Analysis of Farsi Weblogs

A Survey of the Literature

Karine Megerdoomian

August 2008
Analysis of Farsi Weblogs

Karine Megerdoomian
August 2008
# Table of Contents

1   Survey of the Literature 1  
   1.1   Summary 1  
   1.2   Introduction 2  
   1.3   Weblogestan: The Persian Blogosphere 4  
      1.3.1   Characterizing Weblogs 4  
      1.3.2   State of Persian Language Weblogs 5  
   1.4   Blogs, Politics and Society in Iran 7  
      1.4.1   Blogs and Censorship 8  
      1.4.2   The Evolution of a Political Consciousness 9  
      1.4.3   A New Public Space 9  
      1.4.4   Islamist Bloggers 11  
   1.5   Quantitative Research on Iranian Blogs 11  
      1.5.1   Demographic Profile 11  
      1.5.2   Content Analysis 12  
      1.5.3   Social Network Analysis 14  
   1.6   Computational Linguistic Analysis of Persian Blogs 18  
      1.6.1   Linguistic Features 19  
      1.6.2   The Vulgarity Debate 20  
      1.6.3   Computational Systems 22  
   1.7   Future Directions 23  
      1.7.1   Blogspeak: An Emerging Linguistic Genre 24  
         1.7.1.1   Linguistic Features of Netspeak 25  
         1.7.1.2   Social Networks and the Language of Blogs 26  
      1.7.2   Blogs and Individual Differences 27  
         1.7.2.1   Determining Personality Traits 27  
      1.7.3   Age, Gender and Language 28  
      1.7.4   The Intermediary Factor: Blog Topic 29  
      1.7.5   Applications for Persian Blogs 30
List of Figures

Figure 1 – Map of weblogs that include GeoURL data. Source: NITLE Blog Census (map courtesy of Paul Hastings) 3

Figure 2 – The Global Blogosphere: Posts by Language. Source: The Technorati State of the Live Web, April 2007 6

Figure 3 – The Global Blogosphere: Hourly Posts by Language. Source: The Technorati State of the Live Web, April 2007. 6

Figure 4 – Content analysis of Iranian weblogs. Source: Kelly and Etling (2008) 13

Figure 5 – Gender organized by subgroups in Iranian weblogs. Source: Kelly and Etling (2008) 14

Figure 6 – Number of comments on each day of the week. Source: Qazvinian et al (2007a) 15

Figure 7 – Time graph shows number of comments on each day with tagged outliers. Source: Qazvinian et al (2007a) 16

Figure 8 – Venn diagram for different links distribution. Source: Qazvinian et al (2007b) 16

Figure 9 – Social network map of the Iranian blogosphere. Source: Kelly and Etling (2008) 17

Figure 10 – Split of English-language Blogs based on Gender. 29

Figure 11 – Correlation of Gender and Blog Subjects in English-language Blogs. 30
List of Tables

Table 1 – Most discriminating word n-grams for detecting some moods; Source: Mishne (2005) ..............................................................24

Table 2 – Spoken language criteria applied to Blogspeak (adapted from Crystal 2001); Source: Nilsson (2003) ..............................................................26

Table 3 – Written language criteria applied to Blogspeak (adapted from Crystal 2001); Source: Nilsson (2003) ..............................................................27
1 Survey of the Literature

1.1 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 Blogspeak. 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 Blogspeak, especially based on technical or computer-related terms. In addition to all these features, Persian Blogspeak 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 Blogspeak 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 Blogspeak, 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 Blogspeak 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.

1.2 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.
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 charges. 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-thirds of the population is under 30 (the median age is 26) and many are 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

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."
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.

1.3 Weblogestan: The Persian Blogosphere

1.3.1 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.

### 1.3.2 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 Online and The Guardian UK 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 2 (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

---

Figure 2 – The Global Blogosphere: 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).

1.4 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 online 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 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

---

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.
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.

1.4.1 Blogs and Censorship

**Blogs as the new medium.** Nabavi (2004) 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, 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,

---

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.
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.

### 1.4.2 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 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.

### 1.4.3 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, 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.

---

\(^{11}\) 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.
Urban sociologist and geographer Masserat Amir-Ebrahimi’s (2004) article studies the socio-cultural ramifications of blogs in Iran through an investigation of Persian language weblogs, focus groups and interviews with bloggers. 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 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 1.6 for a discussion). In his view, the Internet is a continuum of existing forms: what 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

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)

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 TGC) 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.
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.

### 1.4.4 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 are leading figures in the regime.”

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.

### 1.5 Quantitative Research on Iranian Blogs

#### 1.5.1 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.

1.5.2 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, 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 4). In addition, personal diaries were found most often among the secular/expatriate bloggers and poetry sites.

![Figure 4](image-url)  
*Figure 4 – 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 5). 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, unpolticized 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.
1.5.3 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 dataset\textsuperscript{14}, 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.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{comments_weekday.png}
\caption{Number of comments on each day of the week. \textit{Source: Qazvinian et al (2007a)}}
\end{figure}

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 comments\textsuperscript{15}. 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 6). 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 7).

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.

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

\textsuperscript{15} The dataset is available at http://www.blogscience.org/data.html for research purposes.
Qazvinian et al (2007b) argue for the importance of taking into account comments in SNA. As Figure 8 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.

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 9, 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.

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)

1.6 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.

### 1.6.1 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.

Megerdoomian (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 *gereftæm* vs. the formal *maa gerefte æst* مرا گرفته‌ام vs. ‘has caught me’
- Has shortened verbal stems (especially in the case of the present stem) and inflectional endings.
  ex. informal *migen* می‌گویند vs. formal *miguyænd* ‘they say’
• Includes morphemes that do not exist in the formal language such as the definite article, ex. forushendehe '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 nezeratun نظراتون ‘your views’ vs. formal nezerhayetan 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 mvaqf-œænd16 موافقاند ‘they agree’, the traditional websites use the attached form mvaqfand موافقند. Bloggers writing in formal text also use new spellings such as heta ‘even’ written with an "alef" (ح) instead of the original Arabic writing with an ‘alef maqsurah’ (ح) 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.

1.6.2 The Vulgarity Debate

The so-called behs-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:

  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

16 The tilde (~) in this example represents the shortspace or zero-width nonjoiner (Unicode \u200c) which is used to force a final (i.e., unattached) form on the previous character.
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.
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 Doostdar discusses 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 *roshenfekr* (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.

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

---

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 grandparents 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)

---

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

1.6.3 Computational Systems

Within the frame of the vulgarity debate, much energy has been spent pushing for orthographic standards for online publishing\(^1\), 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 Akbarpour\(^2\), 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 (ُه) and the superlative terin (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 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 affixes\(^3\),

\(^1\) Interestingly, the orthographic standards proposed by intellectual bloggers such as Shokrollahi (http://www.khabgard.com/?id=965590463) 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).


\(^3\) See Section Error! Reference source not found. for a more detailed description of these features of informal text.
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); *haemdighe* (همدیگه) ‘each other’ vs. *haemdigar* (همدیگر) ‘for me’ vs. *berayem* (برایم)

- Colloquial counterparts to words:
  - colloquial *vase* (واسه) ‘for’ vs. formal *beraye* (برایه); *tu* (تُر) ‘in’ vs. *deer* (در)

- Loan words appearing in Romanized form or in the original language:
  - *anlayn* (آنلاین) ‘online’; *chatreum* (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* (کامنتگذار) ‘commenter’; *lagiden* (لادیدن) ‘to blog’

- Interjections:
  - *aaakh!* (آآآخ!) ‘ouch’; *ouuuuh!* (اوووه!) ‘ouch’; *vay* (وَی) ‘well’; *vala* (والا) ‘well’

### 1.7 Future Directions

Recent years have seen a number of workshops and conferences dedicated solely to the study of weblogs. 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 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). 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

---

22 See Error! Reference source not found. for examples of recent and upcoming computational conferences, symposia and workshops related to blogs.

23 The BLOG06 collection is available for research purposes. The dataset and relevant statistics can be found at [http://ir.dcs.gla.ac.uk/test_collections/blog06info.html](http://ir.dcs.gla.ac.uk/test_collections/blog06info.html).
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)

<table>
<thead>
<tr>
<th>Mood</th>
<th>Top words</th>
<th>Top bigrams</th>
<th>Top trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>hungry</td>
<td>hungry, eat, bread, sauce</td>
<td>am hungry, hungry and some food to eat</td>
<td>I am hungry, am finally happened, I am starving ask my mother</td>
</tr>
<tr>
<td>frustrated</td>
<td>n’t, frustrated, frustrating, do</td>
<td>am done, can not problem is to fix</td>
<td>I am done, am tired of stab stab stab, I do not</td>
</tr>
<tr>
<td>love</td>
<td>love, me, valentine, her</td>
<td>I love, love you, love is valentines day</td>
<td>I love you, my god oh, I love him, love you so</td>
</tr>
</tbody>
</table>

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, 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.

#### 1.7.1 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.
1.7.1.1 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 (aaaaahhh!!!, 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 imbibes 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.”

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

### 1.7.1.2 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)**

---

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!
### Spoken Language Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>The Web (in general)</th>
<th>Blogs (in general)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-bound</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Face-to-face</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Loosely structured</td>
<td>Variable</td>
<td>Variable</td>
</tr>
<tr>
<td>Socially interactive</td>
<td>No, with increasing options</td>
<td>Variable</td>
</tr>
<tr>
<td>Immediately revisable</td>
<td>Variable; depends on tools</td>
<td>Yes</td>
</tr>
<tr>
<td>Prosodically rich</td>
<td>Variable</td>
<td>Variable, but limited</td>
</tr>
</tbody>
</table>

### Written Language Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>The Web (in general)</th>
<th>Blogs (in general)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space-bound</td>
<td>Yes, with extra options</td>
<td>Yes</td>
</tr>
<tr>
<td>Contrived</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Visually decontextualized</td>
<td>Yes, but with considerable adaptation</td>
<td>Variable</td>
</tr>
<tr>
<td>Elaborately structured</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Factually communicative</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Repeatedly revisable</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Graphically rich</td>
<td>Yes, but in different ways</td>
<td>Variable, but limited</td>
</tr>
</tbody>
</table>

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.

#### 1.7.2 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.

##### 1.7.2.1 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

   \[ F = 0.5 \times [(\text{nounfrq} + \text{adjfrq} + \text{prepfrq} + \text{artfrq}) - (\text{pronfrq} + \text{verbfrq} + \text{advfrq} + \text{intfrq}) + 100] \]

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.”

### 1.7.3 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.”

1.7.4 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 10 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%).

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 11).
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.

1.7.5 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.

---

Figure 11 – 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.

1.7.5 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.