[~ast]
يم	[ym]	ييم / ئيم	[yym / iym]	ايم	[~aym]
يد	[yd]	ييد / ئيد	[yyd / iyd]	ايد	[~ayd]
ند	[nd]	يند	[ynd]	اند	[~and]

* *æst* cannot be attached to the preceding word.

In the conversational variant, the base form is the one following a consonant. A glide /y/ is inserted if the verbal clitic is attached to a word ending in the vowels /a/ or /u/, and the clitic begins with /i/ (written as [y]).

After a consonant		After vowels /a/ and /u/	
م	[m]	م	[m]
ى	[y]	يى / ئى	[yy/iy]
ه	[h]	س / ست	[s/st]
يم	[ym]	ييم / ئيم	[yym/iym]
ين	[yn]	يين / ئين	[yyn/iyn]
ن	[n]	ن	[n]

In the case of the word-final /e/, several variants exist in the conversational written form. The clitic can be unattached; in this case, an 'alef' [a] is inserted as shown in column 1 below. It may also appear in an attached form for the 1st person singular, and 3rd persons singular and plural (the ones that do not begin with the sound /i/ [y], as shown in column 2. In these instances, similar to the case of the pronominal clitics, the final /e/ sound is assimilated into the initial /æ/ sound of the verbal clitic. So for instance, the traditional (literary) pronunciation of *khæsté-æm* 'I am tired' becomes *khæstæm*. The assimilation of the word-final /e/ is reflected in the orthography in column 3 where the 'silent *he*' is eliminated and the clitic is written in attached form. As the table shows, there are even

more variants for the third person singular form since the final 't' of the verbal clitic may be dropped or maintained, as both pronunciations are attested in spoken Persian.⁵¹

After word-final /e/ (silent <i>he</i>)					
Column 1: detached or isolated form add 'alef'		Column 2: detached form		Column 3: eliminate final 'he' attached form	
خسته‌ام	[xsth~am]	خسته‌م	[xsth~m]	خستم	[xstm]
خسته‌ای	[xsth~ay]				
خسته‌اس	[xsth~as]	خسته‌س / خسته‌ست	[xsth~s/xsth~st]	خستس / خستست	[xsts / xstst]
خسته‌ایم	[xsth~aym]				
خسته‌این	[xsth~ayn]				
خسته‌ان	[xsth~an]	خستن	[xsth~n]	خستن	[xstn]

4.5.3.4 Other complex categories

This section lists a number of words that can form complex tokens in Persian. In each case, the literary variant is introduced first and then its conversational form is presented. Although these elements may be attached in the literary variant, the likelihood of finding these words in the attached form is higher in the conversational written language. It should be noted, however, that none of these words can appear attached to a word-final /e/ (or silent *he*).

- *in* 'this; these' / *an* 'that; those' → *in* / *un* [in conversational] (prefix)

The determiners *in* and *an* (*un* in conversational) may appear attached to the following word as shown.

آیا اینکار ممکن است؟
[Aya *ayn*kar mmkn ast?]
*aya*⁵² *inkar momken æst?*

⁵¹ To illustrate the relative frequency of each orthographic form for 'I am tired', a Google search reveals the following numbers: The traditional orthography *خسته‌ام* [xsth am], which is used in both literary and some conversational texts, has the highest hits as expected. The version with just the clitic (without an intervening 'alef'), *خسته‌م*, is the main one used in conversational texts, while the attached form *خستم* is third in frequency.

خسته‌ام	[xsth am]	→	2,440,000 hits	<i>literary variant</i>
خسته‌م	[xsth m]	→	663,000 hits	<i>conversational - detached</i>
خستم	[xstm]	→	118,000 hits	<i>conversational - attached</i>

A similar informal Google search for 'he/she is tired' results in the following numbers for the various variant forms:

خسته‌است	[xsth ast]	→	3,090,000 hits	<i>literary variant</i>
خسته‌ست	[xsth st]	→	796,000 hits	<i>conversational - detached form 1</i>
خسته‌س	[xsth s]	→	560,000 hits	<i>conversational - detached form 2</i>
خستست	[xstst]	→	624 hits	<i>conversational - attached form 1</i>
خستس	[xsts]	→	620 hits	<i>conversational - attached form 2</i>

⁵² *Aya* is a question particle that indicates an interrogative sentence with a *yes* or *no* response and cannot be translated in English.

this-work possible is?
'Is this (job/action) possible?'

- **hæm** 'also' → **æm** [in conversational] (suffix)

This adverb is usually written as a separate word, but it may appear attached to the preceding (stressed) word. The examples below show two similar sentences found on weblogs, where the first one is written in the literary variant and the second is in the conversational language. As can be seen from these examples, the *hæm* morpheme in the literary version is written as a separate token, while the spoken form *æm* is written attached to the previous word.

Literary Variant

پارسی بلاگ هم هک شد.
parsi blag hæm hæk shod
'Parsi Blog was also hacked.'

Conversational Variant

پرشین بلاگم هک شد!
pershen blagæm hæk shod!
'Persian Blog was also hacked!'

- **ke** 'that' (suffix)

When used as a relativizer 'that', *ke* may attach to the previous word which is normally the head of the relative clause. Although this may happen in the literary variant, it is more common in conversational text.

همون کتابیکه بهم هدیه دادی
[hmvn ktabykh bhm hdyh dady]
hæmun ketabike behem hedia dadi
that book-**that** to-me gift gave.2sg
'that book that you gave me as a gift'

- **væ** 'and' (prefix/suffix)

Although 'vav' is not an attaching character in the Perso-Arabic script, it may appear next to a word without an intervening space.

شورشیان گردآمدند و رفتند ...
[švršyan grd Amdnd vɾftnd]
shureshyan گرد آمدند væɾæftænd
rebels round came **and**-left
'the rebels gathered and left ...'

In conversational text, the 'vav' can appear attached to the preceding character (which never occurs in literary writing and would be considered an error). This is shown below where the 'and' is attached to the preceding words.

بجز از امشب و فرداشبو شبهای دگر توبه کردم که دگر می نخورم
[bjz az amšbv ɸɾdašbv šbhay dgr tvbh krdm kh dgr my nxvrm]
bejoz æz emshæbo ɸærdashæbo shæbhaye degær tobe kærðæm ke degær mey nækhoræm
apart from tonight-**and** tomorrow night-**and** nights-EZ other vow did-1sg that any more wine not-
drink-1sg

'Apart from tonight and tomorrow night and the other nights, I have vowed not to drink wine anymore.'

• **Prepositions** **(prefix)**

Certain prepositions such as *be* 'to' and *bi* 'without' can be attached to the following word, which can be a noun phrase constituent (e.g., noun, pronoun, numeral, adjective, infinitive verb). Similarly, prepositions such as *æz* 'from' or *dær* 'in' (that end in non-attaching characters) may be written without an intervening space separating them from the following word; these should also be treated as attached prepositions for the purpose of a computational analysis. Note that if *be* 'to' [bh] is in the attached form, it loses its final character - the silent 'h' - as shown in the example below.

بنظرشان
[bnzrʃan]
benæzæreshan
be 'to' + *næzær* 'view' + *eshan* 'Poss.3pl'
'in their view'

4.5.4 Verbal Inflection

Persian has a complete verbal inflectional paradigm that consists of simple forms and compound forms (that require an auxiliary verb). The simple forms are divided into two groups depending on whether the "present" or "past" stem of the verb is used in their formation. For instance, the verb *gereftæn* 'to catch' has a "past" stem derived from the infinitival form: *gereft*, and a "present" stem which is usually irregular and difficult to derive from the infinitive: *gir*. Among the simple verbs, the tenses that are formed on the "present" stem are the present indicative, the present subjunctive, the imperative, and the present participle. Those formed on the "past" stem are the simple past, the imperfect, the past participle. Among the compound forms, the future is formed on the "past" stem, while all the other compound forms are based on the past participle, which is itself formed on the "past" stem.⁵³

This section provides a description of the conjugation system for all the tenses in Persian, explaining the subparts used to form each tense. The descriptions of the tenses as well as their usage is adapted from Lazard (1992) and Thackston (1993). Throughout the section, the verb *gereftæn* 'to catch' is used to illustrate the conjugation of the verbs. The following section discusses the morpho-phonological changes on the verbal stems that take place in conversational Persian. The section entitled 'personal inflections' presents the agreement morphemes on the verb. The rest of this presents all the simple and compound tenses in active and passive voice. The various forms of the prefixes are discussed towards the end of the section.

⁵³ "Present" and "past" stems are not necessarily correct terms linguistically. In fact, there is evidence that the so-called past stem does not represent the meaning of past but rather marks some sort of *evidentiality*, such as the expectation by the speaker that an event has taken place or will take place. This could explain, for instance, why it can be used in the formation of the future tense. Nevertheless, we will use these traditional terms for the rest of this section for expository purposes.

4.5.4.1 Stems in conversational Persian

The phonological changes discussed in Section 4.4.1 also apply to the verbal stems, which then affect all the conjugation forms. In addition, there are a number of morphological changes, the most obvious of which is the shortened present stem in Persian verbs.

SHORTENED STEMS

A number of present stems are shortened in conversational Persian. These include:

Verb		Translation	Present Stem (literary)	Present Stem (conversational)
آوردن	<i>aværdæn</i>	Bring	<i>avær</i>	<i>ar</i>
انداختن	<i>ændakhtæn</i>	Throw, Drop	<i>ændaz</i>	<i>ndaz</i>
توانستن	<i>tævanestæn</i>	Be able	<i>tævan</i>	<i>tun</i>
خواستن	<i>khastæn</i>	Want	<i>khah</i>	<i>kha</i>
دویدن	<i>dævidæn</i>	Run	<i>dæv</i>	<i>do</i>
دادن	<i>dadæn</i>	Give	<i>dæh</i>	<i>d</i>
رفتن	<i>ræftæn</i>	Go, Leave	<i>ræv</i>	<i>r</i>
شدن	<i>shodæn</i>	Become	<i>shæv</i>	<i>sh</i>
گذاشتن	<i>gozashtæn</i>	Allow, Put	<i>gozar</i>	<i>zar</i> ⁵⁴
گفتن	<i>goftæn</i>	Say, Tell	<i>gu</i>	<i>g</i>
نشستن	<i>neshestæn</i>	Sit	<i>neshin</i>	<i>shin</i>

Literary	Script/Translit	Convers.	Script/Translit	Translation
<i>mikhahæm</i>	می‌خواهم [myxvahm]	<i>mikham</i>	میخوام [myxvam]	'I want'
<i>bedæhænd</i>	بدهند [bdhnd]	<i>bedæn</i>	بدن [bdn]	'that they give'
<i>miayænd</i>	می‌آیند [my~Aynd]	<i>mian</i>	میان [myan] ⁵⁶	'they are'
<i>mi ændazæd</i>	می‌اندازد [my~andazd]	<i>mindaze</i>	می‌ندازه [my~ndazh]	'he/she/it'

⁵⁴ Although the original spelling of this verb is with the 'zal' character, conversational orthography may use either the 'zal' or the 'ze'. Hence the following two forms are attested in conversational writing: نذار [nDar] and نزار [nzar] 'don't allow'.

⁵⁵ Certain words in Persian contain [xva] but are pronounced *kha*, with a silent [v]. These words are spelled with a *vav* character due to historical reasons. The conversational variant is written sometimes with the silent 'v' and sometimes without it. The following represent the hits on Google for two versions of *mikham* 'I want':

میخوام	[mixvam]	1,410,000 hits
میخام	[mixam]	62,300 hits

⁵⁶ This word can also be written as میان [my~yan] with an extra glide or 'y' inserted between the two vowels.

Note that there are also some person and agreement changes in the conversational form that will be discussed in the following section.

VOWEL ALTERNATION

In particular environments, the vowel may change as shown below. If the vowel is written in the script, the orthography may be different in the conversational variant, otherwise there is no change in spelling.

Literary	Script/Translit	Convers.	Script/Translit	Translation
<i>aværdæn</i>	آوردن [Avrdn]	<i>avordæn</i>	آوردن [Avrdn]	'to bring'
		<i>ovordæn</i>	اوردن [avrdn]	
		<i>ovordæn</i>	اووردن [avvrdn]	
<i>ændakhtæn</i>	انداختن [andaxtn]	<i>endakhtæn</i>	انداختن [andaxtn]	'to throw, drop'
<i>shomardæn</i>	شماردن [šmardn]	<i>shomordæ</i>	شمردن [šmrdn]	'to count'
<i>feshardæn</i>	فشاردن [fšardn]	<i>feshordæn</i>	فشردن [fšrdn]	'to press' ⁵⁷
<i>minevisæm</i>	می نویسم [my~nvysm]	<i>minivisæm</i>	می نویسم [my~nyvysm]	'I am writing'

/AN/ TO /UN/ ALTERNATION

The alternation of the /a/ vowel to /u/ before the nasal /n/ can also be seen in verbal stems.

Literary	Script/Translit	Convers.	Script/Translit	Translation
<i>midanæm</i>	می دانم [my~danm]	<i>midunæm</i>	می دونم [my~dvn̄m]	'I know'
<i>khandim</i>	خواندیم [xandym]	<i>khundim</i>	خوندیم [xvndym]	'we read, we sang'
<i>mandænd</i>	ماندند [mandnd]	<i>mundæn</i>	موندن [mvndn]	'they stayed'

GLOTTAL DELETION

The glottal sound /h/ is dropped from verbal stems such as in *fæhmidæn* 'to understand', which is pronounced *fæ:midæn* in conversational Persian.

Literary	Script/Translit	Convers.	Script/Translit	Translation
<i>mifæhmæm</i>	می فهمم [my~fhmm]	<i>mifæmæm</i>	می فمم [my~fmm]	'I understand'

VOICE AND PLACE ASSIMILATIONS

The alternation of the nasals /n/ to /m/ before a bilabial /b/ is also attested in verb stems where *jonbidæn* 'to move, to hurry' is pronounced as *jombidæn*.

Similarly, devoicing effects can be observed as in *ræghsidæn* which is pronounced *rækhsidæn*, hence the voiced 'gh' sound becomes a voiceless 'kh' when preceding the voiceless sound /s/.

⁵⁷ An informal search on Google gives about 27,300 hits for the conversational form *feshordæm* 'I pressed' as opposed to only 85 hits for the literary version *feshardæn*, indicating that the latter is falling out of use in modern standard Persian.

CLUSTERS

The deletion of /t/ in a /st/ cluster in conversational Persian also applies to verbs. For instance, the past tense verb *shekæstæm* 'I broke (it)' is generally pronounced as *shekæssæm* and it is reflected in the orthography. Similarly, the final 't' in the /ft/ cluster in *ræft* 'he left' is dropped in the conversational form *ræf* and is reflected in the orthography.

SYNCOPE

In certain polysyllabic verbal stems where the first syllable consists of a CV (consonant+vowel) structure, the vowel is dropped when a verbal prefix (*mi-* or *be-*) is added. For instance, the verb *shekæstæn* 'to break' in the present tense is pronounced *mishækænæm* 'I break' in literary Persian. In the conversational variant, however, the vowel /e/ of the first syllable of the stem is dropped and the word is then pronounced as *mishkænæm*. This modification does not affect the orthography at all since the vowel /e/ is not written in the Persian script.

Literary	Conversational	Script/Transliteration (conv.)	Translation
<i>mishækænæm</i>	<i>mishkænæm</i>	میشکنم [myšknm]	'I break'
<i>miferestadim</i>	<i>mifrestadim</i>	میفرستادیم [myfrstady]	'we used to send'
<i>bešenævæm</i>	<i>bešnævæm</i>	بشنوم [bšnm]	'(that) I hear'
<i>mishenasi</i>	<i>mishnasi</i>	میشناسی [myšnasy]	'you know'
<i>beforushim</i>	<i>befrushim</i>	بفروشیم [bfrvšym]	'(that) we sell; let's sell'

Common verbs that follow this pattern are *shekæstæn* 'to break', *shomordæn* 'to count', *shenakhtæn* 'to know', *shenidæn* 'to hear', *shekaftæn* 'to crack', *ferestadæn* 'to send', *forukhtæn* 'to sell', *feshordæn* 'to press', *gozashtæn* 'to allow, to put', *gozæstæn* 'to pass', *shetaftæn* 'to hurry', *neshestæn* 'to sit'⁵⁸, *neveshtæn* 'to write'.

4.5.5 Other verbs:

The verb *tævanestan* 'to be able' has both a shortened stem and a /an/ to /un/ alternation, as the present and past stems change from *tævan* and *tævanest* to *tun* and *tunest*, respectively.

Literary	Script/Translit	Convers.	Script/Translit	Translation
<i>mitævanim</i>	می‌توانیم [mytvany]	<i>mitunim</i>	می‌تونیم [mytvny]	'we can'
<i>tævanesti</i>	توانستی [tvany]	<i>tunesti</i>	تونستی [tvny]	'you were able to'
		<i>tunesi</i>	تونسی [tvnsy]	'you were able to'

The present stem of the verb *shostæn* 'to wash' has the form *shu* in literary Persian but in the conversational form it is generally pronounced *shur*, as in *mi-shuræm* 'I am washing'.

The verb *istadæn* 'to stand, to stop' has different stems and conjugation forms in the conversational variant. For this verb, even the infinitival form is different: *vaysadæn*. The past stem is therefore *vaysad* and the tenses formed on the past are then conjugated

⁵⁸ In the case of *neshestæn* 'to sit', Syncope applies only if the verbal stem *neshin* is used as in *minshinæm* 'I sit'. However, in the conversational speech, the stem is often changed to *shin* in which case this operation does not apply *mishinæm* 'I sit'.

regularly. The present stem is possibly *vays*. In the present tense, however, the *va* part behaves as a particle and the progressive affix *mi-* separates it from the rest of the stem: *va-mi-ssam* 'I am standing, stopping', *va-mi-ssi* or *vay-misti* 'you are standing, stopping', etc. The imperative forms are *vaysa* or *vaysta* 'stand! stop!', *vaysim* or *vassim* 'let's stand/stop', and *vaysin* or *vassin* when referring to a 2nd plural subject. The subjunctive forms do not take the *be-* prefix: *æge vaysæm ...* 'if I stand/stop...'

4.5.5.1 Personal inflections

Three types of personal inflections are used in the literary variant for conjugating the Persian verbal forms.

- **Default written form - literary**

Present Inflection		Imperative Inflection		Past Inflection	
م	[m]	N/A		م	[m]
ی	[y]	∅	[]	ی	[y]
د	[d]	N/A		∅	[]
یم	[ym]	یم	[ym]	یم	[ym]
ید	[yd]	ید	[yd]	ید	[yd]
ند	[nd]	N/A		ند	[nd]

- **Pronunciation - literary**

Present Inflection		Imperative Inflection		Past Inflection	
م	<i>æm</i>	N/A		م	<i>æm</i>
ی	<i>i</i>	∅		ی	<i>i</i>
د	<i>æd</i>	N/A		∅	
یم	<i>im</i>	یم	<i>im</i>	یم	<i>im</i>
ید	<i>id</i>	ید	<i>id</i>	ید	<i>id</i>
ند	<i>ænd</i>	N/A		ند	<i>ænd</i>

For the literary variant, a 'ye' (or the sound /y/) is inserted between the stem and the inflection if the previous stem ends in a vowel, as in *mi-guyæm* 'I am saying' [my~guy^ym] (from the verb *goftæn* 'to say' with present stem *gu*).

Conversational Variant

The conversational forms differ for the 3rd persons singular/plural, and for the 2nd plural:

- **Written form - conversational**

Present Inflection		Imperative Inflection		Past Inflection	
م	[m]	N/A		م	[m]
ی	[y]	∅	[]	ی	[y]

د / ه	[e/d]*	N/A	∅	[]
يم	[ym]	يم	يم	[ym]
ين	[yn]	ين	ين	[yn]
ن	[n]	N/A	ن	[n]

* -e after consonants and -d after vowels.

The table below shows the pronunciation of the conversational variant if the stem ends in a consonant. The main difference is in the 3rd person singular form of the present inflection which is pronounced -e rather than -æd.

• **Pronunciation: after consonants**⁵⁹

Present Inflection		Imperative Inflection		Past Inflection	
م	æm	N/A		م	æm
ى	i	∅		ى	i
ه	e	N/A		∅	
يم	im	يم	im	يم	im
ين	in	ين	in	ين	in
ن	æn	N/A		ن	æn

The following table represents the conversational pronunciation if the stem ends in the vowels /a/ and /u/. As shown, if the inflection starts with the sound /i/, a /y/ may be inserted between the stem and the inflection. Otherwise, the /i/ of the inflection is itself pronounced as /y/. For example, the present tense of the verb *amædæn* 'to come' whose present stem is simply *a*, gets the following forms (here the present stem *a* is bolded):

<i>miyam</i>	'I am coming'
<i>miyayi /miyay</i>	'you are coming'
<i>miyad</i>	'he/she/it is coming'
<i>miyayim /miyaym</i>	'we are coming'
<i>miyayin /miyayn</i>	'you are coming'
<i>miyan</i>	'they are coming'

• **Pronunciation: after vowels /a/ or /u/**

(there are no past stems ending in a vowel)

Present Inflection		Imperative Inflection	
م	m	N/A	
ى / يى	yi/y	∅	
د	d	N/A	
يم / ييم	yim/ym	يم / ييم	yim/ym

⁵⁹ Stems ending with the vowel /o/ tend to follow this pattern, e.g., *mi-do-æm* 'I am running'

بین / ین	<i>yin/yn</i>	بین / ین	<i>yin/yn</i>
ن	<i>n</i>	N/A	

The next two sections present the simple tenses formed on the present and past stems of the verb. The examples are all given in the literary variant of the language. The conversational variant can be derived by substituting the relevant personal inflectional form or the appropriate stem form.

4.5.5.2 Simple tenses on the present stem

- **Present indicative**

The present tense is used to refer to the simple present (i.e., I catch) as well as the present continuous (i.e., I am catching) in English. It is also used in the case of an action that began in the past but which still continues in the present (ex. *æz diruz ta hala minevisæd* 'He/she has been writing since yesterday'). In the conversational variant, the present tense is used for the future as well.

Present = [my] + PresentStem + PresentInflection

ex.

my + gyr + m	→	my~gyrm	(<i>migiræm</i>)	می‌گیرم
my + gyr + y	→	my~gyry	(<i>migiri</i>)	می‌گیری
my + gyr + d	→	my~gyrd	(<i>migiræd</i>)	می‌گیرد
my + gyr + ym	→	my~gyrym	(<i>migirim</i>)	می‌گیریم
my + gyr + yd	→	my~gyryd	(<i>migirid</i>)	می‌گیرید
my + gyr + nd	→	my~gyrnd	(<i>migirænd</i>)	می‌گیرند

- **Present subjunctive**

The subjunctive is used in instances where the realization of the action is not considered certain. It is used in contexts of doubt, desire, wish, possibility, etc. In Persian, the subjunctive appears after the modals in a sentence as in the conversational form *mitunæm beræm* (I can go). The subjunctive prefix is usually optional with compound verbs: *shena konæm* or *shena bokonæm* ((that) I swim) are both okay. In addition, the first and third persons of the present subjunctive are used in questions expressed by "shall I/we?" in English. The first person plural can be used in the sense of "let's" as in *berim* (let's go). The subjunctive is very often used in subordinate clauses whenever there is a nuance of eventuality, obligation, doubt, wish, etc.

Subjunctive = [b] + PresentStem + PresentInflection

The prefix [b] can sometimes be omitted in literary text.

ex.

b + gyr + m	→	bgyrm	(<i>begiræm</i>)	بگیرم
b + gyr + y	→	bgyry	(<i>begiri</i>)	بگیری
b + gyr + d	→	bgyrd	(<i>begiræd</i>)	بگیرد
b + gyr + ym	→	bgyrym	(<i>begirim</i>)	بگیریم
b + gyr + yd	→	bgyryd	(<i>begirid</i>)	بگیرید

b + gyr + nd → bgyrnd (*begirænd*) بگیرند

- **Imperative**

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

Imperative = [b] + PresentStem + ImperativeInflection

ex.

b + gyr + ∅ → bgyr (*begir*) بگیر
 b + gyr + ym → bgyrym (*begirim*) بگیریم
 b + gyr + yd → bgyryd (*begirid*) بگیرید

4.5.5.3 Simple tenses on the past stem

- **Simple Past (Preterite)**

Also known as the Preterite or the Past Absolute, this tense refers to a completed action. It is also used to describe an action which is about to be completed, so for instance, as an answer to "Are you coming?", one might say *amædæm* 'I came' in the meaning of 'I am coming'. In a certain subordinate clause, the simple past is used to indicate an action that will be completed at the moment when the action expressed by the main verb will take place; a present would be used in English (ex. *vækhti namehe resid, ma ro seda kon* 'when the letter arrives, call us').

Simple Past = PastStem + PastInflection

ex.

grft + m → grftm (*gereftæm*) گرفتم
 grft + y → grfty (*gerefti*) گرفتی
 grft + ∅ → grft (*gereft*) گرفت
 grft + ym → grftym (*gereftim*) گرفتیم
 grft + yd → grftyd (*gereftid*) گرفتید
 grft + nd → grftnd (*gereftænd*) گرفتند

- **Imperfect**

The imperfect expresses continuous, habitual or recurrent actions in the past. It is also used as the conditional tense (ex. *ægær mitævanestæm, hætmæn mikhæridæm* 'if I could've, I certainly would have bought it'.)

Imperfect = *my* + PastStem + PastInflection

ex.

my + grft + m → mygrftm (*migereftæm*) می‌گرفتم
 my + grft + y → mygrfty (*migerefti*) می‌گرفتی
 my + grft + ∅ → mygrft (*migereft*) می‌گرفت
 my + grft + ym → mygrftym (*migereftim*) می‌گرفتیم

my + grft + yd → mygrftyd (*migereftid*) می‌گرفتید
 my + grft + nd → mygrftnd (*migereftænd*) می‌گرفتند

4.5.5.4 compound tenses on the past stem

The Compound Forms mainly consist of an optional prefix attached to the Past Participle and combined with the auxiliary *budæn* 'to be' in present or past tense. The past participle itself is formed by combining the past stem of the verb with the /e/ sound (written as a silent *he*). As the participle ends in the 'silent *he*' character, the following element cannot appear attached to it, hence the present auxiliary is written either in detached form (with intervening half-space) or in isolated form (with intervening full-space).

The present tense of the auxiliary *budæn* 'to be' is simply the forms of the copula verb discussed in Section 4.5.3. The past tense of the verb *budæn* 'to be' follows the rules for the derivation of the past tenses; in that sense the verbal inflectional system of Persian is recursive.

The future is formed by combining the auxiliary *khastæn* 'to want' (without a prefix) with the verbal element. The passive used the auxiliary *shodæn* 'to become'.

- **Perfect**

The Perfect or Present Perfect (also known as the Past Narrative) refers to an action that has recently been completed or that started in the past but hasn't yet completed. Its usage is similar to English.

Present Perfect = PastParticiple + *budæn* (present)

where PastParticiple = PastStem + 'h' [grft + h = grfth]

ex.

grfth + am	→	grfth~am	(<i>gerefteæm</i>)	گرفته‌ام
grfth + ay	→	grfth~ay	(<i>gereftei</i>)	گرفته‌ای
grfth + ast	→	grfth~ast	(<i>gerefteæst</i>)	گرفته‌است
grfth + aym	→	grfth~aym	(<i>gerefteim</i>)	گرفته‌ایم
grfth + ayd	→	grfth~ayd	(<i>gerefteid</i>)	گرفته‌اید
grfth + and	→	grfth~and	(<i>gerefteænd</i>)	گرفته‌اند

در عراق اختطاف کنندگان هفت راننده لاری را گروگان گرفته‌اند.

[source: NewsVOA, 30 July 2004]

[dr eraq axTaf knndgan hft ranndh lary ra grvgan **grfth and**]

In Iraq kidnappers seven driver Lari OBJ hostage **taken-have**.

→ 'Kidnappers in Iraq **have taken** seven Lari drivers hostage.'

Conversational Variant

In the conversational language, the present perfect is very different and undergoes two morphophonological changes:

- (i) the past participle ending /e/ (written as silent *he*) is assimilated into the initial /æ/ vowel of the personal inflection;
- (ii) the stress that appears on the past participle ending in the literary form moves to the vowel of the personal inflection.

These forms are pronounced similar to the Simple Past tense except for a change in the stress pattern. For example, the Simple Past of *gereftæn* 'to catch' for the 1st person singular is *gerEftæm* 'I caught' where the main stress falls on the last syllable of the past stem, shown here with capitalization. The Present Perfect for the 1st person singular, however, is *gereftÆm* 'I have caught' with the main stress falling on the inflection (i.e., the last syllable of the whole word). In addition, the *æst* auxiliary of the 3rd person singular is dropped in the conversational language. The full paradigm is shown below for the two tenses in conversational Persian. It should be noted that there is no difference in orthography at all between the two forms except for the 3rd person singular, although the pronunciation is quite distinct throughout.

Present Perfect		Simple Past	
گرفتم	<i>gereftÆm</i>	گرفتم	<i>gerEftæm</i>
گرفتی	<i>gereftI</i>	گرفتی	<i>gerEfti</i>
گرفته	<i>gereftE</i>	گرفت	<i>gerEft</i>
گرفتیم	<i>gereftIm</i>	گرفتیم	<i>gerEftim</i>
گرفتین	<i>gereftIn</i>	گرفتین	<i>gerEftin</i>
گرفتن	<i>gereftEn</i>	گرفتن	<i>gerEftæn</i>

• Pluperfect

This tense is also known as the Past Perfect. The use of the pluperfect is very much the same as in English indicating that at a moment in the past an action was already completed. It is, however, also used as a descriptive tense (ex. *istade bud* '(he/she/it) was standing').

Past Perfect = PastParticiple + *budæn* (preterite)

where PastParticiple = PastStem + [h]

[grfth]

ex.

grfth bvd + m	→	grfth bvdm	(<i>gerefte budæm</i>)	گرفته بودم
grfth bvd + y	→	grfth bvd y	(<i>gerefte budi</i>)	گرفته بودی
grfth bvd + ∅	→	grfth bvd	(<i>gerefte bud</i>)	گرفته بود
grfth bvd + ym	→	grfth bvd ym	(<i>gerefte budim</i>)	گرفته بودیم
grfth bvd + yd	→	grfth bvd yd	(<i>gerefte budid</i>)	گرفته بودید
grfth bvd + nd	→	grfth bvd nd	(<i>gerefte budænd</i>)	گرفته بودند

او به همراه افرادش در زیر زمین مسجد پناه گرفته بودند.

[source: CyrusOnline 2007]

[av bh hmrah afradš dr zyr zmy n msjd pnah grfh bvdnd]

he along with people-his in basement mosque shelter **taken-had.3pl.**
 → 'He and his people **had taken** shelter in the basement of the mosque.'

- **Compound Imperfect**

This tense (also known as Past Narrative Continuous) expresses a past action considered in its duration and which has taken place in a completed past. It is also used when speaking of bygone days. This tense is not used in the conversational variant.

Compound Imperfect = [my] + PastParticiple + *budæn* (present)

where PastParticiple = PastStem + [h]

[grfth]

ex.

my + grfth + am	→	my~grfth~am	(<i>migerefteæm</i>)	می گرفته ام
my + grfth + ay	→	my~grfth~ay	(<i>migereftei</i>)	می گرفته ای
my + grfth + ast	→	my~grfth~ast	(<i>migerefteæst</i>)	می گرفته است
my + grfth + aym	→	my~grfth~aym	(<i>migerefteim</i>)	می گرفته ایم
my + grfth + ayd	→	my~grfth~ayd	(<i>migerefteid</i>)	می گرفته اید
my + grfth + and	→	my~grfth~and	(<i>migerefteænd</i>)	می گرفته اند

رئیس دولت بگوید کدام حزب باج می گرفته است؟

[source: Armin Montazeri weblog; 2007]

[riys dvl t bgvyd kdam Hzb baj **my grfth ast?**]

head government say(subj) which party ransom **was-taking-is?**

→ 'The head of state should say what political party used to take ransom.'

Conversational Variant

In the conversational language, this tense behaves like the present perfect, apart from the additional *mi* prefix. The full paradigm is shown below and contrasted with the basic Imperfect.

Compound Imperfect		Simple Imperfect	
می گرفتم	<i>mi gereftÆm</i>	می گرفتم	<i>mi gerEftæm</i>
می گرفتی	<i>mi gereftl</i>	می گرفتی	<i>mi gerEfti</i>
می گرفته	<i>mi gereftE</i>	می گرفت	<i>mi gerEft</i>
می گرفتیم	<i>mi gereftlm</i>	می گرفتیم	<i>mi gerEftim</i>
می گرفتین	<i>mi gereftln</i>	می گرفتین	<i>mi gerEftin</i>
می گرفتن	<i>mi gereftÆn</i>	می گرفتن	<i>mi gerEftæn</i>

- **Double Compound Past**

This tense is the completed past of the perfect, indicating that an action was already completed. It is similar to the pluperfect, except that where the pluperfect refers to an anterior action with respect to the preterite, the double compound past expresses an anterior action in the context of the perfect tense. Sometimes, this tense replaces the

pluperfect in order to indicate that a fact is not certain but only presumed. There is no equivalent tense in English and it is generally translated like the Pluperfect. This tense is not used in the conversational variant of the language and is very rare in the literary form.

Double Compound Past = PastParticiple + *budæn* (perfect)

where PastParticiple = PastStem + 'h'

[grfth]

and *budæn* (perfect) = *budæn* (past-participle) + *budæn* (present)

ex.

grfth bvdh + am	→	grfth bvdh~am	(<i>gerefte budeæm</i>)	گرفته بوده‌ام
grfth bvdh + ay	→	grfth bvdh~ay	(<i>gerefte budei</i>)	گرفته بودهای
grfth bvdh + ast	→	grfth bvdh~ast	(<i>gerefte budeæst</i>)	گرفته بوده‌است
grfth bvdh + aym	→	grfth bvdh~aym	(<i>gerefte budeim</i>)	گرفته بودهایم
grfth bvdh + ayd	→	grfth bvdh~ayd	(<i>gerefte budeid</i>)	گرفته بودهاید
grfth bvdh + and	→	grfth bvdh~and	(<i>gerefte budeænd</i>)	گرفته بوده‌اند

ریچارد کلارک، رئیس سابق عملیات ضد تروریسم کاخ سفید، مقامات دولت آمریکا را متهم کرده است که خطر شبکه القاعده را جدی نگرفته بوده‌اند.

[source: BBC Persian, 10 April 2004]

Richard Clark, head former operations anti-terrorism White House, authorities government America OBJ accused has-done that danger network Al-Qaeda OBJ serious **not-taken have-had**.

→ 'Richard Clark, former head of anti-terrorist operations in the White House, has accused U.S. government authorities of not taking the danger of the Al-Qaeda network seriously.'
(literally: ... that they **had not taken** seriously the danger of the Al-Qaeda network)

Conversational Variant

In the conversational variant, the auxiliary follows the present perfect conjugation; the inflection is attached to the past participle and receives the main stress. The full paradigm is shown below and contrasted with the Past Perfect or Pluperfect. This tense is extremely rare in the conversational speech, however.

Double Compound Past		Past Perfect	
گرفته بودم	<i>gerefte budÆm</i>	گرفته بودم	<i>gerEfte budæm</i>
گرفته بودی	<i>gerefte budI</i>	گرفته بودی	<i>gerEfte budI</i>
گرفته بوده	<i>gerefte budE</i>	گرفته بود	<i>gerEfte bud</i>
گرفته بودیم	<i>gerefte budIm</i>	گرفته بودیم	<i>gerEfte budim</i>
گرفته بودین	<i>gerefte budIn</i>	گرفته بودین	<i>gerEfte budin</i>
گرفته بودن	<i>gerefte budÆn</i>	گرفته بودن	<i>gerEfte budæn</i>

• Past Subjunctive

Also known as the Compound Subjunctive, this tense has the same modal value as the simple subjunctive but it expresses actions referring to the past. (ex. *momken æst reside bashæd* '(he/she/it) may have arrived.')

Past Subjunctive = PastParticiple + *budæn* (subjunctive)

where PastParticiple = PastStem + [h]

[grfth]

and *budæn* (subjunctive) = PresentStem + PresentInflection

ex.

grfth baš + m	→	grfth bašm	(<i>gerefte bashæm</i>)	گرفته باشم
grfth baš + y	→	grfth bašy	(<i>gerefte bashi</i>)	گرفته باشی
grfth baš + d	→	grfth bašd	(<i>gerefte bashæd</i>)	گرفته باشد
grfth baš + ym	→	grfth bašym	(<i>gerefte bashim</i>)	گرفته باشیم
grfth baš + yd	→	grfth bašyd	(<i>gerefte bashid</i>)	گرفته باشید
grfth baš + nd	→	grfth bašnd	(<i>gerefte bashænd</i>)	گرفته باشند

برنامه ریزان نظامی آمریکا و بریتانیا باید بر سر نوعی قرار گرفته باشند.

[source: BBC Persian; April 2003]

[brnamh ryzan nżamy Amryka v brytanya bayd br sr nvey dv rahy qrar **grfth bašnd**
planners military US and UK must at type-a crossroads placement **have-been-taken**

→ 'American and British military planners must **have been placed** at a crossroads.'

• Future

The future tense expresses an action in the future. The auxiliary used to form the future tense is the verb *khastæn* 'to want'.

Future = *khastæn* (present) + PastStem

where *khastæn* (present) = PresentStem + PresentInflection [note: without the prefix *my*]

ex.

xvah + m grft	→	xvahm grft	(<i>khahæm gereft</i>)	خواهم گرفت
xvah + y grft	→	xvahy grft	(<i>khahi gereft</i>)	خواهی گرفت
xvah + d grft	→	xvahd grft	(<i>khahæd gereft</i>)	خواهد گرفت
xvah + ym grft	→	xvahym grft	(<i>khahim gereft</i>)	خواهیم گرفت
xvah + yd grft	→	xvahyd grft	(<i>khahid gereft</i>)	خواهید گرفت
xvah + nd grft	→	xvahnd grft	(<i>khahænd gereft</i>)	خواهند گرفت

دموکراتهای آمریکا بر روسیه سخت خواهند گرفت

[source: Mehr News; 2006]

[dmvkrathay Amryka br rvsyh sxt **xvahnd grft**
democrats US on Russia hard **will-take**

→ 'U.S. Democrats will be hard on Russia.'

(literally: U.S. Democrats **will take** it hard on Russia)

Conversational Variant

In the conversational language the present tense is used to express the future. However, the future tense can be used in the conversational variant if one wishes to emphasize that an action will be taking place in the future.

4.5.5.5 Progressive

The progressive emphasizes that an action is in the process of taking place. It is formed by combining the modal verb *dashtæn* 'to have' with the verb. This tense is more common in the conversational language than in the literary variant and in fact is the main form used to express the continuous present such as 'I am reading' (vs. the habitual 'I read').

Present Progressive = *dashtæn* (present) + Present-Tense

where *dashtæn* (present) = PresentStem [dar] + PresentInflection [*note*: without the prefix *my*]

ex.

dar + m my + gyr + m → darm my~gyrm (*daræm migiræm*) دارم می‌گیرم

Past Progressive = *dashtæn* (preterite) + Imperfect

ex.

dašt + m my + grft + m → daštm my~grftm (*dashtæm migereftæm*) داشتم می‌گرفتم

4.5.5.6 Passive voice

A complete passive conjugation is formed with the help of the past participle followed by the passive auxiliary verb *shodæn* 'to become', regularly conjugated. The formation of all the tenses is listed below for the passive. Only the third person singular inflection is illustrated for each case, and whenever the conversational form is different, it is listed.

The use of the passive is restricted in Persian although it is more common in literary text. It is usually not used when the sentence can be expressed by the active voice. The passive is used particularly when the agent of the action is not expressed. The compound forms are rarely used in conversational Persian.

INDICATIVE

Present = PastParticiple + *shodæn* [Present]

grfth my~švd (*gerefte mishævæd*) گرفته می‌شود [literary form]

grfth my~šh (*gerefte mishe*) گرفته می‌شه [conversational form]

Simple Past = PastParticiple + *shodæn* [Simple Past]

grfth šd (*gerefte shod*) گرفته شد

Imperfect = PastParticiple + *shodæn* [Imperfect]

grfth my~šd (*gerefte mishod*) گرفته می‌شد

Perfect = PastParticiple + *shodæn* [Perfect]

grfth šdh~ast (*gerefte shodeæst*) گرفته شده‌است [literary form]

grfth šdh (*gerefte shode*) گرفته شده [conversational form]

Pluperfect = PastParticiple + *shodæn* [Pluperfect]

grfth šdh bvd (*gerefte shode bud*) گرفته شده بود

CompoundImperfect = PastParticiple + *shodæn* [CompoundImperfect]

grfth my~šdh~ast (*gerefte mishodeæst*) گرفته می‌شده‌است [literary form]

grfth my~šdh (*gerefte mishode*) گرفته می‌شده [conversational form]

DoubleCompound = PastParticiple + *shodæn* [DoubleCompound]

grfth šdh bvdh~ast (*gerefte shode budeæst*) گرفته شده بوده است [literary form]
 grfth šdh bvdh (*gerefte shode bud*) گرفته شده بوده [conversational form]

Future = PastParticiple + *shodæn* [Future]

grfth xvahd šd (*gerefte khahæd shod*) گرفته خواهد شد

SUBJUNCTIVE

PresentSubjunctive = PastParticiple + *shodæn* [PresentSubjunctive]

grfth švd (*gerefte shævæd*) گرفته شود [literary form]
 grfth šh/grfth bšh (*gerefte she/gerefte beshe*) گرفته شه / گرفته بشه [conversational form]

CompoundSubjunctive = PastParticiple + *shodæn* [CompoundSubjunctive]

grfth šdh bašd (*gerefte shode bashæd*) گرفته شده باشد [literary form]
 grfth šdh bašh (*gerefte shode bashe*) گرفته شده باشه [conversational form]

IMPERATIVE

Imperative = PastParticiple + *shodæn* [Imperative]

grfth šv (*gerefte sho*) گرفته شو

4.5.5.7 Negation

The negation morpheme attaches to the beginning of the conjugated form of the main verb in the active voice (e.g., *nægereftim* 'we didn't catch/get'; *nægerefte budæm* 'I hadn't caught/gotten'; *nækhahæm gereft* 'I will not catch/get'). In the passive voice the negation morpheme appears on the passive auxiliary *shodæn* 'to become' (*gerefte næshode bud* 'had not been caught'). The negative morpheme has the same form in both variants of the language.

Orthographic Variance

- ن - 'n'

Occurs before consonants and it is always written attached to the following word. The negation prefix is generally pronounced /næ/, but it is pronounced /ne/ before the mi prefix (the durative prefix in certain verb tenses). However, since the vowels /æ/ or /e/ are not written in the Persian script, this does not affect the orthography of the morpheme.

Transcription	Persian script/transliteration	Translation
<i>næ</i> ræft	نرفت [nrft]	'(he/she) didn't go'
<i>ne</i> miræft	نمی‌رفت [nmy~rft]	'(he/she) wasn't going'

- نی - 'ny'

Occurs before vowels. It is always written attached to the following word.

Transcription	Persian script/transliteration	Translation
<i>næ</i> yamæd	نیامد [nyamd]	'(he/she) didn't come'

4.5.5.8 Progressive prefix

This prefix is used to form the present indicative, the imperfect and the compound imperfect and has a durative or progressive value. It has the same pronunciation in both literary and conversational variants.

Orthographic Variance

- می - 'mi'

This prefix can appear in either attached, detached or isolated form as shown below for *minevisi* 'you are writing'. The *mi* prefix has the same form before consonants or vowels.

Isolated Form (intervening full space)	Detached Form (intervening half-space)	Attached Form
می نویسی	می نویسی	مینویسی
[my nvysy]	[my~nvysy]	[mynvysy]

4.5.5.9 Subjunctive prefix

This prefix characterizes the subjunctive and imperative. In compound verbs that have the prepositional preverbal elements *dær* or *bær*, this morpheme is omitted. In other compound verbs, it is generally optional. It is always attached to the following element.

Conversational Variant

In conversational language, the prefix is pronounced *be* but in certain contexts it can be pronounced as *bo*, e.g., *boro* 'go', *bodoe* '(that) he runs'. Before vowels it is pronounced as *bi*, e.g., *bia* 'come'. In rare cases, the vowel /e/ may optionally change to /i/ as in *bishin* 'sit' or *binivis* 'write'.

Orthographic Variance

- ب - 'b'

Occurs before consonants.

Transcription	Persian script/transliteration	Translation	
<i>benevis</i>	بنویس [bnvys]	'write!'	(literary/conversational)
<i>binivis</i>	بینویس [bynyvys]	'write!'	(conversational)

- بی - 'by'

Occurs before vowels.

Transcription	Persian script/transliteration	Translation	
<i>biavær</i>	بیاور [byavr]	'bring!'	(literary/conversational)
<i>biar</i>	بیار [byar]	'bring!'	(conversational)

4.5.6 Morphotactics

There is a relative ordering of the affixes that appear on a word. For instance, the order of the main affixes that can appear on nouns is as shown in the Table below (stress is marked with an accent in the pronunciation column). Note that this table does not represent all of the possible morphemes that can appear on the noun. For instance, *hæm* 'also' can be attached as in *heyvunam* (حيونام) = *heyvun* 'animal' + *a* 'plural' + *m* 'also' meaning 'the animals also'.⁶⁰

		Order of Affixes			Example	Transliteration	Pronunciation	Translation
Noun	plural	ezafe	∅	کتاب	ktab	ketáb-e	the book of	
	plural	indefinite relativizer	copula verb	کتابیه (کتابییه که)	ktabyh ktabayyh (ke)	ketáb-i-e ketabá-i-e (ke)	it's a book they're books (that)	
		pron. clitic		کتابمه	ktabamh	ketabá-m-e	they're my books	
	definite			کتابه	ktabh	ketab-é	the book	

The relative ordering of all affixes is not discussed in this report but they can in general be used to restrict possible morphological analyses.

4.6 Syntax

There are a number of features in the conversational variant of Persian that either do not exist in the literary language or are not as common. It was already pointed out that there is no definite marker on nouns in the literary variant in contrast with the conversational language. In addition, conversational text contains more instances of scrambling (permutations of word order), idiomatic expressions, and cultural inferences. Although a more detailed study of the syntactic properties of Persian Blogspeak are needed, this section discusses some of the features that are more commonplace in the conversational language found on blogs.

4.6.1 The verb 'to be'

There are two distinct ways of expressing 'to be' in Persian. Traditionally, the clitic copula verb (ex. *æm* = I am) which was discussed in Section 4.5.3 is used for expressing descriptions and predicates, while the *hæstæm* type is generally used for speaking of existence (i.e., equivalent of 'there is/are'). However, the distinction between the two verbs is less obvious in modern Persian and in many contexts the two verbs are being used interchangeably as in *iruniæm* = *iruni hæssæm* 'I am Iranian'.

⁶⁰ This example can also mean *heyvun* 'animal' + *a* (plural) + *m* 'my' = 'my animals'

4.6.2 The indefinite 'ye'

Traditionally, the word *yek* or *ye* meaning 'a, one' is not used together with the indefinite affix (*i*) on the same noun. However, the two co-occur more often in conversational Persian as in

ye/yek ketab-i one book-INDEF 'a book'

4.6.3 Topicalization

Topicalization refers to the emphasis placed on the element being talked about in a sentence by preposing it to the beginning of the sentence. Examples from English are "Those girls, they giggle when they see me" and "Cigarettes, you couldn't pay me to smoke them"⁶¹. There are a lot more cases of topicalization in conversational text in Persian than in literary writing. In the following example the topicalized elements are highlighted.

اون مرده، ماشین زده تش.
[avn mrdh, mašyn zdh tš]
un mærdə mashin zædætesh
that man car has hit-3sg

'that man, a car has hit him' (i.e., that man, he has been hit by a car)

The topicalized element often appears with the object marker (which has also been called a topic marker in grammar books and in the linguistic literature). In these cases, the topicalized element does not have to be at the beginning of the sentence and it may be moved around more.

امشبو منتظر نباش
[amšbv mntzr nbaš]
emshæbo montæzer næbash
tonight-OBJ waiting not-be.2sg
'as for tonight, don't wait (up)'

4.6.4 The uses of 'ke'

Ke can be used as a relativizer as in *mærdi-ro ke diruz didæm ...* 'the man **that** I saw yesterday...', or it can be used as a conjunction introducing indirect discourse as in *behesh goftæm ke khæstæm* 'I told him/her **that** I was tired'. It is also used in the meaning of 'in order to' or 'so that' as in *behet pul dadæm ke bæram ketab bekhæri* 'I gave you money **so that** you will buy me a book'. Another interesting usage of *ke* is in what has come to be known as the "indifference-*ke*" constructions since these sentences show a certain indifference by the speaker. These constructions are found mainly in conversational language.

رفت که رفت
[rft kh rft]
ræft ke ræft
left.3sg that left.3sg
'so what if he left?'

همین هست که هست
[hmyn hst kh hst]
hæmin hæst ke hæst

⁶¹ Source: the Free Dictionary at www.thefreedictionary.com.

this is.3sg that is.3sg
 'well, that's all there is!'
 خوب گذاشتی که گذاشتی
 [xvb gDašty **kh** gDašty]
khob gozashti ke gozashti
 well put.2sg that put.2sg
 'well, so what if you put it?'

Ke can also be used to mean 'when' as in

بمباران‌ها که شروع می‌شدن ما در می‌رفتیم
 [bmbaranha **kh** šrve my~šdn ma dr my~rftym]
bombaranha ke shoru mishodæn ma dær miræftim
 bombings that start became.3pl we out went.1pl
 'When the bombings started, we would run away.'

Finally, there is an additional use by *ke* that is found mainly in the conversational language where it is used as emphasis. In these emphatic constructions, the *ke* follows the focused element in the sentence, but at the same time it conveys the meaning 'though'. Examples are given below:

من که بهت گفتم
 [mn **kh** bht gftm]
mæn ke behet goftæm
 I that to-you said.1sg
 'But I told you' or 'I told you though'
 ما از ممیزی نمی‌ترسیم که!
 [ma az mmyzy nmy~trsym **kh!**]
ma æz momæyezi nemitærsim ke
 we from Momayezi not-fear.1pl that
 'But we are not afraid of Momayezi!'

4.6.5 Dropped prepositions

In literary Persian, the preposition *dær* 'in' is used to indicate a place as in *dær khane* 'at home' or 'in the house'. In conversational Persian, it is very common to drop *dær* or its conversational equivalent *tu* as in *dadashæm khunæs* (brother-my house-is) to say 'My brother is at home'⁶². Similarly, the preposition *be* 'to' indicating a direction is often dropped as in *mirim xune* (we-go home) as opposed to *mirim be xune* (we-go to home). These constructions are also common in English.

4.6.6 Free word order

The traditional subject-object-verb or in general verb-final word order of literary Persian is not followed as strictly in conversational text. Several examples are shown below where the verb in each sentence is highlighted.

هیچوقت حس استیصال بهم دست نداده بود تهران
 [hyčvqt Hs astySal bhm **dst ndadh bvd** thran]
hichvæxt hese estisal behem dæst nædade bud tehran

⁶² This sentence is ambiguous as it also means 'my brother is a house!'

never feeling abject poverty to-me **hand not-given was** Tehran
 'The feeling of abject poverty **had never hit** me in Tehran'

دیدن دریاچه آلیس آرامش داد بهم
 [dydn dryačh Alys **Aramš dad** bhm]
didæne dæryacheye alis aramesh dad behem
 seeing lake Alice **tranquility gave** to-me
 'Seeing Alice Lake **gave me tranquility.**'

عین خر داره خون میاد ازش
 [eyn xr darh **xvn myad** azš]
eyne khær dare khun miyad æzæsh
 like donkey has **blood coming** from-it
 'It's **bleeding** like crazy'

یک عالمه حرف داشتم در این مورد
 [yk ealmh **Hrf daštm** dr ayn mvrđ]
yek alæme hærf dashtæm dær in moređ
 a lot **word had.1sg** in this subject
 'I **had** a lot **to say** on this subject'

از چشمم انداختشون یه خورده
 [az čšmm **andaxtšvn** ye xvrđh]
æz cheshmæm ændakhteshun ye khurde
 from eye-my **dropped-them** a bit
 'I lost respect for them a bit.'

4.6.7 Other constructions

The conversational variant of the language includes many expressions and syntactic constructions that are very rarely found in literary text. In fact, traditionalists would resist writing down such constructions which are nevertheless used quite commonly in colloquial speech. One such example is reduplication where a word is repeated with its first letter modified, usually used for emphasis or intensification. In English 'pizza-schmizza' is one such example and in Persian conversational text they are relatively common as in *kotæk motæk* 'beating schmeating' or *patogh matogh* 'hangout schmangout'. Another form of reduplication is also often used with adverbs such as *paværchin paværchin* 'on tiptoe', where the word is simply repeated to intensify the effect.

Two other examples of conversational constructions are given below: the first one takes the colloquial expression *dæmet gærm* 'alright!; good for you!' (which literally means 'may your breath be warm') and separates the two parts by placing the whole object in the middle of the two subparts; the second sentence sounds very colloquial in its construction and uses expressions such as *in hærfa* which could be translated as 'this kind of stuff'.

هوراااا، دم همه دخترای ماهی که هر دفعه جلوی در ورزشگاه کتک خوردند ولی واندادند گرم!
 [hvraaaaa, **dm** hmh dxtray mahy kh hr dfeh jlvy dr vrzšgah ktk xvrđnd vly vandadnd **grm**]
*huraaaaa, **dame** hame dokhtæraye mahi ke hærfa dæfe jeloje dære værzeshgah kotæk khordænd væli vanædadænd **gærm!***
 huraaaaay **breath** all girls awesome-Rel that each time front-of gate stadium beating ate.3pl but up not-gave.3pl
warm

'huraaaaay, good for all those awesome girls who got beaten up in front of the stadium gate every time and didn't give up!'

بعد از خیمه شب بازی کلیسا و عروس داماد و بیوس و این حرفها، به منزل برادر داماد دعوت شدیم برای شام
 [bed az xymh šb bazy klysa v ervs damad v bbvs v ayn Hrfha, bh mnzl bradr damad devt šdym bray
 šam]

*bæd æz kheyme shæb bazie kelisa o ærus damad o bebus o in hærfæ, be mænzele bæradære damad
 dævæt shodim bæraye sham*

after puppet-show-of church and bride groom and kiss-kiss and this words, to house-of brother-of groom
 invitation became.1pl for dinner

'After the puppet show at the church and the bride and groom and the kisses and this kind
 of stuff we were invited to the groom's brother's house for dinner.'

4.7 Other Orthographic Issues

We have seen that blogs written in conversational language often do not follow a standard set of orthographic rules and may write morphemes attached or unattached as the author sees fit. Blogs contain ellipses, emoticons and hyperlinks, which require special document segmentation. In addition, spelling errors are much more common than one may encounter in non-blog websites. It should also be noted that even blogs written in literary text display a varying range of orthographic patterns when it comes to affixes in Persian, which differ significantly from the standard rules followed by newsprint media online. In fact, blogs maintained by journalists and intellectuals follow orthographic guidelines that differ from the traditional rules. For instance, while a number of bloggers always write the clitic copula in detached form as in موافقاند [mvafq~and] 'they agree' (pronounced *movafeghænd*), the traditional websites use the attached form موافقند [mvafqnd]. Bloggers writing in literary text also use new spellings to reflect the Persian pronunciation of Arabic loanwords such as حتى [Hta] 'even' (pronounced *hæta*) instead of the traditional حتى [Hty].

The literary blogs of journalists and intellectuals tend to not attach morphemes to words in text, even if this has traditionally been the method of writing in Persian and is recommended by the Persian Language Academy. The following provide examples of morphemes that were previously treated as attached affixes but are now often written as separate morphemes in literary blogs, but a comprehensive overview of these orthographic tendencies has not been attempted in this paper and is left for further study.

Plural morpheme. In traditional text, the *gan* morpheme replaces the 'silent *he*' character, but the [h] is kept in literary blogs.

Literary blogs: نمایندگان [nmayndh~gan] *næmayændegan* 'representatives'

Traditional text: نمایندگان [nmayndgan]

Comparison morpheme. Attached in traditional text but written detached in literary blogs.

Literary blogs: کمتر [km~tr] *kæmtær* 'less'

Traditional text: کمتر [kmtr]

Copula clitic. 3rd person clitic is attached to words ending in 'ye' in traditional text.

Literary blogs: کافیست [kafy~st] *kafist* 'it's enough'

Traditional text: کافیسست [kafyst]

Copula clitic. Attached to words ending in a consonant in traditional text.

Literary blogs: موافق ام [mvafq~am] *movafeghæm* 'I agree, I am in agreement'

Traditional text: موافقم [mvafqm]

Pronominal clitic. Attached to words ending in a consonant in traditional text.

Literary blogs: دندان‌شان [dndan~šan] *dændaneshan* 'their tooth'
 Traditional text: دندانشان [dndanšan]

Literary blogs: نگاه‌هایشان [ngah~hay~šan] *negahhayeshan* 'their glances'
 Traditional text: نگاه‌هایشان [ngah~hayšan]

4.8 Conclusion

This section provides a first detailed account of the language of Persian blogs that contain instances of both literary language and of the very distinct conversational Persian, sometimes mixed within the same blog post. The paper presents a description of the literary variant and then introduces language changes that have led to the conversational variant, with an emphasis on computational analysis. A comprehensive description of phonological changes, patterns in morphology, and orthographic variance is presented. In addition, several issues in the lexicon and the syntactic structures found in Persian-language weblogs are examined. The examples and patterns discussed show that there is a large amount of linguistic and orthographic variance found in Persian blogs. The diverse spellings also give rise to more ambiguity and provide new challenges for analyzing and processing online Persian documents.

References

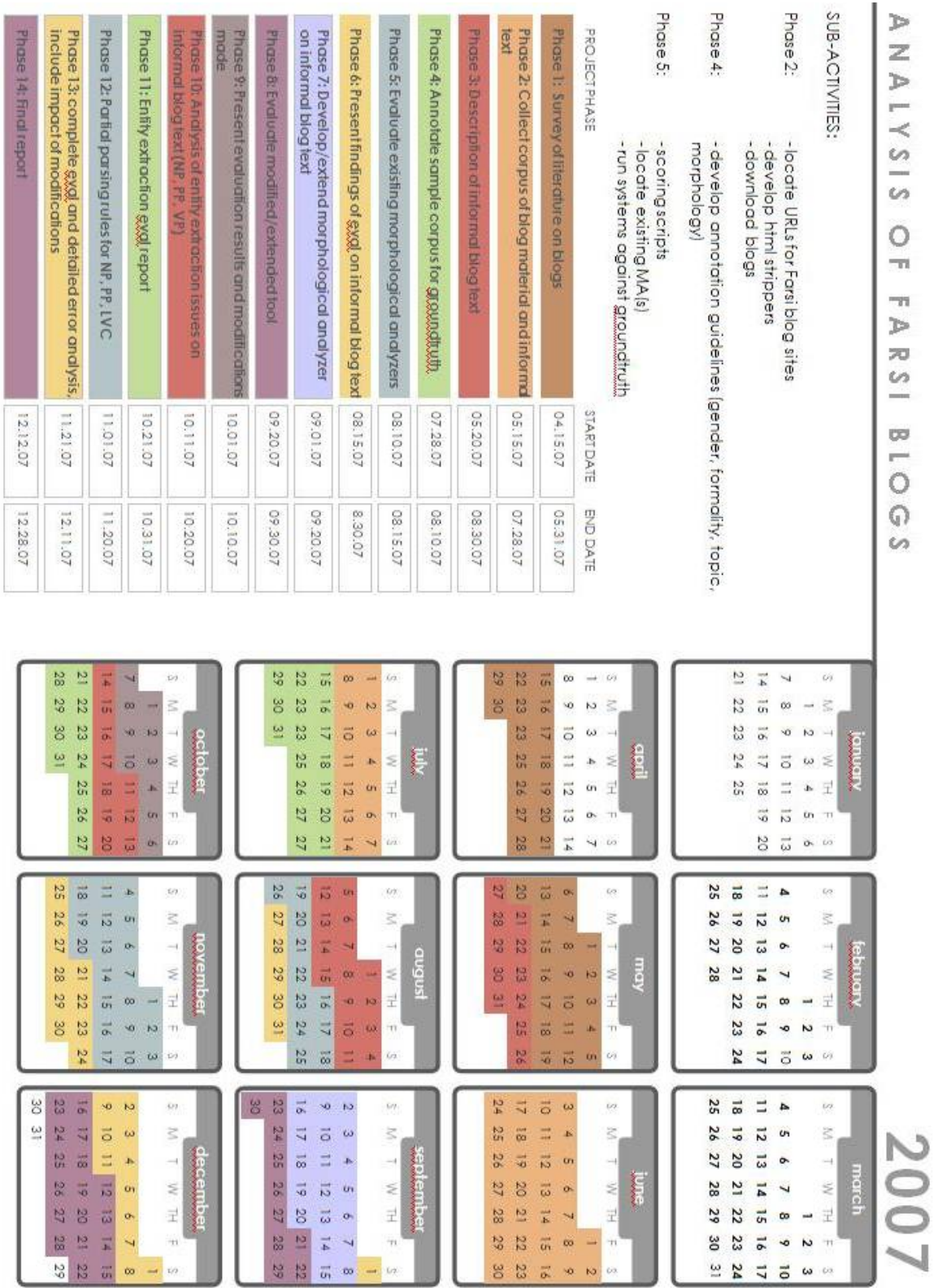
- Adamic, Lada A. (1999). The Small World Web. In *Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, ECDL.
- Adar, Eytan, Lada A. Adamic, and Rajan M. Lukose (2004). Implicit Structure and the Dynamics of Blogspace. In *Workshop on the Weblogging Ecosystem*, 13th International World Wide Web Conference.
- Alavi, Nasrin (2005). *We are Iran: the Persian Blogs*. Brooklyn, NY: Soft Skull.
- Alexanian, Janet A. (2006). Publicly Intimate Online: Iranian Web Logs in Southern California. In *Comparative Studies of South Asia, Africa and the Middle East*, vol. 26, no. 1.
- Amir-Ebrahimi, Masserat (2004). Performance in Everyday Life and the Rediscovery of the "Self" in Iranian Weblogs. In *Bad Jens: Iranian Feminist Newsletter*, 7th edition.
- Amtrup, Jan W., Hamid Mansouri Rad, Karine Megerdooian and Rémi Zajac (2000). Persian-English Machine Translation: An Overview of the Shiraz Project. Memoranda in Computer and Cognitive Science, MCCS-00-319. Computing Research Laboratory, New Mexico.
- Anderson, Jon W. (1999). The Internet and Islam's New Interpreters. In Eickelman, Dale F. and Jon W. Anderson (eds.), *New Media in the Muslim World: The Emerging Public Sphere*. Bloomington: Indiana University Press.
- Anderson, Jon W. (1997). Is the Internet Islam's 'Third Wave' or the 'End of Civilization'? Globalizing Politics and Religion in the Muslim World. In *Virtual Diplomacy*, Feb. 26. Available at <http://www.usip.org/virtualdiplomacy/publications/papers/polrelander.html>
- Beckerman, Gal (2006). The Man who Brought Blogging to Iran. In *CJRDaily*, Feb. 3. Interview available at http://www.cjrdaily.org/the_water_cooler/the_man_who_brought_blogging_t.php
- Beesley, Kenneth R. and Lauri Karttunen (2003). *Finite-State Morphology: Xerox Tools and Techniques*. CSLI Publications, Palo Alto.
- Behrouzan, Orkideh (2005). Persian Blogs against "The Dual Language". In *Anthropology News*, February.
- Behrouzan, Orkideh (2004). The Dual Language. In *Iranian.com*. Electronic document available at <http://www.iranian.com/Diaspora/2004/August/OB/>
- Bourdieu, Pierre (1984). *Distinction: A Social Critique of the Judgment of Taste*. Richard Nice (trans.). Cambridge, Mass: Harvard University Press.
- Brin, Sergey and Lawrence Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107-117.
- CIA World Factbook (2008a). Rank order - Internet users. Available at <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html>
- CIA World Factbook (2008b). Country profile: Iran. Available at <https://www.cia.gov/library/publications/the-world-factbook/geos/ir.html>
- Crystal, David (2001). *Language and the Internet*. Cambridge: Cambridge University Press.
- Derakhshan, Hossein (2001). Chetor yek veblâg-e fârsi besâzim [چطور یک وبلاگ فارسی بسازیم] 'How to create a Persian weblog'. Posted on weblog *Editor: myself* on November 5, 2001. Available at http://i.hoder.com/archives/2001/11/011105_007529.shtml
- Doostdar, Alireza (2004). "The Vulgar Spirit of Blogging": On Language, Culture, and Power in Persian Weblogestan. In *American Anthropologist*, vol. 106, issue 4, pp. 651-662.
- Esmaili, Mohamad (1998). "The *an* to *un* alternation in Persian". Ms., Georgetown University.
- Gao, Liwei (2007). *Chinese Internet Language: A Study of Identity Constructions*. Munich: Lincom GmbH.
- Gao, Liwei (2006). Language Contact and Convergence in Computer-Mediated Communication. In *World Englishes*, vol. 25, no. 2, pp. 299-308.
- Gibson, David, Jon Kleinberg, and Prabhakar Raghavan (1998). Inferring Web Communities from Link Topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia : Links, Objects, Time and Space*, p.225-234.

- Glance, Natalie, Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton, and Takashi Tomokiyo (2005). Deriving Marketing Intelligence from Online Discussion. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*.
- Golkar, Saeid (2005). Politics in Weblogs: A Safe Space for Protest. In *Iran Analysis Quarterly*, vol. 2, no. 3, winter 2005.
- Good, Robin (2005). Group And Multi-User Blog Platforms Compared. Electronic document, available at http://www.masternewmedia.org/news/2005/05/16/group_and_multiuser_blog_platforms.htm
- Gruhl, David, David Liben-Nowell, R. Guha, and Andrew Tomkins (2004). Information Diffusion through Blogspace. In *ACM SIGKDD Explorations Newsletter*, vol.6 no.2, pp.43-52.
- Hale, Constance and Scanlon, Jessie (1999). *Wired Style: Principles of English Usage in the Digital Age*. New York: Broadway Books.
- Halavi, Jordan (2006). The Iranian Weblog Research Project: Survey results – Preliminary Quantitative Data on the Readership of Select Iranian Weblogs. Electronic document, available at <http://www.persianimpediment.org/research/iwrresults.pdf>
- Herring, Susan C. and Paolillo, John C. (2006a). Gender and Genre Variation in Weblogs. In *Journal of Sociolinguistics*, vol. 10, no. 4, pp. 439-459.
- Herring, Susan C., Lois Ann Scheidt, Inna Kouper and Elijah Wright (2006b). Longitudinal Content Analysis of Weblogs: 2003-2004. In M. Tremayne (Ed.), *Blogging, Citizenship, and the Future of Media*. London: Routledge.
- Herring, Susan C., Lois Ann Scheidt, Sabrina Bonus and Elijah Wright (2005). Weblogs as a Bridging Genre. In *Information, Technology & People*, vol. 18, no. 2, pp. 142-171.
- Herring, Susan C., Lois Ann Scheidt, Sabrina Bonus and Elijah Wright (2004). Bridging the Gap: A Genre Analysis of Weblogs. In *Proceedings of the 37th Hawai'i International Conference on System Sciences (HICSS-37)*. Los Alamitos: IEEE Computer Society Press.
- Heylighen, Francis, and Jean-Marc Dewaele (2002). Variation in the Contextuality of Language: An Empirical Measure. In *Foundations of Science*, vol. 7, no. 3, pp. 293-340.
- Hinrichs, Lars (2006) *Codeswitching on the Web: English and Jamaican Creole in E-mail Communication*. Amsterdam: John Benjamins.
- Howard, Philip N, and World Information Access Project. *World Information Access Report - 2008*. 3. Seattle: University of Washington, 2008.
- Imran, Sabiha (2006). Romanized Persian in On-line Communications. Talk presented at the *BASIS Technology Government Users Conference*. Washington, D.C., June 14.
- Jamali, Mohsen. Mining Persian Blogs' Social Network. Master's Thesis, Sharif University, Tehran, Iran.
- Jensen, Peder Are Nøstvold (2004). Blogging Iran – A Case Study of Iranian English Language Weblogs. Master's Thesis, University of Oslo, Norway.
- Kelly, John and Bruce Etling (2008). Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere. Research Publication No. 2008-01, The Berkman Center for Internet and Society at Harvard Law School. April 6. Available at <http://cyber.law.harvard.edu/publications>
- Kleinberg, Jon (1999). Authoritative Sources in a Hyperlinked Environment. In *Journal of the ACM*, vol. 46, no. 5, pp. 604-632.
- Kritikopoulos, Apostolos, Martha Sideri, and Iraklis Varlamis (2006). BlogRank: Ranking Weblogs based on Connectivity and Similarity Features. In *Proceedings of the 2nd International ACM Workshop on Advanced Architectures and Algorithms for Internet Delivery and Application*.
- Kumar, Ravi, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins (2003). On the Bursty Evolution of Blogspace. In *Proceedings of the 12th International World Wide Web Conference*.
- Kurland, Oren and Lillian Lee (2005). PageRank without Hyperlinks: Structural Re-ranking using Links Induced by Language Models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Lazard, Gilbert (1992). *A Grammar of Contemporary Persian*. Mazda Publishers.

- Macdonald, Craig and Iadh Ounis (2006). The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow.
- Mahootian, Shahrzad (1997). *Persian*. Routledge
- Megerdoomian, Karine (2006). Extending a Persian Morphological Analyzer to Blogs. In *Proceedings of the Second Workshop on Persian Language and Computers*. Tehran University, Iran.
- Megerdoomian, Karine (2000). Persian Computational Morphology: A Unification-Based Approach. Memoranda in Computer and Cognitive Science, MCCS-00-320. Computing Research Laboratory, New Mexico.
- Milroy, Lesley (1987). *Language and Social Networks*. 2nd ed. Oxford: Basil Blackwell Ltd.
- Mishne, Gilad and Maarten de Rijke (2006). Capturing Global Mood Levels using Blog Posts. In AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006).
- Mishne, Gilad (2005). Experiments with Mood Classification in Blog Posts. In *Style2005 - 1st Workshop on Stylistic Analysis of Text for Information Access*, at SIGIR 2005.
- Mullen, Tony and Robert Malouf (2006). A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*.
- Nabavi, Ebrahim (2004). Shast hezâr sardebir [شصت هزار سردبیر] 'Six thousand editors'. Article published in *BBC Persian*. Electronic document available at http://www.bbc.co.uk/persian/iran/story/2004/11/041115_mj-en-iran-web-log-anniv.shtml
- Nakajima, Shinsuke, Junichi Tatemura, Yoichiro Hino, Yoshinori Hara, and Katsumi Tanaka (2005). Discovering Important Bloggers based on Analyzing Blog Threads. In *Proceedings of WWW2005 Workshop on the Weblogging Ecosystem*.
- Nilsson, Stephanie (2003a). A Brief Overview of the Linguistic Attributes of the Blogosphere. Ms. Umeå Universitet. Electronic document available at <http://www.eng.umu.se/stephanie/web/blogspeak.pdf>
- Nilsson, Stephanie (2003b). The Function of Language to Facilitate and Maintain Social Networks in Research Weblogs. Ms. Umeå Universitet. Electronic document available at <http://www.eng.umu.se/stephanie/web/LanguageBlogs.pdf>
- NITLE Census (2007). *Blog Census: Languages*. Published by the National Institute for Technology and Liberal Education (NITLE). Available at <http://www.hirank.com/semantic-indexing-project/census/lang.html>
- NITLE Census (2003). *Equal Numbers, Different Interests*. Published by the National Institute for Technology and Liberal Education (NITLE). Available at <http://www.hirank.com/semantic-indexing-project/census/weblog/index.html>
- Nowson, Scott and Oberlander, Jon (2006). Differentiating Document Type and Author Personality from Linguistic Features. In *Proceedings of the Eleventh Australasian Document Computing Symposium (ADCS 2006)*.
- Nowson, Scott (2006). *The Language of Weblogs: A Study of Genre and Individual Differences*. Unpublished Doctoral Thesis, University of Edinburgh.
- Nowson, Scott, Oberlander, Jon and Gill, Alistair J. (2005). Weblogs, Genres and Individual Differences. In *the Proceedings of the 27th Annual Conference of the Cognitive Science Society*.
- Nunberg, Geoffrey (2004). Blogging in the Global Lunchroom. Commentary Broadcast on "Fresh Air" on April 20, 2004. Transcript available at <http://www.ischool.berkeley.edu/~nunberg/lunchroom.html>
- Ó Baoill, Andrew (2004). Conceptualizing the Weblog: Understanding What it is in order to Imagine What it Can Be. In *Interfacing: A Journal of Contemporary Media Studies*.
- OpenNet (2006). Internet Filtering in Iran in 2004-2005: A Country Study, by OpenNet Initiative. Available at <http://www.opennetinitiative.net/studies/iran/>
- Ounis, Iadh, Maarten de Rijke, Craig Macdonald, Gilad Mishne and Ian Soboroff (2006). Overview of the TREC-2006 Blog Track. In the 15th Text Retrieval Conference Proceedings. Available at http://trec.nist.gov/pubs/trec15/t15_proceedings.html

- Qazvinian, Vahed, Abtin Rassolian, and Mohammad Shafiei (2007). A Large-Scale Study on Persian Weblogs. In the Proceedings of Workshop on Text-Mining and Link-Analysis (TextLink 2007), The Twentieth International Joint Conference on Artificial Intelligence.
- Qazvinian, Vahed, Abtin Rassolian, and Jafar Adibi (2007). Observations on Failure in Blogs. In the *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'07)*.
- RSF (2004). Reporters Without Borders report, Internet under Surveillance: Obstacles to the Free Flow of Information Online. Available at http://www.rsf.org/article.php3?id_article=10733.
- Sheykh Esmaili, Kyumars, Mohsen Jamali, Mahmood Neshati, Hassan Abolhassani, and Yasaman Soltan-Zadeh (2006). Experiments on Persian Weblogs. In *WWW2006 Workshop on Weblogging Ecosystem*.
- Sheykh Esmaili, Kyumars, Mohsen Jamali, and Mahmood Neshati (2005). Kashfe ejtemââte mowjud dar veblâghâye fârsi [کشف اجتماعات موجود در وبلاگ های فارسی] 'Exploring Social Networks in Persian Weblogs'. Technical Report, Sharif University, Tehran, Iran.
- Schler, Jonathan, Moshe Koppel, Shlomo Argamon and James Pennebaker (2006). Effects of Age and Gender on Blogging. In *the Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.
- Shokrollahi, Seyyed Reza (2006). Kâsh ferdowsi zende bud! [کاش فردوسی زنده بود!] 'If Only Ferdowsi were Alive!'. Posted on the weblog *khâbgard* on April 30. Available at <http://www.khabgard.com/?id=1146350454>
- Shokrollahi, Seyyed Reza (2003). Be ebtezâl zende-bâd naguyim! [به ابتذال زنده باد نگوییم!] 'Let's not Wish a Long Life to Vulgarity!'. Posted on the weblog *khâbgard* on November 1. Available at <http://www.khabgard.com/?id=-2118994923>
- Sifry, Dave (2007). The Technorati State of the Live Web: April 2007. Available at <http://technorati.com/weblog/2007/04/328.html>
- Sifry, Dave (2006). The Technorati State of the Blogosphere: October 2006. Available at <http://technorati.com/weblog/2006/11/161.html>
- Tavosanis, Mirko (2006). Linguistic Features of Italian Blogs: Literary Language. In *the Proceedings of the Workshop on NEW TEXT: Wikis and Blogs and Other Dynamic Text Sources*, 11th conference of the European Chapter of the Association for Computational Linguistics (EACL-2006). Trento, Italy. Available at <http://www.aclweb.org>
- Tehrani, Hamid (2007). Iranian Muslim Bloggers. In *History News Network*. Posted on November 26. Available at <http://hnn.us/articles/44774.html>
- Thackston, Wheeler M. (1993). *An Introduction to Persian*. Bethesda: IBEX Publishers.
- Tsujimura, Natsuko (2007). Language Change in Progress: Evidence from Computer-Mediated Communication. Paper presented at the 33rd annual meeting of the Berkeley Linguistic Society.
- Walker, Jill (2003). Final Version of Weblog Definition. Posted on weblog *jill/txt* on June 28. Available at http://huminf.uib.no/%7Ejill/archives/blog_theorising/final_version_of_weblog_definition.html
- Warschauer, Mark (2001). Language, Identity, and the Internet. In *Mots Pluriels*, no. 19, October.
- Wikipedia: Blog. <http://en.wikipedia.org/wiki/Blog>
- Yasuda, Norihito, Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki (2006). Identifying Bloggers' Residential Areas. In *AAAI Spring Symposia 2006 on Computational Approaches to Analyzing Weblogs*.

Appendix A Project Schedule



Appendix B Table of Persian Letters and Transliteration

(See notes on next page)

Persian Script	Persian Letter Name	Report Transliteration	Pronunciation	IPA Transcription	Report Transcription
آ	alef ba kolah	A	father	[v]	<i>a</i>
ا	alef	a	cat, bed, boy	[æ][e] [o]	<i>æ, e, o</i>
ب	be	b	bin	[b]	<i>b</i>
پ	pe	p	pin	[p]	<i>p</i>
ت	te	t	ten	[t]	<i>t</i>
ث	se	c	sip	[s]	<i>s</i>
ج	jim	j	gin	[dʒ]	<i>j</i>
چ	che	č	chin	[tʃ]	<i>ch</i>
ح	he	H	hen	[h]	<i>h</i>
خ	khe	x	German bach	[x]	<i>kh</i>
د	dal	d	den	[d]	<i>d</i>
ذ	zal	D	zip	[z]	<i>z</i>
ر	re	r	Spanish arte	[r]	<i>r</i>
ز	ze	z	zip	[z]	<i>z</i>
ژ	zhe	ž	pleasure	[ʒ]	<i>zh</i>
س	sin	s	sip	[s]	<i>s</i>
ش	shin	š	shin	[ʃ]	<i>sh</i>
ص	sat	S	sip	[s]	<i>s</i>
ض	zat	Z	zip	[z]	<i>z</i>
ط	ta	T	ten	[t]	<i>t</i>
ظ	za	z	zip	[z]	<i>z</i>
ع	eyn	e	silent or American English button	[ʔ]	[see notes]
غ	gheyn	Q	French merci	[v]	<i>gh</i>
			--	[G]	
ف	fe	f	fan	[f]	<i>f</i>
ق	ghaf	q	French merci	[v]	<i>gh</i>
			--	[G]	
ک	kaf	k	con	[k]	<i>k</i>
گ	gaf	g	gun	[g]	<i>g</i>
ل	lam	l	land	[l]	<i>l</i>
م	mim	m	man	[m]	<i>m</i>
ن	nun	n	nun	[n]	<i>n</i>
و	vav	v	van, too, boy	[v] [u] [o]	<i>v, u, o</i>
ه	he	h	hen	[h]	<i>h</i>
ی	ye	y	you, jeep	[j] [i]	<i>y</i>
ء	hamze	'	silent or American English button	[ʔ]	[see notes]

 NOTES

Transliteration refers to a one-to-one mapping of Persian characters to a romanized form. The letters correspond directly to the written form of the word.

Transcription provides the pronunciation of the Persian for an English speaker.

Eyn and *hamze*: Originally, these characters represented the "glottal stop" – the sound made when the vocal cords are drawn together interrupting the flow of air from the lungs and then released as pressure builds up below them.⁶³ In English, the glottal stop is found in the break separating the syllables in *uh-oh* or in the pronunciation of American English *button* or *glottal*. In modern conversational Persian, however, the glottal stop is often dropped giving rise to a lengthening of the vowel preceding it (note that the vowel may be an unwritten vowel such as /æ/, /e/, or /o/). Therefore, *eyn* and *hamze* can be pronounced as either (i) the glottal stop (represented by the apostrophe) as in *bæ'd* 'then' and *mæ'mur* 'official, responsible' or (ii) silent but with a lengthening of the preceding vowel as in *bæ:d* 'then' and *mæ:mur* 'official, responsible'.

Gheyn and *ghaf*: In the past, *gheyn* (غ) and *ghaf* (ق) represented different sounds, denoted by [ɣ] and [ɢ], respectively. In modern standard Persian, there is no difference in the pronunciation of غ and ق (both of them representing [ɣ] or [ɢ], depending on their position in the word). The classic pronunciation difference (for غ and ق) is preserved in Afghani Persian (Dari) and Tajiki Persian as well as in the Southern Iranian dialects of Persian such as Yazdi and Kermani. Note that [ɣ] is pronounced like the French 'r' sound whereas [ɢ] (voiced uvular plosive consonant) is a rare sound articulated with the back of the tongue against or near the uvula (in the back of the throat).

The diphthong /ow/ is used in this report to represent the pronunciation of words like روحانی [rvHany] pronounced *rowhani*. The /ow/ sound is similar to the pronunciation of English 'door'.

⁶³ http://en.wikipedia.org/wiki/Glottal_stop

Appendix C Persian Parts of Speech for Morphological Annotation Tasks

Adjective

Modifies the noun; describes it. It can appear in comparative (*bozorg-tar*) or superlative (*bozorg-tarin*) constructions.

Note: when tagging adjectives, disregard affixes like pronouns, "ezafe", etc.

POS	Description	Examples
Adj	Singular adjective	زیبا – شرم آور
AdjPl	Plural adjective	[see note below]
AdjComp	Adjective, singular, comparative	زیباتر
AdjSup	Adjective, singular, superlative	زیباترین
AdjPlComp	Adjective, plural, comparative	خنگترها
AdjPlSup	Adjective, plural, superlative	انقلابیترینها
AdjCop	Singular adjective with copula ⁶⁴	زیباست – خنگه
AdjCC	Singular adjective, comparative, copula	زیباتره
AdjSC	Singular adjective, superlative, copula	زیباترینه
AdjPlCop	Plural adj. with copula	«غیر اخلاقی» هایشانند
AdjPlCC	Plural adj., comparative, copula	حساستران / حساستر هایند
AdjPlSC	Plural adj., superlative, copula	حساسترینیم

Note: AdjPl and NnPl may be hard to distinguish sometimes. In these cases, the context of the sentence and the meaning may be more helpful. For instance, آزادی خواهان could be either one depending on the context and in most dictionaries, these ambiguous noun/adj elements are listed as both.

Adverb (Adv)

Modifies the verb (*tond midævim*, *tænha zendegi mikonæn*) or the adjective (*xeyly xub avaz mixune*). Gives information about the way something is done. Adverbs can indicate time or place (ex. *emruz*, *færda*, *inja*, *birun*⁶⁵).

⁶⁴ copula = the verb 'to be' that attaches to words as in

خرَم، خری، خره / خرست، خریم، خرین / خرید، خرن / خرنند

⁶⁵ Note that some words can be used as either adverb or preposition, such as *birun* or *næzdik*, but they will be used differently in a sentence:

Some common adverbial endings are:

- *æn*:

تقریباً

- *ane*:

عاقلانه، خوشبختانه

Classifier (Cl)

A classifier is used in number constructions to indicate type or measure.

تا، نفر، فروند، دست، دانه -- کیلو، لحظه، نوع

These will always be preceded by a number and usually followed by a noun:

چهار دست لباس – دو کیلو سیب زمینی

Conjunction (Conj)

Here, we include both conjunctions and coordinators.

Coordinators are used to connect two sentences, verbs, nouns, etc.: و، یا

Conjunctions connect two sentences or often start a sentence; in Persian they often contain **که**:
که، هر چند که، برای اینکه، وقتیکه، باوجود اینکه، چون(که)

Note that this **که** is used to connect sentences as in the following examples and it often follows a verb:

امیدوارم بعد از این ماجراهای مسخره بنویسه **که** مالیخولیا بهترین توصیف کننده وضعیت این روزهاست!
آقای مک کورمک گفت **که** ارتباط با ایران از طریق کانال رایان کروکر کانال خواهد بود
بهتره **که** اون هم بسته نشه

Determiner (Det)

Used before a noun to "demonstrate" the entity. Things like 'this', 'that'.

این، آن، همین، همان، چنین، چنان، هر

Foreign Word (FW)

English or other foreign words mixed up in a sentence. These are only words that are written in another script (e.g., 'http' below), not those that are transliterated into Persian.

وبلاگ های وردپرس علاوه بر http، پروتکل https رو هم دارن که حداقل هنوز فیلتر نمی شه کردش

birun-e xane → preposition
næzdik-e mædrese → preposition

ræftim birun → adverb
xeyli næzdik vaysadi → adverb

Note that the prepositions take an "ezafe" and are linked to the following noun; adverbs stand alone.

Noun

Lexical elements that can take a plural (ex. *ketab-ha*), can be used with numbers (*yek ketab*), can be used with a determiner (*an ketab*).

Note: when tagging nouns, disregard affixes like pronouns, "ezafe", etc.

POS	Description	Examples
Nn	Singular noun	آموزش
NnPl	Plural noun	موجودات - مطالب - شهرستان های
NnCop	Noun, singular with copula verb	دانشجوست - کتابه
NnPlCop	Noun, plural with copula verb	بینندگانمانند - کتابامن

Numbers (Num)

Any type of number (e.g., بیست و دو, پنج, ششم, هفتمین, بیست و دو) or digits. This includes all types of numeric expressions such as date or time.

Preposition (Prep)

Placed before a noun to indicate direction, time, space, etc.

Except for a few of them (چون⁶⁶, جز, تا, از, بر, در, با, تا, جز, چون⁶⁶), prepositions take an 'ezafe' in Persian to link to the following nominal element.

بعد از، بدون، علیه، به گزارش، بخاطر

Postposition (Post)

Refers to the object marker را.

Pronoun (Pron)

Replaces the noun, so it can be in a place in the sentential structure that a noun could normally occupy. Pronouns include:

من، تو، آنها، این، خودش، همدیگر، یکدیگر

Proper Noun (Prop)

A noun that indicates a person, place, organization, etc. The proper noun is capitalized in English.

حمید، تهران، دریای مازندران، سازمان ملل، گنکورد، واشنگتن پست

⁶⁶ چون in the meaning of 'like'.

Quantifier (Qua)

Represents a quantity and usually appears before a noun. Things like *all, many, most* are quantifiers. The quantifier can be between the determiner and the noun in Persian as in

این همه آدم ...

Or it can appear without the determiner in which case it may take the *ezafe* as in the two first examples below:

همه آدمها، اغلب مردم، خیلی از حیوانات

Question word (Que)

Words that are used to ask questions. Things like *what, where, when, how*.

آیا، کی، کجا، چقدر، چقدر، چندتا، کدام

Relativizer (Rel)

Used to relate a noun to a phrase that describes it. In Persian the main relativizer is *که* and it is used to form relative clauses.

ترجیح دادم نظرات اون هایی رو که به نظرم نزدیک تر هستن فقط لینک کنم

کلاس هایی که اینجا گذروندم ...

باشیم به دوستان هموفوب و/یا روشنفکری که آزادی رو فقط در چهارچوب عرفی مورد قبول خودشون می پذیرن ...

Sometimes relativizers can appear in the beginning of a phrase such as:

آنچه، هرآنچه، هرکجا، هرکه

Title (Title)

Used to introduce a person; things like *mister, mrs., Seyyed, Haji*, etc.

Interjections and Abbreviations and Acronyms (UH)

Interjections refer to expressions that indicate surprise, anger, etc. (اووووه! آخ! وای، ای ول) or forms used to call or address someone (آهای).

Abbreviations: ق.

Acronyms: Usually proper names that represent the English pronunciation such as BBC (بی بی سی).

Verb

Lexical elements that expresses an action, an event or a state. It can be conjugated.

Note: when tagging verbs, disregard affixes like negation (ننوشتم) or pronouns (نوشتمش).

POS	Description	Examples
Vinf	Infinitive	نوشتن
Vpast	Past tense (conjugated)	نوشتیم، نوشته ایم، نوشته بودند
Vpart	Past participle	نوشته، نوشته شده
Vfut	Future tense (conjugated)	خواهم نوشت
Vpres	Present tense (conjugated) ⁶⁷	می نویسم، بنویسند
Vimpv	Imperative (conjugated)	بنویس!

Compound verbs

For compound verbs, tag each part separately:

- صادر + کرده است → Nn + Vpast
 قرار + می گیرد → Nn + Vpres
 زیر + سؤال + برد → Prep + Nn + Vpast
 آزاد + شدند → Adj + Vpast

Modals

Tag modals as regular verbs and tag the second verb as a past verb (even though it's really only the past stem)

- نمی شه دیدش → Vpres + Vpast

⁶⁷ The present tense includes the subjunctive tense: *ægær benevisænd*

Appendix D Additional References

D.1 Blog sites used in this document

The findings in this technical report have benefited from the following websites:

Persian Online, University of Texas at Austin - http://dev.laits.utexas.edu/persian_online/
 Persian Phonology, Wikipedia - http://en.wikipedia.org/wiki/Persian_phonology
 Voice (Phonetics), Wikipedia - http://en.wikipedia.org/wiki/Voiced_consonant
 Sounds Iranian, a collective blog for Iran-related research with a focus on Iranian blogs - <http://soundsiranian.wordpress.com/>

In addition, the corpus data used for the examples and for the researched performed came mainly from the following sources:

Google Search - <http://www.google.com/> (whereby most examples noted in the report were found)
 Zeitoun - <http://z8un.blogfa.com/>
 Khorshid Khanoom - <http://khorshidkhanoom.com/>
 Man yek zanam - <http://www.hastii.blogspot.com/>
 Khabgard - <http://www.khabgard.com/>

D.2 A sampling of recent and upcoming conferences on weblogs

Euroblog2007: Social Software – A Revolution for Communication? Implications and Challenges for Public Relations, Journalism and Marketing

Organized by EUPRERA, the European Public Relations Research and Education Association. Ghent, Belgium, March 16-17th, 2007. <http://www.euroblog2007.org/>

Towards Genre-Enabled Search Engines: The Impact of NLP

Workshop held in conjunction with RANLP-2007. Borovets, Bulgaria. 30 September, 2007. <http://www.sics.se/use/genre-ws/>

Multi-source Multilingual Information Extraction and Summarization

Workshop held in conjunction with RANLP-2007. Borovets, Bulgaria. 26 September, 2007. <http://www-lipn.univ-paris13.fr/~poibeau/mmies.html>

International Conference on Weblogs and Social Media

March 26-28, 2007. Boulder, Colorado, U.S.A. <http://www.icwsm.org/>. The conference aims to bring together researchers from different subject areas (e.g., computer science, linguistics, psychology, statistics, sociology, multimedia and semantic web technologies) and foster discussions about ongoing research in various areas of NLP on social media.

Text Retrieval Conference (TREC): Blog Track

February-November 2007. Conducted by National Institute of Standards and Technology (NIST). <http://trec.nist.gov>. The purpose of the blog track is to explore information seeking behavior in the blogosphere.

Computational Approaches to Analyzing Weblogs

A 2006 AAAI Spring Symposium. *Proceedings*: Technical Report SS-06-03, published by The AAAI Press, Menlo Park, California, available at <http://www.aaai.org/Library/Symposia/Spring/ss06-03.php>

3rd Annual workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics

May 23, 2006, in Edinburgh, Scotland. <http://www.blogpulse.com/www2006-workshop/index.html>.
The workshop brought together researchers from diverse areas, working in both academic and commercial settings to discuss the increasing technological, social, political and cultural impact of weblogs.

ClickZ Weblog Business Strategies 2003 Conference & Expo

June 9-10, 2003 at Boston, Mass. <http://searchenginestrategies.com/blog/spring03/index.html>
The conference discusses the evolution of blogs from a mere "log" of favorite URLs from the late '80s and '90s to a platform that the business world is taking seriously, presenting trends and analyses, expert opinions, case studies, and "how to" sessions that will help medium to large enterprises add Weblogs into their business strategies.