نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

پیاده سازی یک غلط یاب املایی فارسی تحت وب

احمد یوسفان[۱] و بی بی صدیقه طباطبایی[۲]
[۱] کاشان، بلوار قطب راوندی، دانشگاه کاشان، دانشکده مهندسی، گروه مهندسی کامپیوتر، گروه مهندسی کامپیوتر، yoosofan@kashanu.ac.ir
[۲] دانش آموخته‌ی گروه مهندسی کامپیوتر دانشگاه کاشان، tabatabaeised@gmail.com

چکیده

غلطیاب املایی فارسی یکی از ابزارهای مهمی است که در راستای کمک به نویسنده‌ی یک متن فارسی می‌تواند به او کمک شایانی در یافتن و درست کردن غلط فارسی نوشته شده در یک متن نماید. تا کنون غلطیاب‌های گوناگونی آماده شده و در برخی از ابزارهای نگارش به کار گرفته شده است. با این همه در این زمینه پژوهش هنوز ادامه دارد زیرا کارآیی بسیاری از این ابزارها در حد بالایی نیست و باید الگوریتم‌های تازه‌ای برای بهبود آنها پیشنهاد شود. به کارگیری ایده‌های نو در کنار دیگر کارهای پیشین می‌تواند به بهبود کار غلطیاب‌های کنونی کمک شایانی نموده و برخی از مشکل‌های آنها را برطرف نماید. در این مقاله در آغاز برخی از روش‌های متداول غلط‌یابی بررسی شده است سپس روشی ترکیبی برای غلطیاب املایی فارسی پیشنهاد شده و این پیشنهاد به کمک زبان javascript و php و html پیاده سازی شده است. تحت وب بودن این ابزار پیاده سازی شده کمک می‌کند تا بتوان به سادگی آن را آزمایش نمود و اشکال‌های آن را برطرف نمود.

کلمات کلیدی

غلطیاب املایی، فارسی، پیاده سازی، وب، php، javascript

۱- مقدمه

زبان فارسی دربردارنده گنجینه‌ی بزرگی از زیباترین سروده‌ها و داستانها است. زبان فارسی یکی از پربارترین زبان‌های دنیا است. کتاب‌هایی چون مثنوی معنوی، دیوان حافظ، رباعیّات خیام و ... به زبان‌های گوناگون گیتی برگردانده شده و بارها چاپ شده‌اند. برترین ویژگی این نوشته‌ها، انسانی بودن آنها است بگونه‌ای که همه‌ی انسان‌ها گرایشی درونی به این نوشته‌ها دارند. متأسفانه این درخت تنومند امروزه نیاز به توجه بیشتری دارد زیرا برای دنیای نوین آماده نشده است[۱].

این زبان که از نیمه‌ی سده‌ی سوم، آثاری از آن در دست است و از آن تاریخ به بعد، روز به روز گسترش یافته و آثار بی‌شماری در آن آفریده شده، در حیطه‌ی فرهنگ آسیایی و بلکه جهانی، تبدیل به زبانی شده است که شاهکارهای جهانی آفریده است؛ چنان که کمتر زبانی از زبان‌های زنده‌ی موجود می‌توان سراغ داشت که در این پهنه با زبان فارسی کوس برابری زند. از جمله ویژگی‌های این زبان ، دیرندگی آن است در طول بیش از هزار سال؛ بدین معنی که دگرگونی‌ها در آن، در مقام سنجش با بسیاری از زبان‌های جهان، به نسبت کمتر بوده است؛ به طوری که فارسی زبانان امروز به آسانی می‌توانند شعر فردوسی را بخوانند و بفهمند و از آن بالاتر این که در آثار بسیاری از گویندگان و نویسندگان سده‌های اخیر و حتی معاصر می‌توان کاربردهای کهن هزار سال پیش را دید؛ حتی در کتاب‌های دوره‌ی دبستانی می‌توان شعر رودکی و فردوسی را گنجانید[۲].

رواج صنعت چاپ و ماشین‌های تحریر و رایانه‌ها و شتاب زدگی در نوشتن، بی‌دقتی را در درست و زیبا نوشتن دامن زده است. همچنین داده‌ها و صفحه‌های روی شبکه‌ی جهانی که برخی باید روزانه تغییر کنند، این دردسر را چندین برابر کرده‌اند[۱].

متأسفانه بی‌دقتی و رواج بسیاری از خطاهای نگارشی به ویژه در صفحه‌های شخصی فارسی روی شبکه‌ی جهانی به اندازه‌ای زیاد شده است که بیم آن می‌رود زبان دیرپای فارسی دچار مشکلات گوناگونی گردد و نتواند جایگاه خود را در میان زبان‌های زنده‌ی دنیا حفظ نماید.

امروزه رایانه‌ها برای آماده کردن نوشته‌های گوناگون بسیار به کار برده می‌شوند و بیشتر کسانی که نوشته‌ای را آماده می‌کنند از ابزاری در رایانه برای آماده کردن نوشته‌ی خود کمک می‌گیرند. همانند دیگر زبان‌های دنیا باید ابزارهای گوناگونی به کمک نویسنده‌ی متن فارسی در دنیای کنونی بیاید تا بتواند به او در بهتر نوشتن متن فارسی و همچنین پرهیز از خطا کمک نمایند. همچنین افزون بر گزارش نادرست بودن یک کلمه پیشنهاد یا پیشنهادهایی برای کلمه‌ی جایگزین آن به کاربر بدهد تا او را در درست کردن آن کلمه کمک نماید.

ناهماهنگی‌های گوناگونی در نگارش خط فارسی دیده می‌شود همچنین در نگارش رایانه‌ای متن فارسی قالب‌ها، ابزارها و سیستم عامل‌های گوناگون و روش‌های گوناگون کد کردن نوشته‌ی فارسی دیده می‌شود که در [۷-۳ ، ۱] به برخی از آنها پرداخته شده است. فرهنگستان زبان و ادب فارسی کوشیده است برخی ناهماهنگی‌ها را در خط فارسی برطرف نماید و استاندارد یکسانی برای نگارش خط فارسی پیشنهاد دهد در آدرس persianacademy.ir به روزترین نسخه از این استاندارد با نام دستور خط فارسی گذاشته شده است گرچه بسیاری از نویسندگان و حتی برخی کتاب‌های درسی به خوبی این استاندارد را رعایت نمی‌کنند ولی به هر حال می‌تواند به عنوان پایه‌ای برای درست نویسی متن فارسی به کار گرفته شود.

تا کنون غلطیاب‌های گوناگونی برای زبان فارسی آماده شده است و پروژه‌های گوناگونی در این زمینه انجام شده است که در این میان می‌توان به ابزار آماده شده برای خطایابی فارسی از Microsoft