دانشگاه صنعتی شریف
انجمن زبان‌شناسی ایران

سومین همایش ملی
زبان شناسی رایانشی

۲۸ و ۲۹ آبان ۱۳۹۳
دانشگاه صنعتی شریف

واکاوی توسعه خط و زبان فارسی در شبکه و رایانه

- یحیی تابش، دانشگاه صنعتی شریف
- حامد ملک، دانشگاه صنعتی امیرکبیر

مقاله شماره ۱۰۶۱

مقدمه

سیستم‌های مختلف در شبکه و رایانه اعم از سیستم‌های تجاری و اقتصادی، یا آموزشی و فرهنگی به خط فارسی و ویژگی‌های زبان فارسی برای ذخیره و تبادل اطلاعات وابسته‌اند.

در این رابطه موضوعات زیر را بررسی می‌کنیم:

- استاندارد یونی‌کد و خط فارسی
- صفحه کلید استانداردِ خط فارسی
- فارسی‌سازی نرم‌افزارهای متنِ باز
- چالش‌های آینده

کد مقاله: ۱۰۶۱

استاندارد یونی‌کد و خط فارسی

لزوم فارسی سازی سیستم‌های کامپیوتری از ۱۳۴۱ پس از راه افتادن اولین کامپیوتر آی بی اِم ۱۶۲۰ !

کد مقاله: ۱۰۶۱

استاندارد یونی‌کد و خط فارسی

- دهۀ ۱۹۶۰ مطرح شدن کدگذاری ASCII، ۷ بیتی و بعد ۸ بیتی
- استاندارد سازی خط فارسی در سال ۱۳۵۶ بر مبنای کد اَسکی ۷ بیتی به صورت تک نمادی
- عدم تناظر یک به یک با علایم چاپی و نمایشی، عدم استقبال، تجدید نظر به صورت دو نمادی
- ترویج کامپیوترهای شخصی در دهۀ ۱۳۶۰، کدگذاری‌های متنوع و نا هم‌آهنگ

کد مقاله: ۱۰۶۱

استاندارد یونی‌کد و خط فارسی

- در سال ۱۳۶۷ تدوین استاندارد دو نمادی، استاندارد ۲۹۰۰ ، عدم استقبال به خاطر محدودیت‌ها
- تدوین استاندارد ۳۳۴۲ در سال ۱۳۶۹ ، کد ۸ بیتی فارسی، از جامعیت برخوردار بود، معرفی مفاهیم فاصلهٔ مجازی و اتصال مجازی
- ۱۰ سال به عنوان استاندارد ملی حفظ شد ولی کدهای متنوع و غیر استاندارد هم بسیار رایج بود.

کد مقاله: ۱۰۶۰

استاندارد یونی‌کد و خط فارسی

- ترویج اینترنت و لزوم استاندارد سازی خطوط و نویسه‌ها در سطح جهان، مطرح شدن استاندارد *یونی‌کد*
- استاندارد *یونی‌کد* شیوه‌ای جهانی برای کدگذاری نویسه‌ها و متون
- این استاندارد روشی هماهنگ برای کدگذاری متون چند زبانه مشخص می‌کند که تبادل اطلاعات را در سطوح بین‌المللی میسر می‌سازد

استاندارد یونی‌کد و خط فارسی

- یونی‌کد کدگذاری پیش‌فرض استاندارد اینترنت است
- در کلیه سیستم‌عامل‌ها و زبان‌های برنامه‌نویسی امروزی پشتیبانی می‌شود
- ثبات داده‌ها، امکان تبادل بین‌المللی متون، ساده شدن نرم‌افزارها و کم شدن هزینه‌های تولید، از جمله مزایای یونی‌کد است.

کد مقاله: ۱۰۶۱

استاندارد یونی‌کد و خط فارسی

- *یونی‌کد* شیوه‌ای نیز برای کدگذاری ۸ بیتی متون مشخص کرده است که نویسه‌ها را پس از اِعمال یک تابع خاص به کدشان، به‌صورت دنباله‌هایی که از یک تا چهار بایت دارند نگه می‌دارد، این شیوه با نام UTF-8 شناخته می‌شود.
- به این خاطر که نویسه‌های اسکی را عیناً حفظ می‌کند و در نتیجه، هم برنامه‌های قدیمی راحت‌تر با آن کنار می‌آیند و هم طول پرونده‌های لاتین را زیاد نمی‌کند، بسیار محبوب است.
- در واقع بسیاری از سیستم‌هایی که ادعای پشتیبانی *یونی‌کد* را می‌کنند، تنها UTF-8 را پشتیبانی می‌کنند و پرونده‌های *یونی‌کدی*، عمدتاً در قالب UTF-8 ذخیره شده‌اند.

کد مقاله: ۱۰۶۱

استاندارد یونی‌کد و خط فارسی

• تدوین استاندارد فارسی تحت *یونی‌کد* در سال ۱۳۸۰ به سفارش شورای‌عالی انفورماتیک در دانشگاه صنعتی شریف،

• تصویب استاندارد توسط مؤسسه استاندارد و تحقیقات صنعتی ایران و انتشار استاندارد تحت شمارهٔ ۶۲۱۹ در سال ۱۳۸۱

• هم‌آهنگی با *کنسرسیوم یونی‌کد* و پشتیبانی از خط فارسی توسط *یونی‌کد*

استاندارد یونی‌کد و خط فارسی

- در استاندارد یونی‌کد، نویسه‌های فارسی در بلوک مربوط به خط عربی قرار دارند.
- این بلوک دربرگیرنده نویسه‌های زبان‌هایی است که از خط عربی استفاده می‌کنند:
  - فارسی
  - اردو
  - پشتو
  - کردی

کد مقاله: ۱۰۶۱

استاندارد یونی‌کد و خط فارسی

• در *یونی‌کد* با وجود یکی‌سازی کدهای حروف مشترک، برای حروف فارسی با بار معنایی یا نمایشی متفاوت با حروف عربی، نویسه‌های جداگانه در نظر گرفته شده است.

• کلیهٔ حروف خاص فارسی (پ، چ، ژ، گ) و نیز ک و ی فارسی که با حرف مشابه در عربی تفاوت نمایشی دارند، مکان جداگانه‌ای به خود اختصاص داده‌اند.

• کلیه اِعراب‌های متداول حضور دارند و میان شکل فارسی، اردو و عربی ارقام نیز به‌علت شکل و رفتار متفاوت تفاوت‌هایی منظور شده است.

کد مقاله: ۱۰۶۱

ویژگی‌های فنی
استاندارد یونی‌کد و خط فارسی

- الگوریتم دو جهته

به علت تفاوت جهتِ نوشتن خط‌های فارسی و لاتین و اعداد و علائم، از الگوریتم دو جهته برای ارائه ترتیب قابل نمایش این گونه متون استفاده می‌شود

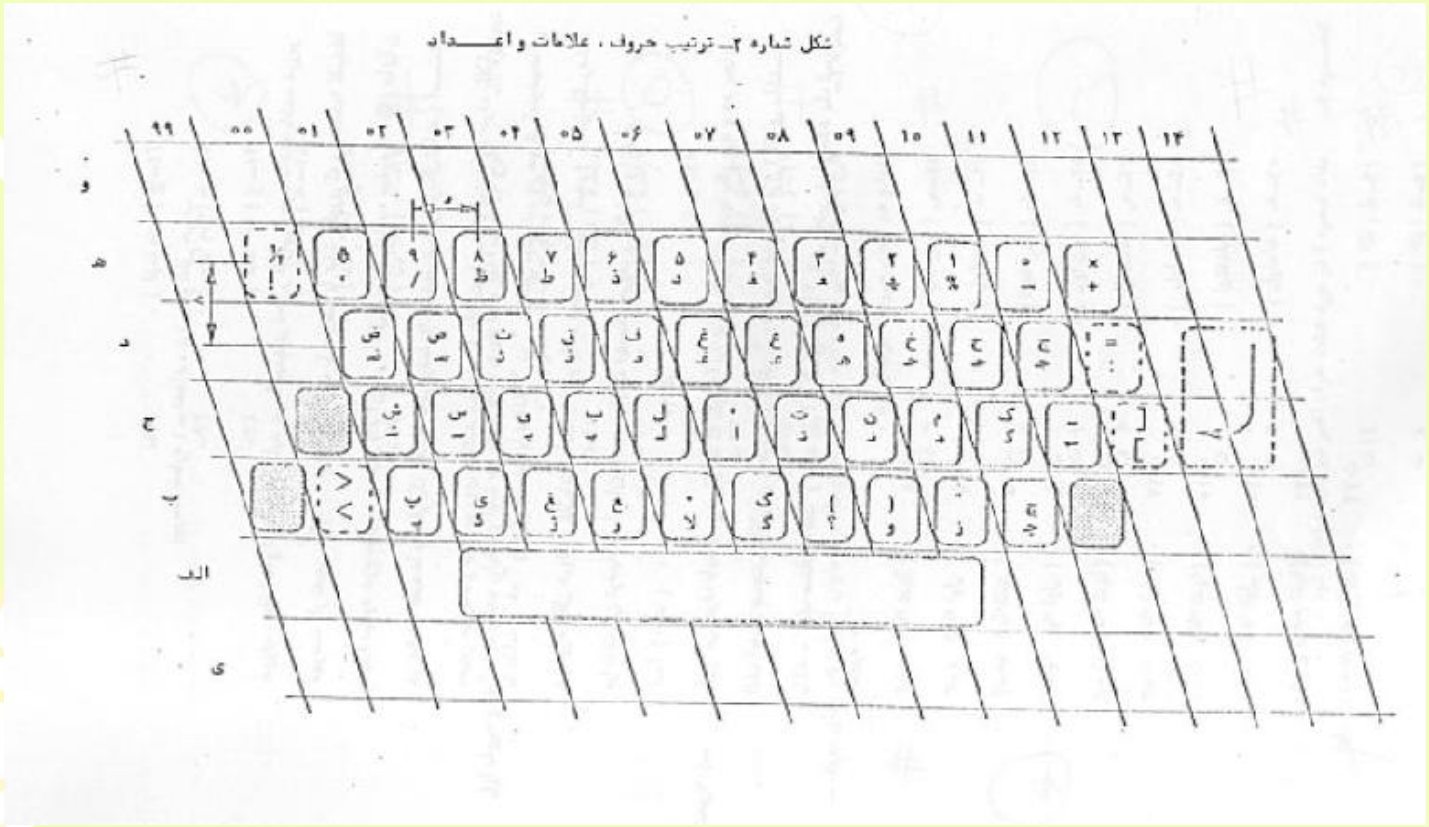
- الگوریتم اتصال

چون حروف فارسی، بسته به حروف قبل و بعد از خود شکل‌های مختلفی می‌گیرند، برای نمایش شکل مناسب از الگوریتم اتصال استفاده می‌شود

کد مقاله: ۱۰۶۱

صفحه کلیدِ استاندارد خط فارسی

بر اساس استاندارد صفحهٔ کلید ماشین تحریر، ۱۳۵۱



کد مقاله: ۱۰۶۱

صفحه کلیدِ استاندارد خط فارسی

- تدوین استاندارد ۲۹۰۱ در سال ۱۳۷۳ برای صفحه کلید فارسی هم‌پای استاندارد ۳۳۴۲:

«طرز قرار گرفتن حروف و علائم زبان فارسی بر روی صفحه کلید کامپیوتر»

- پس از تصویب استاندارد ۶۲۱۹ لزوم تجدید نظر در استاندارد ۲۹۰۱، تدوین استاندارد ۹۱۴۷ در سال ۱۳۸۶:

«فناوری اطلاعات- چیدمان حروف و علائم فارسی بر صفحه کلید رایانه»

- در تدوین استاندارد جدید از استاندارد ۲۹۰۱ که از اصول فنی پیشرفته‌ای برخوردار بود، استفاده کامل به عمل آمد و حداکثر سازگاری با آن در نظر گرفته شد.

کد مقاله: ۱۰۶۱

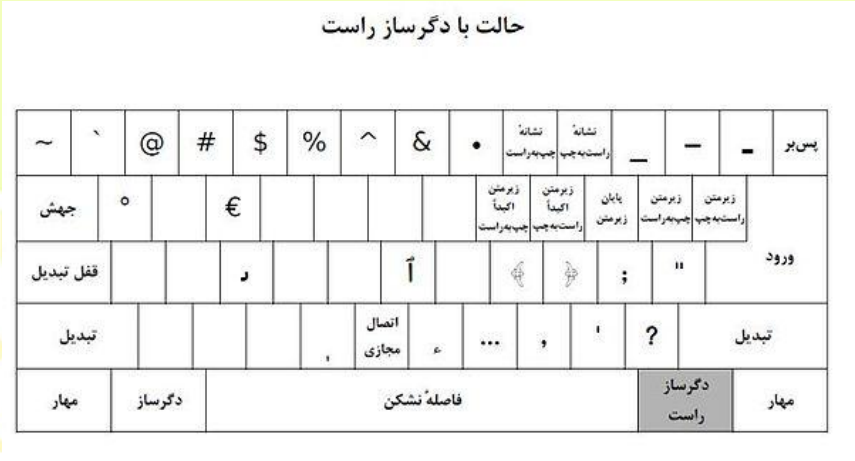
صفحه کلیدِ استاندارد خط فارسی

حالت عادی

اتصال مجازی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۰	-	=	\	پس‌بر
جهش	ض	ص	ث	ق	ف	غ	ع	ه	خ	ح	ج	چ		
قفل تبدیل	ش	س	ی	ب	ل	ا	ت	ن	م	ک	گ		ورود	
تبدیل		ظ	ط	ز	ر	ذ	د	پ	و	.	/		تبدیل	
مهار	دگرساز		فاصله								دگرساز راست		مهار	

کد مقاله: ۱۰۶۱

صفحه کلیدِ استاندارد خط فارسی



کد مقاله: ۱۰۶۱

ویژگی‌های
صفحه کلیدِ استاندارد خط فارسی

- اعداد را می‌توانید به درستی به فارسی بنویسید: ۰۱۲۳۴۵۶۷۸۹
- Shift و Space برابر با *فاصله مجازی* است. مثل: نیم‌بها
- Shift و - دیگر با زیرخط (Underline) تفاوت دارد و کشش فاصله‌است. مثل:

نـــــــــــــــــیم‌بها

- «ی» و «ک» فارسی هستند. (بدون دو نقطه در زیر «ی» و بدون همزه روی «ک»).
- «پ» روی کلید M قرار دارد.
- «ژ» روی Shift و C قرار دارد.

کد مقاله: ۱۰۶۱

نرم‌افزارهای متنِ باز

- جنبش نرم‌افزارهای آزاد (Free Software Movement)
- آغاز توسعه از سال ۱۹۸۴ با سیستم‌عامل GNU
- تعریف نرم‌افزار آزاد:
  - آزادی استفاده از نرم‌افزار با هر هدف
  - آزادی انتشار نرم‌افزار
  - آزادی مطالعه و تغییر آن
  - آزادی انتشار تغییرات انجام شده
- نرم‌افزار آزاد = نرم‌افزار متن‌باز
- امکان محلی‌سازی و پشتیبانی از قابلیت‌های زبان‌های مختلف به صورت قانونی و فنی فراهم است

کد مقاله: ۱۰۶۱

نرم‌افزارهای متنِ باز

- گنو/لینوکس: عموما مجموعه‌ای نرم‌افزاری متن‌باز است که در توزیع های مختلف منتشر می‌شود.
- بیش از ۶۰٪ از رایانه های کارگزار از آن استفاده می‌کنند.
- طی سال‌های اخیر استقبال روزافزونی به استفاده از این سیستم‌عامل در سمت کاربر دیده می‌شود.
- به همین‌دلیل نیز نیاز به محلی‌سازی این سیستم‌عامل از اهمیت بالایی برخوردار است.

کد مقاله: ۱۰۶۱

سیستم عامل لینوکس

- گنو/لینوکس: عموما مجموعه‌ای از نرم‌افزارهای متن‌باز است که در توزیع های مختلف منتشر می‌شود.
- بیش از ۶۰٪ از رایانه های کارگزار از آن استفاده می‌کنند.
- طی سال‌های اخیر استقبال روزافزونی به استفاده از این سیستم‌عامل در سمت کاربر دیده می‌شود.
- به همین‌دلیل نیز نیاز به محلی‌سازی این سیستم‌عامل از اهمیت بالایی برخوردار است.

کد مقاله: ۱۰۶۱

فارسی‌سازی لینوکس

- فارسی‌سازی:
  - پشتیبانی از استاندارد یونی‌کد
  - پشتیبانی از قلم‌های فارسی
  - ترجمه واسط کاربری
  - پشتیبانی از صفحه‌کلید فارسی
  - اصلاح قالب نمایش تاریخ، اعداد، واحدها و غیره
  - تعریف روش مرتب‌سازی
  - اصلاح جهت نمایش متون و واسط کاربری
  - افزودن تقویم شمسی

کد مقاله: ۱۰۶۱

لینوکس شریف

کد مقاله: ۱۰۶۱

سیستم عامل آندروید

- سیستم‌عاملی برای گوشی‌ها، رایانه‌های همراه و یا سامانه‌های نهفته (Embedded Systems)
- متن‌باز
- مبتنی بر لینوکس
- از سال ۲۰۰۷ با پشتیبانی گوگل منتشر شد.
- هم‌اکنون بیش از ۸۵٪ از بازار گوشی‌ها را در اختیار دارد.
- به دلیل باز بودن و عدم انحصار، نصب و توسعه آن برای انواع سخت‌افزارها و گوشی‌ها امکان‌پذیر است.

کد مقاله: ۱۰۶۱

فارسی‌سازی آندروید

- طی چند سال ابتدایی انتشار، پشتیبانی بسیار ضعیفی از زبان فارسی داشت.
- اولین اقدامات توسط بعضی شرکت‌های داخلی برای پشتیبانی از زبان فارسی آغاز شد.
- در سال ۱۳۸۹ *فارسی‌تل* با پشتیبانی کامل از زبان فارسی منتشر شد.
- با همکاری بعضی متخصصین ایرانی، پشتیبانی بسیار خوبی از زبان فارسی در نسخه‌های جدید فراهم شده است.

کد مقاله: ۱۰۶۱

چالش‌های زبان فارسی و نرم‌افزارهای متنِ باز

- اقدامات بسیار خوبی طی یک دهه اخیر برای پشتیبانی از خط و زبان فارسی در نرم‌افزارهای متن‌باز انجام گرفته است.
- هم‌چنان بعضی مشکلات به خصوص در ارتباط با تقویم شمسی و قلم‌های غیر استاندارد وجود دارد.
- چالش‌های پیش رو:
  - تدوین استانداردهای ملی برای گوشی‌های هوشمند
  - توسعه و تهیه پیکره‌های متنی و توجه بیشتر به پشتیبانی از ابزارهای هوشمند در سه حوزه پردازش متن، صدا و تصویر از زبان فارسی

چالش‌های آینده

- توسعه خط و زبان فارسی در رایانه و شبکه هم‌چنان یک ضرورت است، ادامه هم‌کاری با کنسرسیوم *یونی‌کد*، مراقبت از خط و زبان فارسی و توسعه لازم باید مورد توجه قرار گیرد.
- توجه به چالش‌های توسعهٔ نرم‌افزارهای آزاد نیز یک امر ضروری است.
- چالش جدی فعلی در زمینهٔ زبان فارسی و نرم‌افزارهای آزاد، پشتیبانی مناسبِ ابزارها و نرم‌افزارهای هوشمند از زبان فارسی است. هم اکنون به دلیل نبود پیکره‌های مناسب فارسی، پشتیبانی از زبان فارسی در این‌گونه ابزارها و به تبع آن در نرم‌افزارها و سرویس‌های کاربردی ضعیف بوده و از دقت پایینی برخوردار است.

چالش‌های آینده

- ضروری است که به توسعه پیکره‌های متنی توجه شود.
- تهیه پیکره‌های متنی مناسب که از حجم و کیفیت بالایی برخوردار بوده و از نظر قانونی نیز تحت اجازه‌نامه‌های متن باز مانند Creative Commons و یا GFDL باشد، می‌تواند اولین قدم باشد.
- بدون شک، تهیه و انتشار پیکره‌ها، ابزارها و نرم‌افزارهای پردازش زبان فارسی به صورت آزاد نفعی است که نه تنها کاربران، بلکه توسعه‌دهندگان و شرکت‌های تولیدکنندهٔ نرم‌افزار نیز از آن بهره خواهند برد.