ویرایش‌گر متن شریف: سامانهٔ ویرایش و خطایابی املایی زبان فارسی

بهرام وزیرنژاد[۱]، فاطمه سلطان‌زاده[۲]، محسن مهدوی مزده[۳] و مهدی مرادی[۴]

[۱ و ۲ و ۴] آزمایشگاه پردازش زبان و گفتار، گروه زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، تهران، ایران

[۳] دانشگاه آریزونا

چکیده

در مقالهٔ حاضر سامانه‌ای هوشمند ، جهت ویرایش و خطایابی املایی متون فارسی معرفی شده است. هدف از طراحی و ایجاد این سامانه، ویرایش متون فارسی برای کاربردهای پردازش زبان طبیعی است. این سامانه بر مبنای یک رویکرد مهندسی قابل توسعه، از سه زیرسامانه تشکیل شده‌است: ویرایش‌گر متن فارسی معیار، خطایاب املایی خودکار زبان فارسی و ستاک‌یاب واژگان زبان فارسی. این سه بخش با یکدیگر در تعاملند؛ بدین شکل که ابتدا ستاک واژه‌های متن شناسایی می‌شود و درصورت عدم وجود ستاک در فهرست واژه‌های زبان، واژهٔ مذکور به‌عنوان واژه‌ای نادرست شناسایی خواهد شد ؛ سپس خطایاب خودکار فهرستی از واژه‌های جایگزین را پیشنهاد خواهدکرد. در زیرسامانهٔ ویرایش‌گر، متن موجود ویرایش‌شده و متنی یک‌پارچه که منطبق بر معیارهای مصوب فرهنگستان زبان و ادب فارسی است، به کاربر ارائه خواهد شد. نتایج ارزیابی نشان‌دهندهٔ دقتی بسیار خوب در حدود ۹۵٪ در ستاک‌یابی کلمات، ۹۲٪ در ویرایش و ۹۶٪ در خطایابی املایی زبان فارسی است.

واژگان کلیدی : پردازش زبان طبیعی، خطایاب املایی خودکار ، ویرایش‌گر متن فارسی معیار، ستاک‌یاب.

۱- مقدّمه

در جهان امروز کاربرد گستردهٔ رایانه در تسهیل امور انسان امری غیر قابل انکار است. فناوری اطلاعات در کشورهای پیشرفته با سرعتی شگرف در حال گسترش است و حجم چشم‌گیری از اطلاعات مورد نیاز انسان به‌صورت الکترونیک در دسترس قرار دارد. بدیهی است که این حجم چشم‌گیر اطلاعات به مدیریتی مناسب و دقیق نیازمند است. هدف از انجام این طرح ارائهٔ ابزاری برای پردازش اسناد و متون فارسی الکترونیکی است تا از نظر نگارشی یک‌پارچه و بدون خطاهای املایی و تایپی باشند ؛ در این‌صورت کاربران رایانه می‌توانند به‌سهولت از آن‌ها بهره گیرند. از سوی دیگر دارابودن این ویژگی نه تنها برای انسان، بلکه برای پردازش خودکار متون توسط رایانه نیز امری مهم و ضروری تلقی می‌شود. در بسیاری از کاربردهای پردازش زبان طبیعی ، همچون موتورهای جستجوگر، خلاصه‌سازی متون، استخراج واژه‌های کلیدی متون، ترجمهٔ ماشینی، انواع داده‌کاوی و متن‌کاوی، سامانه‌های تبدیل متن به گفتار، دسته‌بندی و خوشه‌بندی متون، سامانه‌های پرسش و پاسخ و بسیاری امور دیگر، حجم عظیمی از اطلاعات توسط رایانه ، به‌طور خودکار مورد بازبینی و تحلیل قرار می‌گیرد. بدیهی است که درصورت عدم وجود یک‌پارچگی در ساختار متون و وجود خطاهای املایی و نگارشی فراوان، پردازش خودکار متون با چالش‌های بسیاری روبه‌رو خواهد بود. در این بخش به چالش‌های پیش رو در پردازش زبان طبیعی به‌خصوص پردازش خط و زبان فارسی خواهیم پرداخت.

در پردازش متون زبان طبیعی با زبان نوشتاری سروکار داریم. این مسئله باعث می شود اگر چه به جهت ازدست‌دادن اطلاعات گویشی ، مانند لحن گوینده، آهنگ صدا، تاکید و درنگ، با مشکلات و ابهاماتی مواجه شویم، ولی در مقابل با شکل محدودتری از زبان کار می‌کنیم. در تلاش برای ساخت یک سامانهٔ پردازش و درک متون فارسی با مسائل و مشکلاتی مواجه می‌شویم که بعضی در بیش‌تر

زبان‌ها بروز کرده و برخی خاص زبان فارسی هستند. همچنین برخی از این پیچیدگی‌ها به طبیعت زبان و نارسایی‌های قواعد زبان‌شناسی مربوط و برخی دیگر برخاسته از مشکلات ایجاد سامانه های هوش مصنوعی است (دبیرخانهٔ شورای عالی اطلاع رسانی، ۱۳۸۸).

نخستین پیچیدگی پردازش زبان طبیعی، ابهام واژگانی است. ابهام واژگانی در کلیهٔ زبان‌های طبیعی چالشی شناخته شده است. یک شکل از ابهام واژگانی در فارسی از آنجا ناشی می‌شود که ترکیب واژه‌ها منجر به تشکیل واژه‌ای می‌شود که ممکن است در اثر بی‌دقتی کاربران، از دید رایانه به دو یا چند شکل مختلف خوانده شود. ، برای مثال در جایی که منظور نویسنده واژهٔ "سیب‌زمینی" است، ممکن است در اثر بی‌دقتی "سیب زمینی" نوشته شود؛ در این صورت رایانه قادر به تشخیص واژهٔ اصلی نخواهد بود. دومین دلیل پیچیدگی پردازش واژگانی زبان‌های طبیعی، ترکیب واژه‌ها با یکدیگر و تولید واژه‌هایی است که حاوی اطلاعاتی مانند مالکیت، جمع یا مفردبودن واژه و غیره است. به‌عنوان مثال واژهٔ "کتاب‌هایشان" را در نظر بگیرید. این واژه در لغت‌نامه وجود ندارد. بدون تحلیل صرفی مناسب از دید رایانه این واژه، واژه‌ای بی‌معنا تلقی می‌شود (کاشفی و دیگران، ۱۳۸۹).

دلیل سوم به ماهیت زایایی زبان‌های طبیعی مربوط می‌شود. قواعد تولید واژه در یک زبان خاص وجود واژه‌هایی را ممکن می‌سازد که هنوز در آن زبان تولید نشده‌اند. احتمال وجود واژه‌ای مانند " حمایتگری" دلیلی بر این مدعاست.

مسئلهٔ دیگر، عدم وجود لغت‌نامه‌ای جامع برای کلیهٔ لغات یک زبان است. برای بسیاری از واژه‌های مورد استفاده توسط انسان مدخلی در واژه‌نامه‌ها وجود ندارد. در روند تولید یک واژه‌نامه هرچند تلاشی عظیم و ستودنی صورت گرفته باشد ؛ اما هیچ واژه‌نامه‌ای را نمی‌توان یافت که حاوی تمامی لغات یک زبان باشد ؛ چون که وجود چنین واژه‌نامه‌ای با ماهیت زبان آدمی که بسیار پویا و تغییرپذیر است در تضاد می‌باشد. حال به بررسی مشکلات پردازش خودکار زبان و خط فارسی می‌پردازیم.

زبان فارسی ویژگی‌هایی دارد که پردازش این زبان را با دشواری‌هایی روبه‌رو می‌کند. از جملهٔ این دشواری‌ها نظام پیچیدهٔ واژه‌سازی[1] و واژه‌پردازی[2] در آن است. به‌لحاظ واژه‌سازی، واژه‌های زبان‌فارسی می‌توانند واژه‌های بسیط (مانند "مرد") و واژه‌های غیربسیط[3] همچون مشتق (مانند "مردانه")، واژه‌های مرکب (مانند "جوان‌مرد") و واژه‌های مشتق‌مرکب (مانند "جوان‌مردانه") باشند. نظام تصریف و واژه‌پردازی این زبان امکان افزودن وندهای[4] متعددی را به کلمات آن فراهم می‌سازد. به حضور وندهای گوناگون در مثال‌هایی مانند "نخورده‌ام"، "کوچکترم"، "دوستانمانند"، "نمی‌دیدمشان" و "بهترینهایمان" توجه نمایید. بدیهی است که این نظام زبانی پیچیده، دشواریی‌های بسیاری را در ارتباط با نحوهٔ اتصال اجزای مختلف کلمه به یکدیگر اعم از ستاک، پیشوند[5] و پسوند[6] موجب می‌شود. در حال حاضر یکی از اساسی‌ترین مشکلات پردازش زبان و خط فارسی، همین مسئلهٔ فاصله‌گذاری و نیم‌فاصله‌گذاری میان کلمات و هم‌چنین میان علایم سجاوندی است. به‌دلیل تعدد وندها در زبان فارسی در زمینهٔ نحوهٔ اتصال اجزای کلمه به یکدیگر مشکلاتی وجود دارد. به واژه‌ای مانند "می‌رفته ام" توجه نمایید که می‌تواند به صورت‌های متفاوتی چون "میرفته ام"، "می رفته ام"، "می رفته‌ام" و غیره نوشته شود. از دیگر مشکلات پیش رو می‌توان به حضور وندهای میانجی در واژه‌هایی مانند "پرندگان"، "کتاب‌هایی"، "خانه‌ای" و غیره اشاره کرد که کار پردازش زبان را دشوار می‌سازد.

مسئله دیگر در پردازش زبان فارسی ناشی از این است که در موارد متعدد ، وندهای تصریفی و اشتقاقی مشترک هستند؛ هم‌چون وند تصریفی "ان" در مواردی مانند "درختان" و وند اشتقاقی "ان" در "خوران". مشکل دیگر در پردازش زبان فارسی مسئلهٔ اعراب است. اعراب در زبان فارسی با آنکه تلفظ می‌شود در اکثر موارد نوشته نمی‌شود. از دید رایانه یک واژهٔ خاص به همراه اعراب و بدون اعراب دو واژهٔ متفاوت تلقی می‌شود. از دیگر چالش‌های بسیار مهم در پردازش متون فارسی حضور "ی" و "ک" و اعداد به شکل عربی و غیراستاندارد در متون فارسی است که می‌بایست آنها را با معادل‌های فارسی این حروف جایگزین کرد. هم‌چنین مشکل دیگر حضور و عدم حضور تنوین در نگارش لغات عربی تنوین‌دار توسط فارسی‌زبانان است که از دید رایانه بسیار مشکل‌آفرین است. علاوه‌براین، می‌توان به تفاوت‌های میان گفتار و نوشتار زبان فارسی اشاره کرد. صورت گفتاری و نوشتاری در زبان فارسی به‌لحاظ سبک

[1] Word-formation
[2] Inflection
[3] Complex Lexeme
[4] Affix
[5] Prefix
[6] Suffix

کلام، صرف واژگان و نحو جمله بسیار متفاوت است. با توجه به اینکه حجم چشم‌گیری از اطلاعات به زبان فارسی در محیط وب در بسیاری از وبلاگ‌ها و وبگاه‌های فارسی زبان به شکل محاوره‌ای نگارش شده است، پردازش رایانشی این متون از چالش‌های مهم در پردازش خط فارسی تلقی می‌شود (کاشفی و دیگران، ۱۳۸۹).

عدم نمایش کسرهٔ اضافه در نوشتار منجر به دشواری‌هایی در تشخیص مرز عبارات اسمی می‌شود و حضور افعال مرکب گوناگون در زبان فارسی پردازش رایانشی متون را از منظر فاصله‌گذاری چالش‌برانگیز می‌کند. هم‌چنین عدم وجود یک معیار قطعی در رسم‌الخط زبان فارسی پردازش این زبان را با مشکلاتی بسیار روبه‌رو می‌کند. به‌عنوان مثال می‌توان به نحوهٔ نگارش غیربسیط در زبان فارسی اشاره کرد که هنوز بر سر جدانویسی، پیوسته‌نویسی و یا تلفیقی از این دو بین زبان‌شناسان اختلاف سلیقه وجود دارد. در مورد یای آخر کلمه در مواردی مانند "خانهٔ" و "خانه‌ی" توافق نظر وجود ندارد و تاکنون معلوم نشده است که کدام‌یک از این دو صورت، نگارش درست محسوب می‌شوند (رسولی و مینایی بیدگلی، ۱۳۸۷).

مشکل مهم دیگر، از منظر خطاهای املایی، در دستور خط فارسی این است که برخی حروف در زبان فارسی دارای صورت آوایی مشابه ولی صورت نوشتاری متفاوت هستند (مانند "ذ"،"ز"،"ظ" و "ض"). مورد عکس این نیز در زبان فارسی وجود دارد. برای مثال واژهٔ "ورود" را در نظر بگیرید. در این واژه حرف "و" بازنمایی یکسان از دو آوای متفاوت ارائه می‌دهد که اولی هم‌خوان "\v\" و دیگری واکهٔ "\u\" است. بدیهی است که این گونه مسائل ، پردازش رایانشی زبان فارسی را با دشواری‌هایی چند روبه‌رو می‌سازد.

مقالهٔ حاضر به‌منظور رفع مشکلات موجود در پردازش خط و زبان فارسی توسط رایانهٔ ، سعی بر آن دارد که به معرفی سامانهٔ ویرایش و خطایابی املایی زبان فارسی بپردازد. ما در این پژوهش برای نیل به متنی یکپارچه و عاری از خطاهای املایی و نگارشی، به اصلاح فاصله‌گذاری و نیم‌فاصله‌گذاری میان کلمات و هم‌چنین علایم سجاوندی، اعداد، اصلاح نویسه‌های عربی و غیراستاندارد و اصلاح خطاهای املایی تنها غیرصورت‌کلمه‌ای پرداختیم. در ادامه به بررسی کارهای پیشین در این راستا می‌پردازیم ؛ سپس در بخش سوم، سامانهٔ طراحی‌شدهٔ را در جهت رفع مشکلات موجود که خود شامل سه زیرسامانهٔ ویرایش‌گر[1] متن فارسی معیار، خطایاب املایی[2] خودکار زبان فارسی و ستاک‌یاب[3] کلمات زبان فارسی است، به‌تفصیل معرفی خواهیم کرد. در بخش‌های چهارم و پنجم نیز به‌ترتیب به ارزیابی سامانه، بررسی ویژگی‌ها و کاستی‌های سامانه و معرفی کارهای آینده خواهیم پرداخت.

۲- کارهای پیشین

مطالعات در زمینهٔ خطایاب املایی از اوایل دههٔ شصت میلادی آغاز شد. پس از آن پژوهش‌های دیگری در این راستا با استفاده از روش‌های یادگیری ماشینی، خطایابی با استفاده از چندوزنی‌ها ، کلیدهای مشابهت، کلیدهای آوایی، روش‌های خطایابی بدون واژه‌نامه انجام یافته است (کاشفی و دیگران، ۱۳۸۹). در سال ۱۹۹۰ مدل کانال نوفه‌ای در خطایابی خودکار معرفی شد. در این مدل کلیهٔ واژگان زبان از کانال نوفه‌ای عبور داده می‌شوند و خروجی کانال با کلمهٔ ورودی مقایسه و محتمل‌ترین واژه‌ها در قالب فهرستی به کاربر ارائه می‌شود (ژورافسکی و مارتین، ۲۰۰۷) . در پژوهشی (شیخ الاسلام و همکاران، ۲۰۱۲)، از مدل کانال نوفه‌ای در طراحی یک خطایاب املایی برای زبان فارسی استفاده شده است. در پژوهشی دیگر (بریل و مور، ۲۰۰۰)، مدل پیچیده‌تری از کانال نوفه‌ای معرفی شده است. در این روش به جای اصلاح حرف به حرف واژه‌ها، زیررشته‌هایی از لغات انتخاب شده و با زیررشته‌ای مناسب جایگزین می‌شود. این روش برای اصلاح خطاهای واحد تا چندگانه مفید خواهد بود. در پژوهشی دیگر (فیلی، ۲۰۱۰)، یک خطایاب برای زبان فارسی ارائه شده است که توانایی تشخیص و تصحیح خطاهای املایی، نحوی و معنایی را داراست. خطایاب مذکور از یک روش مرتب‌سازی ترکیبی با استفاده از شباهت رشته‌ای و بسامد واژه استفاده می‌کند. در این روش جدولی شامل شباهت میان حروف زبان فارسی تولید شده است که با استفاده از آن شباهت میان رشته‌ها محاسبه می‌شود. در طراحی خطایابی املایی خودکار دیگر برای ذخیره‌سازی واژه‌های زبان فارسی، از الگوریتم سریع ساخت MADFA استفاده شده است. بدین‌ترتیب، علاوه‌بر افزایش سرعت، حجم واژه‌ها به‌طور متوسط به یک سوم کاهش یافته و سپس از این مجموعه از کلمات در خطایابی املایی استفاده شده است (صبوریان و دیگران، ۱۳۸۵). در پژوهشی دیگر (فیلی و

[1] Normalizer

[2] Spell Checker

[3] Stemmer

همکاران ۲۰۱۴)، علاوه‌بر خطایابی املایی به کشف و تصحیح خطاهای نحوی نیز پرداخته شده است. در پژوهش مذکور از یک روش ترکیبی[1] در طراحی سامانه خطایابی زبان فارسی استفاده شده است؛ بدین شکل که در طراحی خطایابی املایی از روش‌های آماری[2] و در خطایابی نحوی از روش‌های قاعده‌مند[3] پردازش زبان طبیعی بهره گرفته شده است.

دبیرخانهٔ شورای عالی اطلاع‌رسانی محصولی را تحت عنوان "ویراستیار"، در جهت استانداردسازی متون و نویسه‌ها و خطایابی املایی واژگان فارسی تهیه کرده است. ویراستار به‌صورت متن‌باز در فضای وب قابل دسترسی است[4]. در پژوهشی دیگر (شمس‌فرد، ۲۰۱۰)، سامانه‌ای جهت پیش‌پردازش زبان فارسی شامل ویرایش‌گر، خطایاب املایی متون و تحلیل‌گر صرفی لغات زبان تهیه شده است. در خطایاب املایی مذکور، راهکاری برای تصحیح خطاهای ناشی از حذف فاصلهٔ بین کلمات ارائه شده است.

در زمینهٔ تحلیل ساخت‌واژی و یافتن ریشه[5] و ستاک[6] واژه‌ها نیز کارهایی انجام شده‌است. به‌عنوان مثال، در پژوهشی (مواجی و دیگران، ۱۳۹۰)، از روش ساخت‌بنیاد در تحلیل صرفی لغات زبان فارسی استفاده شده است. در پژوهشی دیگر (یوسفیان و دیگران، ۱۳۸۵) به مسئلهٔ یافتن ریشه افعال زبان فارسی پرداخته شده است. علاوه‌بر این، فولادی و ارومچیان (۱۳۸۵) از روش یادگیری ماشینی بدون نظارت و شاهمیری و دیگران (۱۳۸۳) با کمک شبکه عصبی ،مصنوعی ، تحلیل‌گرهای صرفی دیگری را برای پردازش لغات زبان فارسی طراحی کرده‌اند. از دیگر پژوهش‌ها در این راستا می‌توان به پژوهش انجام‌شده توسط حسامی‌فرد و قاسم‌ثانی (۱۳۸۳)و هم‌چنین به ریشه‌یاب آماری تهیه‌شده توسط محمدی نصیری و دیگران (۱۳۸۳) اشاره کرد.

۳- ویرایش‌گرمتن شریف

در این پژوهش پس از بررسی مشکلات و کاستی‌های پیش‌رو در امر پردازش زبان طبیعی، سامانه‌ای خودکار در ویرایش زبان طبیعی طراحی شده‌است. همان‌طور که پیش‌تر اشاره شد، ویرایش متن، اصلاح انواع خطاهای املایی متن و دستیابی به متنی یک‌پارچه مطابق با شیوه‌نامهٔ مصوب "فرهنگستان زبان و ادب فارسی" گامی مهم در راستای پردازش رایانشی متون فارسی است. هدف از طراحی این سامانه ، رفع مشکلات و ابهامات موجود در متون زبان فارسی جهت آمادگی متن برای کاربردهای متنوع پردازش زبان طبیعی توسط رایانه است. ویرایش‌گر متن شریف در جهت دست‌یابی به این مهم از سه زیرسامانه تشکیل شده است: ۱- زیرسامانه ستاک‌یاب کلمات زبان فارسی، ۲- زیرسامانهٔ خطایاب املایی واژگان زبان فارسی و ۳- زیرسامانهٔ ویرایش‌گر متن فارسی معیار. در شکل (۱) چگونگی نحوهٔ تعامل زیرسامانه‌ها با یکدیگر نمایش‌داده شده است.

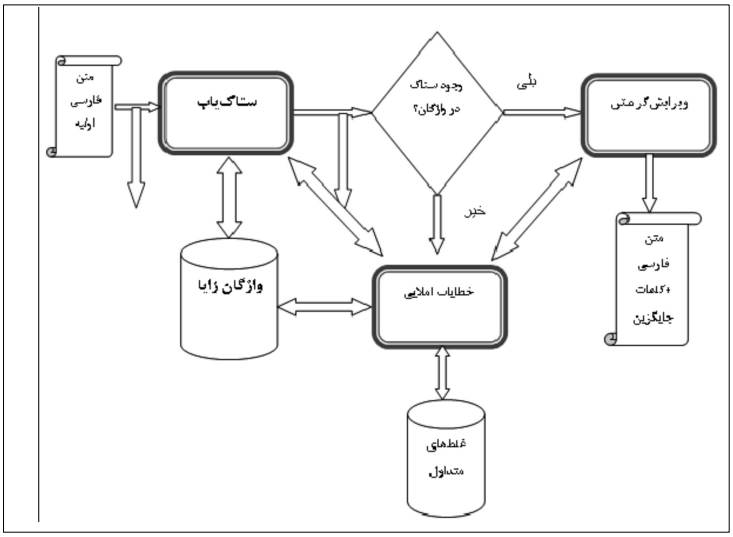
همان‌طور که در شکل (۱) مشاهده می‌کنید ابتدا زیرسامانهٔ ستاک‌یاب، کلیهٔ واژه‌های متن را از دیدگاه صرف ، مورد تجزیه و تحلیل قرارداده و ستاک آن‌ها را شناسایی می‌کند. درصورتی‌که ستاک واژه‌ای در مجموعه واژگان زایای زبان فارسی که مشتمل بر حدود ۴۵۰۰۰ مدخل واژگانی است (اسلامی و دیگران، ۱۳۸۳)، یافت نشد، واژهٔ مذکور توسط زیرسامانهٔ خطایاب املایی خودکار بررسی می‌شود. درصورتی‌که واژهٔ نادرست در فهرست واژه‌های نادرست متداول زبان فارسی باشد، صورت صحیح واژه به‌طور خودکار جایگزین واژهٔ نادرست خواهدشد؛ در غیر این صورت حالات محتمل واژهٔ مذکور توسط الگوریتمی شناسایی شده و بر اساس ترکیبی از میزان نزدیکی به واژه و مدل زبانی تک‌کلمه‌ای‌های[7] محتمل قابل جایگزینی، مرتب‌سازی و به کاربر نمایش داده می‌شود. در صورت وجود واژهٔ مطلوب در فهرست واژه‌های پیشنهادی، کاربر می‌تواند از میان آنها جایگزین مورد نظر را انتخاب کند . پس از آن، واژهٔ مذکور به‌طور خودکار جانشین واژهٔ نادرست پیشین می‌شود.

در زیرسامانهٔ ویرایش‌گر متن معیار فارسی، متن فارسی نخستین به‌لحاظ علایم سجاوندی، فاصله‌گذاری و نیم‌فاصله‌گذاری میان کلمات، استفاده از حروف استاندارد در نگارش و سایر معیارهای حائز اهمیت در ویرایش متن فارسی، مورد بازبینی و تحلیل قرار می‌گیرد و درنهایت متنی یک‌پارچه به کاربر عرضه می‌شود.

سامانهٔ ویرایش‌گر متن شریف در سرویس وب به‌طور رایگان و برخط قابل دسترس کلیهٔ فارسی‌زبانان است[1].

در ادامه قابلیت‌ها و نحوهٔ فعالیت هر زیرسامانه به تفصیل شرح داده می‌شود.

[1] Hybrid
[2] Statistical Approach
[3] Rule-based Approach
[4] http://www.virastyar.ir/
[5] Root
[6] Stem
[7] Unigrams



(شکل – ۱): نمای کلی سامانهٔ ویرایشگر متن شریف.

۳-۱- زیرسامانهٔ ستاک‌یاب کلمات زبان فارسی

صرف شاخه‌ای از زبان‌شناسی است که به مطالعهٔ ساختمان درونی واژه و روابط حاکم بر آن می‌پردازد. از مهم‌ترین اهداف مطالعات صرفی دست‌یابی به نظریه‌ای است که به کمک آن بتوان ساختمان واژه را توصیف کرد و ابزاری را برای شناسایی انواع واژه و قواعد ساخت ، فراهم آورد (شقاقی، ۱۳۸۹).

به‌منظور دست‌یافتن به این هدف، در پژوهشی به توصیف ساختمان تصریفی کلمات زبان فارسی پرداخته شده و ساختارهای تصریفی زبان فارسی معرفی شده است (اسلامی و علیزاده، ۱۳۸۸). ارائهٔ چنین ساختاری نه‌تنها در حوزهٔ مطالعهٔ زبان انسان بلکه در زمینهٔ پردازش زبان طبیعی نیز سودمند است. برای مثال در ویرایش خودکار متون ، نیازمند به دانستن ساختارهای تصریفی زبان فارسی به‌منظور شناسایی ستاک واژه‌های متن هستیم.

در پژوهش مذکور برای شناسایی ستاک اسم، فعل، صفت، قید و سایر اجزای کلام از روشی مبتنی بر قاعد استفاده شده است. برای مثال برای یافتن ستاک اسمی در عبارت "معلمانمانند" ابتدا واژه‌بست ربطی "ند" و پس از آن به‌ترتیب واژه‌بست شخصی "مان" و تکواژ جمع "ان" از انتهای کلمه حذف می‌شود. برای یافتن ستاک فعلی مانند "ندیدمش" ابتدا واژه‌بست شخصی "ش" و سپس شناسهٔ "م" از انتهای آن و سپس تکواژ نفی "ن" از ابتدای واژه حذف می‌شود. لازم به ذکر است که در راستای یافتن ستاک کلمات زبان فارسی، ما در پژوهش خود از قواعد معرفی‌شده در پژوهش نام‌برده استفاده کردیم ،

باید در نظر داشت که پیچیدگی‌های موجود در طراحی ستاک‌یابی که در تعامل با یک خطایاب املایی است ، دوچندان می شود ؛ ستاک‌یاب باید قادر باشد ستاک واژه‌هایی را که به شکلی نادرست نوشته شده‌اند نیز بیابد. برای مثال واژهٔ "می‌نویسم" که به‌اشتباه "می‌نویشم" تایپ شده را درنظر بگیرید ؛ پس از حذف "م" از انتهای آن به‌صورتی مانند "می‌نویش" می‌رسیم ؛ حال این پرسش مطرح می‌شود که در این مثال "می" تکواژ استمراری است یا بخشی از ستاک واژه؟ متأسفانه از این‌گونه ابهام‌ها در تفسیر واژه‌های نادرست به فراوان یافت می‌شود. در این پژوهش سعی ما بر آن بود که با بررسی حالات مختلف، قوانینی برای ستاک‌یاب تعریف کرده و بر مشکلات بالا نایل آییم. پس از آنکه ستاک مورد نظر شناسایی شد، در زیرسامانهٔ خطایاب املایی واژه‌هایی مناسب برای جایگزینی با آن پیشنهاد می‌شود. در ادامه به‌تفصیل به شرح عملکرد این زیرسامانه می‌پردازیم.

۳-۲- خطایاب خودکار املایی زبان فارسی

خطایاب‌های املایی به‌طورمعمول به تصحیح خطاهای حروف‌چینی، نگارشی و خطاهای ناشی از بازشناسی نویسه‌های نوری می‌پردازند. خطاهای حروف‌چینی به خطاهایی هم‌چون درج اشتباه نویسه‌های مجاور در صفحه کلید اشاره دارند. خطاهای نگارشی ناشی از عدم آگاهی نویسنده از قواعد و واژگان زبان ناشی می‌شود و خطاهای

[17] http://81.31.191.11/normalizer2/

ناشی از بازشناسی نویسه‌های نوری نیز به عدم موفقیت رایانه در بازشناسی یک نویسهٔ خاص اشاره دارد. پژوهش‌های صورت‌گرفته بر روی پیکره‌های بزرگ نشان می‌دهد که ۸۰٪ تا ۹۰٪ از خطاهای املایی به‌دلیل چهار نوع خطای عمده روی می‌دهد. این چهار نوع خطا عبارتند از حذف یک حرف، درج یک حرف، جایگزینی یک حرف با حرف دیگر و جابه‌جایی دو حرف مجاور در واژهٔ صحیح (کاشفی و دیگران، ۱۳۸۹). برای مثال واژهٔ "سبز" را در نظر بگیرید ؛ این چهار نوع خطا برای واژهٔ مذکور در جدول زیر آمده است.

(جدول- ۱): انواع خطاهای املایی برای واژهٔ "سبز".

نوع خطا	مثال
درج	سبزر
حذف	سب
جایگزینی	سبر
جابه‌جایی	سزب

ما در این پژوهش چهار نوع خطای درج، حذف، جایگزینی و جابه‌جایی را در خطایابی املایی تحت پوشش قرار داده‌ایم. فرآیند خطایابی در طی فرآیند زیر صورت می‌گیرد. پس از آنکه ستاک واژه توسط ستاک‌یاب کشف و شناسایی شد ، ولی در مجموعهٔ واژگان زایا یافت نشد، واژهٔ مذکور وارد زیرسامانهٔ خطایاب می‌شود. این زیرسامانه از یک روش ترکیبی در خطایابی املایی بهره می‌برد. در طی این پژوهش، فهرستی از واژه‌های نادرست متداول به‌همراه صورت صحیح آنها تدوین شده است. از آنجایی که این گونه از خطاها بسیار متداول است و همگان در صورت صحیح آن‌ها اتفاق نظر دارند، درصورت مواجهه با این خطا رایانه به‌طور خودکار واژهٔ درست را جایگزین واژهٔ پیشین خواهد کرد.

روش دوم مورد استفاده در این خطایاب، روش فاصلهٔ دمرالونشتاین[18] است. درصورتی‌که واژهٔ نادرست در این فهرست جایی نداشته باشد، با اجرای الگوریتم دمرا لونشتاین نزدیکترین واژه‌ها به واژهٔ نادرست یافته می‌شوند. طرح کلی الگوریتم در شکل (۲) آمده است.

این الگوریتم دو واژه را به‌عنوان ورودی می‌گیرد و کوتاه‌ترین فاصلهٔ ممکن این دو واژه را محاسبه می‌کند و به‌عنوان خروجی باز می‌گرداند. در این روند زیررشته‌های متفاوت از دو واژه انتخاب و برای هر حالت انواع خطاهای احتمالی بررسی و حالتی با کوتاه‌ترین فاصله انتخاب می‌شود. در این الگوریتم تشخیص چهار نوع خطای درج، حذف، جایگزینی و جابه‌جایی لحاظ شده است. در اینجا واژه‌هایی مانند "سب، سبزر، سبر" که دارای خطاهای حذف، درج و جایگزینی هستند ، دارای فاصلهٔ یک از واژهٔ "سبز" هستند و واژهٔ "سزب" که دارای غلط جابه‌جایی است، دارای فاصله دو از واژهٔ صحیح است.

```
DamerauLevenshtein (q , l)
{
    f_dl(0 , 0) = 0
    if (q_i == l_j)
        d(q_i , l_j) = 0
    else
        d(q_i , l_j) = 1
    if ((q_i == l_{j-1}) and ( q_{i-1} == l_j))
        t(q_i , l_j) = 0
    else
        t(q_i , l_j) = 2
    f_dl(i , j) = min ((f_dl(i – 1 , j) + 1), (f_dl(i , j - 1) + 1), (f_dl(i – 1, j - 1) + d(q_i, l_j)), (f_dl(i – 2 , j - 2) + t(q_i, l_j)))
    return f_dl(|q| , |l|)
}
```

(شکل- ۲): الگوریتم فاصلهٔ دمرالونشتاین (کاشفی و دیگران، ۱۳۸۹)

در روند خطایابی پژوهش حاضر، ابتدا فاصلهٔ واژهٔ نادرست از کلیهٔ واژگان زایای زبان سنجیده شده و سپس نتایج حاصل از آن ارزیابی شد. پس از آن برای افزایش سرعت خطایابی املایی، محدودیت‌هایی در انتخاب واژه‌ها برای راه‌یابی به فرآیند مقایسه کلمات لحاظ شد ، که در ادامه به آنها اشاره خواهیم‌کرد.

یکی از محدودیت‌ها در انتخاب واژه‌ها جهت استفاده در الگوریتم، مقایسهٔ طول واژهٔ محتمل با طول واژهٔ آغازین است. از آنجایی که در اکثر موارد اختلاف طول واژه‌های جایگزین کوچک‌تر از سه واحد است، در روند انتخاب واژه‌ها تنها کلماتی بررسی می‌شوند که اختلاف طول آن‌ها در این بازه جای داشته‌باشد.

از دیگر محدودیت‌های لحاظ‌شده می‌توان به برچسب مقولهٔ نحوی محتمل برای واژهٔ نادرست اشاره کرد. با توجه به ساختار تصریفی واژهٔ نادرست می‌توان نوع مقولهٔ نحوی این واژه را حدس زد و تنها کلماتی را با آن مقایسه کرد که دارای این مقولهٔ نحوی باشند.

پس از درنظرگرفتن محدودیت‌های انتخاب واژه‌ها و محاسبهٔ فاصلهٔ آن‌ها با کلمهٔ آغازین، باید واژه‌های محتمل در قالب فهرستی به‌ترتیب احتمال صحت واژه به کاربر پیشنهاد شوند. پس از آنکه کاربر یکی از واژه‌های فهرست را انتخاب کرد، آن واژه به‌طور خودکار به جای واژهٔ پیشین درج می‌شود. مرتب‌سازی این فهرست نیز بر اساس ترکیبی از فاصلهٔ واژه و مدل زبانی تک‌کلمه‌ای‌های محتمل (که همان بسامد واژه در پیکرهٔ زبانی است) صورت می‌گیرد. در بخش۴ به نحوهٔ ارزیابی این زیرسامانه اشاره خواهیم‌کرد.

[18] Damerau-Levenshtein Distance

۳-۳- زیرسامانهٔ ویرایش‌گر متن فارسی معیار

همان‌طور که در بخش‌های پیشین اشاره شد، مسائل و مشکلاتی چند ، در خصوص نگارش متون فارسی الکترونیک وجود دارد. به منظور رفع چنین مشکلاتی ، در این بخش به معرفی زیرسامانهٔ ویرایش‌گر متن فارسی معیار می‌پردازیم. هدف از ارائهٔ این زیرسامانه، ویرایش خودکار متون فارسی برای دست‌یابی به متنی خوانا و یکپارچه است.

در سامانهٔ ویرایش‌گر متن فارسی شریف، پس از آن که متن فارسی به‌لحاظ اشکالات املایی بررسی و اصلاح شد، متن تصحیح‌شده به‌عنوان ورودی زیرسامانهٔ ویرایش‌گر متن دریافت می‌شود تا به متنی ویرایش‌یافته و بدون آشفتگی نگارشی تبدیل و به کاربر ارائه شود . چگونگی فعالیت این ویرایش‌گر در ادامه شرح داده شده‌است.

زیرسامانهٔ ویرایش‌گر متن فارسی معیار، ابتدا متن را به جملات و سپس جملات را به کلمات تقسیم می‌کند؛ سپس فواصل میان علایم سجاوندی همچون "."، "("، "،" و واژه‌های متن را اصلاح می‌کند؛ بدین شکل که فواصل اضافی را حذف و فواصل مناسب را درج می‌نماید؛ به‌عنوان مثال "رفتم ." را با "رفتم. " جایگزین خواهد کرد. پس از آن به تصحیح فواصل میان واژه‌ها و همچنین اجزای مختلف واژه می‌پردازد. همان‌طور که پیش‌تر اشاره شد، نظام صرف در زبان فارسی نظامی پیچیده است و واژه‌ها از وندهای تصریفی و اشتقاقی فراوان تشکیل شده‌اند. گاه ممکن است وندهای یک واژه به‌صورت گسسته از آن در جمله ظاهر شوند که این امر منجر به بدخوانی جمله توسط رایانه خواهد شد. برای مثال واژهٔ "صندلی‌های" ممکن است به شکل "صندلی‌های" و یا "می‌دیدم" به‌صورت "میدیدم" و یا "می دیدم" نوشته شود. این سامانه قادر است که بین اجزای مختلف کلمه، نیم‌فاصله را لحاظ کند . برای این منظور ابتدا تحلیل صرفی کلمه صورت می‌گیرد و اجزای مختلف آن چون پیشوند و پسوند و غیره تشخیص داده می‌شود و سپس فاصله‌گذاری مناسب بین اجزا صورت می‌گیرد. هم‌چنین اگر در جایی به جای یک فاصله، چندین فاصله قرار داشته باشد آن را اصلاح کنند ؛ برای مثال "در خانه" را با "در خانه" جایگزین می‌کند. می‌بایست این نکته را توجه داشت گاه ممکن است که وندهایی مانند "تر"، "می"، "ها" و غیره در واژه‌ای خود بخشی از ستاک باشند. به‌عنوان مثال به واژه‌هایی مانند "دختر"، "میناب" و "فارنهایت" توجه‌کنید. این واژه‌ها ممکن است به‌اشتباه تفسیر شوند. برای پیش‌گیری از بروز چنین مشکلاتی ، زیرسامانهٔ ویرایش‌گر متن فارسی معیار حاوی فهرستی از واژه‌های مستثنا است. این زیرسامانه با دراختیار داشتن این واژه‌ها قادر است آنها را به‌درستی تفسیر کند.

از مهم‌ترین مشکلات پردازش زبان فارسی توسط رایانه ، مسئله حضور نویسه‌های "ی" و "ک" و اعداد عربی و غیر استاندارد در متون فارسی است. در زیرسامانهٔ معرفی‌شده این نویسه‌ها با نویسه‌های فارسی متناظر با خود، جایگزین می‌شوند.

از دیگر ویژگی‌های این زیرسامانه، تنظیم فاصلهٔ میان اعداد و همچنین میان اعداد و کلمات فارسی است. هنگامی که کلمات به اعداد چسبیده‌اند، سامانه فاصله مناسب را بین اعداد و کلمات لحاظ خواهد کرد. برای مثال در جملهٔ "امروز علی۵ تا کتاب خرید"، بین واژهٔ "علی" و عدد "۵" فاصلهٔ مناسب درج شده و جمله به‌صورت "امروز علی ۵ تا کتاب خرید"، نمایش داده می‌شود. البته در مواردی که کلمهٔ فارسی تک‌حرفی است، سامانه قادر خواهد بود این موارد تمییز داده و بین حرف فارسی و عدد هیچ‌گونه فاصله‌ای قرار ندهد. جملهٔ "علی در واحد ب۹ زندگی می‌کند" نمونه‌ای از این‌گونه موارد است. از دیگر ویژگی‌های این زیرسامانه، عدم درج فاصله بعد از نقطه هنگامی که نقطه بین اعداد قرار بگیرد؛ برای مثال در جملهٔ "نسخهٔ ۲,۳ این برنامه آماده نیست"، بین دو عدد فاصله‌ای درج نخواهد شد.

علاوه‌بر موارد ذکر شده، این زیرسامانه قادر است مواردی را که کاربر از یک مخفف استفاده می‌کند تشخیص داده و بین اجزای واژهٔ مخفف، فاصله را درج نکند ؛ همانند جملهٔ "آقای م.م.خ با او مخالف است ". همچنین، ایجادنکردن فاصله ، هنگامی که حرف فارسی بعد از حرف انگلیسی بیاید و فقط یک نویسه باشد (همانند "عنصر **j**م را با عنصر **i**م جابه‌جا می‌کنیم") و همین‌طور ایجاد فاصله بین واژه‌های انگلیسی و فارسی متصل به هم (همانند "دیروز آقای Flanders آمد")، از دیگر ویژگی‌های زیرسامانهٔ مذکور است.

به‌طور خلاصه، زیرسامانهٔ ویرایش‌گر متن فارسی معیار با بهره‌گیری از راه‌کارهای معرفی‌شده، متن نوشتاری فارسی را به متنی یک‌پارچه، خوانا و به دور از آشفتگی‌های متداول مبدل خواهد کرد. در بخش ۴ نتایج حاصل از ارزیابی این پژوهش را بیان می‌کنیم.

۴- ارزیابی و گزارش نتایج

در بخش خطایاب املایی، تشخیص و تصحیح به‌صورت دو فرآیند جداگانه در نظر گرفته می‌شوند و لذا ارزیابی آنها نیز

جداگانه صورت می‌گیرد. چون ممکن است ، سامانه‌ای خطاها را خوب شناسایی کند ، ولی در ارائهٔ جایگزین خوب عمل نکند. به همین سبب، برای ارزیابی زیرسامانهٔ خطایاب املایی خودکار از دو معیار دقت[19] و میانگین عکس رتبه[20] استفاده شده است.

در ارزیابی بخش تشخیص خطا ، از معیار دقت استفاده شده است. معیار دقت، بیان‌گر میزان خطاهای تشخیص‌داده‌شده توسط سامانه است که به‌درستی شناسایی شده‌اند (ژورافسکی و مارتین، ۲۰۰۷). فرمول زیر بیان‌گر معیار دقت است که R تعداد خطاهای درست تشخیص‌داده‌شده توسط سامانه و T کل خطاهای موجود در متن است:

$$\text{precision}=\frac{|R|}{|T|} \qquad (۱)$$

لازم به ذکر است که در ارزیابی ویرایش‌گر متن نیز از معیار دقت در تصحیح خطاهای ویرایشی استفاده شده‌است.

در بخش تصحیح خطاها نیز از معیار میانگین عکس رتبه، استفاده شده است. معیار میانگین عکس رتبه، با توجه به عکس رتبهٔ گزینهٔ درست محاسبه می‌شود. به‌عنوان مثال، اگر سامانه چهار کلمه نامزد را به کاربر دهد و دومین گزینه، صحیح باشد، میانگین عکس رتبه $\frac{1}{2}$ خواهد بود. فرمول این معیار در زیر ذکر شده است. در این فرمول، |N| بیان‌گر تعداد کل خطاها و rank نشان‌دهندهٔ رتبهٔ گزینهٔ صحیح برای واژهٔ نادرست است:

$$\text{MRR}=\frac{1}{|N|}\sum_{i=1}^{|N|}\frac{1}{\text{rank}i} \qquad (۲)$$

داده‌های آزمایشی، جملاتی حاوی پانصد واژه دارای خطاهای املایی و حروف‌چینی و نیز خطاهای نگارشی می‌باشند. در زیر نمونه‌ای از دادهٔ آزمون و عملکرد سامانه در رابطه با آن آمده است.

(شکل- ۳): نمونه‌ای از داده آزمون حاوی خطاهای املایی و نگارشی و عملکرد سامانه در رابطه با آن.

نتایج ارزیابی حاکی از این است که زیرسامانهٔ خطایاب از دقتی معادل با ۹۶٪ در شناسایی و تشخیص خطا برخوردار است. همان‌طور که گفته شد، برای سنجش موفقیت سامانه در رتبه‌بندی فهرست واژه‌های جایگزین برای واژهٔ نادرست از معیار میانگین عکس رتبه استفاده کرده‌ایم. ارزیابی رتبه‌بندی این فهرست در دو حالت با لحاظ‌کردن محدودیت‌هایی مانند برچسب مقولهٔ نحوی محتمل برای واژهٔ نادرست و تفاوت طول دو واژهٔ درست و نادرست و بدون لحاظ‌کردن محدودیت‌های بالا انجام شده است و نتایج حاصل از این دو روش با هم مقایسه شده است. در روش رتبه‌بندی بدون درنظرگرفتن محدودیت به میانگین عکس رتبه معادل با ۰/۴۱ و در روش رتبه‌بندی با در نظرگرفتن محدودیت به میانگین عکس رتبه معادل با ۰/۶۱ دست یافته‌ایم. درنتیجه درنظرگرفتن محدودیت‌های بالا ، منجر به رتبه‌بندی نویسه شده است.

برای ارزیابی زیرسامانهٔ ستاک‌یاب کلمات زبان فارسی از معیار دقت استفاده شده است. داده‌های آزمون این بخش ، جملاتی حاوی ۲۰۳۳ کلمه است و زیرسامانهٔ ستاک‌یاب توانسته است در تحلیل و ستاک‌یابی آنها به دقتی معادل با ۹۵٫۸۱٪ دست یابد.

در مورد زیرسامانهٔ ویرایش‌گر، نتایج ارزیابی ، حاکی از این است که این زیرسامانه در اصلاح اشکالات ویرایشی به دقتی معادل با ۹۲٪ دست یافته است.

۵- نتیجه‌گیری

در این مقاله سامانه‌ای جهت ویرایش و خطایابی املایی متون فارسی معرفی شده است. این سامانه که خود از سه زیرسامانهٔ ستاک‌یاب کلمات زبان فارسی، خطایاب املایی خودکار زبان فارسی و ویرایش‌گر متن فارسی معیار تشکیل شده ، قادر است که خطاهای املایی متون فارسی را شناسایی و اصلاح و هم‌چنین می‌تواند متون فارسی را ویرایش کند. برای نیل به اهداف بالا، یک ستاک‌یاب واژه‌ها، مبتنی بر نظام واژه‌سازی و تصریف زبان فارسی، که در تعامل با سایر بخش‌های سامانه است، طراحی شده است. نتایج ارزیابی این پژوهش بیان‌گر این است که سامانه معرفی شده به‌دقتی بسیار خوب در ویرایش و خطایابی املایی متون فارسی دست یافته است.

سامانه شریف یک غلط‌یاب قابل توسعه است که از یک تحلیل‌گر ساخت‌واژی قدرتمند بر پایه نظریه‌های زبان‌شناسی بهره می‌برد. تحلیل‌گر ساخت‌واژی شریف از قابلیت تحلیل صورت کلمه‌های مشتق و مرکب برخوردار است و توانایی شناسایی ریشهٔ کلمات و در مورد افعال تشخیص زمان، شخص و وجه را نیز دارد که می‌تواند در یک سامانه غلط‌یاب هوشمند نقش مؤثری داشته باشد. سامانه‌های مشابهی که در زبان فارسی طراحی و ایجاد شده‌اند ، اگرچه قابلیت تحلیل

[19] Precision
[20] Mean Reciprocal Rank (MRR)

صورت‌کلمه‌ها را با روش‌های کم‌هزینه و سطحی‌تر دارند ، اما فاقد تحلیل‌گری جامع هستند که بتواند صورت‌های پیچیده تصریف و اشتقاق را تحلیل کند از این رو در مواجهه با کلمات پیچیده نظیر "خونسردانه" و یا "روشنفکران" دچار خطا می‌شوند.

سامانه شریف همچنین با بهره‌گیری از یک مدل زبانی ، قابلیت توجه به بافت را دارد و می‌تواند غلط‌هایی را که از نظر نوشتاری ، کلمه‌ای از واژگان هستند با توجه به بافت تشخیص دهند. برای مثال امکان تشخیص غلط "منشور" بجای "منظور" در عبارت "منشور من از این حرف این است." وجود دارد. همچنین سامانه امکان مرتب‌سازی پیشنهادهای جایگزین را با توجه به بافت ارائه می‌کند.

به‌عنوان کارهای آینده می‌توان مدل زبانی مورد استفاده در سامانه را به مدل‌های زبانی مرتبه بالاتر گسترش داد و از آن در جهت رفع انواع دیگری از خطایابی املایی چون خطاهای صورت‌کلمه‌ای بهره جست. با استفاده از چنین مدل زبانی می‌توان خطاهایی را که خود مدخلی از واژه‌ها هستند نیز تشخیص داد. بدین شکل سامانه قادر خواهد بود تا خطاهای املایی چون استفاده نادرست کلمهٔ "حیات" به جای "حیاط" و یا "منظور" به جای "منشور" را نیز تشخیص دهد و اصلاح کند. علاوه‌بر این، در پژوهش حاضر امکان تشخیص خطاهایی که ناشی از عدم درج فاصله بین چند کلمهٔ متوالی است (همانند " اورادیدم ") گنجانده نشده است که می‌توان آن را به‌عنوان کارهای آینده منظور کرد. بدین ترتیب، با افزودن ویژگی‌های جدید به سامانه می‌توان توانایی سامانهٔ حاضر را در خطایابی املایی و نگارشی زبان فارسی ارتقا بخشید و از آن در جهت دست‌یابی به نگارش فارسی معیار بهره جست.

۶- مراجع

اسلامی، محرم، شریفی آتشگاه، مسعود، علیزاده لمجیری، صدیقه، زندی، طاهره. "واژگان زایای زبان فارسی"، مجموعه مقالات نخستین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۳.

اسلامی، محرم و علیزاده، صدیقه. "ساخت تصریفی کلمه در زبان فارسی، زبان و ادب فارسی"، نشریهٔ دانشکدهٔ ادبیات و علوم انسانی دانشگاه تبریز، ۱۳۸۸، شمارهٔ مسلسل ۲۱۱، ایران.

حسامی فرد، رضا، قاسم ثانی، غلامرضا. "طراحی یک الگوریتم ریشه یابی برای زبان فارسی"، مجموعه مقالات نخستین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۳.

دبیرخانه شورای عالی اطلاع رسانی."فاز اول طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی: استخراج نیازمندی‌های ابزار خطایاب املایی در لایهٔ نحو زبان فارسی به پیکره‌های فارسی مورد نیاز"، کد زیر پروژه: پیک متن فارسی-۲- ث، ۱۳۸۸، ویراست اول.

رسولی، محمد صادق، مینایی بیدگلی، بهروز. "روشی جدید در خطایابی املایی زبان فارسی"، دومین همایش داده‌کاوی ایران، دانشگاه صنعتی امیرکبیر، ایران، ۱۳۸۷.

شاهمیری، امیر شهاب، صفابخش، رضا و دژکام، رسول. "تعیین ریشه زبانی واژگان فارسی و عربی به کمک شبکه عصبی مصنوعی"، مجموعه مقالات نخستین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۳.

شقاقی، ویدا ."مبانی صرف". چاپ چهارم، تهران، سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت)، ۱۳۸۹.

صبوریان، محسن، دری نوگورانی، صادق. "طراحی و پیاده‌سازی یک خطایاب فارسی"، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۵.

فولادی، کاظم، ارومچیان، فرهاد. "یادگیری بدون سرپرست ساخت واژهٔ زبان فارسی به کمک بستر توصیف با طول بهینه"، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۵.

کاشفی، امید، نصری، میترا، کنعانی، کامیار. "خطایابی املایی خودکار در زبان فارسی"، تهران، شورای عالی اطلاع‌رسانی، دبیرخانه، ۱۳۸۹.

محمدی نصیری، مجتبی، شیخ اسماعیلی، کیومرث و ابوالحسنی. حسن، "ریشه یاب آماری برای زبان فارسی"، مجموعه مقالات نخستین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۳.

مواجی، وحید، اسلامی، محرم، وزیرنژاد، بهرام. "پارس مورف: تحلیلگر صرفی زبان فارسی"، پردازش داده ها و علائم، ۱۳۹۰، شمارهٔ ۱، پیاپی۱۵.

یوسفیان، احمد، صالحی زارعی، سمیه، مینایی بیدگلی، بهروز. "دشواری‌های ریشه‌یابی فارسی و روشی برای ریشه‌یابی فعل‌های سادهٔ فارسی"، مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه، دانشگاه تهران، تهران، ۱۳۸۵.

Brill, E., and Moore, R. C., "An Improved Error Model for Noisy Channel Spelling Correction", In Proceedings of ACL, 2000, pp. 286-293.

Faili, H., "Detection and Correction of Real-Word Spelling Errors in Persian Language", In the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'-10), August, 2010, pp. 533-536.

Faili, H., Ehsan, N., Montazery, M., and Pilehvar, M. T., "Vafa spell-checker for detecting spelling, grammatical, and real-word errors of Persian language", Literary and Linguistic Computing, 2014.

Jurafsky, Daniel, Martin, James H., Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition, Prentice-Hall, 2007.

Shamsfard, M., Jafari, H.S. and Ilbeygi, M., "STEP-1: A Set of Fundamental Tools for Persian Text Processing", In 8th Language Resources and Evaluation Conference, 2010.

Sheykholeslam, M. H., Minaei-Bidgoli, B., and Juzi, H., "A Framework for Spelling Correction in Persian Language Using Noisy Channel Model", In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012.

بهرام وزیرنژاد عضو هیأت علمی دانشگاه صنعتی شریف است. ایشان دکترای خود را در رشتهٔ مهندسی پزشکی- بیوالکتریک از دانشگاه صنعتی امیرکبیر دریافت کرد. او همچنین طی سال‌های ۲۰۰۷ و ۲۰۱۵ به‌عنوان پژوهش‌گر و استاد مهمان در دانشگاه سیدنی استرالیا مشغول به امور پژوهشی و تدریس بود. از او بیش از چهل مقاله در کنفرانس‌ها و نشریات معتبر داخلی و خارجی به چاپ رسیده است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش زبان طبیعی و گفتار، زبان‌شناسی رایانشی یادگیری ماشینی، هوش مصنوعی و تحلیل داده است .

نشانی رایانامهٔ ایشان عبارت است از:

bahram@sharif.edu

فاطمه سلطان‌زاده در سال ۱۳۸۴ در مقطع کارشناسی رشتهٔ مهندسی کامپیوتر، گرایش نرم‌افزار در دانشگاه شهید چمران اهواز پذیرفته شد. سلطان‌زاده سپس در سال ۱۳۹۰ در مقطع کارشناسی ارشد در رشته زبان‌شناسی‌رایانشی پذیرفته و در این رشته به اخذ رتبه نخست از دانشگاه صنعتی شریف نایل شد . زمینه‌های تخصصی مورد علاقهٔ ایشان پردازش زبان طبیعی، نحو و صرف زبان و زبان‌شناسی شناختی است.

نشانی رایانامهٔ ایشان عبارت است از:

fatemeh.slt@gmail.com

محسن مهدوی مزده، مدرک کارشناسی خود را در رشته مهندسی نرم‌افزار از دانشگاه صنعتی شریف دریافت کرد. وی هم‌اکنون دانشجوی دکتری زبان‌شناسی در دانشگاه آریزونا است. حوزه‌های فعالیت ایشان نحو، نحو فارسی، زبان‌شناسی رایانشی است.

نشانی رایانامهٔ ایشان عبارت است از:

mahdavi@email.arizona.edu

مهدی مرادی، دانش‌آموختهٔ رشته زبان‌شناسی رایانشی از دانشگاه صنعتی شریف است. پردازش زبان طبیعی، مدل‌سازی دانش و ایجاد هستان‌شناسی دانش عرفی از حوزه‌های فعالیت پژوهشی ایشان است.

نشانی رایانامهٔ ایشان عبارت است از:

meh_mor2003@yahoo.com