بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی

مرضیه گل‌تاجی[۱]
سعیده بذرگر[۲]

چکیده

این پژوهش به بررسی مسائلی پرداخته است که پایگاه‌های مقاله‌های فارسی در جستجوی ریختهای مختلف یک کلمه با آن روبرو هستند. برای پاسخگویی به سؤالهای پژوهش، از روش پیمایش مقایسه‌ای استفاده شده است. جامعه پژوهش عبارت است از سه پایگاه مقاله‌های فارسی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»، «جهاد دانشگاهی»، و «پژوهشگاه اطلاعات و مدارک علمی ایران». محققان سیاهه‌ای شامل ۱۷ کلیدواژه را با دقت در متون فارسی انتخاب نموده‌اند که هرکدام از آنها نمایانگر یک مورد از چالشهای زبان فارسی در برخورد با فناوری نوین هستند. سپس کلیدواژه‌ها در جعبه جستجوی پایگاه‌های مذکور وارد و نتایج هر کدام نیز ثبت گردید. این بررسی نشان داد چالشهای ریختیِ شناخته شدهٔ زبان فارسی، تأثیر زیادی بر بازیابی اطلاعات در هر یک از سه پایگاه مورد نظر دارد. همچنین، هیچ‌کدام از این سه پایگاه به شیوه ای جامع و قابل ملاحظه به حل مسائل ریخت شناسی واژگان فارسی نپرداخته اند و هر پایگاه به صورت جداگانه از میان ۱۷ چالش پیش رو تنها به رفع تعداد محدودی از آنها پرداخته است.

کلیدواژه‌ها: بازیابی اطلاعات، ریخت‌شناسی، پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پایگاه *پژوهشگاه اطلاعات و مدارک علمی ایران*، پایگاه جهاد دانشگاهی، زبان فارسی.

---

۱. دانشجوی کارشناسی ارشد علوم کتابداری و اطلاع رسانی دانشگاه شیراز marzieh.goltaji@gmail.com
۲. دانشجوی کارشناسی ارشد علوم کتابداری و اطلاع رسانی دانشگاه شیراز sdbazrgar@yahoo.com

مقدمه

کتابداران از مدتها پیش دریافتند بین تحلیل موضوعی مطالب و زبانشناسی، رابطه‌ای مستقیم و الزامی وجود دارد. این رابطه با پیدایش علم اطلاع‌رسانی و کاربرد رایانه در این رشته، شتاب و اهمیت بیشتری یافت. امروزه اطلاع‌رسانی و اصطلاح‌شناسی شاخه‌هایی از علوم را تشکیل می‌دهند که ارتباطی نزدیک و مداوم بین آنها برقرار است. دانش اطلاع‌رسانی در حوزهٔ فعالیت خود عمدتاً با اطلاعات نوشتاری، که زبان وسیلهٔ اصلی انتقال آن است، سر و کار دارد. هسته اصلی هر زبان ویژه، اصطلاحات علمی یا واژگان آن است. این اصطلاحات برای ارتباط علمی و انتقال صحیح اطلاعات به کار گرفته می شود و چنانچه دچار هرج و مرج و نابسامانی شود، زبان تفهیم و تفاهم و جریان درست اطلاعات مختل می‌گردد. کتابداران و اطلاع‌رسانان که رابط بین تولیدکنندگان و مصرف‌کنندگان اطلاعات هستند، پیش از سایر متخصصان ضرورت استاندارد کردن واژگان علوم را دریافتند و همزمان با توسعه بانکهای اطلاعاتی، به رعایت آن اصرار ورزیدند (مرتضایی، ۱۳۸۱).

ما اکنون در دورانی به سر می بریم که با حجم عظیمی از اطلاعات در موضوعات متنوع روبرو هستیم. این کثرت اطلاعات در محیطهای الکترونیکی و بخصوص وب، گرچه باعث تسهیل دستیابی کاربران به اطلاعات مورد نیاز شده، مستلزم به کارگیری شیوه‌ها و تمهیدات خاص در بازیابی آنهاست.

از آنجا که زبان فارسی، در مواجهه با محیط الکترونیکی، از جهت شیوه نگارشی، دارای مشکلاتی است که بر کیفیت کاوش در محتویات آن تأثیر می‌گذارد، تأثیر برطرف سازی این موانع در طراحی هر پایگاه اطلاعاتی فارسی زبان بر میزان بازیابی رکوردهای مرتبط، چه از لحاظ کمیّت و چه از لحاظ محتوای

رکوردهای بازیابی شده، برکسی پوشیده نیست. از این رو، پژوهش حاضر سعی دارد با در نظر گرفتن عمده‌ترین مشکلات نگارشی در زبان فارسی و نتایج بازیافت هر کدام در سه پایگاه اطلاعاتی «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»، «پژوهشگاه اطلاعات و مدارک علمی ایران» و «پایگاه اطلاعات علمی جهاد دانشگاهی» توانایی این پایگاه‌ها را در شیوه برخورد با این مشکلات و یا برطرف کردن چالشهای مربوط، مقایسه کند.

تعریف ریخت‌شناسی[1]: بررسی ساخت کلمه و اجزای تشکیل‌دهندهٔ آن که شامل پایه، پیشوند، میانوند و پسوند می‌شود.

معایب خطوط رایج جهان

به طور کلی، نقایصی در خطوط رایج جهان وجود دارد که می‌توان آنها را بدین‌گونه طبقه‌بندی کرد:

۱- صداهای یکسان به وسیله حروف مختلفی نوشته می شود. در فارسی صدای «س» به سه صورت (س، ص، ث) و صدای «ز» چهار صورت (ز، ذ، ض، ظ) دارد؛ در زبان فرانسه کلمه «سن» پنج شیوه نگارش دارد که اگر صورتهای جمع را نیز به حساب بیاوریم، ده شکل می‌شود: (saint, ceint, sein, seing, sain).

۲- بسیاری از حروف نوشته می‌شوند، ولی خوانده نمی‌شوند؛ یعنی نشانه‌هایی بی‌فایده در نوشتن به کار می‌رود. در فارسی نوشتن «واو معدوله»[2] و «هاء غیر ملفوظ»

---

1. Morphology.

۱. واو معدوله، واوی است که در این زمان عموماً نوشته می‌شود ولی خوانده نمی‌شود، مانند خواهش. اما در زمانهای قدیم آن را با کیفیت خاصی تلفظ می‌کرده‌اند و چون در هنگام تلفظ ضمه به فتحه عدول می‌کرده است، آن را واو معدوله نامیده‌اند. هنوز در برخی از لهجه‌ها تلفظ آن به صورت قدیم مانده است. پیش از واو معدوله همیشه حرف «خ» و پس از آن یکی از حروف «د. ر. ز. س. ش. ن. و. ه. ی» آمده است.

از این قبیل است. در انگلیسی نمونه این مورد بسیار است مانند High که دو حرف آخر آن به کلی از تلفظ ساقط است، یا K در کلمه know (ناتل خانلری، ۱۳۷۳).

۳- چند صدای مختلف با یک حرف نمایانده می شود:

city/cook or ginder/girl

۴- یک صدای واحد، به صورت مجموعه‌ای از چند حرف نگاشته می‌شود:

shoot/character/nation/physics/coat

۵- بعضی صداها معادل حرفی ندارند. مثلا صدای y قبل از u شنیده می‌شود، ولی نوشته نمی‌شود:

cute/futle/utility

۶- گاهی یک حرف نماینده مجموعه‌ای از صداهاست؛ مانند حرف x در کلمه box که صدای /ks/ می‌دهد (یارمحمدی، ۱۳۶۴).

ویژگیهای زبان و خط فارسی

زبان و نگارش فارسی از ویژگیهای متنوعی برخوردار است که هر کدام از آنها به نوعی می‌تواند بر بازیابی اطلاعات در پایگاه‌های فارسی زبان تأثیر بگذارد. در ادامه، به برخی ویژگیهای شاخص و تأثیرگذار در بازیابی اطلاعات به زبان فارسی اشاره می شود:

۱. سه مصوت کوتاه یعنی حرکات زیر و زبر و پیش(ـَـــِـــُ) را از نگارش حذف می‌کنیم و این باعث می‌شود به جای اینکه از خط و نوشتار پی به معنا ببریم، از معنای کلمه و جایگاه آن در جمله، آن را درست بخوانیم؛ مانند کلمات (کَرَم، کَرَم، کِرِم، کُرم، کِرْم) و (مَلَک، مَلِک، مُلک، مِلک) و یا سه کلمه (حَکَم، حُکم، حِکَم) و نیز ننوشتن مصوتهای کوتاه در داخل متن باعث می‌شود برای تلفظ صحیح، از لاتین کلمات به صورت پانویس متن آورده شود که همین امر باعث اتلاف وقت و انرژی می‌شود. البته، همین لاتین‌نویسی هم قاعده خاصی ندارد و هر

ناشر و نویسنده‌ای سلیقه خاص خودش را برای آوانویسی حروف فارسی به لاتین دارد. به عنوان نمونه، برای نشان دادن حرکت فتحه و الف و آ هیچ‌گونه هماهنگی در کتابها و بخصوص فرهنگهای مختلف دیده نمی‌شود. هر چند برخی معتقدند همین ننوشتن حرکات مزیتی است و موجب تندنویسی می‌شود.

۲. برای یک حرف چند علامت مختلف داریم مانند علامتهای (س،ص، ث) که هر سه در فارسی یکسان خوانده می‌شوند و همچنین (ذ، ز، ض، ظ) و نیز (ت، ط). البته این امر در زبان انگلیسی هم وجود دارد، چنانکه «ف» ممکن است به شکلهای «F. GH. PH.» باشد.

۳. یک علامت را برای دلالت بر چند حرف مختلف استعمال می‌کنیم، مانند «و» که پنج مورد نوشتن دارد یکی برای بیان ضمه در کلمات «خوش» و «تو». دیگر بیان مصوت ممدود یا «واو ماقبل مضموم» مانند «شور» و «او». سوم بیان حرف صامت «واو» در کلماتی چون «آواز» و «والی» و «عفو». چهارم بیان حرف مصوت مرکبی که در کلمات "نو" و "جوشن" و مانند آنهاست. پنجم حرفی که در زبان کنونی خوانده نمی‌شود مانند «واو معدوله» در کلمات «خواهر» و «خواستن».

۴. حرفهایی وجود دارد که در بعضی از کلمات هنگام نوشتن حذف می‌شود؛ مانند «الف» در کلمات «اسحق» و "اسمعیل" و "الله". در عین حال حذف این حروف دائمی نیست و بیشتر به سلیقه نگارشی افراد بستگی دارد؛ مانند اسماعیل یا اسحاق.

۵. نقطه‌هایی متعدد در بالا و پایین حرف که هم سبب دشواری و هم موجب اشتباه در خواندن می‌شود. اهمیت بیش از حد نقطه درخط فارسی هنگام تشخیص نوری کاراکترها [1] (اُ. سی. آر.) اشکال اساسی تولید می‌کند. به عنوان مثال، کلمات

1. optical character recognition(ocr).

زیر را در نظر بگیرید که با یک یا چند نقطه عوض می‌شوند (بُر، بَر، پُر، پَر، تَر، پُز، پَز، بُز، تِز).

۶. خط فارسی از راست به چپ نوشته می‌شود و این امر نیز به نوبهٔ خود مشکلاتی به وجود می‌آورد، از جمله نبود هماهنگی و ایجاد مشکل در نوشتن متون ریاضی و شیمی، نت های موسیقی، دستورات شطرنج؛ خط تصویری یعنی علایم گرافیکی که در کل جهان استفاده می‌شود؛ مانند علایم راهنمایی و رانندگی همگی از چپ به راست نگاشته می شود.

۷. پیوسته‌نویسی و جدانویسی کلمات مرکب که در اکثر موارد به صورت سلیقه‌ای اعمال می‌شود مانند تنوع استفاده از «می» چسبان و غیر چسبان و یا تنوع نحوهٔ به کار بردن «علامتهای جمع ‹ها، ان، جات› ، هم، هیچ، که، (ضمایر شخصی متصل مان، تان، شان)، شناسی، را، چه، چون، تر، ترین، بی (پیشوند نفی)، به، ای (نشانه ندا)، آن و این» در کلمات به صورت پیوسته و یا جداگانه: (آنچه ، آن چه)؛ (همچنانکه، همچنان که)؛ (جنابعالی، جناب‌عالی)؛ (هیچکس، هیچ کس)؛ (میتواند، می‌تواند)؛ (آن ها، آنها) در این مورد کلماتی که پیشوند و یا پسوند دارند نیز در شکلهای مختلف نوشته می‌شوند. برخی از کلمات در دو شکل متصل‌نویسی و منفصل‌نویسی به دو شکل مختلف ظاهر می‌شوند. مانند «علاقمند و علاقه‌مند؛ اندیشمند و اندیشه‌مند».

مصدرها و فعلهای مرکب و اسمهای مشتق از آنها نیز به دو صورت متصل و منفصل نوشته می‌شوند؛ مانند «نگه‌داشتن و نگهداشتن». در جستجوی مطالب از اینترنت این مورد تولید اشکال می کند، چنانکه جستجوی «هیچ کس» نتایج متفاوتی را با جستجوی «هیچکس» می‌آورد و یا جستجوی «کتاب‌شناسی» و «کتابشناسی» در موتور جستجوی گوگل نتایج متفاوتی را ارائه می کند.

۸. سی و دو حرف الفبای فارسی همراه با چهار علامت مد، همزه، تنوین، تشدید به ۱۳۰ شکل مختلف ظاهر می‌شوند و تفاوت این اشکال در اتوماسیون خط فارسی تولید اشکال می‌کند. «تنوع و تعدد نویسگان[1]، یادگیری زبان و خط فارسی را برای آموزگار و آموزنده دشوار و برای نوآموز توانفرسا می‌سازد. تعداد زیاد نویسگان در رابطه با خودکارسازی زبان توسط رایانه مشکلاتی در خصوص تعداد و ترتیب قرار گرفتن نویسگان در جدولهای کد ایجاد می‌کند و طراحان کد در جای دادن این تعداد نویسه در جدولها با مسئله کمبود جا رو به رو هستند. هر چند مشکل جا با کد ۱۶ بیتی حل شده است، اما مسائل دیگری همچنان باقی می‌مانند که احتیاج به برطرف شدن دارند» (محقق زاده و زارعیان،۱۳۸۳).

۹. نوشتن ک و گ (کـ گـ ک گ گ ک) در شکلهای مختلف نیز باعث سردرگمی و عدم جستجوی صحیح می‌شود.

۱۰. در اغلب اوقات یک فاصله اضافی معنای متفاوت و یا متضادی را می‌دهد (مثل مادر ، ما در).

۱۱. سه کرسی مختلف برای حرفهای مختلف الفبا باعث می‌شود در مقایسه با اکثر زبانها تعداد سطرهای هر صفحه به مراتب بیشتر گردد، چون برخی حروف روی خط کرسی قرار می‌گیرند و برخی پایین خط کرسی و برخی بالای خط کرسی مثل (ا ب م).

۱۲. از آنجا که حروف در نوشتن اغلب به صورت چسبیده و پیوسته نوشته می‌شوند، تشخیص حرف به حرف نوشته به وسیلهٔ رایانه را، دچار مشکل می‌کند.

۱۳. در اُ. سی. آر. فارسی همچنین اعداد نیز مشکل‌ساز هستند، چنانکه صفر در فارسی یک نقطه کوچک است که می‌تواند رایانه را به اشتباه بیندازد و نیز اعداد ۱ و ۲ و ۳ بسیار شبیه هم هستند و تفاوتشان در یک دندانه کوچک است.

. . . . . .

۱. نویسگان، جمع نویسه معادل Characters

۱۴. تنوع املایی یا تنوع در رسم‌الخط بعضی از کلمات که همه شکلهای آن نیز درست است مانند «اتاق و اطاق» و یا «امپراتور و امپراطور». و کلماتی که فقط یک شکل آنها صحیح است، ولی شکل ناصحیح آن نیز زیاد استفاده می‌شود، مانند « ذغال و زغال؛ خوشنود و خشنود». البته این جدا از تنوع در مفهوم کلمات است که در دیگر زبانها نیز وجود دارد، یعنی برای بعضی از مفاهیم ممکن است کلمات متنوعی استفاده شود؛ مانند «کامپیوتر و رایانه».

۱۵. به کار بردن همزه درصورتهای مختلف مانند (مسأله، مسئله)؛ (مسئول، مسوول).

۱۶. استفاده از « ا» و « آ » به جای یکدیگر مانند (فرایند و فرآیند).

۱۷. شکلهای مختلف ضبط نامهای بیگانه در فارسی: ورود واژه های بیگانه معمولا از راه ورود پدیده‌های فرهنگی نو در عرصه های مختلف فنی، علمی، اجتماعی، سیاسی و هنری و .... و یا از طریق افراد دو زبانه انجام می‌گیرد که به وام‌گیری زبان معروف است و کم و بیش در تمام زبانها وجود دارد. واژه‌های بیگانه اغلب برای پرکردن خلأ واژه‌های علمی و یا ارتباطی سودمندند، اما وجود آنها مسائلی از قبیل چگونگی ضبط آنها در زبان وام را به وجود می‌آورد. برای ضبط واژه‌های به وام گرفته شده به سبب اختلاف فاحش نشانه‌های الفبای فارسی با نشانه‌های الفبای خارجی، مشکلات جدی وجود دارد. از جمله اینکه الفبای فارسی آوانگار نیست و به همین جهت در ضبط دقیق تلفظ واژه‌های زبان فارسی نیز ناتوان است. این ناتوانی در ضبط واژه‌های بیگانه به مراتب بیشتر است. در مورد برگردان اسامی خارجی به خط فارسی نیز قاعده خاصی وجود ندارد و هر کس بنا بر سلیقه و ذوق خود این کار را انجام می‌دهد، در نتیجه یک کلمه واحد به شکلهای مختلف

نوشته می‌شود. برای مثال (اتومبیل و اتوموبیل)؛ (کلسیم، کلسیوم، کالسیوم) و یا اسم Franklin به صورت (فرانکلین، فرانکلن، فرنکلین، فرنکلن) ضبط شده است.

۱۸. استفاده یا استفاده نکردن از «ی» در کلمات مختوم به «الف» مانند (موسی و موسا).

۱۹. استفاده یا استفاده نکردن از «ء» برای کلمات مختوم به های بیان حرکت در حالت مضاف مانند (خانه مسکونی و خانهء مسکونی و یا خانه‌ی مسکونی).

۲۰. استفاده یا استفاده نکردن از اعراب برای کلمات.

۲۱. انواع مختلف جمع برای یک واژه مفرد: به عنوان مثال، جمع بستن یک واژه با علایم جمع فارسی وعلایم جمع عربی مانند (معلم، معلمین، معلمان، معلم‌ها).

۲۲. تنوینهای زبان عربی نیز از جمله دشواریهای رعایت اصل همخوانی نوشتاری و گفتاری هستند.

۲۳. در نگارش یاء وحدت یا نکره در آخر کلماتی که به هاء مختفی یا غیر ملفوظ ختم می‌شوند، سه نوع املا دیده می‌شود. (خانه‌ای، خانه‌یی، خانهٔ).

۲۴.کلمه‌های عربی در شکلهای گوناگون در زبان فارسی نوشته می‌شوند. (مبدا، مبداء)؛ (ابتدا، ابتداء)؛ (نسبتاً، نسبته، نسبتا) و ....

۲۵. وجود دندانه‌های متعدد در کلمات، خواندن کلمات و بخصوص در او.سی.آر. فارسی اشکال ایجاد می‌کند؛ مانند کلمات: نشستن و استشهاد.

۲۶. حروف فارسی اغلب مشابهند و با اندکی غفلت به جای هم نوشته می‌شوند و مطلب را به کلی دگرگون می‌کنند، مانند (در، رد، ور) (راثی ساربانقلی، ۱۳۸۴).

بنابراین، با در نظر گرفتن موارد فوق می توان چنین استنباط کرد که این ویژگیها در زبان فارسی با وجود اینکه در خواندن متن اشکال کمی به وجود

می‌آورند و هر آشنای به زبان فارسی به راحتی می‌تواند آنها را بخواند، در فناوری امروزه و تجزیه و تحلیل کلمات به کمک رایانه اشکال اساسی تولید می‌کنند و چنانچه قاعده‌ای جامع و مانع برای آنها وضع گردد، بزرگ‌ترین مشکل خط فارسی حل می‌شود. منظور اینکه، برای مثال خواندن سه کلمه «بی‌حوصلگی، بیحوصلگی، بی‌حوصله‌گی» مشکلی ایجاد نمی‌کند. اما در محیط الکترونیکی و شبکه اینترنت برای بازیابی این کلمه باید برای تمام شکلهای آن، جستجو را انجام دهیم (البته اگر از تمام شکلهای نوشتاری آن آگاهی داشته باشیم).

بیان مسئله و اهمیت پژوهش

امروزه روش غالب در جستجوی اطلاعات از پایگاه‌های اطلاعاتی، روش کلیدواژه ای است. اما جستجو به این روش، دشواریهای خاص خود را دارد. چنانچه فردی به دنبال اطلاعاتی در مورد «کتابشناسی» باشد، این کلیدواژه را می تواند به سه شکل بنویسد: «کتابشناسی، کتاب‌شناسی و کتاب‌شناسی». از آنجا که پایگاه‌های اطلاعاتی، نظامهایی تطبیق دهنده هستند، دقیقاً همان کلمه‌ای را بازیابی خواهند کرد که وارد جعبه جستجو شده است. بنابراین، برای هرکدام از این شکلها، تعداد نتایج متفاوتی بازیابی خواهد شد. چنانچه کاربری تنها یک شکل از این سه مورد را به کار ببرد، اطلاعاتی را که به شکلهای دیگر نوشته شده است، از دست خواهد داد. از این رو، در این مقاله سعی خواهد شد تا مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و پایگاه اطلاعاتی جهاد دانشگاهی بررسی و در پایان پیشنهادهایی برای بهبود کارایی پایگاه‌های اطلاعاتی فارسی ارائه شود.

هدفهای پژوهش

۱. بررسی برخی چالشهای ریختی شناخته شده زبان فارسی، بر میزان بازیابی اطلاعات در سه پایگاه اطلاعاتی فارسی مورد بررسی

۲. مقایسه میزان توانایی سه پایگاه در رفع چالشهای ریختی مدنظر.

پیشینهٔ پژوهش

از آنجا که پیشینه‌های یافت شده برای این پژوهش به دو دسته مشکلات نگارشی در زبان فارسی و مشکلات نگارشی در سایر زبانها قابل دسته‌بندی است، هر دو گروه را جداگانه بررسی خواهیم کرد.

الف) پیشینه‌های مرتبط در زبان فارسی

«حرّی» (۱۳۷۲) در مقالهٔ خود با عنوان «کامپیوتر و رسم‌الخط فارسی» بیان می‌دارد که یکی از متغیرهای عمده در ذخیره و بازیابی اطلاعات فارسی، رسم الخط یا شیوه خط فارسی است. گرچه حروف و کلمات به عنوان ورودی و خروجی هر سیستم رایانه‌ای در هر زبان اهمیت دارند، خط فارسی به دلیل ویژگی آن، در رویارویی با رایانه دارای مسائل پیچیده تری است. وی معتقد است پنج مورد که اختصاصاً به مسئله پیوند میان زبان فارسی و رایانه مربوط می‌شود، از این قرار است: هماهنگ کردن حروف، استفاده از تکواژها، استفاده از سیاهه آماده، پیوند ساختگی میان کلمات، هماهنگی رسم‌الخط.

«سمائی و همکاران» (۱۳۷۹) در طرح پژوهشی با عنوان «یکسان‌سازی شیوه رسم‌الخط اسامی ترکیبات شیمیایی در زبان فارسی» تلاش کرده‌اند شیوهٔ نگارش اسامی ترکیبات شیمیایی و بخصوص ترکیبات آلی در زبان فارسی و معضلات مربوط به آن را بررسی و الگوهایی برای یکسان نویسی آنها پیشنهاد کنند.

«نشاط» (۱۳۷۹)، در بررسی خود با عنوان «مسائل رسم الخط فارسی در رویارویی با فناوری نوین اطلاعاتی» سعی دارد با استفاده از شواهد موجود و الزامهای مربوط به زبان نظامهای رایانه‌ای به عنوان وجه غالب فناوریهای نوین و نیز ناسازگاری میان این دو ، تصویری از وضع موجود را عرضه و راه حلهای ممکن را ارزیابی کند.

«مرتضایی» (۱۳۸۱) در مقاله‌ای با عنوان «مسائل زبان و خط فارسی در ذخیره و بازیابی اطلاعات» مشکلات گوناگونی را که در جریان ذخیره و بازیابی اطلاعات و ایجاد پایگاه های اطلاعاتی به زبان فارسی به وجود می‌آید، بررسی کرده است.

«محقق‌زاده و زارعیان» (۱۳۸۳) در مقاله‌ای با عنوان «ارائه راه‌حل برای برخی مسائل اتوماسیون و نگارش فارسی» ضمن برشمردن ایرادهایی که در مورد پردازش خط فارسی به وسیلهٔ رایانه به وجود می‌آید، پیشنهادهایی را برای این مشکل ارائه می‌کند.

«بی‌جن خان» (۱۳۸۳) در مطالعه خود نقش پیکره های زبانی را در نوشتن دستور زبان بررسی و نقدهایی را بر پیکره های زبانی مطرح کرده است. وی به رابطهٔ دستور زبان و پیکره زبانی اشاره و نرم‌افزاری را برای انواع جستجو در پیکره‌ها، تجزیه و تحلیل آماری داده‌ها و در نهایت گزارش‌گیری از داده‌های آماده‌سازی شده، معرفی می‌کند. وی در نتیجهٔ تحقیق خود بیان می‌دارد که با استفاده از این روش علاوه بر اینکه می‌شود ساخت احتمالی نظامهای زبانی را مطالعه کرد، یافته‌های زبانشناسی نظری را هم می‌توان در حوزهٔ دستور زبان در قالب فرضیه های زمانی محک زد.

«راثی ساربانقلی» (۱۳۸۴) در مقاله خود با عنوان «مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت، مطالعه موردی: کاربران مرکز اینترنت دانشگاه

آزاد اسلامی واحد شبستر» مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت توسط کاربران مرکز اینترنت دانشگاه آزاد اسلامی شبستر را بررسی نمود. نتایج پژوهش نشان داد بیشتر مشکل کاربران در جستجو، توجه نکردن به شکلهای مختلف نوشتاری واژه و استفاده نکردن از عملگر OR می‌باشد.

«عبدالهی نورعلی» (۱۳۸۶) در پژوهش خود با عنوان «کندوکاو مسائل ریخت‌شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای وب» به بررسی مسائلی پرداخته است که جستجوگرهای فارسی در کاوش ریخت‌شناسی مختلف یک کلمه با آن روبرو هستند. برای این مهم از سه جستجوگر بین‌المللی گوگل، یاهو و آلتاویستا[1] که امکان جستجو به زبان فارسی را دارند، استفاده شد. نتایج نشان داد هیچ کدام از جستجوگرها، چالشهای زبان شناختی زبان فارسی را در جهت بهبود کاوش، مورد توجه قرار نداده اند. با توجه به دستاوردهای حاصل از پژوهش، الگویی برای ایجاد اصلاحات در شیوهٔ نگارش فارسی ارائه شد تا از این طریق بتوان پردازش متون رایانه ای را تسهیل نمود.

ب) پیشینه‌های مرتبط در سایر زبانها

«هدلاند[2] و دیگران» (2000) ویژگیهای زبان سوئدی را از نظر بازیابی بررسی کردند. آنها مطالعه‌ای مقایسه‌ای بر روی زبانهای سوئدی، فنلاندی و انگلیسی انجام دادند تا میزان ابهامهای لغوی را در این زبانها معیّن کنند. محققان پیشنهاد می‌کنند برچسب‌گذاری ادات سخن جهت بازیابی کلمات هم نگاشت، می‌تواند مفید باشد.

«مونز و دوریکه»[3] (2002) با تمرکز بر اثرات تحلیلهای ریخت‌شناسی همچون ریشه‌سازی و جداسازی کلمات مرکب، کارآیی بازیابی اطلاعات را بررسی کردند. این مطالعه بر روی زبانهای هلندی، آلمانی و ایتالیایی انجام شده است. نتایج نشان

1. Altavista.
2. Hedlund.
3. Monz & De Rijke.

داد بازیابی اطلاعات حدود ۲۵٪ برای زبان آلمانی، ۶۹٪ برای زبان هلندی و ۲۵٪ برای زبان ایتالیایی بهبود یافت.

«درویش»[1] (2002) روشی را برای ایجاد یک تحلیلگر ریخت‌شناسی ارائه می‌دهد. این تحلیلگر توانایی تولید ریشه‌های احتمالاتی یک کلمه را خواهد داشت. در این نظام، قواعد ریشه‌سازی خودکار مورد استفاده قرار گرفته است. محقق برای ارزیابی این نظام، آن را با یک تحلیلگر ریخت‌شناسی عربی موجود در بازار مقایسه کرده است.

«مقداد»[2] (2005) در یک پژوهش، عملکرد سه ابزار جستجوی عمومی را با سه جستجوگر عربی (که اختصاصاً مسائل زبان شناختی عربی را لحاظ می‌کنند) مقایسه کرد. نتایج نشان داد جستجوگرهای عمومی، نظیر آلدوب[3]، آلتاویستا و گوگل در بازیابی مدارک عربی، ناقص عمل می‌کنند. همچنین، نتایج این تحقیق، نیاز به تحقیقات بیشتر در زمینهٔ عملی بودن ابزارهای جدید بازیابی اطلاعات در جستجوگرها را نشان داد.

«تاث»[4] (2006) به بررسی قابلیتهای زبان شناختی جستجوگرهای انگلیسی و مجاری پرداخت. محقق سه ابزار جستجوی انگلیسی به نامهای گوگل، آلتاویستا و آلدوب را با پنج جستجوگر محلی مقایسه نمود. تحلیل داده‌ها بر پایه چند شاخص انجام شد که عبارت بودند از: ریشه‌سازی، بازیابی لهجه‌های مختلف، کوتاه‌سازی و جستجوی مترادفها. نتایج حاکی از آن بود که جستجوگرهای محلی، مسائل زبان مجاری را بهتر از جستجوگرهای انگلیسی مورد توجه قرار داده‌اند. ابزارهای انگلیسی

. . . . . .

1. Darwish.
2. Moukdad.
3. Altheweb.
4. Toth.

زبان، لهجه‌های مختلف زبان مجاری را به خوبی پشتیبانی نمی‌کردند، که این امر به بازیابی ضعیف اطلاعات منجر می‌شد.

پرسشهای پژوهش

۱- چالشهای ریختی شناخته شده زبان فارسی چه تأثیری بر بازیابی اطلاعات در هر یک از سه پایگاه مورد نظر داشته است؟

۲- کدام یک از سه پایگاه مورد نظر، چالشهای ریختی مورد نظر را در الگوریتمهای بازیابی خود مورد توجه قرار داده اند؟

روش‌شناسی پژوهش

این پژوهش با استفاده از روش پیمایش مقایسه‌ای انجام پذیرفته است. داده‌های جدولها نیز بر اساس آمار توصیفی بررسی شده است. زمان گردآوری داده‌ها مهر ۸۷ بود. از آنجا که این سه پایگاه از جمله پایگاه‌های مهمی هستند که مقاله‌های فارسی را نمایه می‌کنند، در پژوهش حاضر مورد بررسی قرار گرفتند.

شیوهٔ اجرای تحقیق بدین شکل است که ابتدا سیاهه‌ای مشتمل بر ۱۷ چالش نگارشی در زبان فارسی با استفاده از متون موجود شناسایی شد. سپس برای هریک از آنها مصداقهای موجود در زبان فارسی انتخاب و بررسی گردید. مصداقها به صورت کلیدواژه هایی در سه پایگاه جستجو شد تا اطمینان حاصل شود دست کم یک پیشینه برای آن چالش وجود داشته باشد. این کلیدواژه ها به عنوان وسیله گردآوری داده ها به کار گرفته شده اند و نتایج هر یک از جستجوها در قالب تعداد رکوردهای یافت شده برای هر واژه در هر سه پایگاه، در جدول شماره ۱ ارائه شده است.

شایان ذکر است، برای اطمینان از اینکه بازیابی‌های هم تعداد یک محتوا دارند، رکوردهای بازیابی شده به صورت گزینشی با هم مقایسه گردید.

در بعضی موارد نیز برای کنترل رخداد یک واژه از کنترل مدارک همپوشان در سه پایگاه استفاده شد؛ بنابراین تا حد امکان از نبود رویداد یک واژه در سه پایگاه اطمینان حاصل شد، لذا مقدار صفر در جدول یک به معنای یافت نشدن رکورد برای ریخت مورد نظر در پایگاه است.

در جدول شماره ۲، نسبت تعداد رخدادهای مختلف واژگان به صورت دو به دو برای هر واژه در هر یک از سه پایگاه محاسبه گردید. همان‌طور که مشاهده می‌شود، در صورت یکسان بودن تعداد نتایج، برچسب «یک» و در غیر این صورت برچسب «غیر از یک» به هر کدام داده شد. بنابراین، مفهوم «یک» در آن جدول به احتمال قوی به معنای یکسان بودن رکوردهای بازیابی شده از دو صورت واژه در پایگاه مورد نظر است. برای مثال، در مرکز منطقه ای اطلاع رسانی علوم و فناوری، برای واژه «محمد» ۱۳۹۸۲ رکورد و برای واژه «محمّد» نیز ۱۳۹۸۲ رکورد بازیابی شد؛ لذا نسبت «یک» میان این دو واژه در این پایگاه برقرار است.

جدول ۱. آمار نتایج بازیابی شده برای هرکدام از مشکلات زبان فارسی به تفکیک پایگاه ها

ردیف	مشکل	مشکل ریخت‌شناسی	پایگاه		
			مرکز منطقه‌ای	ایرانداک	جهاد دانشگاهی
۱	تشدید	محمد	۱۳۹۸۲	۱۲۸۸۱	۸۶
		محمّد	۱۳۹۸۲	۰	۰
۲	همزهٔ پایانی	املا	۸	۲	۶۲
		املاء	۱۲	۱	۱
۳	نشانه‌های جمع	معلمان	۴۰۱	۱۴۰	۱۰۳
		معلمین	۱۸	۰	۱۸
۴	برگرداندن کلمات خارجی	آمریکا	۴۵۰۷	۶۷۶	۳۰
		امریکا	۵۷۲	۳۳	۶

ردیف	مشکل	مشکل ریخت‌شناسی	پایگاه		
			مرکز منطقه‌ای	ایرانداک	جهاد دانشگاهی
۵	**های غیر ملفوظ**	واژگان	۵۰۹	۵۳	۶۵
		واژه گان	۰	۰	۰
۶	**تنوین**	واقعاً	۱۰۷	۰	۱
		واقعا	۱۰۷	۳۴	۳
۷	**همزه متصل به یای وحدت**	رضایی	۱۱۴۹	۷۲۱	۱۱۸۷
		رضائی	۲۲۲	۱۴۰	۶
۸	**استفاده از "ا" و"آ" به جای هم**	درآمد	۹۷۰	۱۷۲	۱۴۶
		درامد	۱	۲	۳
۹	**الف مقصوره**	اسحاق	۹۷	۳۷	۱
		اسحق	۲۰	۱۴	۲
۱۰	**پیوسته نویسی ، بی فاصله نویسی یا جدانویسی ترکیبات**	کتابشناسی	۶۸۳	۴۲	۵
		کتاب شناسی	۱۴۵	۲۲	۲
		کتاب‌شناسی	۶۸۳	۲۲	۰
۱۱	**تای منقوط**	مشکات	۱۳	۲	۹
		مشکوه	۱۰۹۸	۲۶	۲۱
		مشکوۀ	۸	۰	۱
۱۲	**صامت میانجی«ی»**	دو استقامت	۲	۳	۲
		دوی استقامت	۰	۱	۰
۱۳	**تنوع صورتهای درست یک کلمه**	اتاق	۳۵۸۹	۱۰۳	۴۱
		اطاق	۲۱	۴	۴
۱۴	**همزه به صورتهای مختلف**	مسئول	۷۷	۱۳	۱۷
		مسؤول	۰	۱	۱۷
۱۵	**تنوع در تلفظ**	داود	۱۱۵۴	۴۴۱	۵
		داوود	۳۵۶	۱۷۲	۲
۱۶	**خط تیره**	اقتصادی اجتماعی	۴۸۸۱	۲۹۵	۱۳
		اقتصادی - اجتماعی	۱	۲۹۵	۲۲
۱۷	**نقطه بین سرنامها**	اچ آی وی	۱۴	۲۳	۵
		اچ. آی. وی	۰	۲۳	۰

جدول۲. مقایسه نسبت ریختهای مختلف هر واژه در پایگاه های مختلف

ردیف	مشکل	نسبت ریخت کلمات به یکدیگر	نسبت تعداد ریخت بازیابی شده کلمات به تفکیک پایگاه		
			مرکز منطقه‌ای	ایرانداک	جهاد دانشگاهی
۱	**تشدید**	محمد/ محمّد	۱	غیر از یک	غیر از یک
۲	**همزه پایانی**	املا/ املاء	غیر از یک	غیر از یک	غیر از یک
۳	**نشانه های جمع**	معلمان/معلمین	غیر از یک	غیر از یک	غیر از یک
٤	**برگرداندن کلمات خارجی**	آمریکا/امریکا	غیر از یک	غیر از یک	غیر از یک
٥	**های غیر ملفوظ**	واژگان/واژه گان	غیر از یک	غیر از یک	غیر از یک
٦	**تنوین**	واقعاً/واقعا	۱	غیر از یک	غیر از یک
۷	**همزه متصل به یای وحدت**	رضایی/رضائی	غیر از یک	غیر از یک	غیر از یک
۸	**استفاده از "ا" و"آ" به جای هم**	درآمد/درامد	غیر از یک	غیر از یک	غیر از یک
۹	**الف مقصوره**	اسحاق/اسحق	غیر از یک	غیر از یک	غیر از یک
۱۰	**پیوسته نویسی ، بی‌فاصله نویسی یا جدانویسی ترکیبات**	کتابشناسی/کتاب شناسی	غیر از یک	غیر از یک	غیر از یک
		کتاب شناسی/ کتاب‌شناسی	غیر از یک	۱	غیر از یک
		کتابشناسی/کتاب‌شناسی	۱	غیر از یک	غیر از یک
۱۱	**تای منقوط**	مشکات/مشکوه	غیر از یک	غیر از یک	غیر از یک
		مشکوه/مشکوة	غیر از یک	غیر از یک	غیر از یک
		مشکات/مشکوة	غیر از یک	غیر از یک	غیر از یک
۱۲	**صامت میانجی«ی»**	دو استقامت/دوی استقامت	غیر از یک	غیر از یک	غیر از یک
۱۳	**تنوع صورتهای درست یک کلمه**	اتاق/اطاق	غیر از یک	غیر از یک	غیر از یک
۱٤	**همزه به صورتهای مختلف**	مسئول/مسؤول	غیر از یک	غیر از یک	۱
۱٥	**تنوع در تلفظ**	داود/داوود	غیر از یک	غیر از یک	غیر از یک
	خط تیره	اقتصادی - اجتماعی/ اقتصادی- اجتماعی	غیر از یک	۱	غیر از یک
۱٦	**نقطه بین سرنام**	اچ آی وی/اچ. آی. وی.	غیر از یک	۱	غیر از یک
۱۷	**تعداد موارد حل شده**		۳	۳	۱

یافته‌های پژوهش

با استفاده از آمارهای داده شده در جدولهای فوق، می‌توان در پاسخ به سؤالهای پژوهش چنین بیان داشت:

۱- چالشهای ریختی شناخته شده زبان فارسی چه تأثیری بر بازیابی اطلاعات در هر یک از سه پایگاه مورد نظر داشته است؟

باید گفت، بر اساس اطلاعات ارائه شده در جدول شماره ۱، شاهدیم که هر شکل نوشتاری کلمه در زبان فارسی نتایج متعددی را در هر پایگاه اطلاعاتی در پی دارد. به طور مثال، به بررسی تأثیری که الف مقصوره و لحاظ یا عدم لحاظ آن در کلمه «اسحاق» داشته است، خواهیم پرداخت:

طبق آمار به دست آمده از جدول شماره ۱، برای کلمه «اسحاق» در دو شکل نوشتاری مختلف شاهدیم که چنانچه برای نوشتن این کلمه از الف مقصوره استفاده نشود، نتایج به دست آمده در سه پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران، و جهاد دانشگاهی به ترتیب تعداد رکوردهای بازیابی شده ۹۷، ۳۷ و ۱ می باشد و چنانچه در حالتی دیگر برای نوشتن کلمه «اسحاق» از الف مقصوره استفاده شود و شکل نوشتن این کلمه در جعبه جستجوی پایگاه به شکل «اسحق» باشد، نتیجه متفاوتی به دست خواهد آمد، به طوری که در این حالت رکوردهای بازیابی شده در سه پایگاه مدنظر به ترتیب ۲۰، ۱۴، ۲ می‌باشد. بدین ترتیب، متوجه می‌شویم مشکل ریختی الف مقصوره در هر سه پایگاه اطلاعاتی فارسی باعث اختلاف در تعداد رکوردهای بازیابی شده، می شود و چنانچه کلمه «اسحق» را به جای کلمه «اسحاق» در جعبه جستجوی پایگاه اطلاعاتی مرکزمنطقه ای اطلاع رسانی علوم و فناوری وارد کنیم، ۷۷ رکورد و در پایگاه اطلاعاتی پژوهشگاه اطلاعات و مدارک علمی ایران ۲۳ رکورد اطلاعات را از

دست خواهیم داد و بدین طریق این عامل باعث ریزش رکوردهای اطلاعاتی مفید خواهد شد، اما در پایگاه جهاد دانشگاهی با شکل «اسحاق» ۱ رکورد و با شکل «اسحق» ۲ رکورد بازیابی شده است؛ یعنی افزایش تعداد رکورد رخ داده است. بنابراین، شاهدیم که چگونه شکلهای متنوع نوشتار کلمات می تواند باعث کاهش یا افزایش تعداد رکوردهای بازیابی شده در پایگاه‌های فارسی زبان شوند.

۲- کدام یک از سه پایگاه مورد نظر، چالشهای ریختیِ ذکر شده را در الگوریتمهای بازیابی خود مورد توجه قرار داده اند؟

بر اساس اطلاعات موجود در جدول شماره ۲، شاهدیم که هیچ کدام از سه پایگاه فارسی مورد نظر، به شیوه‌ای جامع چالشهای ریخت شناسی زبان فارسی را در جهت بهبود نتایج کاوش مورد توجه قرار نداده‌اند، اگرچه در بعضی موارد تساوی تعداد رکوردهای بازیابی شده در ریختهای مختلف یک واژه را به احتمال قوی می‌توان به منزلهٔ رفع آن چالش خاص در الگوریتم بازیابی پایگاه در نظر گرفت، اما نمونه‌هایی از این دست برای هر پایگاه نسبت به حجم مشکلات ریخت‌شناسی موجود، درصد بسیار اندکی را به خود اختصاص می دهد. به طور مثال، از میان ۱۷ چالش موجود که در جدولهای فوق طرح شد، پایگاه مرکز منطقه ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران، و جهاد دانشگاهی به ترتیب موفق به حل سه، سه و یک مورد از مسائل ریخت شناسی زبان فارسی شدند. پایگاه‌های مذکور از میان تمامی چالشهای ریخت شناسی مطرح شده تنها برای موارد زیر چاره جویی نموده اند:

پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری: تنوین، تشدید، پیوسته‌نویسی و بی‌فاصله‌نویسی؛ پژوهشگاه اطلاعات و مدارک علمی ایران:

جدانویسی و بی فاصله نویسی، خط تیره، نقطه بین سرنام‌ها؛ پایگاه جهاد دانشگاهی: همزه به صورتهای مختلف.

نتیجه‌گیری

بدیهی است، پردازش بهتر و سریع‌تر متون فارسی با استفاده رایانه در زمانه ما یک ضرورت اساسی به نظر می رسد. پایگاه های اطلاعاتی که با استفاده از زبان و شیوه خط کنونی به ذخیره و بازیابی اطلاعات می پردازند، نمی توانند کارایی مطلوبی داشته باشند و این شکلهای متنوع نوشتار کلمات می تواند باعث کاهش یا افزایش تعداد رکوردهای بازیابی شده در پایگاه‌های فارسی زبان شود. بر این اساس، شاهدیم که پایگاه‌های اطلاعاتی فارسی با وجود عمر نسبتاً کوتاه، با مشکلات بسیاری دست به گریبانند، که اگر هر چه زودتر چاره‌اندیشی نشود، با توجه به هجوم اطلاعات دیگر، مهار آن آسان نخواهد بود. نتایج بررسی نشان داد هیچ‌کدام از سه پایگاه فارسی مورد نظر، به شیوه‌ای جامع چالشهای زبانشناختی زبان فارسی را در جهت بهبود نتایج کاوش مورد توجه قرار نداده‌اند. اگرچه در بعضی موارد تساوی تعداد رکوردهای بازیابی شده در ریختهای مختلف یک واژه را می‌توان به منزلهٔ رفع آن چالش خاص در نظر گرفت، اما نمونه‌هایی از این دست برای هر پایگاه نسبت به تعدد مشکلات ریخت‌شناسی موجود، درصد بسیار اندکی را به خود اختصاص می‌دهد. به طور مثال، از میان ۱۷ چالش موجود که در جدولهای فوق طرح شد، پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران، و جهاد دانشگاهی به ترتیب موفق به حل سه، سه و یک مورد از مسائل ریخت‌شناسی زبان فارسی شدند.

پیشنهادها

با توجه به اینکه هر کدام از سه پایگاه مورد بررسی توانسته در سطحی محدود بر برخی از چالشهای ریخت‌شناسی واژگان فارسی فایق آید، و با در نظرگرفتن این نکته که مشکل حل شده هر پایگاه با سایر پایگاه‌ها متفاوت است و این سه پایگاه مورد نظر در کشور ما از پایگاه های علمی معتبر به شمار رفته و هر روز نیز بر تعداد کاربران آنها افزوده می شود، متأسفانه اغلب کاربران از آنچه به هنگام جستجو در این پایگاه ها رخ می‌دهد، آگاهی ندارند. از این‌رو، آگاهی نداشتن و همچنین تنوع صورت نوشتاری، تأثیر زیادی بر از دست دادن مدارک مربوط در هر یک از سه پایگاه اطلاعاتی مورد بررسی دارد. بنابراین، پیشنهاد می‌شود طراحان پایگاه‌های اطلاعاتی فارسی در نشستی پیرامون این موضوع به ارائه تجربه‌ها و یافته‌های خود پرداخته و از دستاوردهای دیگران در این حیطه بهره ببرند. همچنین می توان هنگام طراحی پایگاه‌ها، آنها را به اصطلاحنامه مجهز نمود تا کاربران از ریختهای مختلف واژه به اصطلاح پذیرفته شده راهنمایی شوند. همین‌طور طراحان می توانند تمهیداتی را دربارهٔ چگونگی استفاده از پایگاه و الگوریتمهای مرتبط با مسائل ریختی واژگان به کاربرده شده برای جستجو در اختیار کاربران قرار دهند تا از این طریق آنها راحت‌تر به جستجو بپردازند و بتوانند حداکثر نتایج دلخواه خود را بیابند. به نظر می‌رسد همکاری بین متخصصان زبان شناسی با متخصصان عرصه بازیابی اطلاعات به منظور جهت دهی تحقیقاتی در این زمینه بسیار ضروری است.

منابع

- بی جن خان، محمود (۱۳۸۳). نقش پیکره‌های زبانی در نوشتن دستور زبان، معرفی یک نرم‌افزار رایانه‌ای. *مجله زبان‌شناسی*، ۱۹ (۲)، ۴۸-۶۷.

- راثی ساربانقلی و محمد صابر (۱۳۸۴). مهارت در جستجوی اطلاعات فارسی از اینترنت. *مجله الکترونیکی نما*، ۵ (۱)، بازیابی ۲۲ آبان ۱۳۸۷، از http://www.irandoc.ac.ir/Data/E_J/vol5/rasi.htm

- حری، عباس (۱۳۷۲). کامپیوتر و رسم‌الخط فارسی. *پیام کتابخانه*. ۳ (۱)، ۶-۱۱.

- سمائی، مهدی (۱۳۷۹). *یکسان‌سازی شیوه رسم‌الخط اسامی ترکیبات شیمیائی در زبان فارسی*. طرح پژوهشی، مرکز اطلاعات و مدارک علمی ایران، تهران.

- عبدالهی نورعلی، محمدصادق (۱۳۸۶). *کندوکاو مسائل ریخت شناسی زبان فارسی در بازیابی اطلاعات از جستجو گرهای وب*. پایان نامه کارشناسی ارشد کتابداری و اطلاع رسانی، دانشگاه شیراز، شیراز.

- محقق‌زاده، محمدصادق و کاظم زارعیان (۱۳۸۳). ارائه راه‌حل برای برخی مسایل اتوماسیون نگارش فارسی. *فصلنامه اطلاع‌رسانی*، ۱۹(۳-۴)، ۱-۱۰.

- مرتضایی، لیلا (۱۳۸۱). مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. *فصلنامه اطلاع‌رسانی*، ۱۷ (۱-۲)، ۱-۷.

- ناتل خانلری، پرویز (۱۳۷۳). *زبان‌شناسی و زبان فارسی*. تهران: توس.

- نشاط، نرگس (۱۳۷۹). مسائل رسم‌الخط فارسی در رویارویی با فناوری نوین اطلاعاتی. در *فهرست‌های رایانه‌ای، کاربرد و توسعه*. مجموعه مقالات همایش کاربرد و توسعه فهرست‌های رایانه‌ای در کتابخانه‌های ایران، آبان ۲۷-۲۸، (۴۰۱-۴۰۸). مشهد: دانشگاه فردوسی مشهد.

- یارمحمدی، لطف الله (۱۳۶۴). *درآمدی به آواشناسی*. تهران: مرکز نشر دانشگاهی.

- Darwish, K. (2002). Building a Shallow Arabic Morphological Analyzer in One Day. Annual Meeting of the ACL, Proceedings

of the ACL-workshop on Computational approaches to Semitic languages. Philadelphia, 19-28.
- Hedlund, T., Pirkola, A. ,& Kalervo, J. (2001). Aspects of Swedish morphology and Semantics from the perspective of mono- and cross-language information retrieval. *Information Processing and Management,*37,147-161.
- Retrieved November 3, 2008, from http://www.dcs.shef.ac.uk /nlp/clarity/papers/SWEIR-Hedlund-IPM01.pdf .
- Monz, C. , & De Rijke, M. (2002). Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian. Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross language Evaluation Forum, CLEF 2001, Darmstadt, Germany.
- Moukdad, H. (2005). Lost In cyberspace: How Do Search Engines Handle Arabic Queries?.*The international information & library review*,37(4),237-394.
- Toth, E. (2006). Exploring the Capabilities of English and Hungarian Search Engine for Various Queries. *Libri*, 56, 38-47.