FarsiTEX and the Iranian TEX Community

Behdad Esfahbod
Computing Center
Sharif University of Technology
Azadi Avenue
Tehran, Iran
farsitex@behdad.org
http://behdad.org/

Roozbeh Pournader
Computing Center
Sharif University of Technology
Azadi Avenue
Tehran, Iran
roozbeh@sharif.edu
http://sina.sharif.edu/~roozbeh/

Abstract

FarsiTEX, a localized version of LATEX, is a bilingual Persian/English typesetting package, meeting the minimum requirements of Persian mathematical and technical typography. This paper will describe FarsiTEX, together with its history, future and technicalities, its user community, and the reasons behind its success in Iran, amid its various usage and interoperability problems. It will also draw a general picture of the TEX community in Iran, and tries to describe why the community is still far from achieving its basic typographical needs.

Introduction

The Persian language, in its contemporary form, is a language spoken natively in Iran, Afghanistan, and Tajikistan. The local forms are known as Farsi, Dari, and Tajiki respectively. They all use the same basic vocabulary and grammar, but there are differences in both pronunciations and modern vocabulary. In this paper, we will focus on the form used in Iran, which is the official language of the country.

The modern Persian script, as written in Iran, is a right-to-left script with contextually changing shapes of letters, and a derivative of the Arabic script extended by addition of some letters (Peh, Tcheh, Jeh, and Gaf), and modification of a few others (Kaf and Yeh). The script, with roots in the Arab invasion of Persia in the 7th century and later becoming known as the Perso-Arabic script, had then propagated to the areas currently known as Afghanistan, Pakistan, India, Western China, and then even South East Asia and Java, where many languages are written in it with further extensions to the alphabet, including Urdu, the official language of Pakistan. The Unicode Standard, in its latest version 3.2 (Unicode Editorial Committee, 2002), lists a total of 139 letters in the script, which are derivatives of about 28 basic Arabic letters.

The Persian typography, influenced by major calligraphic practices of the pre-printing era, is actually based on the famous Naskh style, which more than 99% of contemporary texts published in it. The alternate style, Nastaliq, a little harder to read but considered very beautiful by the general public, and widely known as the hardest commonly used script style in the world to implement in computers, has had a recent popularity after its many computer implementations appearing in the 1990s. But after a few years, because of readability problems, the usage of Nastaliq has been trimmed down to mainly school books on Persian literature.

Persian scientific typography, blossoming in the 1950s by publications of Gholamhossein Mosahab (who invented the *Iranic* font style, a back-slanted italic form to go with the right-to-left direction of the script), and Tehran University Press, that developed the means to publish the texts with the maximum achievable quality of the days. The human typesetters used many locally developed methods to extend the imported typesetting machines, nowadays called "match stick methods", many of which used match stick parts to provide the proper spacings needed for mathematical formulas.

This was changed in late 1970s by the new typesetting machines made by Linotype which provided easier mechanisms for typesetting mathematics containing Persian text. The machines helped new publishers like Iran University Press and Fatemi Publishing Institute publish technical books in a much shorter typesetting period, making a large volume of mathematical books appear in the 1980s and early 1990s.

A leap happened in 1992, by appearance of two TEX-based typesetting packages called TEX-e-Parsi and LATEX-e-Farsi. The latter disappeared in a short while mainly because of incompetence, but the former, developed by Dadehkavi Iran with some investment from the two above-mentioned publishing houses, remained in use. TeX-e-Parsi design was highly influenced by the way Knuth had created TEX, doing thorough research on the existing typography of Iran at the time. The company, going bankrupt in 1997 because of high expenditure and limited market, has released the latest version of the package in 1996, based on pre-3.0 TEX and LATEX 2.09 with NFSS, but with various modifications both in the TEX engine and LATEX macros (SCO Unix and MS-DOS were supported as platforms). The package, still being used in a highly-tailored form by the mentioned publishing houses and a few mathematical departments, was unfortunately not affordable by individual authors and students. Thus, it could not help authors doing the document preparation themselves, and needed a special section in each department for typesetting manuscripts.

But the bigger leap was another package called Zarnegar, appearing in early 1990s for high quality typesetting using personal computers, which targeted the main stream of typesetting with various fonts and a visual markup language. Because of the good quality of the output and the reasonable price, the package got highly popular, and is still in wide use, estimated to be the second most popular document preparation software in Iran, after Microsoft Word. Unfortunately, Zarnegar's typesetting quality of mathematics is very poor, which has been a source of many badly-typeset technical books.

FarsiTEX

FarsiTEX started as an academic project by Mohammad Ghodsi in Computer Engineering Department of Sharif University of Technology. The project, known as FaTEX in the first year, started in 1991 as three BSc projects supervised by Ghodsi to provide the foundation (Haghghollahi, 1992; Asghari, 1993; Tajrobekar, 1993). After many experiments, FarsiTEX finally settled on the TEX--XET engine and the MS-DOS platform. The main work was done by Hassan Abolhassani and Mehran Sharghi in two MSc theses, the former working on a macro set with some ideas borrowed from the localized Hebrew version of LATEX 2.09 (Abolhassani, 1994), and latter on a METAFONT family of Persian fonts based on Linotron Badr, which he called Scientific Farsi (Sharghi, 1994). The contextual shaping of the letters was done by a pre-processor, which took input documents in the then widely used Iran System character set, and converted them to an internal code page which used four characters for each letter, each for one of the forms used in the Naskh style.

The system was in limited use by authors for about two years, until early 1996 when Ghodsi gathered a new team to concentrate on a public release of the software under GNU General Public License (Free Software Foundation, Inc., 1991). The team created a new syntax and character set for FarsiTEX input files, and consisted of Kiarash Bazargan, who created ftexed, an MS-DOS text editor based on Borland Turbo Vision, Mohammad Mahdian who wrote ftx2tex to handle the new file format, Roozbeh Pournader who revised the macro set, and Sharghi who revised his own fonts. The first public version appeared in October 1996, as an extension to emTEX distribution which was very popular at the time. Explicitly marked as beta-quality software, FarsiTEX was the first Iranian software released under GPL. A manual (Ghodsi and Pournader, 1997) was distributed with the package as a DVI file, and was also made available on paper for a very small fee.

FarsiTEX, imagined by its authors to have a very limited audience because of its scalability problems and various known bugs, grew rapidly among students and professors of mathematics, computer engineering, and physics all over the country, simply because it was the only affordable option available which was good enough for their basic typesetting tasks. The students, many of them now able to afford a PC at home, needed some software to run themselves. FarsiTEX was also evangelized by the new professors who had just returned to Iran after their studies in an American or European university and knew the value of document preparation by the author. Authors of FarsiTEX, betting on about a hundred users, were amused to find a base sized ten times that number.

The FarsiTEX Project Team, born in 1996 and still breathing amid various inactivity periods, has released many small improvements since that time. Also, it has recently done a few alpha releases of a new system based on MikTEX, which includes a Microsoft Windows editor written almost from scratch (written by Mehrdad Sabetzadeh, Shiva Nejati, and Okhtay Ilghami), a localized version of *MakeIndex* supporting Persian ordering (by Nejati), and a FarsiTEX to HTML conversion program (by Mohammad Bakuii). It is unfortunate that the team has not released a single stable version yet, and the MS-DOS release is now frozen forever.

It may be worth noting that code contributions to the project from outside the project team has been very small, although there has been many serious users. The team members are still wondering about the possible reasons, but mostly blame it on the uncooperative nature of the Iranian people!

During these six years, the project has been financially supported by Sharif University of Technology, Ministry of Science, Research, and Technology, High Council of Informatics of Iran, Statistical Center of Iran, and Science and Arts Foundation.

TEXnical Details and Examples

Just like any other non-Unicode Persian software, FarsiTEX has its own character set, as unfortunately no 8-bit Persian character set has ever been both complete and popular. This character set and its inherent semantics make a special text editor an essential part of the FarsiTEX system, and the same time the major barrier for porting the system to other platforms, like Linux.

The Bidirectional Algorithm Bidirectionality, is the main issue to tackle in any Persian TEX system. The TEX--XET engine is of course capable of typesetting bidirectional text, but only if the directions are known *explicitly*. In other words, TEX--XET has nothing to do with the implicit directionalities of Unicode Bidirectional Algorithm (Davis, 2002) which, given some text in a logical order (a run of text as typed through a keyboard, for example) outputs the text in a visual order (the sequence of characters as should appear on a computer screen or a piece of paper). This mapping is far from trivial in cases that characters of both directionalities mix with *neutral* characters (like punctuations and spaces), or weakly directioned ones (like digits).

In FarsiTEX, the text editor is responsible for converting the logical order to the visual one. The editor manipulates files with the `ftx` extension, which are in a special semi-logical semi-visual bidirectional format designed to be as near as possible to the internal representation of the editor (which is in visual order). This format has simplified the bidirectional algorithm by using two different codes for many neutral characters like space and parentheses, one for each of the left-to-right and right-to-left modes. The idea of having different characters for different directions has been borrowed from the ISIRI 3342 (Institute of Standards and Industrial Research of Iran, 1993), a national Iranian character set standard.

The `ftx` format, while easy to process for the editor, is not suitable for a TEX-like engine, which raises the need for the `ftx2tex` converter, that reorders the visual text in the `ftx` file to the logical order, explicitly marking the directionality using `\InE` (Insert English), `\EnE` (End English), `\InF` (Insert Farsi), and `\EnF` (End Farsi) macros. These macros enable the engine to typeset a text in both directions.

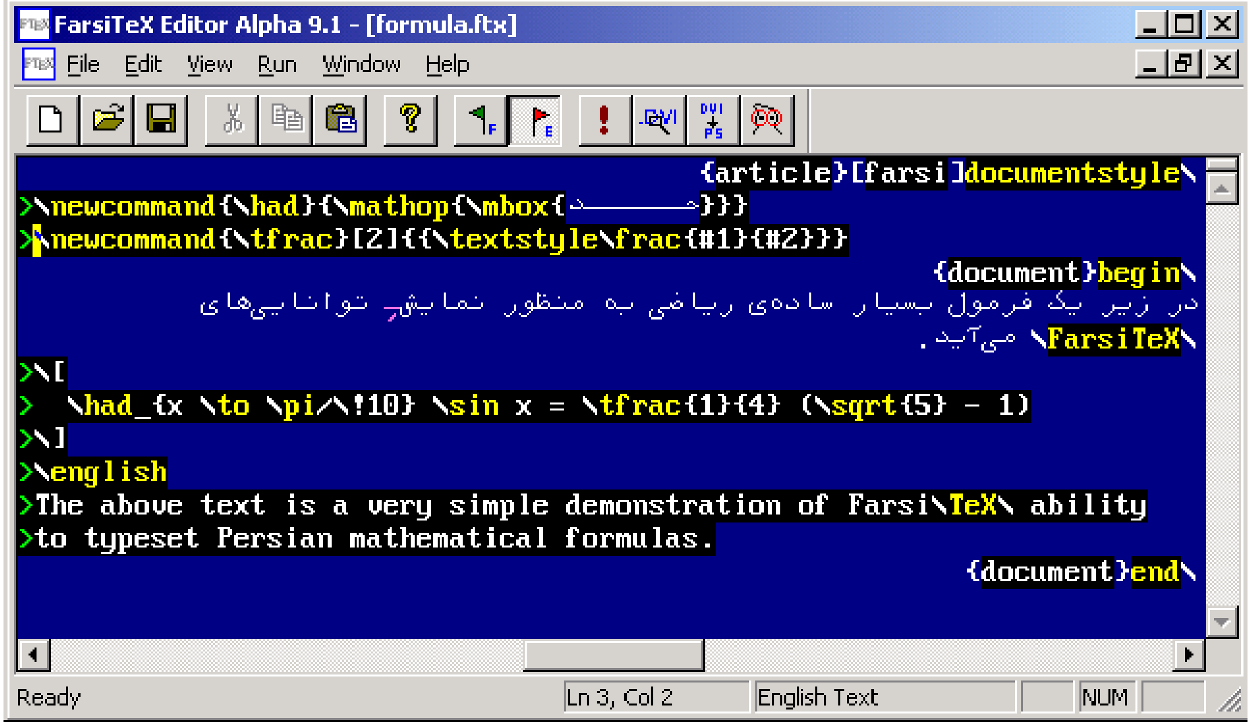
A screenshot of the Microsoft Windows editor, is shown in Figure 1 (FarsiTEX's output can be seen in Figure 2). Two different background colors are used to specify the characters' direction, needed for neutral characters like space, full stop, and parentheses. So, unlike the common bidirectional algorithms, and thanks to the background color, there are no ambiguities in the direction of neutral characters. But the problem of nesting different directionalities still remains.

Joining, Shaping, and Line Justification The Persian script, being a derivative of Arabic, is a cursive script, which means that two adjacent letters may *join* to each other, forming up to four different glyphs for each letter. The `ftx2tex` converter is responsible for detecting the pairs that join (*the joining algorithm*) and selecting the proper glyphs based on joining information (*the shaping algorithm*).

When a typesetter is justifying the lines in a Persian paragraph, it is common to stretch the joining line that appears between two adjacent glyphs. There is no inter-letter spacing in FarsiTEX, and only the joining stem will be stretched. To implement this behavior, the `ftx2tex` converter inserts a *stretchable kashide* character (also known as *tatweel*) character between the two connected letters. This inserted character is defined as an active character expanding to a horizontal glue filled by horizontal rules. A sample of the behavior can be seen in Figure 3.

FarsiTEX Forever

The FarsiTEX Project Team is currently working on a new release with PostScript Type 1 fonts, moved by the serious need of the user community



```
{article}[farsi]documentstyle\
>\newcommand{\had}{\mathop{\mbox{حــــــــــد}}}
>\newcommand{\tfrac}[2]{{\textstyle\frac{#1}{#2}}}
{document}begin\
در زیر یک فرمول بسیار ساده‌ی ریاضی به منظور نمایش توانایی‌های
\FarsiTeX\ می‌آید.
>\[
>  \had_{x \to \pi/\!10} \sin x = \tfrac{1}{4} (\sqrt{5} - 1)
>\]
>\english
>The above text is a very simple demonstration of Farsi\TeX\ ability
>to typeset Persian mathematical formulas.
{document}end\
```

Figure 1: The FarsiTEX editor running under Microsoft Windows. Notice the background color of backslashes and curly braces in the right-to-left lines.

در زیر یک فرمول بسیار ساده‌ی ریاضی به منظور نمایشِ توانایی‌های فارسی‌تک می آید.

$$\underset{x\to\pi/۱۰}{\text{حـــــد}} \sin x = \tfrac{۱}{۴}(\sqrt{۵} - ۱)$$

The above text is a very simple demonstration of FarsiTEX ability to typeset Persian mathematical formulas.

Figure 2: FarsiTEX's output, with the input given in Figure 1. Notice the automatic replacement of European digits (also known as Arabic digits) by Persian ones. The operator appearing before the sine is an alternate form of `\lim`, used in high school mathematics textbooks in Iran.

اَلا یا اَیُهَا السّاقی اَدِر کَأساً وَ ناوِلْها   که عشق آسان نمود اول ولی افتاد مشکل‌ها
به بوی نافه‌ای کاخر صبا زان طره بگشاید   ز تابِ جعدِ مشکینش چه خون افتاد در دل‌ها
مرا در منزلِ جانان چه امنِ عیش چون هر دم   جرس فریاد می‌دارد که بر بندید محمل‌ها
به مِی سجاده رنگین کن گرت پیرِ مغان گوید   که سالک بی‌خبر نبود ز راه و رسمِ منزل‌ها
شبِ تاریک و بیمِ موج و گردابی چنین هایل   کجا دانند حالِ ما سبک‌بارانِ ساحل‌ها
همه کارم ز خودکامی به بدنامی کشید آخر   نهان کِی ماند آن رازی کز او سازند محفل‌ها
حضوری گر همی‌خواهی از او غایب مشو حافظ   مَتیٰ ما تَلْقَ مَنْ تَهویٰ دَعِ الدُنْیا وَ اَهْمِلْها

Figure 3: A sonnet by Hafez, typeset in two columns with text stretched for equal width. This style is necessary for typesetting traditional poems, where justification in shape was a visual reference to the poem's rhyme.

to publish their documents in PDF, and also a Linux text editor, which will make the first teTEX-based Linux release possible. Other plans include Unicode support and integration with Omega, which will need a complete review of the system. The project is being continued in Computing Center, Sharif University of Technology, and can be reached at

`http://www.farsitex.org/`

(which is hosted at SourceForge.net).

References

Abolhassani, Hassan. *Typesetting Persian Documents using TEX*. Master's thesis, Computer Engineering Department, Sharif University of Technology, 1994.

Asghari, Parvaneh. "Scientific design of Traffic fonts using METAFONT". BSc project report, Computer Engineering Department, Sharif University of Technology, 1993.

Davis, Mark. "The Bidirectional Algorithm". Unicode Standard Annex #9, The Unicode Consortium, 2002. Available from `http://www.unicode.org/unicode/reports/tr9/`.

Free Software Foundation, Inc. "GNU General Public License (GPL)". 1991. Available from `http://www.gnu.org/licenses/gpl.html`.

Ghodsi, Mohammad and R. Pournader. *The FarsiTEX Manual*. Computer Engineering Department, Sharif University of Technology, 1997.

Haghghollahi, Jafar. "Designing Persian fonts using METAFONT". BSc project report, Computer Engineering Department, Sharif University of Technology, 1992.

Institute of Standards and Industrial Research of Iran. "ISIRI 3342, Farsi 8-bit Coded Character Set for Information Interchange". 1993.

Sharghi, Mehran. *Scientific Design of Persian Fonts*. Master's thesis, Computer Engineering Department, Sharif University of Technology, 1994.

Tajrobekar, Laleh. "Designing with PostScript in Persian environments". BSc project report, Computer Engineering Department, Sharif University of Technology, 1993.

Unicode Editorial Committee. "Unicode 3.2". Unicode Standard Annex #28, The Unicode Consortium, 2002. Available from `http://www.unicode.org/unicode/reports/tr28/`.