

خطایابی املائی خودکار در زبان فارسی

امید کاشفی، میترا نصری و کامیار کنعانی

همراه با ضمیمه‌های:

مبدل تقویم و مبدل عدد، سینا ایروانیان

مبدل پینگلیش، مهرداد صنوبری

اصلاح علائم نشانه گذاری، کامیار کنعانی

دبیرخانه شورای عالی اطلاع‌رسانی

پاییز ۱۳۸۹

سرشناسه: امید کاشفی، ۱۳۶۳-
 عنوان و نام پدیدآور: خطایابی املایی خودکار در زبان فارسی / امید کاشفی، میترا نصری، کامیار کنعانی.
 مشخصات نشر: تهران: شورای عالی اطلاع‌رسانی، دبیرخانه، ۱۳۸۹.
 مشخصات ظاهری: س، ۱۸۹ ص.: مصور، جدول، نمودار.
 شابک: ۹۷۸-۹۶۴-۸۸۴۶-۳۴-۸
 وضعیت فهرست نویسی: فیبا
 موضوع: فارسی -- داده‌پردازی
 موضوع: خط فارسی -- داده‌پردازی
 موضوع: ویراستاری -- داده‌پردازی
 شناسه افزوده: نصری، میترا، ۱۳۶۱ -
 شناسه افزوده: کنعانی، کامیار، ۱۳۵۷ -
 شناسه افزوده: شورای عالی اطلاع‌رسانی. دبیرخانه
 رده بندی کنگره: ۱۳۸۹ خ ۶۲۳/۲۲ PIRT
 رده بندی دیویی: ۰/۷ ۳۴
 شماره کتاب‌شناسی ملی: ۲۱۵۰۶۷۱

خطایابی املایی خودکار در زبان فارسی

© حق چاپ: ۱۳۸۹، دبیرخانه شورای عالی اطلاع‌رسانی

تالیف: امید کاشفی kashefi@ieee.org

میترا نصری mitra.nasri@ut.ac.ir

کامیار کنعانی kanani@ce.sharif.edu

طرح روی جلد: مورتن زنگه‌پور hootanzangeneh@yahoo.com

نوبت چاپ: اول

شمارگان: ۱۰۰۰ نسخه

ISBN: 978-964-8846-34-8

شابک: ۹۷۸-۹۶۴-۸۸۴۶-۳۴-۸

نشانی: تهران، خیابان شهید مطهری، بین خیابان ستایی و خیابان قائم مقام فراهانی، روبروی اداره امور مالیاتی شمال تهران،

شماره ۳۵۸، کد پستی: ۱۵۸۶۹۹۴۳۱۱

تلفن: ۸۸۸۳۹۸۹۵، شماره: ۸۸۸۳۹۸۹۴، صندوق پستی: ۱۵۸۷۵-۳۴۹۹

نشانی وبگاه: www.scict.ir

تمام حقوق اثر متعلق به دبیرخانه شورای عالی اطلاع‌رسانی است.

فهرست مطالب

صفحه	عنوان
یک	پیشگفتار.....
۱	فصل اول: مقدمه.....
۱	۱-۱- مقدمه
۴	۱-۲- نحوه‌ی سازمان‌دهی مطالب
۵	فصل دوم: چالش‌ها و اشکالات دستورخط فارسی.....
۵	۱-۲- مقدمه
۶	۲-۲- دستور خط فارسی و رایانه
۶	۱-۲-۲- ویژگی‌های خط فارسی در پردازش رایانه‌ای
۹	۲-۲-۲- نویسندگان متون رایانه‌ای
۹	۳-۲-۲- دستور خط فارسی مصوب فرهنگستان
۱۴	۴-۲-۲- نمونه‌هایی از کاربردهای نیازمند به یکسان‌سازی خط فارسی
۱۵	۳-۲- واژگان خاص
۱۹	۴-۲- ترکیب‌ها
۱۹	۱-۴-۲- پیوسته‌نویسی
۲۰	۲-۴-۲- جدانویسی
۲۳	۳-۴-۲- ترکیب‌های اضافی
۲۶	۵-۲- نتیجه‌گیری
۲۹	فصل سوم: چالش‌های خطایابی در زبان فارسی.....
۲۹	۱-۳- مقدمه
۳۰	۱-۱-۳- عناصر تشکیل‌دهنده‌ی ترکیب‌ها در فارسی
۳۲	۲-۳- واژک‌شناسی

صفحه	عنوان
۳۳	۱-۲-۳- صرف واژه‌های غیر فعلی
۵۵	۲-۲-۳- صرف فعل‌ها
۸۲	۳-۲-۳- فاصله گذاری
۸۷	۳-۳- حروف هم‌آوا
۸۸	۴-۳- حروف هم‌شکل
۸۹	فصل چهارم: خطایابی املایی خودکار در زبان فارسی
۸۹	۱-۴- مقدمه
۹۲	۲-۴- پژوهش‌ها و کارهای انجام گرفته پیرامون خطایابی املایی
۹۴	۱-۲-۴- روش فاصله‌ی حروف
۹۵	۲-۲-۴- فاصله‌ی هم‌ینگ
۹۶	۳-۲-۴- فاصله‌ی لای‌نشتاین
۹۷	۴-۲-۴- فاصله‌ی دیرال‌لای‌نشتاین
۹۷	۵-۲-۴- فاصله‌ی وِگنر-فیشیر
۹۸	۶-۲-۴- فاصله‌ی جِرو-وینکلر
۹۹	۳-۴- الگوهای خطاهای املایی
۱۰۱	۴-۴- خطایابی املایی در زبان فارسی
۱۰۲	۱-۴-۴- تشخیص خطا
۱۰۸	۲-۴-۴- تولید پیشنهادات جایگزینی
۱۱۱	۳-۴-۴- رتبه‌بندی پیشنهادات
۱۱۸	۵-۴- ارزیابی
۱۱۸	۲-۵-۴- روش ارزیابی
۱۲۰	۳-۵-۴- نتایج
۱۲۳	۶-۴- پژوهش‌های آتی
۱۲۵	مراجع

ضمیمه اول: مبدل اعداد ۱۲۹

۱-۱ مقدمه ۱۲۹

۲-۱ تبدیل و یکسان‌سازی انواع نوشتارهای رقمی ۱۳۰

۱-۲-۱ انواع صورت‌های نمایش ارقام ۱۳۰

۲-۲-۱ دیگر نویسه‌های مورد استفاده در نوشتار رقمی اعداد ۱۳۱

۳-۲-۱ یکسان‌سازی نوشتارهای رقمی اعداد ۱۳۲

۳-۱ تبدیل اعداد از نوشتار رقمی به نوشتار حرفی ۱۳۲

۱-۳-۱ تشکیل نگاشتی از مقادیر به واژگان ۱۳۲

۲-۳-۱ تبدیل اعداد حداکثر سه رقمی به نوشتار فارسی ۱۳۳

۳-۳-۱ تبدیل اعداد طبیعی در حالت کلی ۱۳۳

۴-۳-۱ تبدیل اعداد حقیقی از نوشتار رقمی به نوشتار حرفی ۱۳۴

۴-۱ تبدیل اعداد از نوشتار حرفی به نوشتار رقمی ۱۳۶

۱-۴-۱ ساختن نگاشت از رشته‌های بسیط اعداد به مقادیر متناظر ۱۳۷

۲-۴-۱ یافتن قطعه‌های متن شامل رشته‌های اعداد ۱۳۷

۳-۴-۱ استخراج اعداد صحیح ۱۳۹

۴-۴-۱ تشخیص اعداد اعشاری با ذکر «اژه‌ی ممیز» آن ۱۴۱

۵-۴-۱ تشخیص اعداد اعشاری در حالت کلی ۱۴۳

۵-۱ نتیجه‌گیری ۱۴۵

ضمیمه دوم: مبدل تقویم ۱۴۷

۱-۲ مقدمه ۱۴۷

۲-۲ تشخیص اعداد طبیعی به کمک عبارت‌های با قاعده ۱۴۸

۳-۲ تشخیص عبارت‌های تاریخ به کمک عبارت‌های با قاعده ۱۴۹

۱۵۱	۴-۲ تشخیص عبارت‌های تاریخ به انگلیسی با عبارت‌های با قاعده
۱۵۲	۵-۲ تشخیص عبارت‌های تاریخ به صورت عددی
۱۵۳	۶-۲ تبدیل تاریخ‌ها از تقویمی به تقویمی دیگر
۱۵۳	۷-۲ تشخیص نوع تقویم
۱۵۵	ضمیمه سوم: مبدل نوشتار فارسی با حروف انگلیسی به فارسی
۱۵۵	۱-۳ مقدمه
۱۵۶	۲-۳ ساختار متون پینگلیش
۱۵۶	۳-۳ نگاشت حروف فارسی و انگلیسی
۱۶۱	۱-۳-۳ تکرار حروف
۱۶۱	۲-۳-۳ استفاده از واژه شکسته
۱۶۱	۳-۳-۳ استفاده از واژگان انگلیسی
۱۶۳	۴-۳-۳ استفاده از حروف و کاراکترهای ویژه در واژه
۱۶۴	۴-۳ مبدل پینگلیش
۱۶۴	۱-۴-۳ کشف و یادگیری الگوهای تبدیل
۱۶۵	۲-۴-۳ استفاده از الگوهای شناسایی شده
۱۶۶	۳-۴-۳ الگوریتم کلی تبدیل واژه پینگلیش
۱۶۶	۵-۳ نتیجه‌گیری
۱۶۹	ضمیمه چهارم: اصلاح علائم نشانه‌گذاری
۱۶۹	۱-۴ مقدمه
۱۷۰	۲-۴ روش تشخیص خطاهای نگارشی
۱۷۱	۳-۴ الگوریتم
۱۷۲	۴-۴ تعریف الگو و ملاحظات مربوط به آن
۱۷۲	۱-۴-۴ ساختار الگوها

صفحه	عنوان
۱۷۳	۲-۴-۴ هم پوشانی
۱۷۳	۳-۴-۴ الگوهای زیر مجموعه
۱۷۳	۴-۴-۴ دور
۱۷۵	۵-۴ عبارت منظم
۱۷۵	۱-۵-۴ نمادها در عبارت های منظم
۱۷۷	۲-۵-۴ گروه بندی یا استخراج زیرالگو
۱۷۹.....	ضمیمه پنجم: واژه نامه ی انگلیسی به فارسی
۱۸۱.....	ضمیمه ششم: واژه نامه ی فارسی به انگلیسی
۱۸۳.....	ضمیمه هفتم: نمایه

فهرست شکل‌ها

صفحه	عنوان
۸۴.....	شکل (۱-۳) نمونه ساخت واژه‌ها و جمله‌ها در زبان فارسی
۸۵.....	شکل (۲-۳) نمونه‌ی هم‌آیی واژه‌ها در زبان فارسی
۸۶.....	شکل (۳-۳) هفت گونه از خطاهای املایی ایجاد شده بر اثر فاصله‌گذاری نادرست
۹۶.....	شکل (۱-۴) الگوریتم کلی فاصله‌ی همینگ
۹۶.....	شکل (۲-۴) الگوریتم کلی فاصله‌ی لوئشتاین
۹۷.....	شکل (۳-۴) الگوریتم کلی فاصله‌ی دیرا-لوئشتاین
۹۸.....	شکل (۴-۴) الگوریتم کلی فاصله‌ی وِگیر-فیشِر
۹۹.....	شکل (۵-۴) الگوریتم کلی فاصله‌ی جِرو
۹۹.....	شکل (۶-۴) الگوریتم کلی فاصله‌ی جِرو-وینکلِر
۱۰۵.....	شکل (۷-۴) قطعات فعلی موجود از مصدر «گفتن» و «آراستن»
۱۰۷.....	شکل (۸-۴) نمونه‌ی ساختار نگهداری واژه‌نامه در حافظه
۱۰۸.....	شکل (۹-۴) فرایند تشخیص خطای املایی در واژه‌های تصریفی غیر فعلی
۱۱۳.....	شکل (۱۰-۴) نمونه‌ی چیدمان صفحه کلید انگلیسی
۱۱۳.....	شکل (۱۱-۴) نمونه‌ی چیدمان صفحه کلید فارسی
۱۱۴.....	شکل (۱۲-۴) الگوریتم محاسبه‌ی فاصله‌ی اقلیدسی میان دو نویسه
۱۱۵.....	شکل (۱۳-۴) الگوریتم محاسبه‌ی فاصله میان نویسه‌ها با پشتیبانی از نویسه‌های ترکیبی
۱۱۶.....	شکل (۱۴-۴) الگوریتم محاسبه‌ی فاصله‌ی میان نویسه‌های بهبود یافته
۱۱۷.....	شکل (۱۵-۴) الگوریتم محاسبه‌ی فاصله‌ی میان دو واژه
۱۱۸.....	شکل (۱۶-۴) الگوریتم محاسبه‌ی امتیاز جایگزینی
۱۲۲.....	شکل (۱۷-۴) مقایسه‌ی روش‌های مختلف رتبه‌بندی

فهرست جدول‌ها

عنوان	صفحه
جدول (۱-۲) ابهام‌های قواعد دستور خط فارسی.....	۱۵
جدول (۲-۲) موارد مبهم قواعد دستور خط فارسی در پیوسته‌نویسی.....	۱۹
جدول (۳-۲) ابهام‌های قواعد دستور خط فارسی در جدا نویسی.....	۲۰
جدول (۴-۲) برخی از قواعد تولید ترکیب‌های چندجزئی.....	۲۵
جدول (۱-۳) عناصر تشکیل دهنده‌ی ترکیب‌ها در فارسی.....	۳۰
جدول (۲-۳) ویژگی‌های پسوند تصریفی نشانه‌ی جمع «ها».....	۳۵
جدول (۳-۳) ویژگی‌های پسوند تصریفی نشانه‌ی جمع «ان».....	۳۶
جدول (۴-۳) ویژگی‌های پسوندهای تصریفی ضمائر ملکی و مفعولی.....	۳۷
جدول (۵-۳) ویژگی‌های پسوندهای تصویفی فعل‌های اسنادی.....	۳۸
جدول (۶-۳) ویژگی‌های پسوند تصریفی «ی» نسبت.....	۳۹
جدول (۷-۳) ویژگی‌های پسوند تصریفی «ی» فکره.....	۴۰
جدول (۸-۳) ویژگی‌های «ی» بدل از کسره.....	۴۱
جدول (۹-۳) ویژگی‌های پسوندهای تصریفی صفات تفصیلی.....	۴۲
جدول (۱۰-۳) ویژگی‌های پسوندهای تصریفی ترتیبی شمارشی.....	۴۳
جدول (۱۱-۳) ویژگی‌های پسوند تصریفی ترتیبی شمارشی.....	۴۴
جدول (۱۲-۳) عبارت‌های پسوندی.....	۴۷
جدول (۱۳-۳) عبارت‌های سازنده‌ی واژه‌ی مشتق-مرکب.....	۵۱
جدول (۱۴-۳) قواعد ساخت واژگان پیشوندی.....	۵۲
جدول (۱۵-۳) علائم اختصاری به کار رفته در بیان الگوهای صرف فعل‌ها.....	۶۱
جدول (۱۶-۳) الگوی صرف فعل ماضی ساده معلوم.....	۶۲
جدول (۱۷-۳) الگوی صرف فعل ماضی ساده مجهول.....	۶۳
جدول (۱۸-۳) الگوی صرف فعل ماضی استمراری معلوم.....	۶۴
جدول (۱۹-۳) نمونه‌ی گویش‌های قدیم فعل ماضی استمراری معلوم.....	۶۴

عنوان

صفحه

جدول (۲۰-۳) الگوی صرف فعل ماضی استمراری مجهول.....	۶۵
جدول (۲۱-۳) نمونه‌ی گویش‌های قدیم فعل ماضی استمراری مجهول.....	۶۵
جدول (۲۲-۳) الگوی صرف فعل ماضی بعید معلوم.....	۶۶
جدول (۲۳-۳) الگوی صرف فعل ماضی بعید مجهول.....	۶۷
جدول (۲۴-۳) الگوی صرف فعل ماضی مستمر معلوم.....	۶۸
جدول (۲۵-۳) الگوی صرف فعل ماضی مستمر مجهول.....	۶۸
جدول (۲۶-۳) الگوی صرف فعل ماضی ساده نقلی معلوم.....	۶۹
جدول (۲۷-۳) الگوی صرف فعل ماضی ساده نقلی مجهول.....	۷۰
جدول (۲۸-۳) الگوی صرف فعل ماضی استمراری نقلی مجهول.....	۷۱
جدول (۲۹-۳) الگوی صرف فعل ماضی استمراری نقلی معلوم.....	۷۲
جدول (۳۰-۳) الگوی صرف فعل ماضی بعید نقلی معلوم.....	۷۲
جدول (۳۱-۳) الگوی صرف فعل ماضی بعید نقلی مجهول.....	۷۳
جدول (۳۲-۳) الگوی صرف فعل ماضی مستمر نقلی معلوم.....	۷۴
جدول (۳۳-۳) الگوی صرف فعل ماضی مستمر نقلی مجهول.....	۷۴
جدول (۳۴-۳) الگوی صرف فعل ماضی التزامی معلوم.....	۷۵
جدول (۳۵-۳) الگوی صرف فعل ماضی التزامی مجهول.....	۷۶
جدول (۳۶-۳) الگوی صرف فعل مضارع اخباری معلوم.....	۷۷
جدول (۳۷-۳) الگوی صرف فعل مضارع اخباری مجهول.....	۷۷
جدول (۳۸-۳) الگوی صرف فعل مضارع مستمر معلوم.....	۷۸
جدول (۳۹-۳) الگوی صرف فعل مضارع مستمر مجهول.....	۷۸
جدول (۴۰-۳) الگوی صرف فعل مضارع التزامی معلوم.....	۷۹
جدول (۴۱-۳) الگوی صرف فعل مضارع التزامی مجهول.....	۸۰
جدول (۴۲-۳) الگوی صرف فعل آینده معلوم.....	۸۰
جدول (۴۳-۳) الگوی صرف فعل آینده مجهول.....	۸۱

عنوان

صفحه

جدول (۳-۴۴) الگوی صرف فعل امر معلوم.....	۸۱
جدول (۳-۴۵) الگوی صرف فعل مضارع امر مجهول.....	۸۲
جدول (۳-۴۶) بسامد گونه‌های مختلف نوشتار واژه‌ها و فاصله‌گذاری میان اجزاء آن‌ها.....	۸۳
جدول (۳-۴۷) حروف هم‌آوای فارسی.....	۸۷
جدول (۳-۴۸) حروف هم‌شکل فارسی.....	۸۸
جدول (۴-۱) نمونه‌ای از انواع خطاهای املائی.....	۹۰
جدول (۴-۲) الگوها و نرخ رخداد خطاهای املائی و حروف چینی در زبان انگلیسی.....	۱۰۰
جدول (۴-۳) الگوها و نرخ رخداد خطاهای املائی و حروف چینی در زبان فارسی.....	۱۰۱
جدول (۴-۴) احتمال رخداد و فاصله‌ی خطاهای املائی تکی در زبان فارسی.....	۱۱۲
جدول (۴-۱) ارزیابی و مقایسه‌ی روش رتبه‌بندی پیشنهادی با روش‌های دیگر.....	۱۲۱

پیشگفتار

باسمه تعالی

بقا و توسعه‌ی فرهنگ ملی در گرو رشد مدام مؤلفه‌های آن است و در این میان افزون بر مؤلفه‌های اخلاقی و عقیدتی، خط و زبان از مهمترین مؤلفه‌های فرهنگ ملی تلقی می‌شود و یکی از عناصر کلیدی و هویت بخش در میان ما ایرانیان است. ادیبان، دانشمندان و فضیای بسیاری در حلقه‌های علمی و فکری این دیار به این زبان سخن گفته و قلم زده‌اند. در این عصر هم روزگاری بود که راهیابی این عنصر هویت بخش در فرهنگ ملی به محیط دیجیتال (اعم از برخط و برون خط) بیشتر به یک آرزو می‌ماند. اما امروز تا حدی از این مرحله گذشته‌ایم و باید به غنا و توانمندی محتوای فارسی موجود در محیط رایانه در رقابت با سایر زبان‌ها بیندیشیم. لذا ضروری است که در برنامه‌ای دراز مدت با دو سویی «پژوهشی» و «کاربردی»، زبان فارسی را به بستری زیبا و پذیرانای مفاهیمی علمی و فنی بدل سازیم و جامعه‌ی فارسی زبان را از این نهر جاری سیراب کنیم. از این رهگذر است که فرهنگ ایرانی- اسلامی ما توانمندتر خواهد شد و مزیت رقابتی خود را در میان فرهنگ‌های دیگر نشان خواهد داد.

فعالیت‌های پردازشی در حوزه‌ی خط و زبان فارسی نمونه‌ای بارز و در عین حال مهم در این راستاست که تا به بار نشستن آن در همه‌ی حوزه‌ها راه پر پیچ و خمی در پیش داریم. دبیرخانه‌ی شورای عالی اطلاع‌رسانی، طی سه سال گذشته، علاوه بر فعالیت‌های جاری خود در حوزه‌ی توسعه‌ی خط و زبان فارسی در فضای دیجیتال، اهتمام خاصی به شناسایی، امکان سنجی و تحلیل مقدماتی پروژه‌های پردازشی این حوزه داشته است. در این مسیر هم از نهادهای علمی و اشخاص حقیقی برای طراحی و پیاده‌سازی این قبیل پروژه‌ها به طور مستقیم و غیر مستقیم حمایت کرده است و این روند هم ادامه خواهد

یافت. اما طبیعی است که به دلایل مختلف فنی، اجرایی و گاهی هم سیاست گذارانه، برخی از این پروژه‌ها جز با ورود مستقیم این شورا به سرانجام نخواهد رسید یا آینده‌ی روشنی برای اجرا و به ثمر نشستن آن متصور نیست. پروژه‌ی پژوهشی-اجرایی خطایاب فارسی (ویراستار) یکی از این قبیل پروژه‌های مهم است که با اقدام و حمایت مستقیم این شورا و همکاری جمعی از محققان به بار نشسته است و اکنون نسخه‌ی اول آن با رویکرد غلطیابی صرفی زبان فارسی عرضه می‌شود.

ضمن پیشکش نرم‌افزار کاربردی «ویراستار ۱»، پژوهشنامه‌ها و پیوست‌های فنی آن به محضر علاقه‌مندان و فارسی‌زبانان سراسر جهان، از ارباب نظر خواستاریم نقدهای عالمانه و راهگشای خود را به دبیرخانه‌ی شورای عالی اطلاع‌رسانی منعکس فرمایند تا در هنگام تجدید نسخه‌ی این نرم‌افزار و ارتقای آن استفاده شود.

مطالعه، طراحی و اجرای این پروژه حاصل تلاش مشترک تعدادی از نهادها، مدیران و پژوهشگران پیشرو در حوزه‌ی «پردازش خط و زبان فارسی» است که مراتب سپاس و قدرشناسی خود را به محضرشان اعلام می‌دارم. امید است با فعالیت هم‌افزای نهادهای علمی و دانشگاهی، شاهد باروری و زرافزون فعالیت‌های صورت گرفته در این بخش باشیم.

دکتر حمید شهریاری

دبیر شورای عالی اطلاع‌رسانی