تحلیل چالشهای پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی

سمیه سادات آخشیک[1]
دکتر رحمت ا... فتاحی[2]

چکیده

مقدمه: ویژگیهای خاصّ دستوری و نگارشی زبان و خط فارسی، دشواریهایی را در ذخیره و بازیابی اطلاعات در محیط رایانه‌ای پدید آورده است. رسم‌الخط فارسی نیز از یک‌سو به علت اختلاف نظر پدیدآورندگان متون و از سوی دیگر پیچیدگیهای ذاتی خود، به‌هنگام ذخیره، جستجو و بازیابی چالشهای متعددی را برای طراحان و نمایه‌سازان پایگاه‌ها، کاربران و پدیدآورندگان منابع به‌وجود آورده است.

روش بررسی: این پژوهش به روش تحلیل محتوا انجام شد. ۱۰۰ عنوان از پایان‌نامه‌های موجود رشته کتابداری و اطلاع‌رسانی به منزله نمونه‌ای از متون فارسی در پایگاه‌های اطلاعاتی پژوهشگاه‌های علوم و فناوری اطلاعات ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و از هر پایگاه ۵۰ عنوان به‌صورت تصادفی انتخاب شد. با استفاده از دستور خط فارسی مصوب فرهنگستان زبان، کلماتی که درست یا نادرست نوشته شده بود، از یکدیگر تفکیک و در مرحله بعد، عنوانهای مورد نظر در هر دو پایگاه و با حالتهای متفاوت کلمات مرکب، جستجو گردید و در نهایت، نتایج بازیابی در پایگاه‌ها، ارزیابی و مقایسه شد.

یافته‌ها: نتایج این بررسی نشان داد ۷۱/۲٪ از کلمات عنوانها به صورت درست و ۲۸/۸٪ نادرست نگارش شده‌اند. همچنین، مشخص شد ۵۱/۶٪ این کلمات، دو جزئی و ۴۷/۵٪ سه جزئی هستند و اغلب نویسندگان پایان‌نامه‌ها، در مورد نحوهٔ نگارش کلماتی که ۲ جزئی و مشتق می‌باشند، دچار خطا شده‌اند. در پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، تنها حالت ثبت شده عنوانها به بازیابی عنوان مورد نظر انجامید و در پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران، تنها ۵۸٪ عنوانها با تغییر رسم‌الخط همچنان بازیابی شدند.

نتیجه‌گیری: این پژوهش نشان داد پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران نسبت به پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، در بازیابی عنوان پایان نامه‌ها در حالتهای مختلف پیوسته

---

۱. دانشجوی دورهٔ دکترای کتابداری و اطلاع‌رسانی، دانشگاه فردوسی مشهد.
somakhshik@gmail.com
۲. استاد گروه کتابداری و اطلاع‌رسانی، دانشگاه فردوسی مشهد. fattahirahmat@gmail.com

و جدا نوشته شده، بهتر عمل می‌کند. همچنین، باید به نویسندگان پایان‌نامه‌ها، استفاده از قواعد یکدست ملی بویژه در نگارش کلمات ۲ جزئی و مشتق تأکید شود.

کلیدواژه‌ها: خط فارسی، ذخیره و بازیابی، پایگاه‌های اطلاعاتی، رسم‌الخط، پیوسته‌نویسی، جدانویسی.

در این جستار کوتاه سعی شده با نگاه به ویژگی پیوسته‌نویسی و جدانویسی واژگان فارسی در محدودهٔ عنوانهای پایان‌نامه‌های کتابداری و اطلاع‌رسانی و بررسی این مشکل در دو پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، وضعیت توجه نویسندگان و همچنین پایگاه‌های مورد نظر به این بخش از رسم‌الخط بررسی و راهکارهایی برای حلّ این مشکلات ارائه شود.

مقدمه

به استناد مرکز آمار جهانی اینترنت، هرچند زبان انگلیسی هنوز هم جزء ده زبان اول دنیای اینترنت است[۱]، تعداد مدارک غیر انگلیسی و کاربران غیر انگلیسی زبان در وب در حال افزایش است. این وضعیت، مطالعه و طراحی سیستمهای بازیابی برای این زبانهای مختلف را ناگزیر ساخته است. چنان‌که «آل احمد و دیگران»[۲] (۲۰۰۸) نیز به این مسئله اشاره کرده‌اند، زبان فارسی به‌عنوان زبان رسمی ایران، افغانستان و تاجیکستان سبب شده منابع زیادی از وب به این زبان تولید شود و کاربران فارسی زبان به دلایل مختلفی در جستجوهای خود از این زبان استفاده کنند، اما به دلیل غالب بودن زبان انگلیسی در اینترنت، جستجو به زبانهای غیرانگلیسی از جمله فارسی، مسائل و مشکلات مختلفی جدا از مشکلات عمومی اینترنت به همراه دارد (راثی، ۱۳۸۴). مشکلات زبان فارسی از یک سو و اهمیت یافتن روزافزون موضوع رایانه و خط و زبان فارسی، که در همه زمینه‌های کاربردی و تحقیقاتی و حتی در زندگی عموم مردم رسوخ یافته، از سوی دیگر، همانطور که «صامتی و بی‌جن‌خان» (۱۳۸۹، نوزده) نیز بیان می‌کنند، سبب شکل‌گیری پژوهشهای زیادی در این حوزه شده است.

---

۱. برای اطلاعات بیشتر نگاه کنید به: http://www.internetworldstats.com/stats7.htm

2. AleAhmad, et al.

دشواریهای زبان فارسی در ارتباط با حوزهٔ ذخیره و بازیابی اطلاعات را می‌توان از نظرگاه‌های مختلفی دسته‌بندی نمود. مقالهٔ حاضر که با دیدگاه ساختاری به مسائل رسم‌الخط فارسی پرداخته است، به‌طور مشخص بر ویژگی پیوسته و جدانویسی کلمات فارسی تأکید دارد. کلماتی که به دو شکل پیوسته و جدا نوشته می‌شوند، هر چند مشکلات کمی در خواندن متن به وجود می‌آورند و هر آشنای به زبان فارسی به راحتی می‌تواند آنها را بخواند، در نظامهای ذخیره و بازیابی اطلاعات، مشکلات زیادی ایجاد می‌کنند. از این رو، نیازمند توجه از سوی پدیدآورندگان متون و منابع و نیز طراحان و نمایه‌سازان پایگاه‌های اطلاعاتی می‌باشند.

پیوسته‌نویسی و جدانویسی در رسم‌الخط فارسی

فرهنگستان زبان و ادب فارسی در باب پیوسته‌نویسی و یا جدانویسی ترکیبات در زبان فارسی، سه فرض را متصور است (دستور خط فارسی، ۱۳۸۸، ص ۳۸) که در ادامه به آنها اشاره شده است. در پژوهش حاضر بر مبنای این دستورالعملهای فرهنگستان عمل شده است.

۱. تدوین قواعدی برای جدانویسی همه کلمات مرکب و تعیین موارد استثنا.
۲. تدوین قواعدی برای پیوسته‌نویسی همه کلمات مرکب و تعیین موارد استثنا.
۳. تدوین قواعدی برای جدانویسی الزامی بعضی از کلمات مرکب و پیوسته‌نویسی بعضی دیگر و دادن اختیار در خصوص سایر کلمات به نویسندگان.

فرهنگستان در تدوین و تصویب دستور خط فارسی، فرض سوم را برگزیده و تنها موارد الزامی جدانویسی و یا پیوسته‌نویسی را مشخص کرده است:

الف) کلمات مرکبی که الزاماً پیوسته نوشته می‌شوند. به عنوان مثال، مرکبهای بسیط‌گونه مانند یکشنبه و کلماتی که جزء دومشان با «آ» آغاز می‌شود و تک هجایی هستند و موارد دیگر که در متن دستور خط فارسی به‌طور کامل توضیح داه شده‌اند.

ب) کلمات مرکبی که الزاماً جدا نوشته می‌شوند. مانند ترکیبهای اضافی، مصدر

مرکب و غیره... .

در عین حال، چنان‌که اشاره شد، نویسندگان، ویراستاران و ناشران آثار فارسی تاکنون از شیوه‌ها و رسم الخط‌های مختلفی استفاده کرده‌اند و متون موجود فارسی با همین گوناگونی در پایگاه‌های اطلاعاتی و در وب ذخیره شده است. به همین سبب، جستجو و بازیابی متون فارسی با چالشهای فراوان همراه است.

ضرورت و هدفهای پژوهش

نظام نحوی یا ساختاری هر زبان، مهم‌ترین شاخص استقلال و تمایز یک زبان از زبانهای دیگر است که بر پایهٔ واژگان زبان شکل می‌گیرد (نوبهار، ۱۳۸۸). زبان فارسی، در مقایسه با سایر زبانهای دنیا، ماهیت متفاوت و ویژه (Oroumchian , et al. , 2007) و نیز نظام ساختاری پیچیده‌ای دارد. به همین دلیل، طراحی سیستمهای ذخیره و بازیابی برای آن نیازمند ملاحظات ویژه‌ای است. این تفاوتها نه تنها در ساختار زبان، بلکه در خط فارسی نیز وجود دارد (دستور خط فارسی، ۱۳۸۸، ص.۱). متأسفانه، نبود استاندارد و تنوع رسم‌الخط و مفاهیم در زبان فارسی (شهیدی و دیگران ۱۳۸۴) سبب پراکندگی سبک و سیاق نگارشی برای این زبان شده است. بی‌توجهی برخی از پدیدآورندگان به این ویژگیهای خط فارسی بویژه در متون و منابع علمی و گاه بی‌توجهی طراحان پایگاه‌های اطلاعاتی و موتورهای جستجو، اغلب به ناکارآمدی این پایگاه‌ها در جستجو و بازیابی منجر شده است. آنچه ضرورت پرداختن به این پژوهش را آشکار می‌سازد، شناسایی مسائل مربوط به پیوسته و جدانویسی در نگارش فارسی و میزان توجه به این مسائل در ذخیره و بازیابی اطلاعات و متون فارسی است. براساس این ضرورت، هدف از پژوهش حاضر، شناسایی کاستیهایی است که از نظر رسم‌الخط فارسی و از جنبه ویژگیهای ترکیب و جدانویسی کلمات در زبان فارسی، در عنوانهای پایان‌نامه‌های کتابداری انعکاس یافته است. همچنین، میزان توجه طراحان و نمایه‌سازان پایگاه‌های اطلاعاتی پژوهشگاه‌های علوم و فناوری اطلاعات ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری به این ویژگیهای کلمات فارسی به منظور تلاش برای بهینه‌سازی این

پایگاه‌های اطلاعاتی، از دیگر هدفهایی است که این پژوهش دنبال می‌کند.

مسئله پژوهش

رسم‌الخط فارسی، چنان‌که «حرّی» (۱۳۷۶) نیز اشاره می‌کند، یکی از متغیرهای عمده در ذخیره و بازیابی اطلاعات به زبان فارسی است و در دهه‌های اخیر نیز مسبب بیشترین اختلاف نظر در مورد شیوه املای کلمات بوده است (شهیدی و دیگران، ۱۳۸۴). دشواریهای حاکم بر نحوهٔ نگارش واژه‌های فارسی، علاوه بر این‌که سبب ناهماهنگی متون می‌شود، برای جستجوگران محیط وب نیز مسائلی را پیش روی می‌نهد. بی‌توجهی کاربران (راثی، ۱۳۸۴)، پدیدآورندگان متون و منابع و نیز طراحان و نمایه‌سازان پایگاه‌های اطلاعاتی فارسی به ویژگیهای پیوسته‌نویسی و جدانویسی واژگان در کنار سایر مسائل رسم‌الخط فارسی، می‌تواند سبب بروز مشکلات زیادی در ذخیره و بازیابی اطلاعات شود. بر این اساس، پژوهش حاضر در پی آن است تا میزان رعایت اصول رسم‌الخط فارسی از جنبهٔ پیوسته‌نویسی و جدانویسی را با محدود نمودن به حوزهٔ کتابداری و اطلاع‌رسانی و صرفاً به پایان‌نامه‌هایی که به‌عنوان نمونه برای این‌کار انتخاب شده‌اند، بررسی کند. همچنین، روشهایی را که ممکن است برخی پایگاه‌های اطلاعاتی در این زمینه اتخاذ نموده باشند، شناسایی و بر مبنای یافته‌های حاصل، ضمن نشان دادن وضعیت حال حاضر، پیشنهادها و راهکارهایی عملی ارائه نماید.

پیشینهٔ پژوهش

بررسی پژوهشهای انجام گرفته در حوزهٔ بازیابی اطلاعات به زبان فارسی بیانگر این است که این مقوله از دیرباز مورد توجه صاحب‌نظران و پژوهشگران علوم کتابداری و اطلاع‌رسانی، رایانه و زبانشناسی بوده است. نگاه به فعالیتهایی که در این زمینه صورت گرفته، از گستردگی مشکلات و دشواریهای زبان فارسی و ابعاد مختلف آن حکایت دارد که در حوزهٔ بازیابی به‌عنوان مسئله رخ نموده و لزوم تلاش برای رفع آنها را

ضروری می‌سازد[۱]. در ادامه، برخی از این پژوهشها در حوزه‌های ریشه‌یابی، پیوسته‌نویسی و جدانویسی و نیز شکلهای مختلف نوشتاری واژگان فارسی، دسته‌بندی و بیان می‌شود.

جدانویسی و پیوسته‌نویسی: مرور پیشینه در این حوزه، نشان از فعالیتهای اندک صورت گرفته دربارهٔ مشکلات جدانویسی و پیوسته‌نویسی دارد. اغلب این پژوهشها، مسائل مطرح در این زمینه را شناسایی نموده‌اند؛ مانند پژوهشی که «شهیدی و همکارانش» (۱۳۸۴) برای یافتن روشی برای رفع چالشهای محتوا کاوی در وبهای فارسی زبان انجام دادند و در نهایت، برخی از مهم‌ترین چالشهای خط فارسی را برشمردند که در بین آنها می‌توان اشاره‌هایی به ویژگیها و مسائل ترکیب و جدانویسی واژگان را نیز ملاحظه نمود. عمده‌ترین راه حلهایی که این پژوهشگران ارائه دادند، عبارت است از: انتخاب مناسب سرعنوانهای موضوعی در وب‌سایتهای فارسی، استمداد از علم اصطلاح‌شناسی در نمایه‌سازی ماشینی، تعریف یک استاندارد برای مفاهیم و رسم‌الخط فارسی در وب، استفاده از مفرد و جمع در نمایه‌سازی و استفاده از یک واسط کاوش فارسی برای رفع چالشهای رسم‌الخطی.

البته در این زمینه، پژوهشی را «کاشفی و همکارانش» (Kashefi, et al., 2010) با عنوان بهینه‌سازی‌یابش مدارک مشابه در بازیابی اطلاعات به زبان فارسی انجام دادند و در آن به شناسایی بیش از ۳۰۰ پسوند و ترکیبهای کلمات و کارآمدی حذف پیشوندها از متون فارسی به هنگام بازیابی آنها پرداختند. در این پژوهش، از چهار روش استفاده شد؛ نمایه‌سازی معانی پنهان، مدل فضای برداری، هم‌آیندی و شینگلینگ[۲]. نتیجه نشان

---

۱. برای اطلاعات بیشتر، نگاه کنید به: نشاط، نرگس (۱۳۷۹). «مسائل رسم‌الخط فارسی در رویارویی با فنّاوری نوین اطلاعاتی». در مجموعه مقالات فهرستهای رایانه‌ای: کاربرد و توسعه. به کوشش رحمت الله فتاحی. مشهد: دانشگاه فردوسی: تهران: مرکز اطلاع‌رسانی جهاد.

۲. الگوریتم شینگلینگ (Shingling) یکی از روشهای موجود در زمینهٔ شناسایی متون تقریباً یکسان است که برای شناسایی کلماتی که حجم زیادی از آنها جزئیات بی‌اهمیت است، به‌کار می‌رود. برگرفته از:
http://www.farsipaper.ir/group/d1fe51a944a24a179e6649b78cc369f3

داد با حذف پیشوندها، میزان بازیابی مدارک مشابه، بهبود و بازیافت این منابع به‌طور قابل ملاحظه‌ای افزایش می‌یابد.

ریشه‌یابی واژگان: ریشه‌یابی، که عبارت است از قرار دادن واژه‌های یک زبان در دسته‌های معنایی یکسان، در بسیاری از زمینه‌های پردازش زبان طبیعی. همچنین پردازش زبان فارسی، مدنظر است. پژوهشی که «رحیم طرقی و همکارانش» (Rahimtoroghi, et al., 2010) در زمینهٔ ریشه‌یابی مبتنی بر قواعد دستوری برای زبان فارسی انجام دادند نیز شاهد این مدعاست. این پژوهشگران، برمبنای قواعد دستور زبان، الگوریتم ریشه‌یابی را طراحی نمودند که از ساختار کلمات و قواعد املایی آنها برای شناسایی ریشه هر کلمه استفاده می‌کند. بر این اساس، ۳۳ قاعدهٔ دستوری شناسایی شد. نتایج نشان داد استفاده از این ریشه‌یاب در سیستمهای بازیابی اطلاعات در مورد زبان فارسی، دقت نتایج بازیابی شده را به میزان ۴/۸٪ افزایش و اندازهٔ فایل نمایه‌سازی شده را تا ۶٪ کاهش می‌دهد.

توجه به ریشه‌یابی گاه در کنار سایر بررسیهای زبانشناختی قرار گرفته است. به‌عنوان مثال، می‌توان به پژوهش «کریم‌پور و دیگران» (Karimpour, et al., 2009) اشاره نمود. دراین پژوهش، از مدل بازیابی Idri و از برچسب‌زن اجزای جملهٔ TNT با استفاده از ۴۰ برچسب پیکرهٔ «بی‌جن‌خان»[1] استفاده شد. بر این اساس، بهبود عملکرد الگوریتمهای بازیابی ارزیابی گردید. همچنین، تأثیر ریشه‌یابی به‌عنوان یکی دیگر از بخشهای کار این پژوهشگران، بررسی شد. **یافته‌های این تحقیق نشان داد هرچند** استفاده از برچسب زنی ارکان سخن ممکن **است تأثیر اندکی در اثر بخشی نتایج** بازیابی شده داشته باشد، زمانی‌که این روش **همراه با ریشه‌یابی به کار می‌رود، دقت** نتایج بازیابی شده به‌میزان قابل توجهی افزایش **می‌یابد.**

---

۱. پیکره‌ای برچسب گذاری شده که برای تحقیقات پردازش **زبان طبیعی در زبان فارسی مناسب** است. این مجموعه از اخبار روزانه و متون رایج، از ۴۳۰۰ موضوع **مختلف جمع‌آوری شده و شامل ۲.۶ میلیون** واژهٔ برچسب‌گذاری شده است. برگرفته از:

http://ece.ut.ac.ir/dbrg/bijankhan

در مورد تأثیر ریشه‌یابی در متون زبان فارسی، پژوهش دیگری توسط «دلامیک و ساووی» (Delamic and Savoy, 2009) انجام گرفته که با هدف ارزیابی راهبردهای مختلف نمایه‌سازی و ریشه‌یابی، استفاده از سیاههٔ واژگان بازدارنده و یک ریشه‌یابی سبک را پیشنهاد می‌کنند. در این پژوهش، از مدلهای بازیابی متعدد از جمله Okapi, DFR, LM و نیز دو مدل کلاسیک فضای برداری یعنی tf idf و نیز Lnu-ltc برای ارزیابی راه حلهای ارائه شده، استفاده گردید. آنچه در این پژوهش مورد توجه است، نگاه ویژه به رسم‌الخط فارسی و پیشنهاد یک ریشه یاب برای این خط است که رایج‌ترین پسوندهای مورد استفاده و حالتهای جمع لغتها را استخراج و در نهایت سیاههٔ واژگان بازدارنده‌ای شامل ۸۸۱ کلمه را پیشنهاد می‌کند که مدیریت و کنترل آنها می‌تواند در بازیابی به زبان فارسی، کمک قابل توجهی باشد.

شکلهای مختلف نوشتاری واژگان: توجه به این‌که واژه‌های فارسی شکلهای مختلف نگارشی دارند، و مسائلی که وجود این اشکال پیش روی بازیابی اطلاعات در وب می‌نهد، در پژوهش «عبدالهی نورعلی» (۱۳۸۶) نیز تأکید شد. وی مسائل ریخت‌شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای وب را بررسی کرد و با استفاده از جستجوگرهای گوگل، آلتاویستا و یاهو، جستجوهایی را به زبان فارسی انجام داد و دریافت که این جستجوگرها، به دشواریهای زبان فارسی در بازیابی اطلاعات نپرداخته و تلاشی برای بهبود نتایج انجام نداده‌اند.

برخی نیز به‌طور مشخص، مسائل زبان و خط فارسی در ذخیره و بازیابی اطلاعات را بررسی کردند. از آن جمله، «مرتضایی» (۱۳۸۱) است که نمونه‌هایی از دشواریهای زبان و خط فارسی را در بازیابی اطلاعات بر می‌شمارد. همچنین، وی به مهم‌ترین دشواریهای زبان فارسی که سبب کندی مراحل ذخیره و بازیابی و نیز کاهش بازیافت می‌شوند نیز پرداخته و معتقد است راهکارهایی از قبیل یکسان سازی واژه‌ها، کاربرد دستورالعمل‌هایی یکدست در تمامی واحدهای چاپ و نشر و هوشمندسازی جستجو می‌تواند به حلّ این مسائل کمک کند.

برخی دیگر نیز به‌عنوان جزئی از پژوهش خود، توجه به این بُعد را نیز از نظر دور

نداشته‌اند. از آن جمله، تحقیقی است که «گل تاجی و بذرگر» (۱۳۸۹) در زمینهٔ بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای علوم اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی انجام دادند و با انتخاب و جستجوی کلیدواژه‌هایی که هرکدام بیانگر نوعی از چالشهای زبان فارسی بود، این کلیدواژه‌ها را در پایگاه‌های موردنظر جستجو کردند. نتایج این پژوهش نشان داد هیچ‌یک از این سه پایگاه، به شیوه‌ای جامع و قابل ملاحظه به حل مسائل ریخت‌شناسی واژگان فارسی نپرداخته‌اند. موارد مورد توجه پایگاه‌های مورد بررسی در این پژوهش، به ترتیب زیر ذکر شده است: پایگاه مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری: تنوین، تشدید، پیوسته‌نویسی و بی‌فاصله‌نویسی؛ پژوهشگاه اطلاعات و مدارک علمی ایران: جدانویسی و بی فاصله‌نویسی، خط تیره، نقطه بین سرنام‌ها؛ پایگاه جهاد دانشگاهی: همزه به صورتهای مختلف.

نگاهی به پیشینهٔ پژوهشهایی که بیان شد، نشان می‌دهد مسائل خط و زبان فارسی در پیوند با ذخیره و بازیابی اطلاعات را می‌توان از ابعاد مختلف بررسی کرد. ویژگیهای خاص حاکم بر نگارش خط فارسی سبب شده تا بررسی دقیقتر هرکدام از آنها و مسائلی که در ذخیره و بازیابی پدید می‌آورند، بیش از پیش اهمیت یابد. به نظر می‌رسد آگاهی از این ضرورت در بین متخصصان حوزه‌های مرتبط، بویژه متخصصان علم کتابداری و اطلاع‌رسانی، به وجود آمده است و زمان آن فرا رسیده تا راه‌حلهایی دقیق و موشکافانه برای هریک از این دشواریها ارائه شود. پژوهش حاضر با این رویکرد و با هدف قرار دادن یکی از این معضلات، که عبارت است از ویژگیهای ترکیب و جدانویسی واژگان فارسی، و به منظور یافتن راهی برای گذر از چالشهای آن در حوزهٔ کتابداری و اطلاع‌رسانی، انجام یافته است.

سؤالهای پژوهش

پژوهش حاضر در پی یافتن پاسخ سؤالهای زیر انجام شده است:

۱. اصول پیوسته‌نویسی و جدانویسی به‌عنوان یکی از مسائل رسم‌الخط فارسی،

تا چه میزان در عنوانهای پایان‌نامه‌های حوزه کتابداری و اطلاع‌رسانی رعایت شده است؟

۲. به لحاظ شکل دستوری، کلیدواژه‌های جدا یا پیوسته نوشته شده، جزء کدام نوع (اسم، صفت، قید، فعل) هستند؟

۳. پایگاه‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری با توجه به ویژگیهای پیوسته‌نویسی و جدانویسی واژگان در عنوانهای پایان‌نامه‌ها چگونه عمل می‌کنند؟

طرح پژوهش

در این پژوهش که به روش تحلیل محتوا انجام شده است، ۱۰۰ عنوان از پایان‌نامه‌های موجود رشته کتابداری و اطلاع‌رسانی به منزلهٔ نمونه‌ای از متون فارسی در پایگاه‌های موردنظر و از هر پایگاه ۵۰ عنوان به‌صورت تصادفی انتخاب شد. به این ترتیب که ابتدا سیاهه‌ای از دانشگاه‌های مجری رشتهٔ کتابداری و اطلاع‌رسانی در مقاطع تحصیلات تکمیلی در ایران جمع‌آوری و پس از آن از طریق جستجوی نام استادان راهنما، به عنوانهای پایان‌نامه‌ها دست یافته شد. همچنین، عنوانهایی که در یک پایگاه یافت می‌شدند، به‌منظور جلوگیری از تکرار یافته‌ها به‌هنگام جستجو در پایگاه دوم، در صورت بازیابی از سیاهه کنار گذاشته شدند. جستجوی اسمها با هدف شناسایی کلمات مرکبی که قابلیت پیوسته و جدانویسی داشتند، انجام شد. تمام کلمات عنوانهای این پایان‌نامه‌ها بررسی و سیاهه‌ای از کلماتی که ویژگی مورد نظر را به لحاظ رسم‌الخطی دارا بودند، فراهم شد. پس از این مرحله، بر اساس دستور خط فارسی مصوب فرهنگستان (۱۳۸۸) کلماتی که درست یا نادرست نوشته شده بودند از یکدیگر تفکیک، و از نظر تعداد اجزا و نوع (مرکب، مشتق و مرکب- مشتق) تحلیل شدند. گفتنی است، تعداد کلماتی که قاعده‌ای برای آنها در فرهنگستان وجود نداشت و در مورد آنها اختیار به نویسنده داده شده بود، بسیار ناچیز بود، با این‌حال، به هنگام جستجو، به‌عنوان کلمهٔ خنثی در نظر گرفته شدند. روایی این کار با مشورت استاد راهنما تأیید گردید. در مرحلهٔ سوم، عنوانهای موردنظر در هر دو پایگاه و با حالتهای متفاوت کلمات مرکب، جستجو شد. نتایج بازیابی در پایگاه‌های اطلاعاتی

پژوهشگاه‌های علوم و فناوری اطلاعات ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، ارزیابی و مقایسه و در نهایت نتایجی حاصل شد که پاسخ سؤالهای تحقیق را شکل داد.

یافته‌های پژوهش

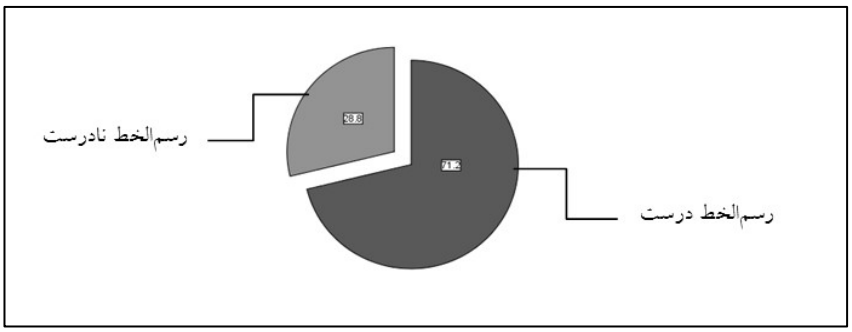
با بررسی عنوانهای مورد نظر، مشخص شد این عنوانها در کل شامل ۱۵۴۷ کلمه می‌باشند. در پی دستیابی به هدفهای پژوهش مبنی بر شناسایی کاستیهای نگارشی از دیدگاه رسم‌الخطی مورد توجه در عنوانهای پایان‌نامه‌های کتابداری و اطلاع‌رسانی، پس از بررسی کلمات و مطابقت آنها با دستور خط فارسی مصوب فرهنگستان زبان، آنهایی که از نظر رسم‌الخطی قابلیت پیوسته و جدانویسی را داشتند، در سیاهه‌ای جداگانه تنظیم شدند که تعداد آنها، ۳۱۶ کلمه، حدود ۲۰/۴٪ کلّ کلمات عنوانها بود. پس از آن، با توجه به قواعد فرهنگستان، به تفکیک کلماتی پرداخته شد که بر این اساس درست و نادرست نوشته شده بودند.

نتایج این بررسی برای یافتن پاسخ سؤال اول این پژوهش، نشان داد نگارش ۲۲۵ کلمه (۷۱/۲٪) درست، ۹۱ کلمه دیگر (۲۸/۸٪) نادرست است. این وضعیت در جدول ۱ نیز آورده شده است.

جدول ۱. فراوانی کلمات عنوانها و دارای ویژگی پیوسته و جدانویسی

تعداد کلّ کلمه‌های عنوانها		کلمه‌های دارای ویژگی مورد نظر		کلمه‌های درست		کلمه‌های نادرست	
فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد
۱۵۴۷	۱۰۰	۳۱۶	۲۰/۴۲	۲۲۵	۷۱/۲	۹۱	۲۸/۸

در شکل ۱ نیز نسبت کلمات درست و نادرست از کل کلماتی که مرکب بودند، نشان داده شده است.



شکل ۱. نسبت کلمات با رسم‌الخط درست و نادرست

در مورد سؤال دوم، بررسی کلمات موردنظر نشان داد اغلب این کلمات (بیش از ۹۵٪) اسم و حدود ۵٪ دیگر، ضمیر می‌باشند. از آنجا که در عنوان فعل به‌کار نمی‌رود و نیز متون علمی بندرت دارای قید یا صفت هستند، نبود چنین کلماتی که ویژگی پیوسته و جدانویسی را نیز داشته باشند، قابل توجیه است. همچنین برای بررسی بهتر، کلمات دارای ویژگی پیوسته و جدانویسی به لحاظ ساختاری نیز تفکیک و به سه دسته تقسیم شدند:

۱- کلمات مشتق: آنهایی هستند که یک جزء آنها معنای قاموسی و اجزای دیگر معنای دستوری دارند.

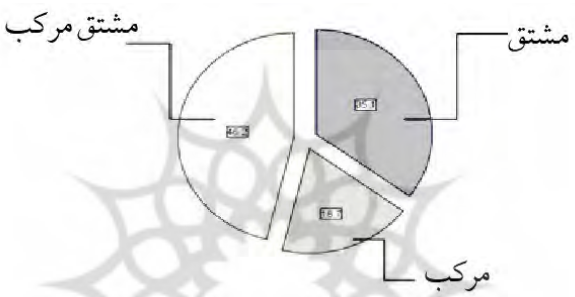
۲- کلمات مرکب: آنهایی هستند که از دو جزء یا بیشتر تشکیل شده‌اند و تمامی اجزا دارای معنای قاموسی می‌باشند.

۳- کلمات مشتق ـ مرکب: آنهایی هستند که دو جزء یا بیشتر از آنها معنای قاموسی و بقیهٔ اجزایشان معنای دستوری دارد.

بر این اساس، حدود ۳۵٪ کلمات، مشتق، بیش از ۱۸٪ مرکب و نزدیک به ۴۲٪ نیز مشتق- مرکب بودند که جدول ۲ و شکل ۲ بیانگر این وضعیت است.

جدول ۲. تفکیک کلمات از نظر ساختاری

نوع کلمه	فراوانی	درصد
مشتق	۱۱۱	۳۵/۱
مرکب	۵۹	۱۸/۷
مشتق ـ مرکب	۱۴۶	۴۶/۲
مجموع	۳۱۶	۱۰۰

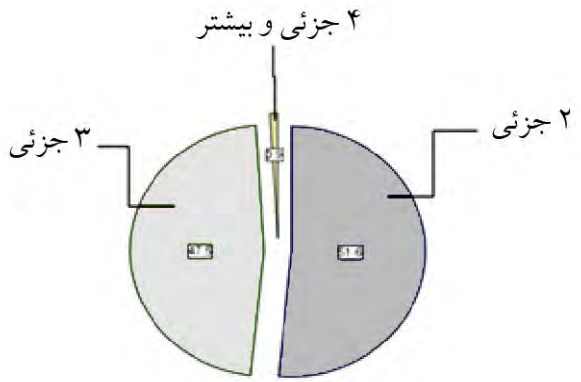


شکل ۱. نسبت کلمات مورد بررسی از نظر ساختاری

همچنین، تعداد اجزای این کلمات نیز بررسی شد. چنان‌که جدول ۳ و شکل ۳ نیز نشان می‌دهند، مشخص شد بیشتر این کلمات، دو جزئی (۵۱/۶٪) و سه جزئی (۴۷/۵٪) هستند و کلمات چهار جزئی، درصد بسیار اندکی از کلمات موردنظر را تشکیل می‌دهند.

جدول ۳. تعداد اجزای کلمات مورد بررسی

جدول اجزای کلمه	فراوانی	درصد
۲ جزئی	۱۶۳	۵۱/۶
۳ جزئی	۱۵۰	۴۷/۵
۴ جزئی و بیشتر	۳	۹/۰
مجموع	۳۱۶	۱۰۰



شکل ۳. نسبت اجزای کلمات مورد بررسی

با بررسی کلمات استخراج شده از عنوانهای بررسی شده، همان‌طور که جدول ۴ نیز نشان می‌دهد، مشخص شد اغلب نویسندگان پایان‌نامه‌ها، در مورد نحوهٔ نگارش کلماتی که ۲ جزئی و مشتق می‌باشند، دچار خطا شده‌اند.

جدول ۴. میزان اشتباه نویسندگان در رسم‌الخط کلمات به تفکیک اجزا و نوع کلمه

تعداد اجزای کلمه	درصد نادرستی (فراوانی نسبی)
۲ جزئی	۳۸/۰۳
۳ جزئی	۱۷/۳۳
۴ جزئی و بیشتر	۶۶/۶۶
نوع کلمه	درصد نادرستی (فراوانی نسبی)
مشتق	۶۷/۵۸
مرکب	۱۳/۵۶
مشتق ـ مرکب	۱۸/۴۹

در مرحلهٔ بعد، به‌منظور دستیابی به هدف دوم این پژوهش و پاسخگویی به سؤال سوم، عنوانهای مربوط به هر پایگاه، با «جستجوی عنوانی» به طور عمدی در حالتهای درست و

نادرست از سوی پژوهشگر جستجو شد؛ به این ترتیب که در عنوانهایی که کلمات به صورت نادرست نوشته شده بود، جستجو با شکل صحیح رسم‌الخطی و همچنین در عنوانهایی که کلمات به‌طور درست به‌کار رفته بود، جستجو با شکل اشتباه رسم‌الخطی نیز انجام شد. گفتنی است، این روش برای هر دو پایگاه اطلاعاتی و در مورد تمامی عنوانها انجام گرفت. چنان‌که جدول۵ نیز نشان می‌دهد، در پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، با اعمال هریک از تغییرات مورد اشاره به هنگام جستجو، عنوان مورد نظر بازیابی نشد و تنها حالت ثبت شدهٔ[1] عنوانها به بازیابی عنوان مورد نظر می‌انجامید. در انجام همین جستجوها در مورد ۵۰ عنوان مورد نظر از پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران، ۲۹ عنوان (۵۸٪) با تغییر رسم‌الخط (درست به نادرست و برعکس) همچنان بازیابی شد، اما ۲۱ عنوان دیگر که ۴۲٪ باقیمانده را تشکیل می‌داد، با این تغییرات، بازیابی نشدند. علت تغییر نوع رسم‌الخط از درست به نادرست و برعکس، این بود که ممکن است کاربر هنگام جستجوی عنوانی در هر حالتی به صورت پیوسته یا ترکیبی، واژه را جستجو کند و پایگاه‌های اطلاعاتی باید توانایی جستجوی مورد نظر از سوی کاربر را بدون توجه به میزان اطلاع وی از نحوهٔ نگارش صحیح کلمات، داشته باشند.

بر این اساس، می‌توان عملکرد ذخیره و بازیابی پایگاه اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران را در مقایسه با پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، مناسب‌تر قلمداد نمود؛ هرچند یافته‌ها نشان داد این پایگاه نیز در زمینه ذخیره‌سازی و بازیابی کلمات فارسی با ویژگیهای پیوسته و جدانویسی، یکپارچه عمل‌نکرده است. نتایج حاصل از جستجوهای عنوانی در دو پایگاه، در جدول ۵ آورده شده است.

---

۱. حالت ثبت شده، نحوهٔ درج عنوان پایان‌نامه در پایگاه مربوط است. حین این پژوهش، عنوانهایی که با غلطهای املایی و تایپی ثبت شده بودند در پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری نیز وجود داشت که نگارنده ناگزیر این عنوانها را با همان اشتباه‌های ثبتی جستجو نمود.

جدول ۵. نتایج حاصل از جستجوی عنوانهای پایان‌نامه‌ها در حالتهای مختلف رسم‌الخطی در دو پایگاه مورد بررسی

پایگاه / نتیجه	پژوهشگاه علوم و فناوری اطلاعات ایران		مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری	
	فراوانی	درصد	فراوانی	درصد
بازیابی نشده	۲۱	۴۲	۵۰	۱۰۰
بازیابی شده	۲۹	۵۸	۰	۰
مجموع	۵۰	۱۰۰	۵۰	۱۰۰

بحث و نتیجه‌گیری

پیچیدگیهای رسم‌الخط فارسی، از یک‌سو سبب‌ساز آشفتگیهایی در ذخیره و بازیابی (صامتی و بی‌جن خان، ۱۳۸۹، ص. نوزده؛ شهیدی و دیگران، ۱۳۸۴ و Oroumchian, et al., 2007) و همچنین جستجوی اطلاعات به زبان فارسی در اینترنت شده و از سوی دیگر به دلیل تأثیرهای هم‌فرسایی مشکلات بر یکدیگر، چالشهای این حوزه را چند برابر ساخته است. به‌عنوان نمونه، تأثیر ترکیب و جدانویسی را بر مرزبندی و تعیین حدود کلمه می‌توان مثال زد. اگر در رسم‌الخط فارسی، مطابق قواعد استاندارد عمل نشود، مشکل مرزبندی کلمات فارسی دو چندان می‌شود، زیرا به دلیل مشکلات عدم شناسایی مرز دقیق کلمات، چالشهای عمده‌ای برای ریشه‌یابی کلمات و الگوریتمهای ریشه‌یابی پدید می‌آید که با استفاده از دستورالعملهای استاندارد برای ترکیب و جدانویسی کلمات، بخشی از این چالشها حل و در صورت آشفتگی رسم‌الخط، مشکلات دیگری که به آنها اشاره شد، افزون خواهد شد. این پژوهش با هدف شناسایی بخشی از مشکلات خط فارسی که بر ذخیره و بازیابی اطلاعات از پایگاه‌های اطلاعاتی تأثیر می‌گذارند، در محدوده کوچکی انجام شد. در بازنگری دوباره نتایج این پژوهش با بخشی از پیشینه که ارتباط نزدیکتری با موضوع دارند، می‌توان به نتایج قابل

توجهی رسید.

نتایج پژوهش حاضر با پژوهش «عبداللهی نورعلی» (۱۳۸۶) همخوان است. در آن پژوهش نشان داده شد که به مسائل ریخت‌شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای گوگل، یاهو و آلتاویستا پرداخته نشده است و در این‌جا مشخص شد که یک نمونه از این مسائل ریخت‌شناسی، یعنی پیوسته و جدانویسی، در پایگاه‌های اطلاعاتی فارسی نیز مورد بی‌توجهی قرار گرفته است. به‌عبارت دیگر، به مسائل ریخت‌شناسی زبان فارسی نه تنها در جستجوگرهای غیر فارسی، بلکه در پایگاه‌های اطلاعاتی فارسی نیز پرداخته نشده است. البته در سالهای اخیر، حرکتهایی از سوی طراحان جستجوگرهای وب مبنی بر پیشنهاد عبارتهای جستجو و نیز پیشنهاد شکلهای مختلف نگارشی کلمه و عبارت مورد جستجو، صورت گرفته است که تا اندازه‌ای می‌تواند برخی از مشکلات خط فارسی را از بین ببرد و پایگاه‌های اطلاعاتی فارسی زبان نیز می‌توانند از این ایده‌ها بهرهٔ لازم را ببرند.

چنان‌که نتایج این پژوهش نشان داد، جستجو در حالتهای مختلف پیوسته و جدای واژگان عنوانی هرچند در پایگاه‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران به طور کامل به جامعیت بازیابی نمی‌انجامد، همراه نمودن کلمهٔ مورد نظر با تعداد بیشتری از واژه‌های عنوان از سوی جستجوگر، در برخی موارد به بازیابی عنوان مورد نظر می‌انجامد. بر مبنای این یافته‌ها که در بخش قبل نیز شرح داده شد، هرچند نتایج پژوهش «گل تاجی و بذرگر» (۱۳۸۹) در مورد بی‌توجهی برخی پایگاه‌های اطلاعاتی فارسی به مسائل ریخت‌شناسی زبان فارسی تأیید می‌شود، نتایج پژوهش حاضر نشان از آن دارد که برخلاف آنچه در پژوهش «گل تاجی و بذرگر» آمده است، پایگاه‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات ایران و نیز مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، به ویژگی پیوسته و جدانویسی کلمات توجه نشان نداده‌اند.

همچنین، این پژوهش تأییدی است بر آنچه «مرتضایی» (۱۳۸۱) مبنی بر تأثیر استاندارد نبودن شکل نوشتاری کلمات در عدم مطلوبیت و جامعیت جستجو، ذکر می‌کند.

با توجه به مشکلات نگارش خط فارسی که به برخی از آنها اشاره شد، ضرورت اندیشیدن در مورد راهکارهای برطرف کردن آن بویژه در محیطهای الکترونیکی جدید، بیش از پیش آشکار است. بدیهی است، نمی‌توان به بهانهٔ این دشواریها، خط غنی فارسی را به همین شکل از وب کنار گذاشت، بلکه باید موشکافانه ابعاد مختلف خط و نیز زبان فارسی را بررسی و راه حلهای مناسبی طراحی نمود. در مورد مشکل ترکیب و جدانویسی، مانند سایر پیچیدگیهای این خط، نمی‌توان یک راهکار منحصر ارائه داد. تلفیقی از آنچه در ادامه آمده است، می‌تواند به رفع بخشی از پیچیدگیها کمک کند. این راهکارها در دو بخش قابل ارائه است.

۱- راهکارهایی برای رعایت فراگیر قواعد یکدست ملی

ـ **توجه و حساسیت نویسندگان و پدیدآورندگان متون و منابع به زبان فارسی، در رعایت قواعدی که فرهنگستان زبان و ادب فارسی تدوین نموده است**. چنان‌که اشاره شد، رعایت این قواعد، دستِ کم در مواردی که قاعده‌ای مشخص و از پیش تعیین شده وجود دارد، همان‌طور که «حرّی» (۱۳۷۲) نیز خاطر نشان می‌سازد، نه تنها کاری پایه‌ای است، بلکه به سبب یکدستی حاکم، به‌هنگام طرح‌ریزی در نمایه‌سازی و طراحی الگوریتمهای ذخیره و بازیابی، به حلّ عالمانه‌تر مسائل نیز خواهد انجامید. البته، باید توجه داشت حتی با فرض اینکه رعایت این قواعد، ضمانت اجرایی لازم را داشته باشد، **بی‌قاعده بودن برخی حالتهای رسم‌الخطی و اختیار نویسندگان، بخش عمده‌ای از مشکلات را حل نشده باقی می‌گذارد**. البته باید توجه داشت، با توجه به اینکه زبان فارسی در کشورهایی مانند افغانستان و تاجیکستان هم کاربرد دارد، با رعایت قواعد رسم‌الخط ملی به‌نظر می‌رسد برخی مشکلات برای سایر جستجوگران فارسی زبان در خارج از ایران که با این قواعد نا آشنا هستند، همچنان باقی خواهد ماند.

ـ **پیش‌فرض نهادن جدانویسی در مواردی که اختیار به نویسندگان داده شده است**. چنان‌که در ابتدای مقاله اشاره شد، سه مفروضه برای مواجهه با مشکلات

نوشتاری خط فارسی قابل طرح است. سومین آنها، یعنی«تدوین قواعدی برای جدانویسی الزامی بعضی از کلمات مرکب و پیوسته‌نویسی بعضی دیگر و **دادن اختیار در خصوص سایر کلمات به نویسندگان**»، هرچند با ارائه قواعد ـ و البته با شرط رعایت آنها از سوی نویسندگان ـ کمک قابل توجهی به یکدستی خط فارسی و رفع مشکلات جستجو و بازیابی می‌کند، با توجه به اینکه راه‌حلهای بینابینی ارائه داده و موارد زیادی را به نویسنده می‌سپارد، به ابهام و چند دستگی در این زمینه منجر می‌گردد. حتی با تصور اینکه تمام پدیدآورندگان متون و منابع در وب مطابق با قواعد استاندارد رسم‌الخط فارسی بنویسند، باز هم زمانی‌که انتخاب شکل نگارش کلمه رسماً به سلیقهٔ نویسنده سپرده شود، مشکلات پردازشی زبان آغاز خواهد شد. این دشواریها نه تنها در مورد کلماتِ با ویژگیهای ترکیب و جدانویسی وجود دارد، بلکه سایر چالشهایی را که نگارش خط فارسی با آن مواجه است، شامل شده و بر ابهام و پیچیدگی نمایه‌سازی، جستجو و بازیابی اطلاعات به زبان فارسی می‌افزاید. رویکرد پیشنهادی پیش‌فرض نهادن جدانویسی در مواردی که اختیار به نویسندگان داده شده است، می‌تواند برخی از مشکلات پیوسته و جدانویسی را برطرف کند. نمونه‌هایی از این مشکلات، عبارتند از: شروع شدن جزء دوم با الف، هم مخرج بودن جزء اول با حرف آغازین جزء دوم، نامأنوس بودن کلمه در حالت پیوسته‌نویسی، بسامد زیاد جزء آغازین کلمه و ابهام در اجزای ترکیب به هنگام پیوسته‌نویسی.

۲- راهکارهای ذخیره و پردازش واژگان

ـ در این زمینه، متخصصان زبانشناسی، علوم رایانه و علوم کتابداری و اطلاع‌رسانی می‌توانند انواع رویکردها و روشهای پردازش هوشمند واژگان فارسی را برگزینند که به ذخیره و پردازش بهینه به قصد بازیابی جامع‌تر و در عین حال دقیق‌تر که کاستیهای ناشی از نبود یکدستی در جدانویسی و پیوسته‌نویسی است، کمک کند. برخی راه حلها در قسمت پایانی پژوهش حاضر ارائه شده است. استفاده از یافته‌های پژوهشی و رویکردهای عملی که در مورد سایر زبانها بویژه زبان عربی اعمال شده، می‌تواند به این

هدف کمک کند.

پیشنهادهای پژوهش

پیشنهادهای پژوهش به تفکیک در دو بخش آمده‌اند: پیشنهادهای عملی و پیشنهادهای پژوهشی.

الف) پیشنهادهای عملی

ـ پیشنهاد می‌شود نویسندگان و پدیدآورندگان منابع، خود را ملزم به رعایت قواعد رسم‌الخط فارسی کنند[۱]. به نظر می‌رسد پایگاه‌های اطلاعات علمی فارسی، نمایه‌سازی را بر اساس کلمات استخراج شده از متون انجام می‌دهند. بر همین اساس، رعایت این قاعده‌ها حداقل در مورد عنوانها، چکیده و کلیدواژه‌های متون علمی، ضرورت بیشتری دارد. این کار گذشته از آن‌که تلاشی برای حفظ پویایی و یکدستی خط فارسی به‌شمار می‌رود، برای طراحان و نمایه‌سازان پایگاه‌های اطلاعاتی مشکلات کمتری را پدید می‌آورد.

ـ به نمایه‌سازان پایگاه‌های اطلاعاتی فارسی زبان و بخصوص پایگاه‌های اطلاع‌رسانی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و پژوهشگاه علوم و فناوری اطلاعات ایران توصیه می‌شود با بهره‌مندی از نتایج پژوهشهای انجام شده در شورای عالی اطلاع‌رسانی ایران در زمینه خط و زبان فارسی، الگوریتمهای نمایه‌سازی خود را متناسب سازند و در جهت بهینه‌سازی نتایج جستجو و کمک به کاربران پایگاه، از امکانات کمکی مانند قابلیتهای پیشنهاد واژگان[۲] استفاده کنند.

ـ به پایگاه‌های اطلاعاتی توصیه می‌شود برای بازیابی کلماتی که ویژگیهای ترکیب

---

۱. دستور خط فارسی مصوب فرهنگستان زبان و ادب فارسی را می‌توانید در http://www.persianacademy.ir/fa/das.aspx مشاهده نمایید.

۲. این قابلیت هم اکنون در برخی موتورهای جستجو از جمله گوگل و یاهو و نیز پایگاه‌های اطلاعاتی وجود دارد.

و جدانویسی را دارند، از الگوریتمهای N-Geram استفاده کنند.

ب) پیشنهادهای پژوهشی

ـ انجام پژوهشی به روش تحلیل محتوا در زمینهٔ بسامد شکلهای مختلف جدانویسی و پیوسته نویسی در حوزه‌های موضوعی مختلف در متون زبان فارسی تا مشخص شود شکل رایج در هر حوزهٔ موضوعی چگونه است و چه راه حلی را می‌توان برای ذخیرهٔ بهتر واژگان زبان فارسی در پیش گرفت.

ـ **تفکیک مهم‌ترین چالشهای سطوح آوایی، واژگانی و ساختاری در زبان و خط فارسی و انجام پژوهشهایی مشابه برای یافتن مشکلات موجود در پایگاه‌های اطلاعاتی.**

ـ **شناسایی و دسته‌بندی نوع واژگان مورد جستجو در پایگاه‌های اطلاعاتی فارسی به‌منظور بررسی پربسامدترین اشتباه‌های رایج کاربران به هنگام پرس و جو در این پایگاه‌ها با هدف طراحی نظامی هوشمند برای بازیابی.**

ـ **شناسایی مشکلات مشابه رسم‌الخط فارسی و عربی به‌منظور مقایسهٔ میزان توجه، استفاده از راهکارهای احتمالی و نیز الگوبرداری از پایگاه‌های اطلاعاتی زبان عربی.**

منابع

- حری، ع. (۱۳۷۲). کامپیوتر و رسم‌الخط فارسی. ***مجله پیام کتابخانه***. تاریخ بازیابی: ۱۳۹۰/۹/۳. قابل بازیابی در:

www.noormags.com/view/fa/articlepage/396231

- دستور خط فارسی (۱۳۸۸). مصوب فرهنگستان زبان و ادب فارسی. تهران: فرهنگستان زبان و ادب فارسی (نشر آثار).
- راثی، م. (۱۳۸۴). مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت، مطالعه موردی کاربران مرکز اینترنت دانشگاه آزاد اسلامی واحد شبستر. تاریخ بازیابی: ۱۳۹۰/۹/۲۵. قابل بازیابی در:

http://www.aqlibrary.org/index.php?module=TWArticles&file=index&func=view_pubarticles&did=885&pid=10

- شهیدی، م، م، صدیقی و ک، زمانی‌فر (۱۳۸۴). روشی برای رفع چالش‌های محتواکاوی در وب‌های فارسی زبان. تاریخ بازیابی: ۱۳۹۰/۹/۳. قابل بازیابی در:

www4.irandoc.ac.ir/etela-art/21/shahidi.pdf

- صامتی، ح و م، بی‌جن‌خان (۱۳۸۹). پیشگفتار. زبان فارسی و رایانه: برگزیده مقالات کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران، کنفرانس مهندسی برق ایران، همایش زبانشناسی اسران، کارگاه زبان فارسی و رایانه (تا خرداد ۱۳۸۶). تهران: سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت).
- عبدالهی نورعلی، م. (۱۳۸۶). کندوکاو مسائل ریخت‌شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای وب. پایان‌نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی، دانشگاه شیراز.
- گل تاجی، م و س، بذرگر (۱۳۸۹). بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه **اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی**. تاریخ بازیابی: ۱۳۹۰/۹/۳. قابل بازیابی در:

http://www.aqlibrary.ir/index.php?module=TWArticles&file=index&func=view_pubarticles&did=885&pid=10

- مرتضایی، ل (۱۳۸۱). مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. ***فصلنامه اطلاع‌رسانی***. دوره ۱۷شماره۲و۱؛ پاییز و زمستان ۱۳۸۰.
- نوبهار، (۱۳۸۸). آیا باید جدا نوشت؟ تاریخ بازیابی: ۱۳۹۰/۹/۳. قابل بازیابی در:

www.aicit.org/jcit/ppl/11_april.pdf

- AleAhmad, A., Amiri, H., Rahgozar, M., Oroumchian, F. (2008). Experiments with English-Persian Text Retrieval .Retrieved: Retrieved 9 July 2012. Available in: khorshid.ut.ac.ir/~a.aleahmad/Files/inews22.pdf
- Dolamic, L., Savoy , J., (2009). Persian Language, is Stemming Efficient. Retrieved 9 July 2012 Available in: http://www.uni-weimar.de/medien/webis/research/events/tir-09/tir09-papers-final/dolamic09-persian-language-is-stemming-efficient.pdf.
- Kashefi, O., Mohseni, N., Minaei, B. (2010). Optimizing Document Similarity Detection in Persian Information Retrieval. *Journal of Convergence Information Technology*. Retrieved 9 July 2012. Available in: www.aicit.org/jcit/ppl/11_april.pdf
- Karimpour, R., (2008) .Using Part of Speech Tagging in Persian Information Retrieval. Retrieved 9 July 2012. Available in: www.clef-campaign.org/2008/.../Karimpour-paperCLEF2008.pdf
- Oroumchian, F., AleAhmad, A., Hakimian, P., Mahdikhani., F., (2007). F N-Geram and Local Context Analysis for Persian Text Retrieval. Retrieved 9 July 2012. Available in: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04555345
- Rahimtoroghi, E., Faili, H., Shakeri, A., (2010). A Structural Rule-based Stemmer for Persian. Retrieved 9 July 2012 Available in: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5734090