

<b>Title:</b>	Notice of Publication: ISIRI 6219:2002, Information Technology—Persian Information Interchange and Display Mechanism, using Unicode
<b>Source:</b>	Institute of Standards and Industrial Research of Iran (ISIRI)
<b>Date:</b>	2002-12-26
<b>Action:</b>	For your information

The Iranian national standard ISIRI 6219, titled *Information Technology—Persian Information Interchange and Display Mechanism, using Unicode*, was published on 2002-11 in Persian. The Persian title is:

فناوری اطلاعات - تبادل و شیوهی نمایش اطلاعات فارسی بر اساس یونی‌کُد

The standard has normative references to both ISO/IEC 10646-1:2000 and the Unicode Standard, Version 3.2, and specifies requirements of encoding Persian texts using ISO/IEC 10646 and Unicode.

ISIRI 6219:2002 has been prepared and drafted by the High Council of Informatics of Iran (<http://www.shci.ir/>), a liaison member of the Unicode Consortium, with contributions from Sharif University of Technology (<http://www.sharif.edu/>) and Science and Arts Foundation (<http://www.science-arts.org/>).

The text of the standard, excluding the covers, is attached to this document. A paper copy may be acquired at the price of 4125 IRR from:

Sales Office  
 Institute of Standards and Industrial Research of Iran  
 PO Box 31585-163  
 Karaj  
 IRAN  
 Fax and Phone: +98 (261) 280-7045

فناوریِ اطلاعات – تبادل و شیوهی نمایشِ اطلاعاتِ  
فارسی بر اساسِ یونی کُد

نسخه‌ی نهایی

**Information Technology – Persian  
Information Interchange and  
Display Mechanism, using Unicode**

Final Version

کمیسیون فنی استاندارد «فناوری اطلاعات – تبادل و شیوه‌ی نمایش اطلاعات  
فارسی بر اساس یونی‌کُد»

رئیس

تابش، یحیی      دانشگاه صنعتی شریف

اعضا

اسفهد میرحسین‌زاده سرابی، سید بهداد      دانشگاه صنعتی شریف

پورنادر، روزبه      دانشگاه صنعتی شریف

خانبان، علی اصغر      دانشگاه لندن

علمدار میلانی، امید      دانشگاه صنعتی شریف

دبیر

پناهی، زهرا      دانشگاه صنعتی شریف

فصیحی، مریم      مؤسسه‌ی استاندارد و تحقیقات صنعتی ایران

فهرست مندرجات . . . . .	صفحه
پیشگفتار . . . . .	پ
مقدمه . . . . .	ث
۱ هدف و دامنه‌ی کاربرد . . . . .	۱
۲ مراجع الزامی . . . . .	۲
۳ اصطلاحات و تعاریف . . . . .	۳
۱-۳ متن . . . . .	۳
۲-۳ خط . . . . .	۳
۳-۳ نویسه . . . . .	۴
۴-۳ مجموعه نویسه . . . . .	۴
۵-۳ شکل . . . . .	۴
۶-۳ متن ساده . . . . .	۴
۷-۳ کدگذاری کردن . . . . .	۴
۴ نمادها . . . . .	۴
۵ نویسه‌های مورد استفاده در متون فارسی . . . . .	۵
۱-۵ نویسه‌های کنترلی . . . . .	۶
۲-۵ علائم نقطه‌گذاری مشترک . . . . .	۷
۳-۵ علائم نقطه‌گذاری فارسی . . . . .	۹
۴-۵ ارقام و علائم ریاضی . . . . .	۹
۵-۵ حروف اصلی فارسی . . . . .	۱۰
۶-۵ حروف فرعی . . . . .	۱۳
۷-۵ نشانه‌های فارسی . . . . .	۱۴
۸-۵ نویسه‌های ممنوع . . . . .	۱۵
۹-۵ نویسه‌های منسوخ . . . . .	۱۶

۱۷	پیوست الف	الگوریتم دوجهته
۱۸	پیوست ب	الگوریتم اتصال
۱۸	ب-۱	رده‌ی اتصال
۱۹	ب-۲	الگوریتم
۲۰	ب-۳	گروه اتصال
۲۲	ب-۴	لیگاتورها
۲۳	پیوست پ	قالب‌های تبادل داده‌ها
۲۴	پیوست ت	سטר بندی و پاراگراف بندی
۲۵	پیوست ث	نرمال سازی و هم‌ارزی
۲۶	پیوست ج	واژه‌نامه
۲۸	پیوست چ	کد نویسه‌ها
۳۱	پیوست ح	نام نویسه‌ها

## پیشگفتار

استاندارد «فتاوری اطلاعات – تبادل و شیوهی نمایش اطلاعات فارسی بر اساس یونی کُد» که پیشنویس آن توسط «شورای عالی انفورماتیک کشور» در کمیسیون‌های مربوطه تهیه و تدوین شده و در دو جلسه‌ی کمیته‌ی ملی استاندارد رایانه و فرآوری داده‌ها مورخ ۱۳۸۰/۱۲/۱۹ و ۱۳۸۱/۲/۱۸ مورد تأیید قرار گرفته است، اینک به استناد بند یک ماده‌ی ۳ قانون اصلاح قوانین و مقررات مؤسسه‌ی استاندارد و تحقیقات صنعتی ایران، مصوب بهمن‌ماه ۱۳۷۱ به‌عنوان استاندارد ملی ایران منتشر می‌شود.

برای حفظ همگامی و هماهنگی با تحولات و پیشرفت‌های ملی و جهانی در زمینه‌ی صنایع، علوم، و خدمات، استانداردهای ملی ایران در مواقع لزوم تجدید نظر خواهند شد و هرگونه پیشنهادی که برای اصلاح یا تکمیل این استاندارد ارائه شود، در هنگام تجدید نظر در کمیسیون‌های فنی مربوطه مورد توجه قرار خواهد گرفت. بنابراین برای مراجعه به استانداردهای ایران باید همواره از آخرین تجدید نظر آن‌ها استفاده کرد.

در تهیه و تدوین این استاندارد سعی شده است که ضمن توجه به شرایط موجود و نیازهای جامعه، در حد امکان بین این استاندارد و استانداردهای بین‌المللی و استانداردهای ملی کشورهای صنعتی و پیشرفته هماهنگی ایجاد شود.

منابع و مراجعی که برای تهیه‌ی این استاندارد به کار رفته به شرح زیر است:

1. The Unicode Consortium, The Unicode Standard, Version 3.2.0, defined by:

*The Unicode Standard, Version 3.0*, Addison-Wesley, 2000, as amended by the

*Unicode Standard Annex #27: Unicode 3.1*

(<http://www.unicode.org/unicode/reports/tr27/>)

and by the *Unicode Standard Annex #28: Unicode 3.2*

(<http://www.unicode.org/unicode/reports/tr28/>).

2. ISO 10646-1:2000 Information Technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane.

3. Dave Ragget, Arnaud Le Hors, Ian Jacobs, "HTML 4.01 Specification", World Wide Web Consortium, December 1999.
4. Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, "Extensible Markup Language (XML) 1.0", World Wide Web Consortium, Second Edition, October 2000.
5. Martin J. Dürst, François Yergeau, Richard Ishida, Misha Wolf, Asmus Freytag, Tex Texin, "Character Model for the World Wide Web 1.0", World Wide Web Consortium, Working Draft, April 2002.

۶. استاندارد ملی ایران ۳۳۴۲: سال ۱۳۷۲ کُد تبادلِ اطلاعاتِ ۸ بیتی فارسی.

۷. استاندارد ملی ایران ۲۹۰۰: سال ۱۳۶۷ کُد تبادلِ اطلاعاتِ به زبانِ فارسی.

۸. استاندارد ملی ایران ۸۲۰: سال ۱۳۵۱ حروفِ فارسی در ماشین‌های تحریر.

۹. دستورِ خطِ فارسی، فرهنگستانِ زبان و ادبِ فارسی، ۱۳۷۸.

۱۰. شیوه‌نامه، مرکزِ نشرِ دانشگاهی، ویرایشِ دوم، ۱۳۷۲.

۱۱. نتایجِ پروژه‌های تحقیقاتیِ گروه «فارسی در شبکه»، مرکزِ محاسبات، دانشگاهِ صنعتیِ شریف، تهران، ۱۳۷۷ تا ۱۳۸۱.

## مقدمه

### آشنایی با استاندارد یونی کد

استاندارد یونی کد (Unicode) شیوه‌ای جهانی برای کُدگذاری نویسه‌ها و متون است. این استاندارد روشی هماهنگ برای کُدگذاری متون چندزبانه مشخص می‌کند که تبادل اطلاعات را در سطوح بین‌المللی میسر می‌سازد. یونی کد کُدگذاری پیش‌فرض استانداردهای اینترنت، از قبیل HTML و XML است و در کلیه سیستم‌عامل‌ها و زبان‌های برنامه‌سازی امروزی پشتیبانی می‌شود. ثبات داده‌ها، امکان تبادل بین‌المللی متون، ساده‌شدن نرم‌افزارها و کم‌شدن هزینه‌های تولید، از جمله مزایای یونی کد برای صنعت فن‌آوری اطلاعات است.

یونی کد از مجموعه نویسه‌های محدود ۸ بیتی بسیار فراتر رفته و با ظرفیت بیش از یک میلیون نویسه، امکان کُدگذاری کلیه زبان‌های نوشتاری دنیا را فراهم می‌کند. به علاوه، برای انتخاب خط و زبان متن، نیازی به استفاده از کُد‌های کنترلی ندارد. یونی کد رفتار یکسانی با نویسه‌های الفبایی، نویسه‌های اندیشه‌نگار، و نمادها و نشانه‌ها دارد، که امکان استفاده از آن‌ها را در اختلاط با یکدیگر فراهم می‌کند. یونی کد، علاوه بر تعیین کُد عددی و نام برای هر نویسه که در استانداردهای مشابه معمول بوده است، اطلاعات بیشتری را نیز که برای پردازش و نمایش متون لازم است تأمین می‌کند، که از آن جمله می‌توان به جهت نویسه و ویژگی‌های الفبایی اشاره کرد.

یونی کد سه قالب برای تبادل و ذخیره‌سازی اطلاعات فراهم می‌کند: UTF-8 برای بسترهای موجود ۸ بیتی (مناسب برای محیط‌های مبتنی بر استاندارد ASCII، از جمله اینترنت)، UTF-16 برای محیط‌های ۱۶ بیتی، و UTF-32 برای محیط‌های ۳۲ بیتی. علاوه بر این، استاندارد یونی کد در تخصیص کُد به نویسه‌ها کاملاً با استاندارد بین‌المللی ISO/IEC 10646 هماهنگ و معادل است. در واقع، هر کاربردی که از استاندارد یونی کد پی‌روی کند، با استاندارد ISO/IEC 10646 نیز کاملاً سازگار است.

برای اطلاعات بیشتر، به فصل ۱ استاندارد یونی کد مراجعه کنید.



## ساختار عمومی استاندارد یونی کد

استاندارد یونی کد، به هر نویسه عددی یکتایی از ۰ تا ۱۱۱۴'۱۱۱۴ اختصاص می‌دهد. این محدوده به ۱۷ صفحه‌ی ۶۵'۵۳۶ نویسه‌ای تقسیم می‌شود. صفحه‌ی اول صفحه‌ی پایه نام دارد و اکثر نویسه‌های مورد استفاده در زبان‌های زنده‌ی دنیا را در بر می‌گیرد. یونی کد بیش از صد هزار نویسه را نیز برای استفاده‌ی خصوصی مشخص می‌کند که می‌تواند برای ذخیره‌سازی داخلی، یا با توافق طرفین برای تبادل اطلاعات به کار رود. یونی کد به هر نویسه نام یکتا و مشخصی تخصیص می‌دهد که معنا یا شکل نویسه را مشخص می‌کند. به علاوه، برای هر نویسه ویژگی‌های الزامی یا اطلاعاتی‌ای را مشخص می‌کند که معنای آن نویسه را معین می‌کنند.

## شیوه‌ی تهیه‌ی این استاندارد

استاندارد حاضر برای تبادل خط فارسی (صورت نوشتاری زبان فارسی) تهیه شده است و قصد مشخص کردن شیوه‌ی مرجعی برای نگارش، یا محدود کردن دایره‌ی نویسه‌های این خط را ندارد. بلکه تلاش شده است کلیه‌ی نویسه‌های مورد استفاده‌ی روزمره در متون فارسی یا متون شامل نقل قول‌های مذهبی، در صورتی که در استاندارد یونی کد موجود باشند، در این استاندارد ذکر شوند. کمیته‌ی فنی این استاندارد نهایت تلاش ممکن را برای اطمینان از سازگاری کامل این استاندارد با استاندارد یونی کد انجام داده است.

یادآوری – پس از اولین جلسه‌ی کمیته‌ی ملی استاندارد برای تصویب استاندارد حاضر، نسخه‌ی 3.2 استاندارد یونی کد در تاریخ ۱۳۸۱/۱/۷ منتشر شد. استاندارد حاضر با نسخه‌ی اخیر استاندارد یونی کد نیز کاملاً سازگار است. نسخه‌ی اخیر، به درخواست شورای عالی انفورماتیک کشور، نویسه‌ای نیز برای علامت «ریال»، با کد U+FD۴C، در نظر گرفته است. شیوه‌ی صحیح استفاده از این نویسه در ضمیمه‌ای بر استاندارد حاضر منتشر خواهد شد.

# فناوریِ اطلاعات – تبادل و شیوهی نمایشِ اطلاعاتِ فارسی بر اساسِ یونی‌کُد

## ۱ هدف و دامنه‌ی کاربرد

هدف از تدوین این استاندارد تعیین شیوهی استفاده‌ی صحیح از دو استاندارد یونی‌کُد و ISO/IEC 10646 برای متون فارسی و قسمت‌های فارسی متون چندزبانه است که در سیستم‌های کامپیوتری، یا هرگونه سیستم دیگری که توانایی‌های پردازشی مورد نیاز در این استاندارد را داشته باشد، به کار می‌رود.

این استاندارد در نمایش، انتقال، تبادل، پردازش، ذخیره‌سازی، ورود، و ارائه‌ی صورت نوشتاری زبان فارسی و نمادهای لازم برای آن به کار می‌رود.

این استاندارد، شیوهی صحیح تبادل و نمایش اطلاعات فارسی را بر اساس استانداردهای همگام یونی‌کُد و ISO/IEC 10646 مشخص می‌کند.

این استاندارد:

- نام، معنی و کُد متناظر با نویسه‌های مورد استفاده در خط فارسی را مشخص می‌کند،
- شیوهی نمایش نویسه‌ها را در متون دوجهته، و شیوهی اتصال حروف فارسی را مشخص می‌کند،
- قالب‌های مختلف یونی‌کُد و ISO/IEC 10646 را برای تبادل داده‌ها مشخص می‌کند،
- شیوهی معین کردن انتهای سطرها و بندها را مشخص می‌کند،
- شیوهی مقایسه‌ی رشته‌های نویسه‌ای را از نظر هم‌ارزی مشخص می‌کند.

بعضی از مسائلی که این استاندارد به آن‌ها نمی‌پردازد

استاندارد حاضر به موارد زیر نمی‌پردازد:

- شیوه‌های واردسازی داده‌ها
- مرتب‌سازی عبارات فارسی و چندزبانه

- شیوه‌ی سطرشکنی و سطربندی متون
- شیوه‌ی ویرایش، درج، و حذف زیرمتن‌ها
- فشرده‌سازی متون، یا مبادله‌ی آن‌ها به صورت کم‌حجم
- مشخص کردن زبان متون و زیرمتن‌ها

یادآوری ۱ – نهایت تلاش ممکن صورت گرفته است تا آن‌چه که این استاندارد معین می‌کند، برخلاف استانداردهای یونی‌کد و ISO/IEC 10646 نباشد. در صورتی که ثابت شود پی‌روی از قسمت مشخصی از این استاندارد، کاربردها را با آن دو استاندارد ناسازگار می‌کند، یا تغییر دو استاندارد فوق‌الذکر در آینده باعث ناسازگاری شود، آن قسمت (و فقط آن قسمت) از این استاندارد باطل بوده و آن‌چه که در آن دو استاندارد مشخص شده جای‌گزین قسمت ناسازگار می‌شود. در چنین صورتی، برای سازگارشدن مجدد، ضمیمه‌هایی بر این استاندارد منتشر خواهد شد.

یادآوری ۲ – پی‌روی از این استاندارد هیچ‌گونه ناسازگاری‌ای با استفاده از یونی‌کد برای خط‌های دیگر از جمله عربی، اردو، کردی و ... ایجاد نمی‌کند. به علاوه، حروف مشترک این خطوط از کدهای مشترک استفاده می‌کنند. به عنوان مثال، حرف الف در همه‌ی این خط‌ها از کد یکسانی استفاده می‌کند.

## ۲ مراجع الزامی

مدارک الزامی زیر حاوی مقرراتی است که در متن این استاندارد به آن‌ها ارجاع داده شده است. بدین ترتیب آن مقررات، جزئی از این استاندارد محسوب می‌شود. در مورد مراجع دارای تاریخ چاپ و/یا تجدیدنظر، اصلاحیه‌ها و تجدیدنظرهای بعدی این مدارک مورد نظر نیست. مع‌هذا بهتر است کاربران ذی‌نفع این استاندارد، امکان کاربرد آخرین اصلاحیه‌ها و تجدیدنظرهای مدارک الزامی زیر را مورد بررسی قرار دهند. در مورد مراجع بدون تاریخ چاپ و/یا تجدیدنظر، آخرین چاپ و/یا تجدیدنظر آن مدارک الزامی ارجاع داده شده مورد نظر است.

استفاده از مراجع زیر برای کاربرد این استاندارد الزامی است:

1. The Unicode Consortium, *The Unicode Standard*, available from  
<http://www.unicode.org/>

2. Mark Davis, “The Unicode Standard Annex #9, The Bidirectional Algorithm”, available from <http://www.unicode.org/unicode/reports/tr9/>
3. Mark Davis, “The Unicode Standard Annex #13, Unicode Newline Guidelines”, available from <http://www.unicode.org/unicode/reports/tr13/>
4. Mark Davis, Martin Dürst, “The Unicode Standard Annex #15, Unicode Normalization Forms”, available from <http://www.unicode.org/unicode/reports/tr15/>
5. François Yergeau, “UTF-8, a transformation format of ISO 10646”, RFC 2279, January 1998, available from <http://www.ietf.org/rfc/rfc2279.txt>
6. Paul Hoffman, François Yergeau, “UTF-16, an encoding of ISO 10646”, RFC 2781, February 2000, available from <http://www.ietf.org/rfc/rfc2781.txt>
7. Mark Davis, “Unicode Standard Annex #19, UTF-32”, available from <http://www.unicode.org/unicode/reports/tr19/>

## ۳ اصطلاحات و تعاریف

در این استاندارد اصطلاحات و/یا واژه‌ها با تعاریف زیر به کار می‌رود:

### ۱-۳ متن

در این استاندارد عموماً منظور از «متن» متن کدشده یا ذخیره‌شده روی کامپیوتر است. در برابر text به کار می‌رود.

### ۲-۳ خط

«خط» مجموعه‌ای از نمادها است که می‌توان با آن کلمات یک یا چند زبان را نشان داد. در برابر script به کار می‌رود.

### ۳-۳ نویسه

«نویسه» کوچکترین واحد متن نوشته شده است، مستقل از شکل آن. در برابر character به کار می رود.

### ۳-۴ مجموعه نویسه

«مجموعه نویسه» مجموعه ای از نویسه هایی است که برای ارائه ی اطلاعات نوشتاری استفاده شوند. در برابر character set به کار می رود.

### ۳-۵ شکل

«شکل» صورت نمایشی نویسه در یک زمینه ی خاص است. نویسه ها می توانند شکل های متعددی داشته باشند. در برابر glyph به کار می رود.

### ۳-۶ متن ساده

«متن ساده» متنی است که شامل اطلاعات ساختاری یا ارائه ای نیست. در برابر plain text به کار می رود.

### ۳-۷ کدگذاری کردن

«کدگذاری کردن» اختصاص یک به یک کدها به نویسه ها است. در برابر encode به کار می رود.

## ۴ نمادها

در متن این استاندارد از نمادهای زیر استفاده شده است:

عدد یا کُد متناظر با نویسه های یونی کد به شکل  $U+n$  مشخص می شود، که در آن  $n$  یک عدد چهار تا شش رقمی در مبنای شانزده است، و از ارقام لاتین 0 تا 9، و حروف لاتین A تا F (جایگزین ۱۰ تا ۱۵) استفاده می کند. عدد  $n$  نباید با صفر شروع شود، مگر این که کمتر از چهار رقم داشته باشد. مثلاً:  $U+0001$ ،  $U+0012$ ،  $U+0123$ ،  $U+1234$ ،  $U+12345$  و  $U+102345$ . در جدول ها ممکن است برای اختصار  $U+$  حذف شود.

مثال -  $U+066B$  کُد یونی کد نویسه ی «ممیز فارسی» است.

محدوده‌ای از نویسه‌های یونی‌کد به شکل  $U+y..U+x$  یا  $x..y$  مشخص می‌شود، که در آن  $x$  و  $y$  اولین و آخرین نویسه‌های محدوده‌اند و نقطه‌ها نمایانگر محدوده‌ی پیوسته‌ای از نویسه‌ها، که شامل دو نویسه‌ی اول و آخر فهرست نیز می‌شود.

مثال -  $U+0900..U+097F$  شامل ۱۲۸ کُد یونی‌کد است.

دنباله‌ی دو یا چند کُد یونی‌کد با ویرگول لاتین جدا شده و به شکل  $\langle U+x, U+y, \dots, U+z \rangle$  مشخص می‌شود. ترتیب نویسه‌ها در نمادگذاری فوق از چپ به راست است. نماد  $U+$  اختیاری است. استاندارد حاضر به کلیه‌ی نویسه‌هایی که تعریف می‌کند نامی یکتا اختصاص داده است. این نام‌ها لزوماً ترجمه‌ی دقیق نام انگلیسی نویسه‌های استانداردهای یونی‌کد و ISO/IEC 10646 نیستند، بلکه بر اساس کاربرد آن نویسه‌ها در کاربردهای فارسی انتخاب شده‌اند. در این نام‌ها فقط از حروف و نشانه‌های زبان فارسی استفاده شده است.

## ۵ نویسه‌های مورد استفاده در متون فارسی

این بخش نویسه‌هایی را در بر می‌گیرد که در این استاندارد معنای مشخصی به آن‌ها تخصیص داده شده است. اگر کاربردی از نویسه‌ای که در این بخش آمده پشتیبانی کند، باید این نویسه را دقیقاً بر مبنای معنای ذکر شده در این استاندارد تفسیر یا تولید کند.

پشتیبانی این نویسه‌ها اجباری است، مگر نویسه‌هایی که با علامت ستاره مشخص شده‌اند. پشتیبانی نویسه‌های ستاره‌دار اختیاری است، ولی در صورت پشتیبانی شدن، آن‌ها نیز باید بر مبنای معنای ذکر شده در این استاندارد تفسیر یا تولید شوند.

در صورتی که کاربردها نیاز به نویسه‌های دیگری نیز داشته باشند، این استفاده باید دقیقاً بر اساس معنای تعریف شده در استاندارد یونی‌کد صورت گیرد.

یادآوری ۱ - از آن‌جا که استاندارد ISO/IEC 10646 معنای چندان مشخصی به نویسه‌ها

تخصیص نمی‌دهد، سازگار بودن با آن استاندارد کافی نیست و استفاده از نویسه‌های دیگر باید با استاندارد یونی‌کد نیز سازگار باشد.

یادآوری ۲ - شکل مشخص شده برای نویسه‌ها در این استاندارد فقط جنبه‌ی اطلاعاتی دارد و

مگر در مواردی که خلاف آن ذکر شده باشد، نویسه‌ها مجازند بسته به قلم مورد استفاده، به هر شکلی که نمایانگر آن نویسه باشد، نمایش داده شوند. حتی ممکن است کاربردها برای نمایش نویسه‌ها از خطی مانند بریل که شباهتی به خط فارسی ندارد استفاده کنند.

## ۱-۵ نویسه‌های کنترلی

جدول ۱ - نویسه‌های کنترلی

کُد	نام نویسه	علامت اختصاری
000A	سطر بعد	LF
000D	سر سطر	CR
200C	فاصله‌ی مجازی	ZWNJ
200D	اتصال مجازی	ZWJ
200E	نشانه‌ی چپ‌به‌راست	LRM
200F	نشانه‌ی راست‌به‌چپ	RLM
2028*	جداکننده‌ی سطرها	LS
2029*	جداکننده‌ی بندها	PS
202A*	زیرمتن چپ‌به‌راست	LRE
202B*	زیرمتن راست‌به‌چپ	RLE
202C*	پایان زیرمتن	PDF
202D*	زیرمتن اکیداً چپ‌به‌راست	LRO
202E*	زیرمتن اکیداً راست‌به‌چپ	RLO
FEFF	نشانه‌ی ترتیب بایت‌ها	BOM

یادآوری ۱ - در صورتی که متن در قالب UTF-8 باشد، برای جداکردن سطرها و بندها باید بسته به بستر کاربرد از LF، CR، یا (LF, CR) استفاده شود. استفاده از LS و PS در متون با قالب UTF-8 مجاز نیست. برای اطلاع از شیوه‌ی صحیح استفاده از این نویسه‌ها، به پیوست ت مراجعه کنید.

یادآوری ۲ – نویسه‌های «فاصله‌ی مجازی» و «اتصالِ مجازی» در الگوریتمِ اتصالِ فارسی به کار می‌روند. برای اطلاعاتِ بیشتر به پیوستِ ب مراجعه کنید.

یادآوری ۳ – نویسه‌های LRO، PDF، RLE، LRE، RLM، LRM و RLO در الگوریتمِ دوجهته به کار می‌روند. برای اطلاعاتِ بیشتر به پیوستِ الف مراجعه کنید.

یادآوری ۴ – نویسه‌ی BOM باید برای تمییز متونی که در صورتِ عدمِ وجود این نویسه ممکن است اشتباه پردازش شوند، به کار برده شود. استفاده از این نویسه در ابتدای پرونده‌های UTF-16 و UTF-32 توصیه می‌شود ولی در ابتدای پرونده‌های UTF-8 که ترتیب بایت‌ها معنی ندارد شدیداً نهی می‌شود. استفاده از این نویسه برای مقاصدِ دیگر مجاز نیست. برای اطلاعاتِ بیشتر به پیوستِ پ مراجعه کنید.

## ۲-۵ علائمِ نقطه‌گذاریِ مشترک

### جدول ۲ – علائمِ نقطه‌گذاریِ مشترک

شکلِ نمایشی	نامِ نویسه	کُد
	فاصله	0020
.	نقطه	002E
:	دونقطه	003A
!	علامتِ تعجب	0021
...	سه‌نقطه‌ی افقی	2026*
-	خطِ تیره	2010*
—	تیره‌منها	002D
	خطِ عمودی	007C
/	خطِ اریب	002F
\	خطِ اریبِ وارو	005C
*	ستاره	002A



ادامه‌ی جدول ۲ - علائم نقطه‌گذاری مشترک

شکلِ نمایشی	نامِ نویسه	کُد
)	پرانتز باز	0028
(	پرانتز بسته	0029
]	کروشه باز	005B
[	کروشه بسته	005D
}	آکولاد باز	007B
{	آکولاد بسته	007D
»	گیومه باز	00AB
«	گیومه بسته	00BB

یادآوری ۱ - نویسه‌ی «خطِ اریب» عمدتاً برای جداسازی اجزای تاریخ یا به‌عنوان علامتِ کسر در داخلِ متن به کار می‌رود. این نویسه با «ممیزِ فارسی» (U+066B) تفاوت دارد:

$$1/4 = \frac{1}{4} = 0.25$$

$$1/4 = 1 + \frac{4}{10}$$

کاربردها موظفند از شکل‌های نمایشی مختلفی برای این دو نویسه استفاده کنند (مگر در مواردی که محدودیت‌های خاصِ نمایشی وجود دارد، مانند دستگاه‌های تلفن همراه).

یادآوری ۲ - نویسه‌های جفتی، از قبیل پرانتزها و قلاب‌ها، بسته به موقعیت خود در متن، شکل‌های مختلفی می‌پذیرند. مثلاً «پرانتز باز» (U+0028) در متون راست‌به‌چپ به شکل «(» و در متون چپ‌به‌راست به شکل «(» ظاهر می‌شود. مشروح این رفتار در پیوستِ الف آمده است.

یادآوری ۳ - نویسه‌ی «تیره‌منها» فقط در مواردی به کار می‌رود که تفکیکِ «خطِ تیره» (U+2010) از «علامتِ منها» (U+2212) ممکن نباشد، مثلاً هنگامی که داده‌ها از قالبِ دیگری که این دو نویسه را متمایز نمی‌داند به قالبِ یونی‌کد تبدیل شده باشند. در صورتِ مشخص بودن معنای نویسه، باید از نویسه‌های دقیق یعنی U+2010 یا U+2212 استفاده کرد.

### ۳-۵ علائم نقطه گذاری فارسی

جدول ۳ - علائم نقطه گذاری فارسی

شکل نمایشی	نام نویسه	کُد
،	ویرگول فارسی	060C
؛	نقطه ویرگول فارسی	061B
؟	علامت سؤال فارسی	061F
-	کشیدگی فارسی	0640

### ۴-۵ ارقام و علائم ریاضی

جدول ۴ - ارقام و علائم ریاضی

شکل نمایشی	نام نویسه	کُد
۰	رقم فارسی صفر	06F0
۱	رقم فارسی یک	06F1
۲	رقم فارسی دو	06F2
۳	رقم فارسی سه	06F3
۴	رقم فارسی چهار	06F4
۵	رقم فارسی پنج	06F5
۶	رقم فارسی شش	06F6
۷	رقم فارسی هفت	06F7
۸	رقم فارسی هشت	06F8
۹	رقم فارسی نه	06F9
/	ممیز فارسی	066B
،	جداکننده‌ی هزارهای فارسی	066C

ادامه‌ی جدول ۴ - ارقام و علائم ریاضی

شکلِ نمایشی	نامِ نویسه	کُد
%	درصدِ فارسی	066A
+	علامتِ به‌اضافه	002B
-	علامتِ منها	2212*
×	علامتِ ضرب	00D7
÷	علامتِ تقسیم	00F7*
<	علامتِ کوچکتر	003C
=	علامتِ مساوی	003D
>	علامتِ بزرگتر	003E

یادآوری - «علامت کوچکتر» و «علامت بزرگتر»، بسته به موقعیت خود در متن، شکل‌های مختلفی می‌گیرند. مشروح این رفتار در پیوست الف آمده است. شکل این نویسه‌ها در جدول فوق با توجه به زمینه‌ی معمول آن‌ها، یعنی در میان اعداد، آمده است.

۵-۵ حروف اصلی فارسی

جدول ۵ - حروف اصلی فارسی

شکلِ نمایشی	نامِ نویسه	کُد
ء	حرفِ فارسیِ همزه	0621
آ	حرفِ فارسیِ آ	0622
ا	حرفِ فارسیِ الف	0627
أ	حرفِ فارسیِ الف با همزه‌ی بالا	0623
ب	حرفِ فارسیِ ب	0628
پ	حرفِ فارسیِ پ	067E

ادامه‌ی جدول ۵ - حروف اصلی فارسی

شکلِ نمایشی	نامِ نویسه	کُد
ت	حرفِ فارسیِ تِ	062A
ث	حرفِ فارسیِ ثِ	062B
ج	حرفِ فارسیِ جیمِ	062C
چ	حرفِ فارسیِ چِ	0686
ح	حرفِ فارسیِ حِ	062D
خ	حرفِ فارسیِ خِ	062E
د	حرفِ فارسیِ دالِ	062F
ذ	حرفِ فارسیِ ذالِ	0630
ر	حرفِ فارسیِ رِ	0631
ز	حرفِ فارسیِ زِ	0632
ژ	حرفِ فارسیِ ژِ	0698
س	حرفِ فارسیِ سینِ	0633
ش	حرفِ فارسیِ شینِ	0634
ص	حرفِ فارسیِ صادِ	0635
ض	حرفِ فارسیِ ضادِ	0636
ط	حرفِ فارسیِ طا	0637
ظ	حرفِ فارسیِ ظا	0638
ع	حرفِ فارسیِ عینِ	0639
غ	حرفِ فارسیِ غینِ	063A
ف	حرفِ فارسیِ فِ	0641
ق	حرفِ فارسیِ قافِ	0642

ادامه‌ی جدول ۵ - حروف اصلی فارسی

شکل نمایشی	نام نویسه	کُد
ک	حرف فارسی کاف	06A9
گ	حرف فارسی گاف	06AF
ل	حرف فارسی لام	0644
م	حرف فارسی میم	0645
ن	حرف فارسی نون	0646
و	حرف فارسی واو	0648
ؤ	حرف فارسی واو با همزه‌ی بالا	0624
ه	حرف فارسی هـ	0647
ی	حرف فارسی ی	06CC
ئ	حرف فارسی ی با همزه‌ی بالا	0626

یادآوری ۱ - بعضی از نویسه‌های جدول فوق را می‌توان به صورت دو نویسه نیز مبادله کرد. مثلاً «حرف فارسی آ» را می‌توان هم به صورت U+0622 و هم به صورت {U+0627, U+0653} مبادله کرد. در این موارد، شکل تک‌نویسه‌ای مرجح است. برای اطلاع دقیق‌تر در این باره، به پیوست ۳ مراجعه کنید.

یادآوری ۲ - حروف فارسی شکل‌های مختلفی به خود می‌پذیرند، مثلاً «حرف فارسی عین» به شکل‌های «ع»، «ع»، «ع» و «ع» دیده می‌شود. این شکل‌ها در الگوریتم اتصال فارسی تعیین می‌شوند. این الگوریتم در پیوست ۳ تشریح شده است. شکل‌هایی که در جدول فوق آمده‌اند فقط جنبه‌ی اطلاعاتی دارند.

جدول ۶ - حروف فرعی

شکل نمایشی	نام نویسه	کُد
اِ	حرف الف با همزه‌ی پایین	0625*
آ	حرف الف وصل	0671*
ك	حرف کاف عربی	0643*
ة	حرف ت گرد	0629
ي	حرف ی عربی نقطه‌دار	064A*
ی	حرف ی عربی بی نقطه	0649*

یادآوری ۱ - استفاده از «حرف کاف عربی» به جای «حرف فارسی کاف» و استفاده از «حرف ی عربی نقطه‌دار» یا «حرف ی عربی بی نقطه» به جای «حرف فارسی ی» به هیچ عنوان مجاز نیست. تنها در صورتی می‌توان از این نویسه‌ها استفاده کرد که شکل خاص آن‌ها مورد نظر بوده، یا متن به زبان عربی، اردو و امثال آن‌ها باشد. کاربردها موظفند این نویسه‌ها را به شکل درست آن‌ها نمایش دهند. «حرف ی عربی نقطه‌دار» هیچ‌گاه نباید بدون نقطه نمایش داده شود.

«حرف کاف عربی» در شکل‌های اول و وسط مانند «حرف فارسی کاف» است، اما در شکل‌های آخر و تنها بدون سرکش و به همراه علامتی شبیه همزه ظاهر می‌شود. «حرف ی عربی نقطه‌دار» در شکل‌های اول و وسط مانند «حرف فارسی ی» است، اما در شکل‌های آخر و تنها با دونقطه در زیرش ظاهر می‌شود. «حرف ی عربی بی نقطه» در شکل‌های آخر و تنها مانند «حرف فارسی ی» است، اما در شکل‌های اول و وسط بدون نقطه ظاهر می‌شود.

جدول ۷ - نشانه‌های فارسی

شکلِ نمایشی	نام نویسه	کُد
◌َ	زیرِ فارسی (فتحه)	064E
◌ِ	زیرِ فارسی (کسره)	0650
◌ُ	پیشِ فارسی (ضمه)	064F
◌̇	دوزیرِ فارسی (تنوینِ نصب)	064B
◌̈	دوزیرِ فارسی (تنوینِ جر)	064D
◌̉	دوپیشِ فارسی (تنوینِ رفع)	064C
◌̊	تشدیدِ فارسی	0651
◌̋	ساکنِ فارسی	0652
◌̌	مدِ فارسی	0653*
◌̍	همزه‌ی فارسیِ بالا	0654
◌̎	همزه‌ی فارسیِ پایین	0655*
◌̏	الفِ مقصوره‌ی فارسی	0670

یادآوری ۱ - نویسه‌های فوق خاصیتِ ترکیب‌شونده دارند و باید بر حسبِ مورد، بالا یا زیرِ نویسه‌ی قبل از خود نمایش داده شوند. در موردِ تأثیر این نویسه‌ها بر الگوریتمِ اتصال، به پیوستِ ب مراجعه کنید.

یادآوری ۲ - در صورتی که نویسه‌های «همزه‌ی فارسیِ بالا» و «همزه‌ی فارسیِ پایین» روی «حرفِ فارسیِ ی»، یا «حرفِ یِ عربیِ نقطه‌دار» بیایند، نویسه‌ی کرسی نقطه‌های خود را از دست می‌دهد.

یادآوری ۳ - کاربردها می‌توانند برای نمایشِ ترکیبِ نشانه‌ها از شکل‌های خاص استفاده کنند. مثلاً برای ترکیبِ «تشدیدِ فارسی» و «زیرِ فارسی» بهتر است به جای نمایش دادنِ «زیرِ فارسی» در زیرِ حرفِ کرسی، آن را در زیرِ «تشدیدِ فارسی» نمایش داد.

## ۵-۸ نویسه‌های ممنوع

این نویسه‌ها نباید در متون فارسی استفاده شوند. استفاده از آن‌ها در متون زبان‌های دیگر مانند عربی و اردو باید بر اساس تعریف موجود در استاندارد یونی‌کد صورت گیرد.

جدول ۸ - نویسه‌های ممنوع

کُد	نام نویسه	شکل نمایشی
06C0	حرف و اردو با همزه‌ی بالا	ۀ
0660	رقم صفر عربی	۰
0661	رقم یک عربی	۱
0662	رقم دو عربی	۲
0663	رقم سه عربی	۳
0664	رقم چهار عربی	۴
0665	رقم پنج عربی	۵
0666	رقم شش عربی	۶
0667	رقم هفت عربی	۷
0668	رقم هشت عربی	۸
0669	رقم نه عربی	۹

یادآوری ۱ - نام نویسه‌های جدول فوق استاندارد نیست و فقط جنبه‌ی اطلاعاتی دارد. این استاندارد به نویسه‌های جدول فوق نامی اختصاص نمی‌دهد.

یادآوری ۲ - نویسه‌ی U+06C0 نباید به هیچ عنوان برای متون فارسی استفاده شود. برای نوشتن عباراتی مثل «خانه ما» باید از نویسه‌ی «حرف فارسی ه» (U+0647) به همراه «همزه‌ی فارسی بالا» (U+0654) استفاده شود. کاربردها می‌توانند در صورتی که در متون فارسی به این نویسه برخوردند آن را بسته به مورد با {0647, 0654} یا {0647, 0654, 200C} جای‌گزین کنند. در صورتی که متن یا زیرمتن به زبان فارسی نباشد، این جای‌گزینی نباید صورت بگیرد.



یادآوری ۳ – استفاده از ارقام عربی (U+0660..U+0669) مگر در مواردی که کاربرد بخواهد میان ارقام فارسی و عربی تمایز قائل شود مجاز نیست. کاربردهایی که بخواهند ارقام عربی را پشتیبانی کنند بهتر است میان شکل ارقام چهار، پنج، و شش فارسی و عربی تمایز قائل شوند. باید دقت شود که ارقام فارسی و عربی از لحاظ جهت‌پذیری در الگوریتم دوجهته‌ی یونی‌کد تفاوت دارند.

#### ۵-۹ نویسه‌های منسوخ

کلیه‌ی نویسه‌هایی که در استاندارد یونی‌کد به‌عنوان منسوخ مشخص شده‌اند، در این استاندارد نیز منسوخ تلقی می‌شوند. کاربردها نباید این نویسه‌ها را تولید کنند، و در صورت برخوردن به آن‌ها می‌توانند از آن‌ها چشم‌پوشی کرده، یا آن‌ها را براساس آخرین نسخه‌ی استاندارد یونی‌کد تفسیر کنند.

# پیوست الف

## الگوریتم دوجهته

### (الزامی)

به علت تفاوت جهت نوشتن خط‌های فارسی و لاتین، و از آن‌جا که متون فارسی شامل اعداد و علائم ریاضی، یا متون چندزبانه، در هنگام پردازش با ابهام روبه‌رو می‌شوند، استاندارد یونی‌کد نویسه‌های این متون را به ترتیب معنایی، یعنی ترتیبی که نویسه‌ها از ذهن خواننده‌ی متن می‌گذرند کدگذاری می‌کند. الگوریتم دوجهته برای تبدیل این ترتیب به یک ترتیب قابل نمایش به کار می‌رود. در کاربردهای مبتنی بر این استاندارد، نویسه‌ها باید به ترتیب معنایی مبادله شوند. بنابراین برای نمایش اطلاعات فارسی، ممکن است لازم باشد رشته‌ی نویسه‌های ورودی به ترتیب دیداری تبدیل شود. شیوه‌ی انجام این تبدیل، باید دقیقاً از ضمیمه‌ی ۹ استاندارد یونی‌کد پی‌روی کند. کاربردهایی که از ضمیمه‌ی فوق‌الذکر پی‌روی نکنند، با این استاندارد سازگار نیستند.

یادآوری ۱ – شکل بعضی از نویسه‌ها، از جمله «پرانتر باز»، به نویسه‌های اطراف خود بستگی دارد. این نویسه‌ها در صورتی که در متون با جهت مخالف قرار گیرند، به اصطلاح قرینه می‌شوند. کاربردها باید قرینه‌سازی را پشتیبانی کنند. مشروح این رفتار در ضمیمه‌ی ۹ استاندارد یونی‌کد آمده است.

یادآوری ۲ – در کدگذاری متون دوجهته، مواردی پیش می‌آید که ترتیب دیداری ضمنی‌ای که از نویسه‌ها به دست می‌آید مطلوب نیست. در این حالت‌ها می‌توان از نویسه‌های کنترل جهت جدول ۱ بخش ۵-۱ استفاده کرد. این نویسه‌ها فقط برای تصحیح ترتیب نمایش متن به کار می‌روند و باید در پردازش‌های دیگر (مانند مرتب‌سازی متن یا جستجوی عبارات) نادیده گرفته شوند. برای اطلاعات بیشتر به ضمیمه‌ی ۹ استاندارد یونی‌کد مراجعه کنید.

## پیوست ب

### الگوریتم اتصال

#### (الزامی)

از آنجا که حروف فارسی، بسته به حروف قبل و بعد از خود اشکال مختلفی می‌گیرند، در صورتی که کاربردها بخواهند نویسه‌ها را با خط فارسی نمایش دهند، برای نمایش اطلاعات فارسی و انتخاب شکل مناسب، باید از الگوریتم مشخص شده در این پیوست استفاده کنند. این الگوریتم، حداقل تغییر شکل مورد نیاز را برای نمایش متون فارسی مشخص می‌کند، ولی ممکن است، بسته به کاربرد، از الگوریتم پیچیده‌تری نیز استفاده شود (مثلاً در کاربردهایی که متن را با خط نستعلیق نمایش می‌دهند). الگوریتم اتصال باید، با در نظر گرفتن نویسه‌های شفاف، پس از الگوریتم دوجهته انجام شود (یا خروجیش با حالتی که این الگوریتم پس از الگوریتم دوجهته انجام می‌شود یکسان باشد).

#### ب-۱ رده‌ی اتصال

هر نویسه، در یکی از رده‌های اتصال دسته‌بندی می‌شود. این رده‌ها، شیوه‌ی تغییر شکل نویسه و تأثیر آن را بر نویسه‌های دیگر مشخص می‌کنند. این رده‌ها به شرح زیرند:

- راست‌وصل: نویسه‌های دوشکلی از قبیل «آ»، «الف»، «دال»، «ر»، «واو»، و «ت‌گرد». با حرف R مشخص می‌شوند.
- دووصل: نویسه‌های چهارشکلی از قبیل «ب»، «جیم»، «سین»، و «صاد». با حرف D مشخص می‌شوند.
- واصل: نویسه‌های ایجادکننده‌ی اتصال، از قبیل «کشیدگی» و «اتصال مجازی». تفاوت این نویسه‌ها با نویسه‌های «دووصل» این است که تغییر شکل نمی‌دهند. با حرف C مشخص می‌شوند.

• فاصل: نویسه‌های قطع‌کننده‌ی اتصال، شامل «فاصله‌ی مجازی» و کلیه‌ی نویسه‌های غیر ترکیب‌شونده که در دسته‌بندی فوق قرار نمی‌گیرند، از قبیل «همزه»، فاصله‌ها، ارقام، علائم نقطه‌گذاری، و حروف خط‌های لاتین، یونانی و غیره. با حرف U مشخص می‌شوند.

• شفاف: نویسه‌های شفاف نسبت به اتصال، شامل نویسه‌های ترکیب‌شونده و کنترلی، از قبیل «زیر»، «دوزیر»، «سکون»، «تشدید»، «الف مقصوره»، و «نشانه‌ی راست‌به‌چپ». با حرف T مشخص می‌شوند.

در این پیوست، از اصطلاح «متصل‌به‌چپ» برای نویسه‌های «دووصل» و «واصل»؛ و از اصطلاح «متصل‌به‌راست» برای نویسه‌های «راست‌وصل»، «دووصل»، و «واصل» استفاده می‌شود. رده‌ی نویسه‌ها، باید بر اساس پرونده‌ی ArabicShaping.txt از پرونده‌های داده‌ای یونی‌کد، که آخرین نسخه‌ی آن در نشانی اینترنتی

<http://www.unicode.org/Public/UNIDATA/ArabicShaping.txt>

در دسترس است تعیین شود.

یادآوری – نویسه‌های «فاصله‌ی مجازی» و «اتصال مجازی» برای تغییر شکل نویسه‌ها به‌کار می‌روند. از این دو نویسه برای ممانعت از اتصال دو حرف مجاور (مثلاً در کلمه‌ی «خانه‌ها»)، یا انتخاب شکلی غیر از شکل معمول حروف (مثلاً در «ه.ش.»، به‌عنوان مخفف «هجری شمسی») استفاده می‌شود.

## ب-۲ الگوریتم

برای نویسه‌ها، بسته به رده‌ی اتصالشان، تا چهار شکل تعیین می‌شود. این شکل‌ها در اصطلاح «اول»، «وسط»، «آخر»، و «تنها» نامیده می‌شوند. نویسه‌های «راست‌وصل» فقط دو شکل «آخر» و «تنها» را می‌گیرند.

الگوریتم اتصال به شرح زیر است:

۱. نویسه‌های «شفاف» رفتار اتصالی نویسه‌های پایه را تغییر نمی‌دهند. (بنابراین از این به بعد، منظور از نویسه‌ی سمت راست، اولین نویسه‌ی غیر شفاف سمت راست خواهد بود؛ همین‌طور در مورد نویسه‌ی سمت چپ.)

۲. اگر نویسه‌ای «راست‌وصل» باشد، و نویسه‌ی سمتِ راستش «متصل‌به‌چپ» باشد، به شکل «آخر» در می‌آید.

۳. اگر نویسه‌ای «دووصل» باشد، نویسه‌ی سمتِ راستش «متصل‌به‌چپ» باشد، و نویسه‌ی سمتِ چپش «متصل‌به‌راست» باشد، به شکل «وسط» در می‌آید.

۴. اگر نویسه‌ای «دووصل» باشد، نویسه‌ی سمتِ راستش «متصل‌به‌چپ» باشد، و نویسه‌ی سمتِ چپش «متصل‌به‌راست» نباشد، به شکل «آخر» در می‌آید.

۵. اگر نویسه‌ای «دووصل» باشد، نویسه‌ی سمتِ راستش «متصل‌به‌چپ» نباشد، و نویسه‌ی سمتِ چپش «متصل‌به‌راست» باشد، به شکل «اول» در می‌آید.

۶. در صورتی که هیچ‌یک از حالت‌های فوق برقرار نباشند، نویسه به شکل «تنها» در می‌آید.

یادآوری ۱ – در صورتی که نویسه‌ای اولین نویسه‌ی غیر «شفاف» سطر یا بند خود باشد، نویسه‌ی سمتِ راستش «فاصل» فرض می‌شود. همین‌طور، در صورتی که نویسه‌ای آخرین نویسه‌ی غیر «شفاف» سطر یا بند خود باشد، نویسه‌ی سمتِ چپش «فاصل» فرض می‌شود.

یادآوری ۲ – از آن‌جا که این الگوریتم پس از الگوریتم دوجهته انجام می‌شود، نویسه‌های سمتِ راست و چپ بر اساس ترتیب دیداری تعیین می‌شوند.

### ب-۳ گروه اتصال

هر یک از حروفِ فارسی‌ای که شکل‌های مختلف می‌گیرند، بسته به شکلِ ظاهری در گروه‌های اتصال دسته‌بندی می‌شوند. این گروه‌ها نیز باید بر اساس پرونده‌ی ArabicShaping.txt از پرونده‌های داده‌ای یونی‌کد، که آخرین نسخه‌ی آن در نشانی اینترنتی

<http://www.unicode.org/Public/UNIDATA/ArabicShaping.txt>

در دسترس است تعیین شود.

بر اساس آخرین نسخه‌ی در دسترس در زمانِ تدوین این استاندارد، حروفِ شکل‌پذیری که در بخش‌های ۵-۵ و ۶-۵ آمده‌اند در این گروه‌ها قرار می‌گیرند:

جدول ۹ - فهرست گروه‌های اتصال حروف

نام گروه	نام لاتین گروه	رده‌ی اتصال	نویسه‌ها
الف	ALEF	راست وصل	«آ»، «الف»، «الف با همزه‌ی بالا»، «الف با همزه‌ی پایین»، «الف وصل»
ب	BEH	دو وصل	«ب»، «پ»، «ت»، و «ث»
ح	HAH	دو وصل	«جیم»، «ج»، «ح»، و «خ»
دال	DAL	راست وصل	«دال»، و «ذال»
ر	REH	راست وصل	«ر»، «ز»، و «ژ»
سین	SEEN	دو وصل	«سین»، و «شین»
صاد	SAD	دو وصل	«صاد»، و «ضاد»
طا	TAH	دو وصل	«طا»، و «ظا»
عین	AIN	دو وصل	«عین»، و «غین»
ف	FEH	دو وصل	«ف»
قاف	QAF	دو وصل	«قاف»
کاف عربی	KAF	دو وصل	«کاف عربی»
گاف	GAF	دو وصل	«گاف»، و «گاف»
لام	LAM	دو وصل	«لام»
میم	MEEM	دو وصل	«میم»
نون	NOON	دو وصل	«نون»
واو	WAW	راست وصل	«واو»، و «واو با همزه‌ی بالا»
ه	HEH	دو وصل	«ه»
تِ گرد	TEH MARBUTA	راست وصل	«تِ گرد»
ی	YEH	دو وصل	«ی»، «ی با همزه‌ی بالا»، «ی عربی نقطه‌دار»، و «ی عربی بی نقطه»

یادآوری - فهرست فوق فقط جنبه‌ی اطلاعاتی دارد. کاربردها موظفند به اطلاعات موجود در

پرونده‌ی ArabicShaping.txt مراجعه کنند.

## ب-۴ لیگاتورها

حروف فارسی می‌توانند بسته به قلم مورد استفاده، اشکال چندحرفی‌ای به نام لیگاتور بسازند. مثلاً ترکیب «لام» و «الف» می‌تواند به شکل «لا»، و ترکیب «ف» و «ی» می‌تواند به شکل «فی» بیاید. بعضی از لیگاتورها اختیاری و بعضی دیگر اجباری‌اند. لیگاتورهای اجباری، لیگاتورهایی هستند که حرف اولشان از گروه اتصال «لام» (LAM) و حرف دومشان از گروه اتصال «الف» (ALEF) باشد. لیگاتورهای اختیاری، لیگاتورهای دیگر هستند که بسته به قلم نمایشی ممکن است به شکل لیگاتور نمایش یابند. کاربردها موظفند در نمایش متون، لیگاتورهای اجباری را به شکل لیگاتور نمایش دهند، مگر در مواردی که جلوه‌های بصری خاص مورد نظر باشد، یا دستگاو نمایش محدودیت‌های ویژه‌ای داشته باشد.

برای اعمال این لیگاتورها، از الگوریتم زیر استفاده می‌شود:

۱. نویسه‌های «شفاف» رفتار لیگاتوری نویسه‌های پایه را تغییر نمی‌دهند.
۲. هر دنباله‌ی دوتایی از نویسه‌ها که نویسه‌ی سمت راستش در گروه «لام» و به شکل «وسط»، و نویسه‌ی سمت چپش در گروه «الف» و به شکل «آخر» باشد، لیگاتوری از دسته‌ی «لام‌الف» را به شکل «آخر» تشکیل می‌دهد.
۳. هر دنباله‌ی دوتایی از نویسه‌ها که نویسه‌ی سمت راستش در گروه «لام» و به شکل «اول»، و نویسه‌ی سمت چپش در گروه «الف» و به شکل «آخر» باشد، لیگاتوری از دسته‌ی «لام‌الف» را به شکل «تنها» تشکیل می‌دهد.
۴. هر گاه بین دو یا چند نویسه که به طور پیشفرض به هم متصل می‌شوند ولی لیگاتور نمی‌شوند، یک یا چند نویسه‌ی «اتصال مجازی» قرار گیرد، در صورت موجود بودن شکل لیگاتوری نویسه‌ها در قلم مورد استفاده برای نمایش، باید از شکل لیگاتوری استفاده شود.
۵. هر گاه بین دو یا چند نویسه، نویسه‌ی «فاصله‌ی مجازی» قرار گیرد، باید شکل عادی نویسه‌ها مورد استفاده قرار گیرد. مثلاً اگر دنباله‌ی «اتصال مجازی، فاصله‌ی مجازی، اتصال مجازی» بین «لام» و «الف» قرار گیرد، این دو حرف نباید لیگاتور شوند، بلکه باید به شکل «لا» نمایش یابند.

## پیوست پ

### قالب‌های تبادل داده‌ها

#### (الزامی)

در استاندارد یونی‌کد می‌توان از قالب‌های متعددی برای تبادل اطلاعات استفاده کرد. ولی استاندارد حاضر فقط به قالب‌های اصلی، یعنی UTF-8، UTF-16، و UTF-32، می‌پردازد. قالب‌های مشتق، مثلاً UTF-16LE، حالت خاصی از قالب اصلی نظیرشان (در این مثال UTF-16) فرض می‌شوند. کاربردهایی که در قالب‌های یونی‌کدی خروجی می‌دهند، یا ورودی قبول می‌کنند، موظفند در صورت استفاده از قالب‌های UTF-8، UTF-16، و UTF-32، به ترتیب از RFC 2279، RFC 2781، و ضمیمه ۱۹ استاندارد یونی‌کد پی‌روی کنند. در صورتی که کاربردها از قالب UTF-8 استفاده می‌کنند، بهتر است نویسه‌ی U+FEFF را در ابتدای خروجی تولید نکنند، ولی بهتر است در صورتی که این نویسه در ابتدای ورودی‌های در قالب UTF-8 بیاید، آن را به عنوان علامت مشخص‌کننده در نظر گرفته، و پردازشش نکنند.

یادآوری – کاربردها بهتر است یکی از صورت‌های نرمال مشخص شده در ضمیمه ۱۵ استاندارد یونی‌کد را انتخاب کرده و خروجی‌های خود را در آن قالب تولید کنند. (برای اطلاعات بیشتر به پیوست ت مراجعه کنید.)



## پیوست ت

### سטר بندی و پاراگراف بندی

#### (الزامی)

کاربردها موظفند نویسه‌های جداکننده‌ی سطرها و بندها را بر اساس توصیه‌های ضمیمه‌ی ۱۳ استاندارد یونی‌کد تفسیر کنند.

علاوه بر محدودیت‌های مشخص شده در ضمیمه‌ی فوق‌الذکر، در صورتی که کاربردی بخواهد متنی در قالب UTF-8 تولید کند، نباید از نویسه‌های «جداکننده‌ی سطرها» (U+2028) و «جداکننده‌ی بندها» (U+2029) استفاده کند. بلکه باید از علامت مخصوص جداکردن سطرها در بستر کاربرد استفاده کند که معمولاً U+000D، U+000A، یا (U+000D، U+000A) است.

## پیوست ث

### نرمال‌سازی و هم‌ارزی

#### (اطلاعاتی)

از آنجا که دنباله‌ای از حروف و نمادها می‌تواند به روش‌های مختلفی به رشته‌ای از نویسه‌ها تبدیل شود (مثلاً کلمه‌ی «مؤمن») را می‌توان هم با نویسه‌ی «واو با همزه‌ی بالا»، و هم با دنباله‌ی نویسه‌های «واو، همزه‌ی بالا» کُدگذاری کرد)، کاربردها بهتر است به‌منظور هماهنگی خروجی‌های خود، آن‌ها را به‌صورت یکی از صورت‌های نرمال یونی‌کد، که در ضمیمه‌ی ۱۵ استاندارد یونی‌کد توصیف شده است تولید کنند.

استاندارد حاضر، به‌عنوان قالب مرجح در تبادل داده‌ها، صورت نرمال C (Normalizaion Form C) و قالب UTF-8 را توصیه می‌کند. این انتخاب به علت فراگیر بودن این قالب، و توصیه‌شدن آن در استانداردهای کنسرسیوم World Wide Web، از جمله HTML و XML، صورت گرفته است. در «صورت نرمال C»، نویسه‌هایی که می‌توانند به چند شکل مختلف کُدگذاری شوند، به شکل تک‌نویسه‌ای کُدگذاری می‌شوند. به‌علاوه، ترتیب واحدی برای حالت‌هایی که چند علامت ترکیب‌شونده روی یک حرف کرسی قرار می‌گیرند تعیین می‌شود. از طرف دیگر، در صورتی که کاربردها صورت‌های نرمال مختلفی را پشتیبانی می‌کنند، بهتر است رشته‌های «هم‌ارز» را تشخیص دهند. برای اطلاع بیشتر، به فصل‌های ۲ و ۳ استاندارد یونی‌کد مراجعه کنید.

یادآوری – کاربردها می‌توانند لایه‌های بالاتری از «هم‌ارزی» را نیز پشتیبانی کنند، مثلاً هم‌ارزی ضعیف بین نویسه‌هایی مانند «کاف» و «کافِ عربی»، یا بین رشته‌نویسه‌های «ی، همزه‌ی بالا» و «یِ عربی نقطه‌دار، همزه‌ی بالا». این استاندارد به این گونه هم‌ارزی‌ها نمی‌پردازد.

## پیوست ج

### واژه‌نامه

#### (اطلاعاتی)

right-joining.....	راست وصل	presentation.....	ارائه
string.....	رشته	information.....	اطلاعات
embedding.....	زیرمتن	Cursive Joining Algorithm.....	الگوریتم اتصال
subtext.....	زیرمتن	Bidirectional Algorithm.....	الگوریتم دوجهته
data.....	داده	transfer.....	انتقال
insert.....	درج	ideographic.....	اندیشه‌نگار
device.....	دستگاه	octet/byte.....	بایت
dual-joining.....	دو وصل	platform.....	بستر
storage.....	ذخیره‌سازی	block.....	بلوک
conformant.....	سازگار	paragraph.....	بند
line.....	سطر	processing.....	پردازش
line-breaking.....	سطر شکنی	file.....	پرونده
transparent.....	شفاف	support.....	پشتیبانی
glyph.....	شکل	conform.....	پی‌روی کردن
final form.....	شکل آخر	interchange.....	تبادل
initial form.....	شکل اول	visual order.....	ترتیب دیداری
isolated form.....	شکل تنها	logical order.....	ترتیب معنایی
presentation form.....	شکل نمایشی	left-joining.....	چپ وصل
medial form.....	شکل وسط	delete.....	حذف
Basic Multilingual Plane.....	صفحه‌ی پایه	letter.....	حرف
sign.....	علامت	script.....	خط

character set .....	مجموعه‌نویسه	non-joining .....	فاصل
environment .....	محیط	compression .....	فشرده‌سازی
sort .....	مرتب‌سازی	encoding .....	قالب
symbol .....	نماد	transformation format .....	قالبِ تبادل
display .....	نمایش	mirroring .....	قرینه‌سازی
document .....	نوشتار	font .....	قلم
..... نویسه‌ی ترکیب‌شونده/بی‌عرض		application .....	کاربرد
combining/non-spacing character .....		code .....	کُد
private use character .....	نویسه‌ی خصوصی	to encode .....	کُدگذاری کردن
base letter/character .....	نویسه‌ی کرسی	ligature .....	لیگاتور
control character .....	نویسه‌ی کنترلی	text .....	متن
entry .....	واردسازی	plain text .....	متن ساده
join-causing .....	واصل	left join-causing .....	متصل‌به‌چپ
edit .....	ویرایش	right join-causing .....	متصل‌به‌راست

## پیوست چ

### کد نویسه‌ها

### (اطلاعاتی)

این پیوست، فهرست کدهای نویسه‌های استاندارد حاضر، در استاندارد یونی‌کد است که به ترتیب الفبایی نام نویسه مرتب شده است.

0623.....	حرف فارسی الف با همزه‌ی بالا	007B.....	آکولاد باز
0628.....	حرف فارسی ب	007D.....	آکولاد بسته
067E.....	حرف فارسی پ	200D.....	اتصال مجازی
062A.....	حرف فارسی ت	0670.....	الف مقصوره‌ی فارسی
062B.....	حرف فارسی ث	202C.....	پایان زیرمتن
062C.....	حرف فارسی جیم	0028.....	پراتز باز
0686.....	حرف فارسی چ	0029.....	پراتز بسته
062D.....	حرف فارسی ح	064F.....	پیش فارسی (ضمه)
062E.....	حرف فارسی خ	0651.....	تشدید فارسی
062F.....	حرف فارسی دال	002D.....	تیره‌منها
0630.....	حرف فارسی ذال	2029.....	جداکننده‌ی بندها
0631.....	حرف فارسی ر	2028.....	جداکننده‌ی سطرها
0632.....	حرف فارسی ز	066C.....	جداکننده‌ی هزارهای فارسی
0698.....	حرف فارسی ژ	0625.....	حرف الف با همزه‌ی پایین
0633.....	حرف فارسی سین	0671.....	حرف الف وصل
0634.....	حرف فارسی شین	0629.....	حرف ت گرد
0635.....	حرف فارسی صاد	0622.....	حرف فارسی آ
0636.....	حرف فارسی ضاد	0627.....	حرف فارسی الف

064D	دوزیرِ فارسی (تنوینِ جر)	0637	حرفِ فارسیِ طا
003A	دونقطه	0638	حرفِ فارسیِ ظا
06F5	رقمِ فارسیِ پنج	0639	حرفِ فارسیِ عین
06F4	رقمِ فارسیِ چهار	063A	حرفِ فارسیِ غین
06F2	رقمِ فارسیِ دو	0641	حرفِ فارسیِ ف
06F3	رقمِ فارسیِ سه	0642	حرفِ فارسیِ قاف
06F6	رقمِ فارسیِ شش	06A9	حرفِ فارسیِ کاف
06F0	رقمِ فارسیِ صفر	06AF	حرفِ فارسیِ گاف
06F9	رقمِ فارسیِ نه	0644	حرفِ فارسیِ لام
06F8	رقمِ فارسیِ هشت	0645	حرفِ فارسیِ میم
06F7	رقمِ فارسیِ هفت	0646	حرفِ فارسیِ نون
06F1	رقمِ فارسیِ یک	0648	حرفِ فارسیِ واو
064E	زیرِ فارسی (فتحه)	0624	حرفِ فارسیِ واو با همزه‌ی بالا
0650	زیرِ فارسی (کسره)	0647	حرفِ فارسیِ هـ
202D	زیرمتنِ اکیداً چپ‌به‌راست	0621	حرفِ فارسیِ همزه
202E	زیرمتنِ اکیداً راست‌به‌چپ	06CC	حرفِ فارسیِ ی
202A	زیرمتنِ چپ‌به‌راست	0626	حرفِ فارسیِ ی با همزه‌ی بالا
202B	زیرمتنِ راست‌به‌چپ	0643	حرفِ کافِ عربی
0652	ساکنِ فارسی	0649	حرفِ یِ عربیِ بی‌نقطه
002A	ستاره	064A	حرفِ یِ عربیِ نقطه‌دار
000D	سرِ سطر	002F	خطِ اریب
000A	سطرِ بعد	005C	خطِ اریبِ وارو
2026	سه‌نقطه‌ی افقی	2010	خطِ تیره
0020	فاصله	007C	خطِ عمودی
200C	فاصله‌ی مجازی	066A	درصدا فارسی
005B	کروشه باز	064C	دویشِ فارسی (تنوینِ رفع)
005D	کروشه بسته	064B	دوزیرِ فارسی (تنوینِ نصب)

2212 .....	علامتِ منها	0640 .....	کشیدگیِ فارسی
0653 .....	مدِ فارسی	00AB .....	گیومه باز
066B .....	ممیزِ فارسی	00BB .....	گیومه بسته
FEFF .....	نشانه‌ی ترتیبِ بایت‌ها	003E .....	علامتِ بزرگتر
200E .....	نشانه‌ی چپ‌به‌راست	002B .....	علامتِ به‌اضافه
200F .....	نشانه‌ی راست‌به‌چپ	0021 .....	علامتِ تعجب
002E .....	نقطه	00F7 .....	علامتِ تقسیم
061B .....	نقطه‌ویرگولِ فارسی	061F .....	علامتِ سؤالِ فارسی
060C .....	ویرگولِ فارسی	00D7 .....	علامتِ ضرب
0654 .....	همزه‌ی فارسیِ بالا	003C .....	علامتِ کوچکتر
0655 .....	همزه‌ی فارسیِ پایین	003D .....	علامتِ مساوی

## پیوست ح

### نام نویسه‌ها

### (اطلاعاتی)

این پیوست، فهرست نام‌های نویسه‌های تعریف‌شده در استاندارد حاضر، به ترتیب کد نویسه در استاندارد یونی‌کد است.

007B	آکولاد باز	000A	سطر بعد
007C	خط عمودی	000D	سر سطر
007D	آکولاد بسته	0020	فاصله
00AB	گیومه باز	0021	علامت تعجب
00BB	گیومه بسته	0028	پرانتز باز
00D7	علامت ضرب	0029	پرانتز بسته
00F7*	علامت تقسیم	002A	ستاره
060C	ویرگول فارسی	002B	علامت به اضافه
061B	نقطه ویرگول فارسی	002D	تیره‌منها
061F	علامت سؤال فارسی	002E	نقطه
0621	حرف فارسی همزه	002F	خط اریب
0622	حرف فارسی آ	003A	دونقطه
0623	حرف فارسی الف با همزه‌ی بالا	003C	علامت کوچکتر
0624	حرف فارسی واو با همزه‌ی بالا	003D	علامت مساوی
0625*	حرف الف با همزه‌ی پایین	003E	علامت بزرگتر
0626	حرف فارسی ی با همزه‌ی بالا	005B	کروشه باز
0627	حرف فارسی الف	005C	خط اریب وارو
0628	حرف فارسی ب	005D	کروشه بسته



0649*	حرفِ یِ عربیِ بی نقطه	0629	حرفِ تِ گرد
064A*	حرفِ یِ عربیِ نقطه دار	062A	حرفِ فارسیِ تِ
064B	دوزیرِ فارسی (تنوینِ نصب)	062B	حرفِ فارسیِ ثِ
064C	دوپیشِ فارسی (تنوینِ رفع)	062C	حرفِ فارسیِ جیم
064D	دوزیرِ فارسی (تنوینِ جر)	062D	حرفِ فارسیِ حِ
064E	زیرِ فارسی (فتحه)	062E	حرفِ فارسیِ خِ
064F	پیشِ فارسی (ضمه)	062F	حرفِ فارسیِ دال
0650	زیرِ فارسی (کسره)	0630	حرفِ فارسیِ ذال
0651	تشدیدِ فارسی	0631	حرفِ فارسیِ رِ
0652	ساکنِ فارسی	0632	حرفِ فارسیِ زِ
0653*	مدیِ فارسی	0633	حرفِ فارسیِ سین
0654	همزه‌یِ فارسیِ بالا	0634	حرفِ فارسیِ شین
0655*	همزه‌یِ فارسیِ پایین	0635	حرفِ فارسیِ صاد
066A	درصدِ فارسی	0636	حرفِ فارسیِ ضاد
066B	ممیزِ فارسی	0637	حرفِ فارسیِ طا
066C	جداکننده‌یِ هزارهایِ فارسی	0638	حرفِ فارسیِ ظا
0670	الفِ مقصوره‌یِ فارسی	0639	حرفِ فارسیِ عین
0671*	حرفِ الفِ وصل	063A	حرفِ فارسیِ غین
067E	حرفِ فارسیِ پِ	0640	کشیدگیِ فارسی
0686	حرفِ فارسیِ چِ	0641	حرفِ فارسیِ فِ
0698	حرفِ فارسیِ ژِ	0642	حرفِ فارسیِ قاف
06A9	حرفِ فارسیِ کاف	0643*	حرفِ کافِ عربی
06AF	حرفِ فارسیِ گاف	0644	حرفِ فارسیِ لام
06CC	حرفِ فارسیِ یِ	0645	حرفِ فارسیِ میم
06F0	رقمِ فارسیِ صفر	0646	حرفِ فارسیِ نون
06F1	رقمِ فارسیِ یک	0647	حرفِ فارسیِ هِ
06F2	رقمِ فارسیِ دو	0648	حرفِ فارسیِ واو

خط تیره	2010*	رقم فارسی سه	06F3
سه نقطه‌ی افقی	2026*	رقم فارسی چهار	06F4
جداکننده‌ی سطرها	2028*	رقم فارسی پنج	06F5
جداکننده‌ی بندها	2029*	رقم فارسی شش	06F6
زیرمتن چپ به راست	202A*	رقم فارسی هفت	06F7
زیرمتن راست به چپ	202B*	رقم فارسی هشت	06F8
پایان زیرمتن	202C*	رقم فارسی نه	06F9
زیرمتن اکیداً چپ به راست	202D*	فاصله‌ی مجازی	200C
زیرمتن اکیداً راست به چپ	202E*	اتصال مجازی	200D
علامت منها	2212*	نشانه‌ی چپ به راست	200E
نشانه‌ی ترتیب بایت‌ها	FEFF	نشانه‌ی راست به چپ	200F

یادآوری — علامت ستاره در سمت راست کد نویسه به معنای اختیاری بودن آن نویسه است.