Persian Computing
with
Unicode

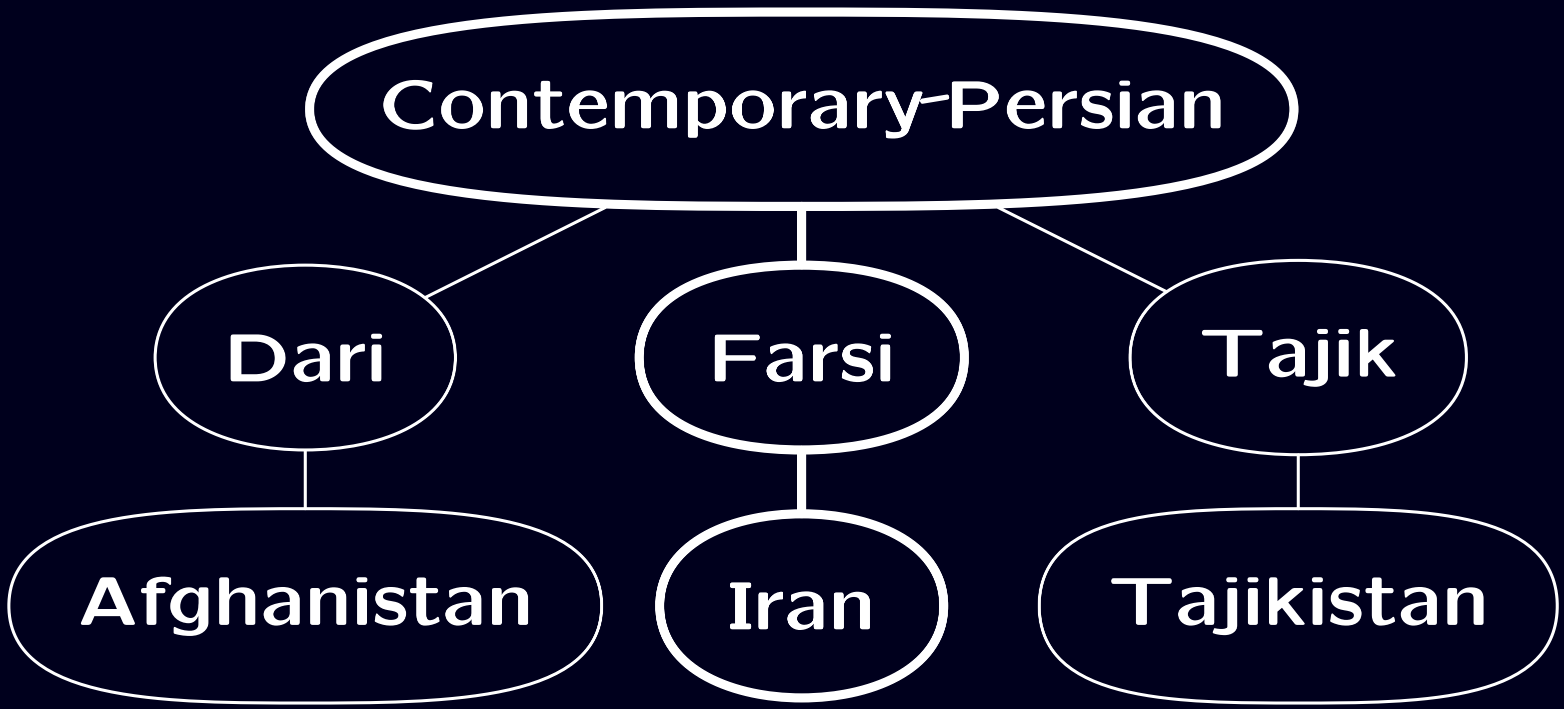
~~Behdad Esfahbod~~ no visa
unicode@behdad.org

Amir Youssefi
youssefi@cs.rpi.edu

The FarsiWeb Project
http://www.farsiweb.info/

March 31, 2004

What is Persian?



Persian in Computers

There are three relevant national standards:

- ISIRI 3342:1992 Farsi 8-bit Coded Character Set for Information Interchange (deprecated)
- ISIRI 2901:1994 Keyboard Layout for Farsi: Characters in Computer
- ISIRI 6219:2002 Information Technology -- Persian Information Interchange and Display Mechanism, using Unicode

Modern Persian Script

- Based on Arabic Script (⟨U+0600..U+06FF⟩ block): With some extra letters, some modified letters

- But with completely different semantics and typographical habits

- Is a bidirectional script: Is written from right to left, except for numbers

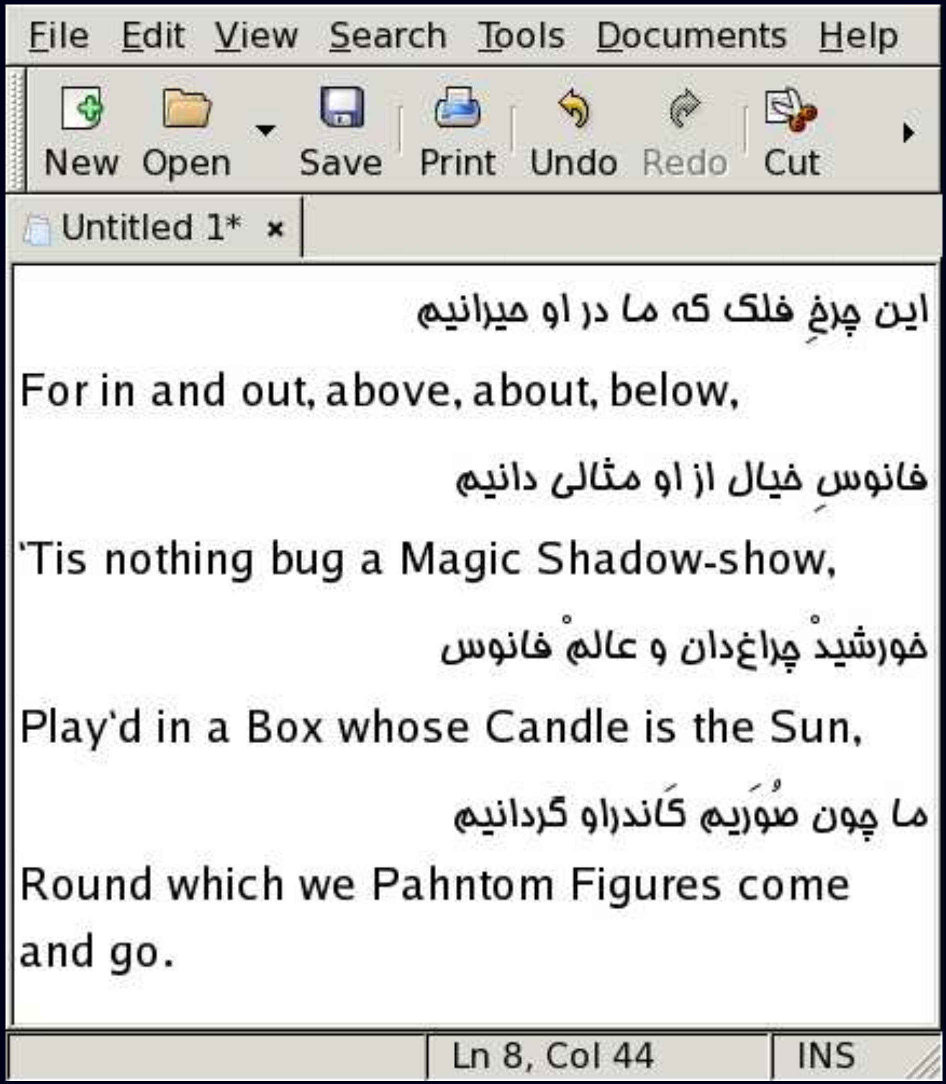
- Needs cursive joining: Two adjacent letters may be *joined*, forming 1, 2, or 4 glyphs for each character: (for example س, ـس, ـسـ, سـ)

Arabic Script Rendering

Input text	Logical order	س ل ا م
After Bidirectional Algorithm	Visual order	م ا ل س
After Arabic Joining Algorithm	Glyph list	م ﺎ ﻠ ﺳ
After Ligation	Glyph list	م ﻼ ﺳ
When Rendered	Output	سلام

With enough care, it is possible, to apply the above algorithms in a different order, and get the same result.

A Bidirectional Document (gedit/Pango)



Alphabet

- Extra letters:
  U+067E *Peh* (پ), U+0686 *Tcheh* (چ),
  U+0698 *Jeh* (ژ), U+06AF *Gaf* (گ)

- Modified letters:

Character	Isol	Fina	Medi	Init
U+0643 *Arabic Letter Kaf* (Arabic Kaf)	ك	ـك	ـكـ	كـ
U+06A9 *Arabic Letter Keheh* (Persian Kaf)	ک	ـک	ـکـ	کـ
U+064A *Arabic Letter Yeh* (Arabic Yeh)	ي	ـي	ـيـ	يـ
U+06CC *Arabic Letter Farsi Yeh* (Persian Yeh)	ی	ـی	ـیـ	یـ

Alphabet (continued)

- Three shapes of composed Hamza Above:
  U+0623 *Alef with Hamza Above* (أ),
  U+0624 *Waw with Hamza Above* (ؤ),
  U+0626 *Yeh with Hamza Above* (ئ)

- Never used characters:
  U+0649 *Alef Maksura* (ى): Like Yeh, but no dots at all
  U+06C0 *Heh with Yeh Above* (ۀ): Should **never** be used instead of ⟨*Heh, ZWNJ, Farsi Yeh*⟩ or ⟨*Heh, Hamza Above*⟩ sequence

Special Characters

- U+0640 *Arabic Tatweel*, for a longer joining stem

  کتاب → کـتاب

- U+200C *Zero Width Non-Joiner*, to prevent joining

  کتابها → کتاب‌ها

- U+200D *Zero Width Joiner*, to choose a joined glyph when it would not join naturally

  ه.ش → ه‍.ش

- U+200E *Left-to-Right Mark*, U+200F *Right-to-Left Mark*, and other bidirectional control chars (⟨U+202A..U+202E⟩)

Numbers

- ⟨U+06F0..U+06F9⟩ *Extended Arabic-Indic Digits*:

  ۰ ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹

  instead of ⟨U+0660..U+0669⟩ *Arabic-Indic Digits*:

  ٠ ١ ٢ ٣ ٤ ٥ ٦ ٧ ٨ ٩

- U+066C *Arabic Thousands Separator* and U+066B *Arabic Decimal Separator*:

  ۹٬۸۷۶٬۵۴۳٫۲۱۰

- Western numerals in Latin context. Persian numerals everywhere else (page numbers, section numbers, . . .)

Other Characters

- Harakat (Vowel) Non-spacing Marks: ـَ ـِ ـُ ـً ـٍ ـٌ ـّ ـٰ . . .

- Arabic Punctuation Marks:
  U+060C *Arabic Comma* (،),
  U+061B *Arabic Semicolon* (؛),
  U+061F *Arabic Question Mark* (؟),
  U+066A *Arabic Percent Sign* (٪),

- ⟨U+00AB, U+00BB⟩ *Double Angle Quotation Marks* (( ))

- Shared Punctuation Marks:
  Latin full stop, exclamation mark, parenthesis, square brackets, . . .

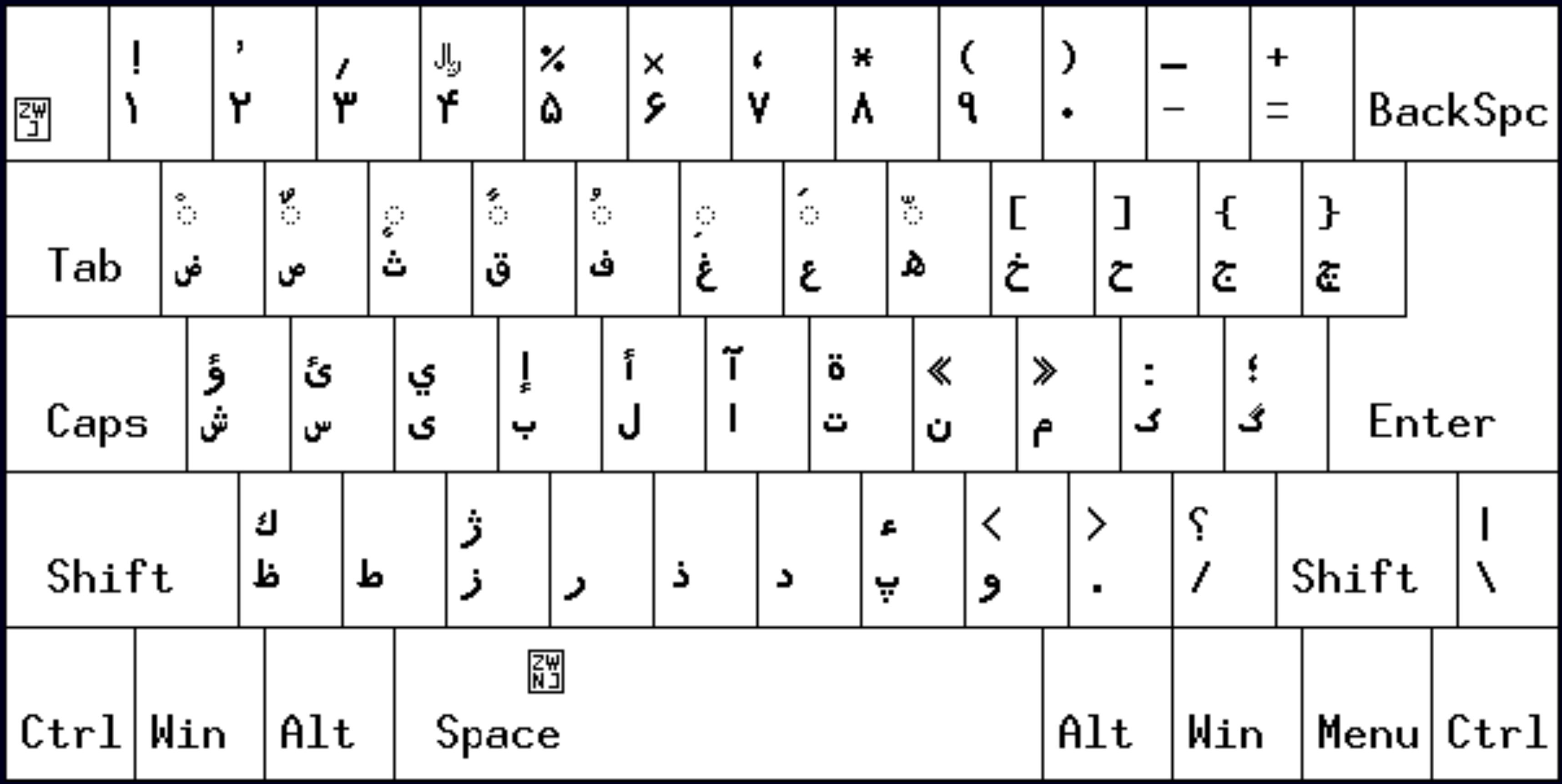
												،	060D	060E	060F
0610	0611	0612	0613	0614	0615						؛				؟
	ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ					
ـ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	ً	ٌ	ٍ	َ	ُ
ِ	ّ	ْ	ٓ	ٔ	ٕ	ٖ	0657	0658							
٠	١	٢	٣	٤	٥	٦	٧	٨	٩	٪	٫	٬	٭	ٮ	ٯ
ٰ	ٱ	ٲ	ٳ	ٴ	ٵ	ٶ	ٷ	ٸ	ٹ	ٺ	ٻ	ټ	ٽ	پ	ٿ
ڀ	ځ	ڂ	ڃ	ڄ	څ	چ	ڇ	ڈ	ډ	ڊ	ڋ	ڌ	ڍ	ڎ	ڏ
ڐ	ڑ	ڒ	ړ	ڔ	ڕ	ږ	ڗ	ژ	ڙ	ښ	ڛ	ڜ	ڝ	ڞ	ڟ
ڠ	ڡ	ڢ	ڣ	ڤ	ڥ	ڦ	ڧ	ڨ	ک	ڪ	ګ	ڬ	ڭ	ڮ	گ
ڰ	ڱ	ڲ	ڳ	ڴ	ڵ	ڶ	ڷ	ڸ	ڹ	ں	ڻ	ڼ	ڽ	ھ	ڿ
ۀ	ہ	ۂ	ۃ	ۄ	ۅ	ۆ	ۇ	ۈ	ۉ	ۊ	ۋ	ی	ۍ	ێ	ۏ
ې	ۑ	ے	ۓ	۔	ە	ۖ	ۗ	ۘ	ۙ	ۚ	ۛ	ۜ		۞	۟
۠	ۡ	ۢ	ۣ	ۤ	ۥ	ۦ	ۧ	ۨ	۩	۪	۫	۬	ۭ	06EE	06EF
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹	ۺ	ۻ	ۼ	۽	۾	06FF

Keyboard Layout

ISIRI 2901:1994, features:

- All letters and punctuation marks
- Reasonable placement
- Persian digits
- Regular and shifted keys only
- Some empty slots

Iranian Standard Layout

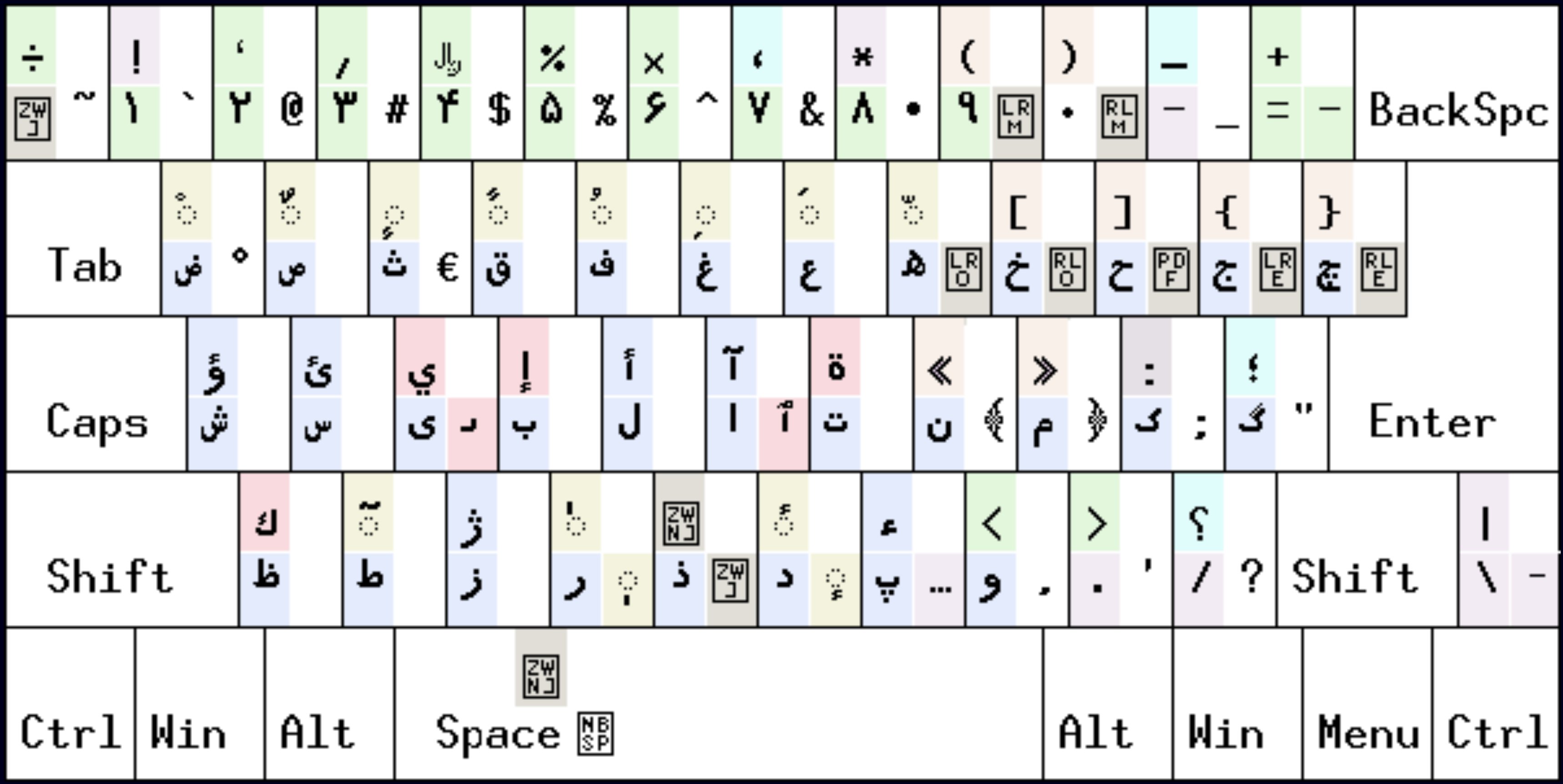


Keyboard Layout (continued)

Proposed update, features:

- Fully backward-compatible with ISIRI 2901:1994
- Unicode 4.0 repertoire, and complete support for ISIRI 6219:2002
- Support quoting Arabic text
- Uses AltGr to add required but rarely used characters
- Adds all ASCII punctuation marks, useful for editing XML, . . .
- Adds bidirectional control characters

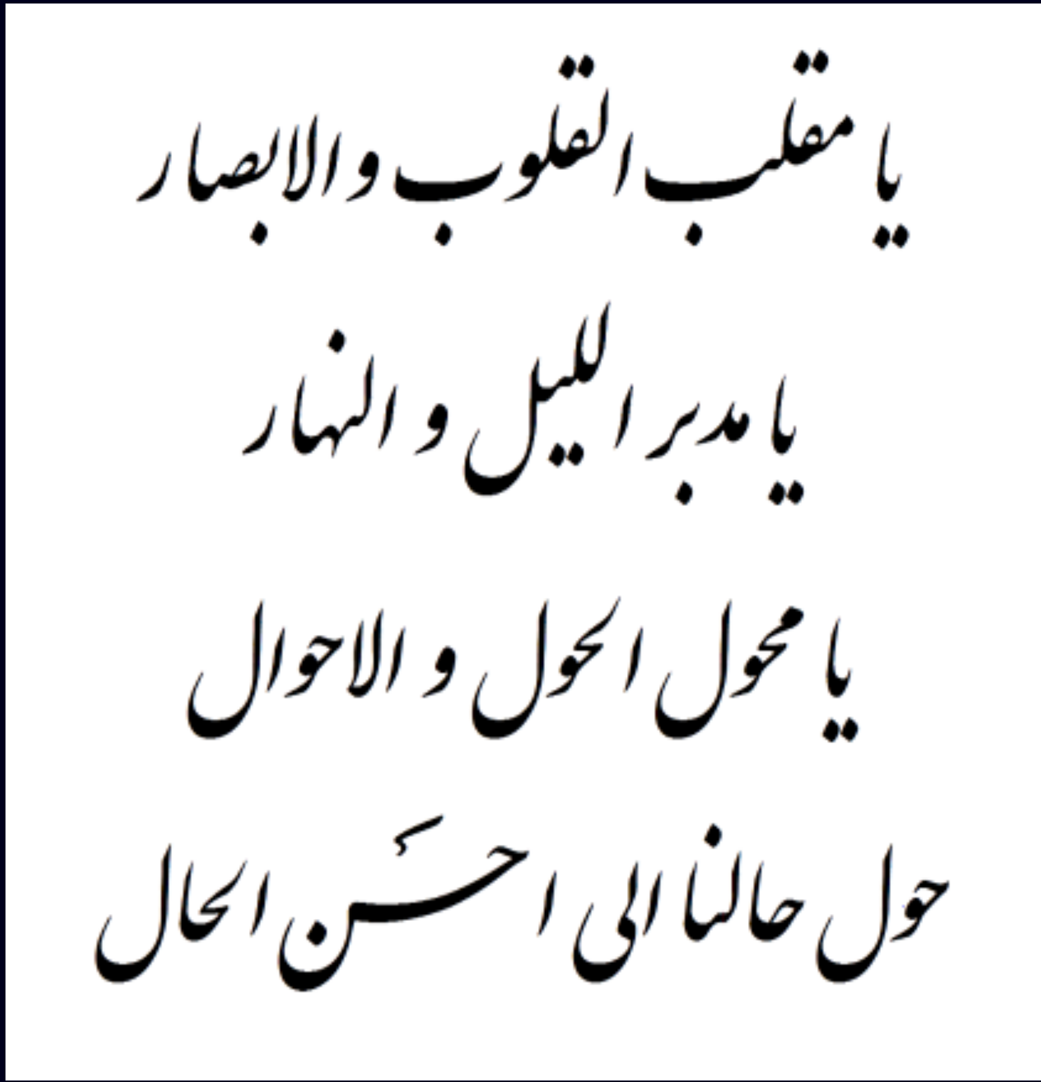
Proposed Iranian Layout



Fonts

- Microsoft fonts are Arabic
- Tahoma is the best looking one
- Persian fonts are not Unicode compatible yet
- The only ligature: Lam-Alef
- Nastaliq is desired, but not possible yet

The Nastaliq Style



Date and Time

- Three calendars in use!
  Gregorian, to synchronize with the rest of the world
  Jalali, the official calendar
  Islamic, for some holidays and ceremonies

- Islamic calendar depends on moon-sighting once a year

- Week starts Saturday

- Business weekdays from Saturday to Thursday

- 24-hour preferred in media

- No AM/PM equivalent

Collation

- Like Arabic basically
- ن < هـ < و < ي → ن < و < هـ < ى
- Some L2 equal pairs:

  ک $<_3$ ك   ى $<_3$ ي   ة $<_3$ ت

  ۴ $<_3$ ٤   ۵ $<_3$ ٥   ۶ $<_3$ ٦
- Traditional rules: Hamza variants are L2 equal:

  أ $<_3$ ؤ $<_3$ ئ
- Modern rules: L2 equal with their base letter:

  ا $<_3$ أ   و $<_3$ ؤ   ى $<_3$ ئ

Loose Searching

ZWNJ ≃ Space	*ZWNJ ≃ empty string*
حجة الاسلام ≃ حجت الاسلام	دایره المعارف ≃ دایرة المعارف
ک‌تاب ≃ کتاب	خانهٔ ≃ خانه‌ی
ي ≃ ی	ك ≃ ک

سؤال ≃ سوال   پائیز ≃ پاییز   تأخیر ≃ تاخیر

مسئول ≃ مسؤول   مؤمن ≄ مأمن   مسأله ≃ مسئله

ملک ≃ مِلک   مِلک ≄ مُلک   مُلک ≃ ملک

٩ ≃ ۹ ...   ٦ ≃ ۶   ٥ ≃ ۵   ٤ ≃ ۴   ... ٠ ≃ ۰

Last Notes to Application Developers

- Typesetting Persian paragraphs:
  Justified lines
  No inter-letter spacing
  No word hyphenation
  Almost no inter-word spacing
  Use Tatweel instead: کتاب من → کـــتـــاب مـــن

- All text fields Right-to-Left

- Persian numbers

- Right-to-Left layout

- Beware: Right and Left are swapped!

An Old-style Persian Poem (MS Word)

زان پیش که حسبِ حال گویم	از صانعِ ذوالجلال گویم
آن خالقِ ماه و خور که چون گوی	زو چرخ فتاده در تکاپوی
زو گوی سپهر مستدیر است	چوگانِ هلال گوشه‌گیر است
از حکمتِ اوست در زد و گیر	چوگانِ قضا و گوی تقدیر
از ماه برین بلند ایوان	گه گوی نموده گاه چوگان
هر ذرّه ز ماه تا بماهی	بر وحدتِ او دهد گواهی
صنعش که ز مهر عالم افروخت	بر جیبِ سپهر گوی زر دوخت
این گوی درست‌زر که مهرست	در چرخ ز گردشِ سپهرست
از شرق بغرب داده راهش	کانجاست محلِّ حالگاهش

A Right-to-Left Dialog (Gtk+/Pango)

A Persian Paragraph (Mozilla)

۲

این سفیدباشی از خیره‌سری است که
دم مار می‌گزد یا نشاطِ عیش کرده
است که گاه می‌نماید و گه می‌رباید؟

چرا وقتی پشتِ ماه خمیده شد و تازی به ماری که
از درخت عرعر بالا می‌رفت پارس کرد و او روی
پنجه‌ها گردن کشید تا ناله‌ی وصل کند خار گزی به
ساقم خلید تا حواسم پرتِ زقزق شود و نتوانم
شاهد نمه‌عرقی باشم که می‌گویند در ربع مسکون
تنها بر منخرین دختر شاهِ سمنگان می‌نشیند وقتی
شبِ چهارده از ایوان به کیوان می‌نگرد.

در این شب و مرتع و اسارت باقی به همین قیاس
گذشت تا سرانجام که دم به تله داد چنان شیهه‌ای
کشید که حتی الاغ‌های آن اطراف هم دانستند که
اسبِ سفید تهمینه دیگر از خیره‌سری دم مار را
نخواهید گزید.

A Persian Table (Mozilla)

پورمقدم، یارعلی، ۱۳۳۰ -

یادداشت‌های یک اسب / یارعلی پورمقدم. - تهران: آرویج، ۱۳۸۰.

۴۵ص.

ISBN 964-7174-6-8

فهرستنویسی بر اساس اطلاعات فیپا.

PIR۷۹۹۲/ی۴۷و۱۷	۸فا۶۲/۳
۱۳۸۰	ی۷۵۶پ
	۱۳۸۰

کتابخانه ملی ایران م۸۰-۲۶۰۵۶

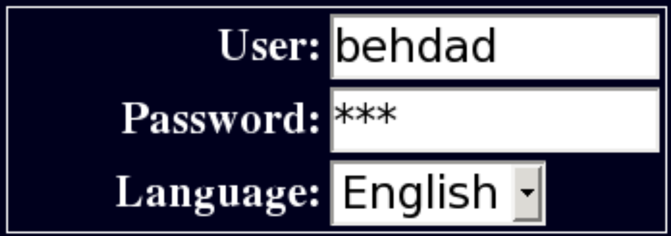
- یادداشت‌های یک اسب
- یارعلی پورمقدم
- طرح روی جلد: هومن خطیبی
- چاپ اول: ۱۳۸۰
- لیتوگرافی: پام مهر ۷۵۰۰۹۳۰
- چاپ: چکاد
- تیراژ: ۳۵۰۰
- قیمت: ۴۵۰ تومان
- شابک: ۹۶۴-۷۱۷۴-۶-۸
- انتشارات آرویج: خیابان شریعتی، بالاتر از سه‌راه طالقانی، خیابان شهید کارگر، پلاک ۱۲، تلفن: ۷۵۲۵۱۶۵، تلفکس: ۷۵۳۷۰۷۶

A Bidirectional HTML Dialog (Mozilla)

The HTML/CSS markup generating the two forms below is exactly the same, except for:

- The Persian form has `lang=fa` once for the `body` tag.
- The Persian form has `direction: rtl` once for the `body` tag.
- In the English form, the labels all have `align: right`, in the Persian one they have `align: left`.

Unfortunately there is no way in CSS to drop the annoying `align` difference.



A Persian Web Page with Two Nice Persian Fonts (Mozilla)

Current Status -- Microsoft Windows

- Render correctly
- Shipped fonts work
- Keyboard layout is terrible
- No Persian digits support
- No Iranian calendar
- Locale data is wrong in places
- No interface translation
- Not trivial to enable Persian support

Current Status -- Linux

- Important systems support rendering
- No good fonts yet
- Standard keyboard layout
- No Persian digits support yet
- KDE claims Iranian calendar support
- Some interface translation done
- Not trivial to enable Persian support

Current Status -- MacOS

- Supports rendering
- No good fonts
- Legacy and standard keyboard layouts

References and Resources

- The Unicode Standard at `http://www.unicode.org/`
- Institute of Standards and Industrial Research of Iran at `http://www.isiri.com` (documents in Persian)
- The FarsiWeb Project at `http://www.farsiweb.info/`
- PersianComputing list at `http://lists.sharif.edu/mailman/listinfo/persiancomputing`
- Typing Persian Word Documents with Windows Tutorial at `http://students.washington.edu/irina/persianword/persianwp.htm`