

تاریخ دریافت: ۸۶/۱۲/۵

تاریخ پذیرش: ۸۸/۷/۲۱

چالش‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب

چکیده

شیوه‌های گوناگون در نوشتن یک واژه، یکی از مشکلاتی است که در بازیابی مدارک مربوطه از طریق ابزارهای جستجو وجود دارد. در زبان فارسی نیز برخی از واژه‌ها به ریخت‌های متفاوتی نوشته می‌شوند. به همین دلیل این پژوهش به بررسی مسائلی پرداخته است که کاوشگران فارسی در کاوش ریخت‌های مختلف یک واژه با آن روبه‌رو هستند. برای پاسخگویی به سؤال‌های پژوهش از دو روش پیمایش مقایسه‌ای و اسنادی استفاده شده است. جامعه پژوهش شامل سه موتور کاوش گوگل، یاهو و آلتاویستا است که امکان جستجو به زبان فارسی را فراهم کرده‌اند. با مطالعه و دقت در متون فارسی، سیاهه‌ای شامل ۱۷ کلیدواژه انتخاب شد که هر کدام نمایان‌گر یک مورد از چالش‌های زبان فارسی در بازیابی اطلاعات هستند. پژوهشگران کلیدواژه‌ها را وارد فیلد جستجوی ابزارهای کاوش انتخابی کرده و تعداد بازیافت‌ها برای هر کدام از موتورهای کاوش را ثبت کردند.

تجزیه و تحلیل داده‌ها در دو سطح آمار توصیفی و استنباطی انجام شد. یافته‌ها نشان داد که موتورهای کاوش وب، شیوه‌های نگارش زبان فارسی را به منظور بهبود کاوش، مورد توجه قرار نداده‌اند. همچنین رابطه معناداری بین شکل واژه و نوع ابزار جستجو وجود دارد.

کلیدواژه‌ها: بازیابی اطلاعات، موتورهای کاوش، شیوه نگارش، زبان فارسی

sadeqabdolahi@yahoo.com

jowkar2003@yahoo.com

۱ کارشناس ارشد کتابداری و اطلاع‌رسانی

۲ عضو هیأت علمی دانشگاه شیراز

مقدمه

در زمانه‌ای که سر می‌بریم که در آن اطلاعات به شکل الکترونیکی تولید شده و انتقال می‌یابد. رشد علمی، فنی و فرهنگی در گرو برقراری ارتباط زبانی و کلامی با دنیای الکترونیکی دانش و فرهنگ است که وب نام دارد و این جز با تقویت کیفی زبان ملی میسر نمی‌شود. لیکن زبان فارسی، در تلاقی با جهان الکترونیکی، به‌ویژه از بعد شیوه نگارش^۱، دارای معضلاتی است که کاوش در محتویات آن را دچار مشکل می‌کند. این مقوله مستلزم تمهیداتی چند است تا زبان فارسی را از زبان شعر و ادب و عرفان، به زبانی مناسب با پهنه الکترونیکی دادوستد دانش، تبدیل کند (صدیقی، ۱۳۸۳).

پیرایش بر روی زبان‌های دنیا خیلی پیشتر از این آغاز شده است. گسترش استانداردهای آماده‌شده برای زبان، ساده کردن و کم کردن قاعده‌های پیچیده و استثناها در زبان روزمره و یکسان کردن گفتار و نوشتار از کارهایی است که بر روی بسیاری از زبان‌ها انجام شده است. برای نمونه، در زبان انگلیسی، زبان‌شناسان بسیاری از قاعده‌های این زبان را پیراسته‌اند و یادگیری این زبان را ساده کرده‌اند (یوسفان نجف آبادی، ۱۳۸۲).

در حال حاضر وبلاگ^۲‌های ایرانی از نظر فراوانی، جزو پنج کشور برتر جهان هستند. اما ویژگی‌های منفی آنها به این شرح است:

۱- شیوه نگارش وبلاگ‌ها بیشتر غیراستاندارد و متغیر است.

۲- نوشته‌های وبلاگ‌ها به نسبت حاوی غلط‌های املایی و نگارشی زیادی است، هرچند که بیشتر وبلاگ‌های مهم و پرخواننده، نگارش قابل قبولی دارند.

۳- شیوه نگارش وبلاگ‌ها، تابع محدودیت‌های محیط الکترونیکی و عدم تطبیق آن با الزام‌های خط فارسی است.

از طرف دیگر، همه نویسندگان وبلاگ‌ها، بنا به اهمیتی که زبان فارسی به عنوان زبان رسمی ما دارد و به حکم مسئولیتی که به عنوان صاحب رسانه دارند، باید خود را موظف بدانند که برای حفظ سلامت زبان فارسی در رسانه خود تلاش کنند. برای این منظور لازم است نویسندگان وبلاگ‌ها، بر کاربرد

^۱ رسم الخط

^۲ Weblog

زبان در وبلاگ خود نظارت کنند و استانداردهای نگارشی زبان و خط فارسی را رعایت کنند.

مسئله و اهمیت پژوهش

در قرن بیست و یکم اطلاعات به سرعت به سمت رقومی شدن پیش می‌رود. وب، بزرگترین مرجع اطلاعات در عصر ما محسوب می‌شود. اما شیوه نگارش فارسی باعث بروز چالش‌های جدی در امر نمایه‌سازی این زبان شده است. مسائل مربوط به خط، یکی از جنبه‌های مهم برنامه‌ریزی زبان^۱ است. ایجاد خط، انتخاب خط مشترک، تغییر خط و اصلاح آن از رایج‌ترین شکل‌های برخورد با خط به حساب می‌آید. گفتار و نوشتار دو بستر برای تحقق زبان هستند که گفتار، بازتاب طبیعی‌تر آن است، اما نوشتار به دلیل آنکه صورت ثابتی دارد و تحولات زبان را منعکس نمی‌کند، نیاز به برنامه‌ریزی زبانی و اصلاح دارد (اسلامی، ۱۳۸۱). چنانچه الگویی مناسب برای رفع این چالش‌ها ارائه شود، بخشی از مسائلی که امروز گریبان‌گیر زبان فارسی است، رفع خواهد شد. از جمله این مسائل می‌توان به موارد زیر اشاره کرد: جستجوی بهینه در وب، ایجاد پایگاه‌های اطلاعاتی به زبان فارسی، ایجاد نظام هم‌آهنگ اطلاع‌رسانی در کشور و مسائل دیگر (حری، ۱۳۷۲).

امروزه روش غالب در جستجوی اطلاعات از موتورهای کاوش وب، روش کلیدواژه‌ای است. اما جستجو به این روش، دشواری‌های خاص خود را دارد. چنانچه فردی به دنبال اطلاعاتی در مورد "آب گرمکن" باشد، این کلیدواژه را می‌تواند به چهار شکل بنویسد: "آب گرم کن، آبگرم کن، آب گرمکن و آبگرمکن". بنابراین موتورهای کاوش وب، برای هر کدام از این شکل‌ها، تعداد بازیافت‌های متفاوتی بازیابی خواهند کرد. چنانچه کاربری تنها یک شکل از این چهار مورد را به کار ببرد، اطلاعاتی که به اشکال دیگر نوشته شده است را از دست خواهد داد. سؤال مهمی که در اینجا مطرح می‌شود این است که: چگونه می‌توان بر این مسئله فایق آمد؟ آیا باید دست به اصلاح شیوه نگارش فارسی زد، یا اینکه نظام‌هایی پیشرفته و سازگار با این شیوه نگارش طراحی کرد؟

پیشینه پژوهش

ابزارهای جستجو اساساً بر مبنای زبان انگلیسی طراحی شده‌اند و کشورهای غیر انگلیسی‌زبان چالش‌های مشابهی با آن‌ها دارند. در پژوهش‌های انجام شده در خارج از کشور، پژوهشگران به مقایسه و

¹ Language Planning

ارزیابی موتورهای کاوش بین‌المللی و محلی پرداخته‌اند. ریشه‌سازی^۱، کوتاه‌سازی^۲ و جستجوی مترادف‌ها از جمله معیارهایی است که این پژوهشگران از آنها در ارزیابی‌های خود به کار برده‌اند. اما از آن جایی که بسیاری از این امکانات در ابزارهای کاوش فارسی وجود ندارد، پژوهشگران ایرانی را به سمت ارزیابی‌های متفاوت‌تری سوق داده است. پژوهشگران ایرانی معیارهایی مانند عملگرهای بولی، میزان پیوند به یک موتور کاوش، حجم پایگاه اطلاعاتی، رتبه بندی بازیافت‌ها، نمایش اطلاعات، واسط کاربری، روز آمد بودن اطلاعات، سرعت بازیابی اطلاعات، نمایه سازی اطلاعات را در پژوهش‌های خود به کار برده‌اند.

الف) پیشینه پژوهش در خارج از ایران

هدلاند و دیگران (Hedlund et al., 2000) به بررسی ویژگی‌های زبان سوئدی از دیدگاه بازیابی اطلاعات پرداختند. مشکلی که این پژوهشگران با آن مواجه بودند بازیابی ضعیف اطلاعات به زبان سوئدی بود. این زبان، ویژگی‌های منحصر به فردی دارد. از آن جمله می‌توان به مذکر و مؤنث بودن نام‌ها و نیز فراوانی استفاده از واژه‌های هم‌نگاشت اشاره کرد. آنها مطالعه‌ای مقایسه‌ای بر روی زبان‌های سوئدی، فنلاندی و انگلیسی انجام دادند تا میزان ابهام‌های واژگانی در این زبان‌ها را شناسایی کنند. پژوهشگران پیشنهاد می‌کنند که برچسب‌گذاری ادات سخن^۳ برای بازیابی واژه‌های هم‌نگاشت، می‌تواند مفید باشد.

سروکا (Sroka, 2000) نسخه‌های لهستانی چند ابزار جستجوی بین‌المللی را به همراه چند موتور کاوش محلی مورد سنجش قرار داد. مهم‌ترین معیار این سنجش، دقت^۴ ابزار کاوش بود که بر اساس ربط قضاوتی ۱۰ نتیجه نخست هر کاوش محاسبه شد. پژوهشگر تعداد بازیافت‌ها و زمان صرف شده برای یک کاوش را در مورد هر کدام از موتورهای کاوش ثبت کرد. در نتیجه این پژوهش "پالسکی اینفوسیک"^۵ به عنوان بهترین ابزار کاوش انتخاب شد.

مونز و دوریجکه (Monz & De Rijke, 2002) با تمرکز روی اثرات تحلیل‌های ریخت‌شناسی همچون ریشه‌سازی و جداسازی واژه‌های مرکب، به بررسی کارآیی بازیابی اطلاعات پرداختند. این مطالعه بر روی زبان‌های هلندی، آلمانی و ایتالیایی انجام شد. یافته‌ها نشان داد که بازیابی اطلاعات در حدود ۲۵٪ برای زبان آلمانی، ۶۹٪ برای زبان هلندی و ۲۵٪ برای زبان ایتالیایی بهبود داشته است.

¹ Stemming

² Truncation

³ Part of speech

⁴ Precision

⁵ Polski Infoseek

بارایلان و گتمان (Bar-Ilan & Gutman, 2002) توانایی ابزارهای جستجو را در مورد زبان‌های غیرانگلیسی مورد بررسی قرار دادند. ۴ زبان روسی، فرانسوی، مجاری و عبری جامعه این پژوهش به شمار می‌آیند. برای هر کدام از این زبان‌ها، ۳ موتور کاوش عمومی یعنی "آلتاویستا"^۱، "فست"^۲ و "گوگل"^۳ به همراه چند موتور کاوش محلی (مخصوص هر کدام از این زبان‌ها) آزمایش شد. این بررسی نشان داد که موتورهای کاوش عمومی وب، ویژگی‌های زبانی زبان‌های غیر انگلیسی را در جستجوی اطلاعات نادیده می‌گیرند.

مقداد (Moukdad, 2005) در پژوهشی، عملکرد ۳ ابزار جستجوی عمومی را با ۳ موتور کاوش عربی (که به طور خاص مسائل زبان‌شناختی عربی را لحاظ می‌کنند) مورد مقایسه قرار داد. یافته‌ها نشان داد که موتورهای کاوش عمومی، نظیر "آلدوب"^۴، "آلتاویستا" و "گوگل" در بازیابی مدارک عربی، ناقص عمل می‌کنند. همچنین یافته‌های این پژوهش، نیاز به پژوهش‌های بیشتر در زمینه عملی بودن ابزارهای جدید بازیابی اطلاعات در موتورهای کاوش را نشان داد.

تاث (Toth, 2006) به بررسی قابلیت‌های زبان‌شناختی موتورهای کاوش انگلیسی و مجاری پرداخت. پژوهشگر ۳ ابزار جستجوی انگلیسی "گوگل"، "آلتاویستا" و "آلدوب" را با ۵ موتور کاوش محلی مورد مقایسه قرار داد. تحلیل داده‌ها بر پایه چند شاخص انجام شد که عبارت بودند از: ریشه‌سازی، بازیابی لهجه‌های مختلف، کوتاه‌سازی و جستجوی مترادف‌ها. یافته‌ها حاکی از آن بود که موتورهای کاوش محلی، مسائل زبان مجاری را بهتر از موتورهای کاوش انگلیسی مورد توجه قرار داده بودند. ابزارهای انگلیسی‌زبان، لهجه‌های مختلف زبان مجاری را به خوبی پشتیبانی نمی‌کردند، که این امر منجر به بازیابی ضعیف اطلاعات می‌شد.

ب) پیشینه در ایران

کوشا (۱۳۸۱) با استفاده از معیارهای مستند به ارزیابی جداگانه و نیز تجزیه و تحلیل مقایسه‌ای ابزارهای کاوش دارای واسط جستجوی فارسی پرداخت. شش ابزار کاوش برگزیده از طریق ۲۷ معیار مرتبط با قابلیت‌های جستجو و بازیابی اطلاعات با یکدیگر مورد مقایسه قرار گرفتند. موتورهای کاوش

¹ Altavista

² Fast

³ Google

⁴ All the web

انتخابی عبارت بودند از: "گوگل"، "ایران کلیک"^۱، "ایران هو"^۲، "ایران مهر"^۳، "پارسیک"^۴، و "اپن دایرکتوری"^۵. نتیجه پژوهش نشان داد که از نظر امکانات جستجو و بازیابی اطلاعات، ابزار کاوش "گوگل" در رتبه نخست و راهنمای موضوعی "ایران هو" در رتبه دوم قرار دارند. به منظور بررسی عامه پسند و رایج بودن ابزارهای کاوش، پژوهشگر تعداد صفحه‌ها یا وب سایت پیوند داده شده به آنها را مورد مقایسه قرار داد. نتیجه نشان داد که رابطه مستقیمی میان رایج و عامه پسند بودن ابزارهای کاوش مورد مطالعه با توانایی‌های جستجوی اطلاعات آنها وجود ندارد.

یوسفان نجف آبادی (۱۳۸۲) در پژوهشی با عنوان "یک نظام بازیابی متنی برای زبان فارسی بر پایه معانی پنهان"، نظامی را طراحی کرده است که با استفاده از نمایه گذاری معانی به بازیابی اطلاعات متنی زبان فارسی می‌پردازد. کارآیی این نظام با ریشه‌یابی و بدون ریشه‌یابی با استفاده از یک مجموعه اسناد گردآوری شده به این منظور و به کمک معیارهای دقت و بازیافت مورد ارزیابی قرار گرفته است. برای کمک به یافتن فهرست واژه‌های سراسری و ریشه‌یابی، یک زبان برنامه‌نویسی ساده، طراحی و بر پایه قاعده‌های زبان فارسی، روشی نوین برای شناسایی خودکار فعل‌های فارسی پیشنهاد شد.

رائی ساربانقلی (۱۳۸۴) در پژوهش خود به بررسی مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت به کمک کاربران مرکز اینترنت دانشگاه آزاد اسلامی شبستر پرداخت. یافته‌های پژوهش نشان داد که ۷۷٪ کاربران از جستجوی پیشرفته گوگل استفاده می‌کنند. بیشتر مشکل کاربران در جستجو، عدم توجه ایشان به شکل‌های مختلف نوشتاری واژه و عدم استفاده از عملگر "OR" بود. این پژوهش، رابطه‌ی معناداری بین گذراندن دوره‌های آموزشی و نیز مدت استفاده کاربران از اینترنت با مهارت آنها را نشان داد.

مطالعه پژوهش‌های انجام شده نشان داد که موتورهای کاوش در بازیابی منابع بر اساس شکل‌های مختلف یک واژه توانمند نیستند. با توجه به آنچه در مقدمه و پیشینه پژوهش بحث شد، از آنجایی که بسیاری از مدارک که با املاءهای مختلف یک کلمه در محیط اینترنت وجود دارند، نمی‌توانند توسط موتورهای کاوش بازیابی شوند، ضروری است که از طریق پژوهش بتوان راهکارهای لازم را شناسایی

¹ Iran click

² Iranhoo

³ Iran mehr

⁴ Parseek

⁵ Open directory

کرد. بنابراین هدف این پژوهش شناسایی راه‌ها و روش‌هایی است که از مشکلات بازیابی مدارک که بر اساس شیوه نگارش واژه‌ها به وجود می‌آید جلوگیری کرد. در این راستا سؤال‌های پژوهشی زیر طراحی شده است:

۱. کدام‌یک از ویژگی‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از وب مشکل ایجاد می‌کند؟
۲. آیا ابزارهای کاوش بین‌المللی (با در نظر گرفتن شیوه نگارش فارسی) نتایج جستجوی یکسانی برای شکل‌های مختلف یک کلمه به دست می‌دهند؟
۳. آیا رابطه معناداری بین شکل واژه‌ها و نوع ابزار جستجو وجود دارد؟

طرح پژوهش

برای پاسخگویی به سؤال نخست پژوهش، روش اسنادی انتخاب گردید. برای پاسخگویی به سؤال‌های دوم و سوم پژوهش از روش پیمایش مقایسه‌ای استفاده شد. با کمک روش پیش گفته مشکلاتی که شکل‌های مختلف واژه در ابزارهای کاوش انتخابی ایجاد می‌کنند، مورد بررسی و تجزیه و تحلیل قرار گرفت.

بررسی ادبیات پژوهش نشان داد که هفت موتور کاوش بین‌المللی یعنی آلتاویستا، اکسایت^۱، گوگل، هات بات^۲، اینفوسیک^۳، لایکاس^۴ و یاهو^۵ به عنوان پر استفاده‌ترین ابزارهای کاوش در دنیا شناخته شده‌اند. در این میان تنها "گوگل"، "یاهو" و "آلتاویستا" امکان جستجو به زبان فارسی را فراهم کرده‌اند. بنابراین، این سه موتور کاوش به عنوان بستر برای پیشبرد این پژوهش در نظر گرفته شدند. ابزارهای کاوش بین‌المللی، از عنکبوت^۶ یا خزنده به منظور شناسایی و نمایه‌سازی صفحه‌ها یا سایت‌های وب در زبان‌های مختلف از جمله زبان فارسی استفاده می‌کنند. این روش نوعی نمایه‌سازی خودکار می‌باشد و می‌تواند صفحه‌های فارسی را در قالب یونی‌کد^۷ شناسایی و در پایگاه خود ذخیره کنند.

¹ Excite

² Hotbot

³ Infoseek

⁴ Lycos

⁵ Yahoo

⁶ Spider

⁷ Unicode

با بررسی پژوهش‌های فارسی، سیاهه‌ای شامل ۱۷ کلیدواژه به صورت تعمدی به عنوان نمونه انتخاب شد. این کلیدواژه‌ها هر کدام نمایانگر یک مورد از چالش‌های زبان فارسی در بازیابی اطلاعات هستند. این کلیدواژه‌ها که به عنوان وسیله گردآوری داده‌ها شناخته می‌شوند، عبارتند از:

موسی یا موسا	اتاق یا اطاق
املاء یا املا	توراه یا تورات
باغها یا باغ‌ها	عطایی یا عطائی
موحدی یا موحدی	مورچه گان یا مورچگان
پتاسیم یا پتاسیوم	خانه من یا خانه‌ی من
زبان شناس یا زبانشناس	مسؤول یا مسؤل
مسئله یا مسأله	دقیقاً یا دقیقن
شمشیرباز یا شمشیرباز	پرتو آفتاب یا پرتوی آفتاب
	فرایند یا فرآیند

به منظور انجام کاوش‌ها، نخست به بخش جستجوی پیشرفته ابزار کاوش وارد شده، سپس در قسمت زبان‌ها، زبان فارسی به عنوان پیش فرض جستجو انتخاب شد. هر یک از شکل‌های مختلف واژه را وارد فیلد جستجوی ابزارهای کاوش انتخابی کرده، سپس تعداد یافته‌های حاصل از جستجو توسط هر یک از موتورهای کاوش ثبت شد. اگرچه موتور کاوش "آلتاویستا"، امکان جستجو به زبان فارسی را فراهم کرده‌است، اما در قسمت جستجوی پیشرفته، امکان انتخاب زبان به کاربران داده نمی‌شود. این مشکل باعث شد که به هنگام بازیابی برخی از کلیدواژه‌ها، تعدادی از بازیافت‌ها به زبان عربی ارائه شود. مشخص است که صافی زبان^۱ در اینجا درست عمل نکرده‌است.

تجزیه و تحلیل داده‌ها

در این پژوهش، تجزیه و تحلیل داده‌ها در دو سطح توصیفی و استنباطی انجام شد. توصیف داده‌ها با آماره‌های توصیفی (جدول فراوانی و نمودار درصد) صورت گرفت. برای بررسی معناداری

^۱ Language filter

رابطه بین نوع موتور کاوش و شکل واژه از آماره‌های استنباطی (آزمون خی دو^۱ و ضریب فی^۲) استفاده شده است.

سؤال ۱. کدامیک از ویژگی‌های شیوه نگارش زبان فارسی در بازیابی اطلاعات از وب مشکل ایجاد می‌کند؟

در پاسخ به این سؤال باید به طور کلی شرایطی که موجب نگارش یک واژه به ریخت‌های گوناگون می‌شود را مورد مطالعه قرار داد. برخی از مطالعه‌های صورت گرفته (مانند مرتضایی، ۱۳۷۶) نیز به این امر توجه کرده‌اند. نتیجه بررسی نشان داد که عامل‌هایی مانند مهارت استفاده از حالت‌های نوشتاری یا دستوری سایر زبان‌ها مانند عربی، سلیق‌های مختلف در نگارش واژه‌ها و یا برگردان آنها به فارسی صورت‌های مختلف یک واژه را در نگارش تشکیل می‌دهد. نمونه‌هایی در این رابطه به شرح زیر می‌باشد:

۱. برگردان واژه‌های خارجی مانند پستالوزی/پستالزی، پتاسیم یا پتاسیوم
۲. نشانه‌های جمع مانند مدارس/ مدرسه‌ها، استادان/ استاداها
۳. پیوسته‌نویسی یا جدانویسی مانند مردم شناسی/ مردمشناسی، روان‌شناسی/ روانشناسی
۴. تنوین مانند اصلاً/ اصلن، دقیقاً/ دقیقن
۵. کسره اضافه مانند اسب سواری/ اسب سواری
۶. صامت میانجی "ی" مانند پرتوی آفتاب/ پرتو آتاب
۷. همزه برای واژه‌های مختوم به "ها" بیان حرکت مانند جامه من/ جامه من/ جامه‌ی من
۸. همزه پایانی مانند انشاء/ انشاء، املاء/ املا
۹. همزه پایانی متصل به "یا" وحدت یا نکره مانند کرسی "ی" / "کرسى" (عطایی/ عطائی)
۱۰. به کار بردن همزه به صورت‌های مختلف مانند مسئله/ مسأله، مسؤل/ مسؤول

¹ χ^2

² Phi coefficient

۱۱. الف مقصوره که گاه به صورت الف مانند اسماعیل، هارون و جز آن نوشته می شود و گاه به همان شکل عربی مانند عیسی

۱۲. تنوع املائی مانند پتر کبیر/ پطر کبیر، اختاپوس/ اختاپوٹ

۱۳. استفاده از "ا" و "آ" به جای یکدیگر مانند فرآهم/ فراهم، برآیند/ براینند

۱۴. کلمات خاص در پیوسته نویسی و جدانویسی مانند علاقمند/ علاقه بندی، اندیشمند/ اندیشه مند

۱۵. "تای" منقوط مانند صلاة/ صلات، مشکوة/ یا مشکات

۱۶. نشانه تشدید مانند معین/ معین، علی/ علی

سؤال ۲. آیا ابزارهای کاوش بین المللی (با در نظر گرفتن شیوه نگارش فارسی) نتایج جستجوی یکسانی برای شکل های مختلف یک کلمه به دست می دهند؟

۱. شیوه برگردان واژه های خارجی

جدول ۱. آمار باز یافت ها برای مقوله شیوه برگردان واژه ها خارجی (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یاهو	گوگل		پتاسیم	شکل واژه
۱۵۳۳۰۰	۲۵۴۰۰	۲۳۹۰۰	۱۰۴۰۰۰	فراوانی	پتاسیم	
%۱۰۰	%۱۶/۶	%۱۵/۶	%۶۷/۸	درصد		
۷۷۵	۳۴۲	۳۴۰	۹۳	فراوانی	پتاسیوم	
%۱۰۰	%۴۴/۱	%۴۳/۹	%۱۲	درصد		
۱۵۴۰۷۵	۲۵۷۴۲	۲۴۲۴۰	۱۰۴۰۹۳	فراوانی	جمع کل	
%۱۰۰	%۱۶/۷	%۱۵/۷	%۶۷/۶	درصد		

همان طور که در جدول نشان داده شد، به طور کامل دو نتیجه متفاوت برای این دو کلیدواژه بازیابی شده است. اطلاعات بیشتری با کلید واژه "پتاسیم" ذخیره شده است و انتخاب کلیدواژه "پتاسیوم" باعث از دست رفتن این اطلاعات می شود. همچنین در هیچ یک از موتور های کاوش، تمهیدی برای

بازیابی صفحه‌های دارای کلیدواژه "پتاسیم" در هنگام جستجوی کلیدواژه "پتاسیوم" اندیشیده نشده است.

۲. شیوه نگارش نشانه‌های جمع

جدول ۲: تعداد بازیافت‌ها برای مقوله شیوه نگارش نشانه‌های جمع (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		باغ‌ها	شکل واژه
۳۵۶۰۰	۲۸۵۰۰	۲۷۰۰	۴۴۰۰	فراوانی	باغ‌ها	
%۱۰۰	%۸۰/۱	%۷/۶	%۱۲/۴	درصد		
۸۲۹۰۰	۱۹۰۰	۱۶۶۰۰	۶۴۴۰۰	فراوانی	باغها	
%۱۰۰	%۲/۳	%۲۰	%۷۷/۷	درصد		
۱۱۸۵۰۰	۳۰۴۰۰	۱۹۳۰۰	۶۸۸۰۰	فراوانی	جمع کل	
%۱۰۰	%۲۵/۷	%۱۶/۳	%۵۸/۱	درصد		

چنانچه کاربری کلیدواژه "باغها" را انتخاب کند، بیشتر اطلاعات موجود که با کلیدواژه "باغ‌ها" ذخیره شده است را از دست می‌دهد. از طرف دیگر انتخاب "باغ‌ها" نیز، ریزش کاذب به بار می‌آورد، چرا که موتورهای کاوش، هر فاصله خالی بین واژه‌ها را همچون عملگر^۱ And در نظر می‌گیرد. بنابراین، صفحه‌هایی بازیابی می‌شوند که در آن کلمه "ها" به‌تنهایی آمده است. پس هیچکدام از موتورهای کاوش تمهیدی برای این مسئله نیندیشیده‌اند.

۳. پیوسته‌نویسی و یا جدانویسی ترکیب‌ها

جدول ۳: تعداد بازیافت‌ها برای مقوله پیوسته‌نویسی و یا جدانویسی ترکیب‌ها (n=3)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		زبان شناس	شکل واژه
۵۸۳۰۰۰	۲۰۸۰۰۰	۲۰۲۰۰۰	۱۷۳۰۰۰	فراوانی	زبان شناس	
%۱۰۰	%۳۵/۷	%۳۴/۶	%۲۹/۷	درصد		
۲۸۹۱	۱۰۹۰	۹۴۲	۸۵۹	فراوانی	زبان‌شناس	

^۱ Operator

٪۱۰۰	٪۳۷/۷	٪۳۲/۶	٪۲۹/۷	درصد	زبان شناس
۲۲۳۶۹۵	۲۰۸۰۰	۲۰۲۰۰۰	۸۵۹	فراوانی	
٪۱۰۰	٪۹/۳	٪۹۰/۳	٪۰/۴	درصد	جمع کل
۸۰۹۵۵۰	۲۲۹۸۹۰	۴۰۴۹۴۲	۱۷۴۷۱۸	فراوانی	
٪۱۰۰	٪۲۸/۴	٪۵۰	٪۲۱/۶	درصد	

همان‌طور که از جدول برمی‌آید، این ترکیب را به سه صورت می‌توان نوشت: جدانویس (زبان شناس)، پیوسته‌نویس (زبان‌شناس) و بی‌فاصله‌نویس (زبان‌شناس) (کابلی، ۱۳۷۳). موتورهای کاوش "آلتاویستا" و "یاهو" برای شکل‌های جدانویس و بی‌فاصله‌نویس دو نتیجه یکسان به بار آورده‌اند، در حالی که موتور کاوش "گوگل" شکل پیوسته‌نویس و بی‌فاصله‌نویس را یکی محسوب کرده‌است. چنین عملکردی به روبات یا عنکبوت موتورهای کاوش برمی‌گردد که هر کدام طبق پیش‌فرض‌هایی که برای آنها تعریف شده‌است، واژه‌ها را شناسایی کرده و در پایگاه خود، نمایه می‌کنند. عملکرد مطلوب آن است که موتورهای کاوش برای هر سه شکل، نتیجه یکسانی به بار آورند.

۴. مشکل تنوین

جدول ۴: تعداد بازیافت‌ها برای مقوله تنوین (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یاهو	گوگل		شکل واژه	دقیقاً
۲۰۲۸۰۰۰	۸۷۵۰۰۰	۶۸۸۰۰۰	۴۶۵۰۰۰	فراوانی		
٪۱۰۰	٪۴۳/۱	٪۳۳/۹	٪۲۲/۹	درصد		
۱۶۷۱۰	۲۲۳۰	۲۱۸۰	۱۲۳۰۰	فراوانی	دقیقن	
٪۱۰۰	٪۱۳/۳	٪۱۳	٪۷۳/۶	درصد		
۲۰۴۴۷۱۰	۸۷۷۲۳۰	۶۹۰۱۸۰	۴۷۷۳۰۰	فراوانی	جمع کل	
٪۱۰۰	٪۴۳	٪۳۳/۷	٪۲۳/۳	درصد		

در بعضی صفحه‌کلیدها، تنوین در جای همیشگی خود قرار نمی‌گیرد، این امر ماشین‌نویس‌ها را سر در گم می‌کند. به‌ناچار، تنوین نصب را که بیشتر در قیدها (مانند واقعاً، فوراً، جداً) کاربرد دارد به همان صورتی که خواننده می‌شوند (واقعن، فورن، جدن) به کار می‌برند. بر اساس نتیجه‌ایی که از جدول ۴ به

دست آمده است، می‌توان این‌طور استنباط کرد که این پراکندگی در کلیدواژه‌هایی که دارای تنوین هستند موجب از دست رفتن بخشی از اطلاعات موجود می‌گردد.

۵. کسره اضافه

جدول ۵: تعداد بازیافت‌ها برای مقوله کسره اضافه (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		شمشیرباز	شکل واژه
۴۷۰۰۰	۱۵۰۰۰	۱۵۷۰۰۰	۱۶۳۰۰۰	فراوانی		
%۱۰۰	%۳۱/۹	%۲۳/۴	%۳۴/۷	درصد		
۳۷۸۷۰۰	۱۵۰۰۰	۱۵۷۰۰۰	۷۱۷۰۰	فراوانی	شمشیرباز	
%۱۰۰	%۳۹/۶	%۴۱/۵	%۱۸/۹	درصد		
۸۴۸۷۰۰	۳۰۰۰۰	۳۱۴۰۰۰	۲۳۴۷۰۰	فراوانی	جمع کل	
%۱۰۰	%۳۵/۳	%۳۷	%۲۷/۷	درصد		

در دو صورت، با وارد کردن یا نکردن نشانه کسره اضافه در نوشتار، موتور کاوش گوگل دو نتیجه متفاوت برای ترکیب **شمشیرباز** و **شمشیرباز**، به دست می‌دهد، اما آلتاویستا و یاهو با آن یکسان عمل کرده‌اند.

۶. صامت میانجی "ی"

جدول ۶: تعداد بازیافت‌ها برای مقوله صامت میانجی "ی" (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		پرتو آفتاب	شکل واژه
۲۲۹۴۰۰	۶۹۶۰۰	۷۸۸۰۰	۸۱۰۰۰	فراوانی		
%۱۰۰	%۳۰/۳	%۳۴/۴	%۳۵/۳	درصد		
۱۳۵۲۰	۱۰۳۰۰	۱۵۷۰	۱۶۵۰	فراوانی	پرتوی آفتاب	
%۱۰۰	%۷۶/۲	%۱۱/۶	%۱۲/۲	درصد		
۲۴۲۹۲۰	۷۹۹۰۰	۸۰۳۷۰	۸۲۶۵۰	فراوانی	جمع کل	
%۱۰۰	%۳۲/۹	%۳۳/۱	%۳۴	درصد		

جوینده اطلاعات در حین کاوش، شاید هیچ‌گاه به این نکته نیندیشد که گذاشتن یا نگذاشتن میانجی "ی" چه تغییر زیادی در تعداد بازیافت‌ها خواهد گذاشت.

۷. استفاده یا عدم استفاده از "همزه"، برای واژه‌های مختوم به "های" بیان حرکت، در حالت مضاف

جدول ۷: تعداد بازیافت‌ها برای مقوله استفاده یا عدم استفاده از "همزه" (n=3)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یاهو	گوگل		شکل واژه	
۱۴۲۷۰	۹۲۱۰	۲۲۶۰	۲۸۰۰	فراوانی	خانه من	شکل واژه
%۱۰۰	%۶۴/۵	%۱۵/۸	%۱۹/۶	درصد		
۷۶۲۰	۲۵۶۰	۲۲۶۰	۲۸۰۰	فراوانی	خانه من	
%۱۰۰	%۳۳/۶	%۲۹/۷	%۳۶/۷	درصد		
۳۶۲۰۰۰۰	۱۱۰۰۰۰۰	۱۲۳۰۰۰۰	۱۲۹۰۰۰۰	فراوانی	خانه‌ی من	
%۱۰۰	%۳۰/۴	%۳۴	%۳۵/۶	درصد		
۳۶۴۱۸۹۰	۱۱۱۱۷۷۰	۱۲۳۴۵۲۰	۱۲۹۵۶۰۰	فراوانی	جمع کل	
%۱۰۰	%۳۰/۴۵	%۳۳/۹	%۳۵/۶	درصد		

نتیجه غیر منتظره این است که ترکیب "خانه‌ی من"، بازیافت‌های بسیار بیشتری از دو ترکیب دیگر به دست داده است، در حالیکه نوشتن دو ترکیب دیگر، برای ماشین‌نویس آسان‌تر است.

۸. "همزه" پایانی

جدول ۸: تعداد بازیافت‌ها برای مقوله "همزه" پایانی (n=2)

جمع کل	موتور کاوش			آماره	متغیر		
	آلتاویستا	یاهو	گوگل		شکل واژه		
۵۳۶۲۰۰	۴۹۱۰۰۰	۱۰۹۰۰	۳۴۳۰۰	فراوانی	املا	شکل واژه	
%۱۰۰	%۹۱/۶	%۲	%۶/۴	درصد			
۲۱۶۵۸۷	۲۱۴۰۰۰	۱۶۱۰	۹۷۷	فراوانی	املاء		
%۱۰۰	%۹۸/۸	%۰/۷	%۰/۵	درصد			
۷۵۲۷۸۷	۷۰۵۰۰۰	۱۲۵۱۰	۳۵۲۷۷	فراوانی	جمع کل		
%۱۰۰	%۹۳/۷	%۱/۷	%۴/۷	درصد			

این ناهماهنگی در شیوه نگارش "همزه" پایانی، به راحتی موجب از دست دادن اطلاعات با ارزش می‌شود. به طور نمونه در موتور کاوش آلتاویستا، فردی که شکل دوم را به کار می‌برد، حدود ۲۸۰ هزار نتیجه جستجو را از دست می‌دهد.

۹. "همزه" پایانی متصل به "یای" وحدت یا نکره

جدول ۹: تعداد بازیافت‌ها برای مقوله "همزه" پایانی متصل به "یای" وحدت یا نکره (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		شکل واژه	
۲۲۴۰۰۰	۱۰۹۰۰۰	۹۶۶۰۰	۱۸۴۰۰	فراوانی	عطائی	شکل واژه
%۱۰۰	%۴۸/۷	%۴۳/۱	%۸/۲	درصد		
۲۸۷۹۰۰	۱۰۹۰۰۰	۹۶۶۰۰	۸۲۳۰۰	فراوانی	عطایی	
%۱۰۰	%۳۷/۹	%۳۳/۶	%۲۸/۶	درصد		
۵۱۱۹۰۰	۲۱۸۰۰۰	۱۹۳۲۰۰	۱۰۰۷۰۰	فراوانی	جمع کل	
%۱۰۰	%۴۲/۶	%۳۷/۷	%۱۹/۷	درصد		

نکته جالب در این کاوش، این است که موتورهای جستجوی آلتاویستا و یاهو برای هر دو شکل، نتیجه یکسانی داشته‌اند. اما مشخص نیست که چرا در واژه‌های دیگری که با همزه به کار می‌روند (مانند مسئله/ مسأله و مسئول/ مسؤول)، این اتفاق نیفتاده است.

۱۰. به کاربردن "همزه" به صورت های مختلف

جدول ۱۰-۱: تعداد بازیافت‌ها برای مقوله به کاربردن "همزه" به صورت های مختلف (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		شکل واژه	
۵۶۹۶۰۰۰	۳۸۱۰۰۰۰	۱۵۳۰۰۰۰	۳۵۶۰۰۰	فراوانی	مسئول	شکل واژه
%۱۰۰	%۶۶/۹	%۲۶/۹	%۶/۳	درصد		
۴۶۷۰۰۰۰	۱۶۸۰۰۰۰	۱۵۸۰۰۰۰	۱۴۱۰۰۰۰	فراوانی	مسؤول	
%۱۰۰	%۳۶/۰	%۳۳/۸	%۳۰/۲	درصد		
۱۰۳۶۶۰۰۰	۵۴۹۰۰۰۰	۳۱۱۰۰۰۰	۱۷۶۶۰۰۰	فراوانی	جمع کل	
%۱۰۰	%۵۳/۰	%۳۰/۰	%۱۷/۰	درصد		

جدول ۲-۱۰: تعداد بازیافت‌ها برای مقوله به کاربردن "همزه" به صورت‌های مختلف (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یاهو	گوگل		مسئله	شکل واژه
۵۱۸۰۰۰۰	۲۰۷۰۰۰۰	۱۳۷۰۰۰۰	۱۷۴۰۰۰۰	فراوانی		
%۱۰۰	%۴۰/۰	%۲۶/۴	%۳۳/۶	درصد		
۴۵۳۲۰۰۰	۳۶۳۰۰۰۰	۷۷۲۰۰۰	۱۳۰۰۰۰	فراوانی	مسئله	
%۱۰۰	%۸۰/۱	%۱۷/۰	%۲/۹	درصد		
۹۷۱۲۰۰۰	۵۷۰۰۰۰۰	۲۱۴۲۰۰۰	۱۸۷۰۰۰۰	فراوانی	جمع کل	
%۱۰۰	%۵۸/۷	%۲۲/۱	%۱۹/۳	درصد		

این دو کلیدواژه، از واژه‌های پربسامد در فارسی هستند و نتیجه به خوبی عدم یکدستی در نگارش "همزه" را نشان می‌دهند.

۱۱. تنوع استفاده از "الف" مقصوره

جدول ۱۱: تعداد بازیافت‌ها برای مقوله تنوع استفاده از "ی" در واژه‌های عربی مختوم به "ا" (n=3)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یاهو	گوگل		موسی	شکل واژه
۴۵۲۹۰۰۰	۳۴۷۰۰۰۰	۵۰۶۰۰۰	۵۵۳۰۰۰	فراوانی		
%۱۰۰	%۷۶/۶	%۱۱/۲	%۱۲/۲	درصد		
۴۲۳۵۰۰۰	۳۴۸۰۰۰۰	۵۰۶۰۰۰	۲۴۹۰۰۰	فراوانی	موسی	
%۱۰۰	%۸۲/۲	%۱۱/۹	%۵/۹	درصد		
۱۷۹۸۶	۱۳۹۰۰	۳۳۵۰	۷۳۶	فراوانی	موسا	
%۱۰۰	%۷۷/۳	%۱۸/۶	%۴/۱	درصد		
۹۰۹۱۰۲۰	۱۱۱۱۷۷۰	۶۹۶۳۹۰۰	۱۰۱۵۳۵۰	فراوانی	جمع کل	
%۱۰۰	%۱۲/۲	%۷۶/۶	%۱۱/۲	درصد		

حرف "ی"، روی بعضی صفحه‌کلیدها به شکل عربی آن نوشته می‌شود (یعنی به صورت ی). داده‌های جدول ۱۱ بیانگر این است که موتور کاوش "گوگل" برای دو شکل موسی و موسی، به طور

کامل دو نتیجه متفاوت به دست می‌دهد. این نکته باید به کاوشگران آموزش داده شود که برای نوشتن حرف "ی" به شکل فارسی، باید از کلید شیفِت به همراه کلید دیگری - که در هر رایانه‌ای متفاوت است - استفاده کنند.

۱۲. تنوع املائی بعضی واژه‌ها که همه درست هستند

جدول ۱۲. تعداد بازیافت‌ها برای مقوله تنوع املائی بعضی واژه‌ها که همه درست هستند (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یا هو	گوگل			
۳۸۳۰۰۰۰	۱۰۶۰۰۰۰	۱۰۳۰۰۰۰	۱۷۴۰۰۰۰	فراوانی	اتاق	شکل واژه
%۱۰۰	%۲۷/۷	%۲۶/۹	%۲۷/۷	درصد		
۳۱۲۹۰۰	۹۸۷۰۰	۸۹۲۰۰	۱۲۵۰۰۰	فراوانی	اطاق	
%۱۰۰	%۳۱/۵	%۲۸/۵	%۳۹/۹	درصد		
۴۱۴۲۹۰۰	۱۱۵۸۷۰۰	۱۱۱۹۲۰۰	۱۸۶۵۰۰۰	فراوانی	جمع کل	
%۱۰۰	%۲۸/۰	%۲۷/۰	%۴۵/۰	درصد		

کاوشگری که می‌خواهد به سرعت به اطلاعات دسترسی پیدا کند، کمتر حضور ذهن دارد که تمام صورت‌های املائی را به خاطر بیاورد.

۱۳. استفاده از "ا" و "آ" به جای هم

جدول ۱۳: تعداد بازیافت‌ها برای مقوله استفاده از "ا" و "آ" به جای هم (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یا هو	گوگل			
۱۳۷۵۰۰۰	۴۱۵۰۰۰	۴۰۳۰۰۰	۵۵۷۰۰۰	فراوانی	فرآیند	شکل واژه
%۱۰۰	%۳۰/۲	%۲۹/۳	%۴۰/۵	درصد		
۱۲۰۸۰۰۰	۳۷۶۰۰۰	۳۶۰۰۰۰	۴۷۲۰۰۰	فراوانی	فرآیند	
%۱۰۰	%۳۱/۱	%۲۹/۸	%۳۹/۱	درصد		
۲۵۸۳۰۰۰	۷۹۱۰۰۰	۷۶۳۰۰۰	۱۰۲۹۰۰۰	فراوانی	جمع کل	
%۱۰۰	%۳۰/۶	%۲۹/۵	%۳۹/۸	درصد		

کاوشرگان از میان دو شکل بالا، آن صورتی را انتخاب می‌کنند که بیشتر به نوشتن آن عادت کرده‌اند. در نتیجه، صفحه‌های زیادی را که ممکن است حاوی اطلاعات با ارزشی باشند، از دست می‌دهند.

۱۴. واژه‌های خاص در پیوسته‌نویسی و جدانویسی

جدول ۱۴: تعداد بازیافت‌ها برای دو کلیدواژه "مورچه گان" یا "مورچه گان" (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یا هو	گوگل		شکل واژه	
۲۰۴۰	۷۰۰	۶۷۸	۶۶۲	فراوانی	مورچه گان	شکل واژه
%۱۰۰	%۳۴/۳	%۳۳/۲	%۳۲/۵	درصد		
۱۴۲۷۰	۱۷۴۰	۱۶۳۰	۱۰۹۰۰	فراوانی	مورچه گان	
%۱۰۰	%۱۲/۲	%۱۱/۴	%۷۶/۴	درصد		
۱۶۳۱۰	۲۴۴۰	۲۳۰۸	۱۱۵۶۲	فراوانی	جمع کل	
%۱۰۰	%۱۵/۰	%۱۴/۲	%۷۰/۹	درصد		

داده‌های جدول بیانگر این است که شکل نوشتاری مورچه گان بازیافت‌های بیشتری در بر داشته است و تمایل برای حذف "های بیان حرکت" در بین نویسندگان وجود دارد.

۱۵. "تای" منقوط

جدول ۱۵: تعداد بازیافت‌ها برای مقوله "تای" منقوط (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	یا هو	گوگل		شکل واژه	
۲۲۲۱۰	۲۱۸۰۰	۳۱۷	۹۳	فراوانی	توراة	شکل واژه
%۱۰۰	%۹۸/۲	%۱/۴	%۰/۴	درصد		
۲۰۷۲۰۰	۷۰۳۰۰	۶۵۱۰۰	۷۱۸۰۰	فراوانی	تورات	
%۱۰۰	%۳۳/۹	%۳۱/۴	%۳۴/۷	درصد		
۲۲۹۴۱۰	۹۲۱۰۰	۶۵۴۱۷	۷۱۸۹۳	فراوانی	جمع کل	
%۱۰۰	%۴۰/۱	%۲۸/۵	%۳۱/۳	درصد		

در این مورد، تعداد کمی نتیجه برای کلمه "توراة" آمده است (به استثنای "آلتاویستا"). پس می‌توان گفت که تمایل نویسندگان به نوشتن "تای" منقوط، به صورت «ت» است. اما همچنان صفحه‌هایی وجود دارند که با "ة" ذخیره شده‌اند و ممکن است دارای اطلاعات خوبی باشند.

۱۶. نشانه "تشدید"

جدول ۱۶: تعداد بازیافت‌ها برای مقوله نشانه "تشدید" (n=2)

جمع کل	موتور کاوش			آماره	متغیر	
	آلتاویستا	ياهو	گوگل		موحدی	شکل واژه
۱۳۷۷۰۰	۴۷۰۰۰	۴۶۷۰۰	۴۴۰۰۰	فراوانی		
%۱۰۰	%۳۴/۱	%۳۳/۹	%۳۲/۰	درصد		
۱۱۶۴۰۰	۴۸۰۰۰	۴۵۴۰۰	۲۳۰۰۰	فراوانی	موحدی	
%۱۰۰	%۴۱/۲	%۳۹/۰	%۱۹/۸	درصد		
۲۵۴۱۰۰	۹۵۰۰۰	۹۲۱۰۰	۶۷۰۰۰	فراوانی	جمع کل	
%۱۰۰	%۳۷/۴	%۳۶/۲	%۲۶/۴	درصد		

بیشتر برای سرعت در کار ماشین‌نویسی، "تشدید" نوشته نمی‌شود. این امر علاوه بر این که در هم‌نگاشت‌ها تولید اشکال می‌کند، طبق جدول ۱۶ موجب بازیابی نشدن صفحه‌های اینترنتی به نسبت یکسان با کلیدواژه تشدیددار می‌شود.

سؤال ۳. آیا رابطه معناداری بین شکل واژه‌ها و نوع ابزار جستجو وجود دارد؟

میزان χ^2 مشاهده شده در سطح معناداری " $\alpha=۰/۰۵$ " نشان می‌دهد که بین شکل واژه و ابزار جستجو رابطه معناداری وجود دارد. بنابراین می‌توان نتیجه گرفت که به کار بردن یک شکل خاص از کلیدواژه و نیز استفاده از یک ابزار جستجوی خاص، در بازیابی اطلاعات اثرگذار است. برای نمونه جستجوی واژه "مسئله" در موتور کاوش "آلتاویستا" بازیافت بیشتری نسبت به جستجوی واژه "مسأله" دارد، اما جستجو در موتور جستجوی یاهو به طور کامل نتیجه‌ای متفاوت به دست می‌دهد.

همچنین مقدار ضریب Phi به دست آمده بیانگر میزان ارتباط بین دو متغیر است. ضریب "فی"، معیاری برای ارزیابی هم‌آیندی بین دو متغیر است. دامنه مقادیر این معیار که جهت هم‌آیندی را هم نشان می‌دهد، از -۱ تا +۱ است. از این معیار می‌توان در ارزیابی هم‌آیندی میان متغیرهای اسمی چندین

مقوله‌ای استفاده کرد (کورتز، ۱۳۷۴: ۳۴۵).

جدول ۱۷: آزمون خی دو برای بررسی رابطه معناداری بین شکل واژه و ابزار جستجو

متغیر	درجه آزادی	χ^2	سطح معناداری	ضریب فی
پتاسیم/پتاسیوم	۲	۱۰۹۸۴۶۵	۰/۰۰۱	۰/۰۸۴
زبان‌شناس/ زبان شناس/ زبان‌شناس	۳	۲۰۴۳۵۸۶۴	۰/۰۰۱	۰/۵۰۲
دقیقاً/ دقیق	۲	۲۳۸۰۶۷۱۰	۰/۰۰۱	۰/۱۰۰
شمشیر باز/ شمشیر باز	۲	۲۵۹۹۵۴۸۳	۰/۰۰۱	۰/۱۷۵
پرتو آفتاب/ پرتوی آفتاب	۲	۱۲۱۵۶۳۹۳	۰/۰۰۱	۰/۲۲۴
باغ‌ها/ باغها	۲	۱۱۲۴۰۹۲۰	۰/۰۰۱	۰/۰۶۵
خانه من/ خانه من/ خانه‌ی من	۳	۷۹۰۲۹۲۵	۰/۰۰۱	۰/۵۰۲
املا/ املاء	۲	۱۴۰۴۴۰۸۴	۰/۰۰۱	۰/۱۳۷
عطایی/ عطائی	۲	۳۳۰۸۷۲۶۰	۰/۰۰۱	۰/۲۵۴
مسئول/ مسئول	۲	۱۹۴۵۴۷۰۹۵۰	۰/۰۰۱	۰/۴۴۸
مسئله/ مسأله	۲	۱۹۴۵۴۷۰۹۵۰	۰/۰۰۱	۰/۳۶۳
موسی/ موسی/ موسا	۳	۱۰۶۸۹۹۷۰۴	۰/۰۰۱	۰/۱۱۰
اتاق/ اطاق	۲	۳۷۵۷۱۰۴	۰/۰۰۱	۰/۰۳۰
فرایند/ فرآیند	۲	۵۷۲۸۵۱	۰/۰۰۱	۰/۰۱۵
مورچه‌گان/ مورچگان	۲	۱۶۶۹۸۶۱	۰/۰۰۱	۰/۳۲۰
تورات/ توراة	۲	۳۴۴۴۰۱۴۸	۰/۰۰۱	۰/۳۸۷
موحدی/ موحدی	۲	۴۸۵۹۶۳۴	۰/۰۰۱	۰/۱۳۸

بحث و نتیجه گیری

۱. شیوه نگارش فارسی باعث بروز چالش‌های جدی در امر نمایه سازی این زبان شده است. در بسیاری موارد چند شکل نگارشی، برای یک واژه، درست شمرده شده است. این چندگونگی شکل واژه‌ها، برای رایانه قابل درک نیست. چراکه رایانه واژه‌ها را تنها به همان صورتی که ذخیره کرده است می‌شناسد و بازیابی می‌کند. بنابراین در مقابل سایر شکل‌های نوشتاری، آن را اصطلاح دیگری محسوب کرده و در هنگام جستجوی اطلاعات آن را بازیابی نمی‌کند.

۲. هیچکدام از موتورهای کاوش، چالش‌های شیوه‌های نگارش فارسی را به منظور بهبود نتیجه کاوش، مورد توجه قرار نداده‌اند. همان‌طور که سروکا (Sroka, 2000)، بارایلان و گتمان (Bar-Ilan & Gutman, 2002)، مقداد (Moukdad, 2005) و تاث (Toth, 2006) در بررسی‌های خود نشان دادند، موتورهای کاوش عمومی وب، ویژگی‌های زبانی زبان‌های غیر انگلیسی را در جستجوی اطلاعات نادیده می‌گیرند. این یافته‌ها نگران‌کننده است، زیرا این موتورهای کاوش (برای مثال "گوگل") در کشور ما، بسیار مورد توجه می‌باشند و بیشتر کاربران، از آنچه به هنگام جستجو در این ابزارها از دست می‌دهند، آگاهی ندارند.

۳. بین شکل واژه و ابزار جستجو رابطه معناداری وجود دارد.

بنابراین، به کار بردن یک شکل خاص از کلیدواژه و نیز استفاده از یک ابزار جستجوی خاص، در بازیابی اطلاعات اثرگذار است.

ابزارهای کاوش اینترنت، مهم‌ترین فناوری حاضر برای دسترسی به اطلاعات در محیط وب به شمار می‌آیند. در حال حاضر ابزارهای کاوش مختلفی در جهان ظهور پیدا کرده‌اند. لیکن ابزارهای جستجویی که امکان جستجوی اطلاعات به زبان فارسی را ارائه می‌دهند، محدود می‌باشند. از طرف دیگر، امکانات و قابلیت‌های آنها برای بازیابی کارآمد و مناسب اطلاعات متفاوت است. بی‌شک ایجاد یک ابزار کاوش قوی ملی، تحت نظارت سازمان‌های رایانه‌ای و انجمن‌های زبان‌شناسی و منطبق با نیازهای اطلاعاتی کاربران اینترنت در ایران می‌تواند میزان دقت در کاوش اطلاعات را بالا ببرد.

از طرفی به دلیل وجود مشکل‌هایی در خط فارسی، همواره برای استفاده از نرم‌افزارهایی چون OCR^۱ برای وارد کردن متن توسط پویشگر^۲ به رایانه و همچنین استفاده از واژه پردازها و ابزارهای مورد استفاده در آنها مانند غلط‌یاب‌های دستوری و املائی محدودیت‌هایی وجود دارد. پردازش، خلاصه‌سازی و بازیابی اطلاعات از محتوای متن، تحلیل آن و استفاده از نرم‌افزارهای تبدیل متن به گفتار و برعکس، همه از مواردی هستند که به دلیل محدودیت‌های خاص خط فارسی، استفاده از آنها به صورت کامل انجام نمی‌پذیرد؛ پایگاه‌های اطلاعاتی که با استفاده از شیوه خط کنونی به ذخیره و بازیابی اطلاعات می‌پردازند، نمی‌توانند کارایی مطلوب داشته باشند. همچنین عواملی که بدان اشاره شد، سبب کندی مراحل ذخیره و بازیابی اطلاعات شده و به نسبت بازیافت اطلاعات را کاهش می‌دهند. پایگاه‌های اطلاعاتی مدارک

^۱ Optimal Character Recognizer

^۲ Scanner

فارسی با وجود عمر کوتاه‌شان با مشکل‌های بسیاری درگیر هستند. در صورتی که برای این مشکل‌ها چاره‌ای اندیشیده نشود، با توجه به حجم فزاینده اطلاعات، مهار آنها آسان نخواهد بود.

پیشنادهایی برای رفع چالش

۱. بیشتر واژه‌پردازهای لاتین، به نظام غلط‌یاب املائی مجهز هستند که کار تصحیح متن الکترونیکی را به صورت خودکار انجام می‌دهد. پیشنهاد می‌شود روش و الگوریتم‌های غلط‌یاب فارسی در برنامه پژوهشی استادان زبان‌شناس فارسی قرار گیرد تا تولیدکنندگان نرم‌افزارهای واژه‌پرداز فارسی بتوانند با به کارگیری این روش‌ها، نرم‌افزارهای خود را به غلط‌یاب املائی خودکار مجهز کنند.

۲. در واژه‌پردازهای پیشرفته با توجه به اصول دستور زبان، امکان تصحیح دستوری متن الکترونیکی گنجانیده شود. برای دستیابی به نظام مشابه در زبان فارسی، لازم است استانداردهای نوشتار فارسی برای رایانه تدوین شود.

۳. نظام‌های بازیابی اطلاعات در زبان انگلیسی از امکانات ریشه‌سازی استفاده زیادی می‌کنند. در اینگونه نظام‌ها، با وارد کردن یک واژه به عنوان کلیدواژه، نظام به‌طور خودکار تمامی مشتقات واژه را نیز جستجو می‌کند. برای نمونه اگر جستجوی "کتاب" مدنظر ما باشد، واژه‌هایی نظیر "کتابخانه"، "کتابداری"، "کتاب فروش" و مانند آنها نیز بازیابی می‌شود. پورتر^۱ یکی از توانمندترین نظام‌های ریشه‌یابی در زبان انگلیسی است. این نظام بر پایه دسته‌بندی واژه‌ها به کمک واج‌ها و هجاها بنا نهاده شده است. برای ایجاد چنین نظامی در زبان فارسی، باید متخصصان زبان‌شناسی و رایانه همکاری نزدیکی با هم داشته باشند.

۴. در جهان کنونی بازنگری در شیوه نگارش فارسی را باید به شکل متفاوتی نسبت به گذشته انجام داد. اگر در گذشته ادیبان به تنهایی برای این امر تصمیم می‌گرفتند، هم اکنون تمامی کسانی که به نحوی با خط سر و کار دارند باید در تصمیم‌گیری دخالت داشته باشند. "فرهنگستان زبان و ادب فارسی"، می‌تواند افرادی شامل گروه‌های زیر را مأمور این کار کند:

- نویسندگان، شاعران، مترجمان، روزنامه‌نگاران و تمامی افرادی که به کار نوشتن مشغول هستند.

¹ Porter

- ویراستاران و نسخه‌پردازان؛ که تا کنون بار یکدست کردن شیوه نگارش متون بر دوش آنها بوده است.
- زبان‌شناسان؛ که پیشنهادها را با توجه به ساختار و کاربرد زبان، ارزیابی کنند.
- خوشنویسان و طراحان و به ویژه طراحان حروف؛ در طرح‌های موجود احساس می‌شود که توجه خاصی به زیبایی‌شناسی خط فارسی نشده است، در حالی که توجه به این امر، که بخشی از میراث فرهنگی و هویت ملی ما به شمار می‌آید کاری الزامی به نظر می‌رسد.
- برنامه‌نویسان رایانه؛ که راهگشای قابلیت‌های بی‌شمار رایانه برای فارسی‌زبان‌ها خواهند بود.
- کتابداران و اطلاع‌رسانان، که کار ترجمه نیازهای اطلاعاتی کاربران به زبان رایانه بر عهده آنها است و به طور مستقیم با این چالش‌ها سر و کار دارند.

کتابنامه

- اسلامی، محرم (۱۳۸۱). دشواری‌های پردازش رایانه‌ای خط فارسی. نشر دانش، ۱۹(۳).
- فرهنگستان زبان و ادب فارسی (۱۳۸۳). دستور خط فارسی. تهران: فرهنگستان زبان و ادب فارسی.
- حری، عباس (۱۳۷۲). کامپیوتر و شیوه نگارش فارسی، پیام کتابخانه، ۳(۱).
- راثی ساربانقلی، محمد صابر (۱۳۸۴). بررسی مشکلات جستجو و بازیابی اطلاعات به زبان فارسی از اینترنت با مطالعه موردی بر روی کاربران مرکز اینترنت دانشگاه آزاد اسلامی واحد شبستر. پایان نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی، دانشگاه آزاد اسلامی، واحد تهران شمال.
- سرمد، زهره (۱۳۷۶). روش‌های تحقیق در علوم رفتاری. تهران: آگه.
- صدیق بهزادی، ماندانا (۱۳۷۱). ناهماهنگی ضبط نام‌های بیگانه در فارسی. فرهنگ، ۱۳.
- صدیقی، محسن (۱۳۸۳). روشی برای رفع چالش‌های محتوا کاوی وب‌های فارسی. نما، ۴(۲). بازیابی در ۱۳۸۷/۱۲/۱۰ از http://www.irandoc.ac.ir/data/e_j/vol4/shahidi.htm
- کابلی، ایرج (۱۳۷۳). پیوسته‌نویسی، جدانویسی یا بی‌فاصله‌نویسی. آدینه، (۹۶).
- کورتز، نورمن (۱۳۷۴). مقدمه‌ای بر آمار در علوم اجتماعی. (ترجمه حبیب‌الله تیموری). تهران: نشر نی.

کوشا، کیوان (۱۳۸۱). معیارهای ارزیابی ابزارهای کاوش اینترنت: مطالعه مقایسه‌ای بر روی ابزارهای کاوش وب با واسط جستجوی فارسی. نشریه الکترونیکی کتابدار. بازیابی در ۱۳۸۷/۱۲/۱۰ از

<http://www.ketabdar.org/magazine/detailarticle.asp?number=25>

مرتضایی، لیلا (۱۳۷۶). مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. فصلنامه اطلاع رسانی. ۱۷(۱-۲).

منصورفر، کریم (۱۳۷۴). روش‌های آماری. تهران: دانشگاه تهران. موسسه چاپ و انتشارات.

یوسفان نجف آبادی، احمد (۱۳۸۲). یک سیستم بازیابی اطلاعات متنی برای زبان فارسی بر پایه نمایه‌گذاری معانی پنهان. پایان نامه کارشناسی ارشد مهندسی کامپیوتر، دانشگاه شیراز.

Bar-Ilan, J. & Gutman, T. (2002). How Do Search Engines Handle Non-English Queries?: A Case Study Retrieved February 2, 2009 from Sciencedirect database.

Hedlund, T et al. (2000). Aspects of Swedish Morphology and Semantics from the Perspective of Mono- and Cross-Language Information Retrieval. Retrieved January 23, 2009 from Elsevier database.

Monz, C. & De Rijke, M. (2002). Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian. In proceedings of *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation forum, CLEF 2001*. September 3-4 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.

Sroka, M. (2000). Web Search Engines for Polish Information Retrieval: Questions of Search Capabilities and Retrieval Performance. *Library Review*, (32), 87-98.

Toth, E. (2006). Exploring the Capabilities of English and Hungarian Search English for Various Queries. *Library Review*, (56), 38-47.