مروری بر دشواریهای زبان فارسی در محیط دیجیتال و تأثیرات آنها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات

دکتر هاجر ستوده[1]
زهره هنرجویان[2]

چکیده

هدفهای پژوهش: زبان فارسی، به سبب ویژگیهای خاصّ آن و در عین حال نهادینه نشدن سبک نگارش استاندارد، در رویارویی با محیطهای الکترونیکی، با دشواریهایی روبه‌روست که تأثیری بسزا بر اثربخشی بازیابی اطلاعات می‌گذارد. پژوهش حاضر می‌کوشد تا با بررسی متون و پیشینه‌های موجود، چالشهای نگارش فارسی، تأثیر آنها بر اثربخشی بازیابی اطلاعات، و پیشنهادهای ارائه شده در جهت رفع این دشواریها را مورد بحث و بررسی قرار دهد.

اهمیت پژوهش: با تحلیل و مرور جامع متونی که دربارهٔ چالشهای نگارش فارسی در محیطهای دیجیتال نگارش یافته است، می‌توان دانشی را که تاکنون در این باره گرد آمده است به تصویر کشید و کاستیها و پیشرفتهای به دست آمده در این زمینه را آشکار ساخت.

روش پژوهش: روش پژوهش حاضر، متن‌پژوهی با رویکرد تحلیل محتواست که از روشهای پژوهش کیفی به شمار می‌آید. «پاراگراف»، «جمله» و «کلمه» به عنوان واحد تحلیل انتخاب شد، زیرا ممکن بود هر دشواری یا راهکار تنها در یک کلمه یا عبارت مورد اشاره قرار گرفته یا در جمله یا پاراگراف شرح داده شده باشد.

یافته‌ها: آثار مورد بررسی، بیش از ۴۰ دشواری نگارشی را در رابطه با جستجو و بازیابی اطلاعات فارسی ذکر کرده‌اند. این گونه‌گونی نگارشی به نایکدستی و تطور بسیار در نگارش فارسی می‌انجامد که می‌تواند اثربخشی بازیابی را بویژه از منظر کاهش دقت یا ریزش کاذب و نیز کاهش جامعیت بازیابی، متأثر سازد. در نتیجه، ضروری است در طراحی الگوریتمهای سامانه‌های جستجو و بازیابی

---

۱. عضو هیئت علمی دانشگاه شیراز sotudeh@shirazu.ac.ir
۲. دانشجوی کارشناسی ارشد کتابداری و اطلاع‌رسانی دانشگاه شیراز
z.honarjooyan@gmail.com

فارسی، به‌هنجارسازی تنوعات و چنددستیهای نگارشی و دستوری مد نظر قرار گیرد. تدوین استاندارد نگارش فارسی، استفاده از سیاهه‌های از پیش تعیین شده، تجهیز پایگاه اطلاعاتی به اصطلاحنامه و فرهنگهای املایی، و تدوین دستنامه یا راهنمای جستجو، از جمله راهکارهای ارائه شده است. این راهکارها با وجود جامع نبودن، کم و بیش اثربخش به نظر می‌رسند.

نتیجه‌گیری: از آنجا که راهکارهای انسانی، نیازمند مشارکت فعّالانه و آموزش نویسندگان متون (تایپیستها و کاربران) است و از روندی بلندمدت و هزینه‌بر برخوردار است، حرکت به سوی راهکارهای خودکارسازی پردازش متن و نمایه‌سازی، ضروری است.

کلیدواژه‌ها: زبان فارسی، بازیابی اطلاعات، نگارش، املا.

مقدمه

خواندن و نگارش فارسی به دلیل ویژگیهای خاصّ این زبان، در پاره‌ای موارد با دشواریهایی همراه است که در رویارویی با رایانه، دو چندان می‌گردد. ورود ناگهانی رایانه به گستره‌ای وسیع از فعالیتهای مختلف اجتماعی، فرهنگی، اقتصادی و فنی، مجال آن را به صاحب نظران نداده است که راهکاری بنیانی و جامع برای مقابله با چالشهای شیوهٔ نگارش بیندیشند و به کار گیرند (حری،۱۳۷۲). نبود استاندارد شیوهٔ نگارش جامع و مورد قبول همگان، به نایکدستی و ناهماهنگی درونده‌ی اطلاعات در پایگاه‌های اطلاعاتی، وب‌سایتها، وب‌لاگها و دیگر منابع دیجیتالی انجامیده که آن نیز به نوبهٔ خود جستجوی فارسی را با مشکلاتی چند همراه ساخته است. این دشواریها بویژه در دنیای وب و با رشد سریع انتشارات الکترونیکی فارسی بر وب، چشمگیر بوده است. شیوه‌نامه‌ای که فرهنگستان ادب و زبان فارسی در سالهای اخیر برای یکدستی نگارش فارسی ارائه کرده نیز نتوانسته است از این دشواریها بکاهد، زیرا این شیوه‌نامه به دلیل ناهماهنگی درونی، هدف قرار دادن عامه مردم و در نتیجه کاهش دقت و پرهیز از وضع قانون برای برخی استثناها، وضع قانون برای پیوسته یا جدانویسی برخی کلمات مرکب و واگذار کردن سایر موارد به سلیقهٔ نویسندگان و در نهایت نپرداختن به همهٔ دشواریهای نگارشی، مورد انتقاد بوده است (طرح جامع پیکره زبان...، ۱۳۸۸؛ فرهنگستان زبان و ادب فارسی، ۱۳۸۳؛ سرمستانی، ۱۳۸۸؛ اشرف‌زاده، ۱۳۸۱). از سوی دیگر، الزام‌آور نبودن به کارگیری این دستورها باعث می‌شود پذیرش و نهادینه شدن

این سبک، فرایندی بسیار بلندمدت، اگر نگوییم ناشدنی، باشد.

مسئلهٔ پژوهش

دسترسی آسان به انبوهی از اطلاعات، دستاورد حضور اطلاعات در محیطهای الکترونیکی بخصوص وب است. در کنار این مزیّت، مسئله بازیابی اثربخش اطلاعات رخ می‌نماید. اثربخشی بازیابی زمانی حاصل می‌شود که نیاز کاربر هرچه بیشتر و بهتر برآورده گردد؛ بدین معنا که شمار بیشتری از مدارک با درجه ربطِ هرچه بیشتر با موضوعِ مورد نظر وی بازیابی گردد. اهمیت این مسئله زمانی که اطلاعات به زبانی چون فارسی مورد نیاز باشد، دوچندان می‌گردد. زیرا شیوهٔ نگارش زبان فارسی، به سبب ویژگیهای خاصّ آن و در عین حال نداشتن سبکی استاندارد، در رویارویی با محیطهای الکترونیکی، با دشواریهایی روبه‌روست که تأثیری بسزا بر اثربخشی بازیابی اطلاعات می‌گذارد.

به طور کلی، مطالعات در این حوزه بر سه محور کلی متمرکز است: ۱) آزمایش تأثیر تکنیکها یا ابزارهای خاصّ بر اثربخشی بازیابی ۲) طراحی و آزمایش تکنیکها، الگوریتمها یا ابزارهای خاص ۳) بررسی دشواریهای نگارش فارسی و تأثیر آنها بر اثربخشی بازیابی اطلاعات. آخرین محور، در دو دستهٔ تحقیقاتی و نظری مد نظر قرار گرفته است. در این میان، مطالعات نظری از اهمیتی بنیادین برخوردارند، زیرا شناسایی دشواریهای نگارش فارسی در مطالعات بازیابی اطلاعات عمدتاً بر پایهٔ آرا و نظریات صاحب نظران در این گونه تحقیقات بنیان می‌شود. از این رو، موفقیت طراحی الگوریتمها و سامانه‌های بازیابی اطلاعات فارسی در لحاظ کردن همهٔ قواعد زبانشناختی و نگارشی، به جامعیت و قوّت اعتبار این آثار بستگی خواهد داشت. تحلیل و مرور جامع متونی که در این باره به رشته تحریر درآمده است، ضمن ارائه اطلاعات درباره دیدگاه صاحب‌نظران این حوزه، دانشی را که تاکنون در این باره گرد آمده است به تصویر می‌کشد و نقاط تاریک و روشن آن را آشکار می‌کند و طراحان سامانه‌ها و پایگاه‌های اطلاعات فارسی را با مقتضیات جستجو و بازیابی به این زبان آشناتر

می‌سازد. با توجه به اهمیت این امر، بررسی حاضر که به روش متن‌پژوهی انجام می‌گیرد، می‌کوشد تا با مرور آثار و پژوهشهای پیشین، دشواریهای نگارش فارسی را شناسایی کند و تأثیر این دشواریها را بر بازیابی مؤثر اطلاعات بسنجد. در پایان نیز پیشنهادهای ارائه شده برای رفع این دشواریها را مورد بحث و بررسی قرار می‌دهد.

هدفهای پژوهش

پژوهش حاضر می‌کوشد تا هدفهای زیر را محقق سازد:

۱- شناسایی دشواریهای زبان فارسی در ذخیره و بازیابی اطلاعات در محیطهای دیجیتالی

۲- بررسی میزان اهمیت دشواریهای زبان فارسی به لحاظ فراوانی آنها در ادبیات مربوط

۳- شناسایی راهکارهای ارائه شده به منظور کاهش یا رفع این دشواریها

۴- تحلیل میزان اثربخشی راهکارهای ارائه شده در پژوهشهای مورد بررسی.

روش پژوهش

روش پژوهش حاضر، متن‌پژوهی با رویکرد تحلیل محتواست. برای یافتن آثار پیرامون دشواریهای ذخیره و بازیابی اطلاعات به زبان فارسی در محیطهای دیجیتالی، در تاریخ ۲۰ اسفند ۱۳۸۹ جستجویی در منابع کتابخانه‌ای، پایگاه‌های اطلاعاتی و نیز منابع وبی صورت گرفت. منابع شناسایی شده، پس از بررسی اولیه به جهت اطمینان از ربط با مسئلهٔ در دست مطالعه، به منظور تحلیل محتوا مورد مطالعه قرار گرفت. از آنجا که ممکن بود هر دشواری یا راهکار تنها مورد اشاره قرار گرفته یا در جمله یا پاراگراف شرح داده شده باشد، پاراگراف، جمله و کلمه به عنوان واحد تحلیل انتخاب شد.

روش گردآوری اطلاعات

به منظور شناسایی پژوهشهای انجام شده در زمینهٔ دشواریهای ذخیره و بازیابی اطلاعات به زبان فارسی، راهبردهای جستجویی متشکل از سه گروه اصطلاحات ناظر

بر دشواریها، ذخیره و بازیابی اطلاعات و زبان فارسی تدوین شد: ۱) «مشکلات»، «دشواری»، «سختیها»، «مسائل»، و «چالشها»؛ ۲) «ذخیره اطلاعات»، «ذخیره‌سازی اطلاعات»، «بازیابی اطلاعات»، «سازماندهی اطلاعات» و ۳) «فارسی». به منظور شناسایی جامع آثار، از فهرست منابع در پایان آثار نیز استفاده شد. در نهایت، آثار بسیاری به زبان فارسی و انگلیسی شناسایی و متن کامل آنها تحلیل شد. بررسی این منابع نشان داد تنها ۱۶ اثر به طور بنیادین مشکلات نگارش فارسی را به طور ویژه از منظر ذخیره و بازیابی در محیط دیجیتالی مد نظر قرار داده‌اند. لازم به ذکر است، شماری از پژوهشها بر معایب نگارش فارسی به طور مطلق متمرکز شده‌اند و به هدف بررسی دشواریها از منظر بازیابی اطلاعات به رشته تحریر درنیامده‌اند (برای نمونه، نگاه کنید به ۶ –۱۹). در پژوهش حاضر، این گونه آثار مد نظر قرار نگرفت.

پیشینهٔ پژوهش

پژوهشها پیرامون زبان فارسی

چنان که بیان شد، پژوهشهای بسیاری در زمینهٔ بازیابی اطلاعات در زبان فارسی انجام شده است که شمار اندکی از آنها به طور بنیادین و جامع به بررسی چالشهای نگارش فارسی پرداخته‌اند. با توجه به آنکه این دسته آثار در بخش یافته‌ها معرفی خواهند شد، از مرور آنها در این بخش خودداری می‌شود.

«سمایی» (۱۳۷۹) به بررسی حالات مفرد و جمع در زبان فارسی پرداخت. «راثی ساربانقلی» (۱۳۸۴) با بررسی مشکلات جستجو و بازیابی اطلاعات فارسی در اینترنت در یکی از واحدهای دانشگاه آزاد نشان داد کاربران به شکلهای مختلف نوشتاری توجهی ندارند و از عملگر«OR» استفاده نمی‌کنند. «عبداللهی» (۱۳۸۶) با بررسی چالشهای ریخت‌شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای گوگل، یاهو، و آلتاویستا نشان داد هیچ یک از جستجوگرهای مذکور، چالشهای زبان‌شناختی فارسی را به منظور بهبود کاوش مورد توجه قرار نداده‌اند. در نهایت، الگویی برای ایجاد اصلاحات در شیوهٔ نگارش فارسی ارائه شد. «گل‌تاجی و بذرگر» (۱۳۸۹) با

بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی‌ایران و جهاد دانشگاهی، نشان دادند چالشهای ریختی شناخته شدهٔ زبان فارسی، تأثیر بسیاری بر بازیابی اطلاعات در هر یک از سه پایگاه مورد نظر دارد. همچنین، هیچ یک از این سه پایگاه به شیوه‌ای جامع و قابل ملاحظه به حل مسائل ریخت‌شناسی واژگان فارسی نپرداخته‌اند.

پژوهشهای بسیاری به طراحی و/ یا آزمایش تکنیکها و راهکارهایی برای خودکارسازی و بهبود اثربخشی بازیابی فارسی پرداخته‌اند. برای نمونه، «یوسفان و همکاران» از طریق تحلیل ریخت‌شناختی، به ریشه‌یابی برخی واژه‌های عربی در زبان فارسی پرداخته‌اند (یوسفان و همکاران، ۲۰۱۰). برخی دیگر ریشه‌یاب فارسی طراحی کرده‌اند (تقوا و همکاران، ۲۰۰۵؛ مهراد و برنجیان، ۲۰۱۱؛ تشکری و همکاران، ۲۰۰۲؛ موسوی میانگاه، ۲۰۰۶؛ برنجکوب و همکاران، ۲۰۰۹). «موسوی میانگاه» (۲۰۰۷) کوشیده است راهکاری برای مشکل تکثّر معانی واژگان فارسی به هنگام ترجمهٔ ماشینی ارائه دهد. در پژوهشهای دیگر، ساخت هستی‌شناسی فارسی[1] به روشهای مختلف از جمله بازمهندسی اصطلاحنامه و بر پایهٔ ویکی‌پدیا به منظور افزایش دقت بازیابی بررسی و/یا آزمایش شده است (شهیدی و همکاران، ۱۳۸۴؛ خسروی و وظیفه‌دوست، ۱۳۸۶؛ فرهودی و همکاران، ۲۰۰۹). «کیوان و همکاران» (۲۰۰۶) برنامهٔ پرشیانت را برای کاربردهای مختلف در پردازش زبان طبیعی فارسی طراحی کردند. «ایرانپور مبارکه و مینایی بیدگلی» (۲۰۰۹) تکنیکی جدید را برای ریشه‌یابی افعال در متون فارسی ارائه کرده‌اند که می‌تواند در پیش‌پردازش زبانشناختی و متن‌کاوی از جمله به منظور برچسب‌زنی ادات سخن و تشخیص مرز جملات به کار رود. در پژوهشی دیگر، روشی برای نمایه‌سازی چندنویسه‌ای[2] متون فارسی پیشنهاد شده است (دانش و همکاران، ۲۰۱۱). کارآیی فنون پردازش زبان طبیعی در بازیابی چندزبانه در تحقیقی

1. Persian Ontology.
2. N-gram indexing.

دیگر مورد آزمایش قرار گرفت (علیزاده و فتاحی، ۲۰۱۰). همچنین، نشان داده شد که برچسب‌گذاری ادات سخن تنها زمانی تأثیری قابل توجه بر اثربخشی بازیابی دارد که با ریشه‌یابی همراه باشد (کریم‌پور و همکاران، ۲۰۰۹). راهکارهایی اکتشافی برای بهبود صحت نتایج برچسب‌زنی ادات سخن نیز آزمایش شده و تأثیر مثبت آنها بویژه برای واژه‌های ناشناخته تأیید شده است (محترمی و همکاران، ۲۰۰۸). آزمایش تکنیکهای نمایه‌سازی چندنویسه‌ای و گسترش پرسش نشان داده است تکنیک نمایه‌سازی چهارنویسه‌ای مبتنی بر مدل فضای برداری، نتیجه‌ای قابل قبول و تکنیک گسترش پرسش «تحلیل محتوای محلی»[1] بهترین نتیجه را برای بازیابی فارسی به همراه خواهد داشت (آل احمد و همکاران، ۲۰۰۷). در پژوهشی دیگر، راهکارهایی برای ریشه‌یابی زبان فارسی و همچنین بومی‌سازی یا سفارشی کردن بخشهایی از موتور جستجو که متأثر از ساختار زبان است، ارائه شده است (طرح جامع پیکره زبان...، ۱۳۸۸). «گزنی» (۱۳۸۵) سامانه‌ای را برای استخراج خودکار عبارتهای کلیدی از متون فارسی به منظور به‌کارگیری در طراحی سامانه‌های بازیابی طراحی کرده است. در پژوهشی دیگر، تکنیکی برای انتخاب مفهوم درست اصطلاحات پرسش در بازیابی انگلیسی– فارسی پیشنهاد شده است که در آن احتمالات ترجمه بر پایهٔ گرافهای مفاهیم اصطلاحات پرسش محاسبه می‌شود (تیموریان و همکاران، ۲۰۰۹). ساخت پیکرهٔ متن[2] فارسی برای به کارگیری در پژوهشهای بازیابی اطلاعات، در کانون توجه دسته‌ای از پژوهشها بوده است. برای نمونه، ساخت پیکرهٔ موضوعی فارسی، مجموعهٔ افعال فارسی و پیکرهٔ متن استاندارد «همشهری» را می‌توان نام برد (خلیفه سلطان و همکاران، ۲۰۱۰الف؛ ۲۰۱۰ب؛ آل احمد و همکاران، ۲۰۰۹).

به طور کلی، مرور پژوهشها نشان می‌دهد پیشرفتها و دستاوردهای بزرگی در حوزهٔ بازیابی اطلاعات فارسی حاصل شده است. با این حال، شمار پژوهشهایی که به طور ویژه و بنیادین به مشکلات نگارش فارسی در الگوریتمهای بازیابی، خواه در پایگاه‌های

1. Local Context Analysis.
2. Persian Corpus.

اطلاعاتی تخصصی، خواه در موتورهای کاوش عمومی‌وب بپردازند، اندک است. از این رو، نیاز شدیدی برای انجام پژوهشهای بیشتر، به منظور شناسایی جامع‌تر و عمیق‌تر چالشهای نگارش فارسی، میزان تأثیر آنها بر اثربخشی بازیابی اطلاعات، ارزیابی راهکارهای پیشنهادی و ارائهٔ راهکارهای جدید، وجود دارد.

پژوهشها پیرامون دیگر زبانها

تحقیقات بسیاری دربارهٔ سبک نگارش و تأثیر آن بر بازیابی اطلاعات در زبانهای دیگر انجام شده است. برای نمونه، نشان داده شده است که برچسب‌زنی ادات سخن می‌تواند از ابهامهای لغوی کلمات هم‌نگاشت در زبان سوئدی بکاهد (هدلاند و همکاران، ۲۰۰). در پژوهشی دیگر، تأثیر مثبت تحلیلهای ریخت‌شناختی مانند ریشه‌سازی و جداسازی کلمات مرکب، بر نتایج بازیابی در زبانهای هلندی، آلمانی و ایتالیایی تأیید شده است (مونتس و دی ریژکه، ۲۰۰۲). همچنین، نتایج بازیابی در موتورهای کاوش عمومی‌که مسائل زبان‌شناختی و ریخت‌شناختی لهجه‌ها یا زبانهای غیرانگلیسی مانند روسی، فرانسوی، مجاری، عربی و عبری را لحاظ نمی‌کنند، مناسب نیست (بارایلان و گاتمن، ۲۰۰۲؛ مقداد، ۲۰۰۵؛ مقداد و لارج، ۲۰۰۱؛ مقداد و سویی، ۲۰۰۵). «لازارینیس و همکاران» (۲۰۰۹) با مروری بر آثار پیرامون دشواریهای بازیابی به زبانهای غیرانگلیسی، پرسشهای فرارو و راهکارهای ممکن برای رفع آنها و همچنین زمینه‌های پژوهشهای آینده را شرح داده‌اند.

جدول ۱. دشواریهای ذخیره و بازیابی رایانه‌ای به زبان فارسی و توزیع آنها در متون

ردیف	چالش	فراوانی متون	ردیف	چالش	فراوانی متون
۱	تشدید (معیّن/ معین)	۵	۲۳	گوناگونی معادلهای علمی	۲
۲	همزهٔ پایانی (املاء/ املا)	۳	۲۴	(عدم) استفاده از «ء» بعد از« های» بیان حرکت در حالت مضاف (خانهٔ مردم / خانه مردم)	۴

ردیف	چالش	فراوانی متون	ردیف	چالش	فراوانی متون
۳	تنوع شیوهٔ دگرنویسی (امریکا / آمریکا)	۷	۲۵	تنوع نگارش یای وحدت نکره بعد از «های» مختفی (خانه‌ایی / خانه‌یی/ خانهٔ)	۴
۴	های غیر ملفوظ (مورچگان/مورچه‌گان)	۲	۲۶	عدم تمایز حروف بزرگ و کوچک در ابتدای جمله	۱
۵	همزهٔ متصل به «یای» وحدت (عطایی/ عطائی)	۳	۲۷	شباهت اعداد (صفر و نقطه / ۱ و ۲ و ۳)	۳
۶	استفاده از «آ» و «ا» به جای هم (درآمد/ درامد)	۵	۲۸	تعدد حروف دندانه‌دار (پیشینیان)	۴
۷	تنوع حروف (اطاق/ اتاق)	۶	۲۹	تعدد نقطه‌های حروف (ث ش پ)	۵
۸	الف کوتاه (تقوی/ تقوا)	۷	۳۰	شباهت شکل حروف (ک گ / ت ث / ر ز)	۵
۹	تای نقطه‌دار (مشکوه/ مشکات / مشکوة)	۳	۳۱	ناتوانی در نشان دادن تلفظهای باستانی و میانه، گویشها و لهجه‌ها	۲
۱۰	«ی» صامت میانجی (پرتوی آفتاب/ پرتو آفتاب)	۲	۳۲	یکسانی نشانهٔ واژهٔ بستهای ربطی فعل «بودن» و «م» مالکیت (پدرم =پدر من / پدر هستم)	۱
۱۱	خط تیره (اقتصادی اجتماعی/ اقتصادی-اجتماعی)	۱	۳۳	یکسانی علامت نکره و اسم ساز و صفت ساز (اجتماعی: اجتماع+ی نکره؛ اجتماعی بودن)	۱
۱۲	نقطه در سرنامها (اچ. آی. وی/ اچ‌آی‌وی)	۱	۳۴	آرایش آزاد سازه‌های جمله (دیروز من کتاب خریدم/ من دیروز کتاب خریدم)	۱
۱۳	پیوسته‌نویسی (سرهم یا با نیم‌فاصله) یا جدا نویسی (کتاب شناسی / کتابشناسی/ کتاب‌شناسی)	۱۳	۳۵	فقدان پایانه‌های تصریفی نمایانگر حالت کلمه در جمله (این کار- خانه را خراب کرد. این کارخانه- را خراب کرد. این- کارخانه را خراب کرد.)	۱
۱۴	تنوع نشانه‌های جمع (عاقلان/ عقلا / عاقلها)	۸	۳۶	اختیاری بودن فاعل ([علی] به مدرسه رفت)	۱
۱۵	تنوین (واقعا/واقعاً/ واقعن)	۴	۳۷	اشتقاق صفر و تغییر مقولهٔ واژگانی کلمه‌ها (انتخابها در شرایطی بد بود/ بد و خوب را تشخیص داد.)	۱

ردیف	چالش	فراوانی متون	ردیف	چالش	فراوانی متون
۱۶	فاصلهٔ بین حروف یک واژه به اشتباه یا به عمد (دوا زده/ دوازده؛ کدگذاری/کـــــد گذاری)	۷	۳۸	واژه‌های به وام گرفته یا ترجمه شده (کامپیوتر/ رایانه)	۱
۱۷	املاهای مختلف همزه (مسئول/ مسؤول)	۶	۳۹	مترادف‌ها (درست/ صحیح)	۱
۱۸	تفاوت در آوا / اعراب (مَرد/ مُرد، دیر (زمان) / دیر [صومعه])	۸	۴۰	اسامی‌عامیانه، تجاری، مشهور یا علمی	۱
۱۹	تعدد شکلهای یک حرف (عـ ـعـ ـع ع)	۸	۴۱	کسرهٔ اضافه (پدر او را تحسین کرد/ پدرِ او را تحسین کرد)	۳
۲۰	یکسانی تلفظ برخی حروف (س ص ث)	۶	۴۲	آوانویسی به جای ترجمه (سورس/ منبع)	۱
۲۱	نوشتن «ک» و «گ» با سرکش و بی آن (ک/ ک)	۳	۴۳	همنام‌ها و هم‌آواها شیر (ماده نوشیدنی، حیوان، ابزار)	۱
۲۲	نگارش از راست به چپ	۷			

یافته‌های پژوهش

چالشهای نگارش فارسی در محیط دیجیتال

جدول ۱ مشکلات نگارش زبان فارسی در بازیابی اطلاعات و همچنین شمار پژوهشهایی را که به هر مشکل پرداخته‌اند، برای درک بهتر اهمیت هریک از دیدگاه نویسندگان، گرد آورده است. آثار مورد بررسی روی هم رفته ۴۳ چالش نگارشی را نام برده‌اند. چالشهای دیگری نیز ذکر شده بود که به نظر نمی‌رسد تأثیر مستقیم بر بازیابی اطلاعات داشته باشد. مانند وجود «و» ناخواندنی در کلماتی چون «خواهش»، «خواندن» یا تلفظهای مختلف یک حرف (مثل خوش/ او / والی). این دو ویژگی تنها در

برنامه‌های تشخیص و پردازش صوت[1] یا در صورت ضعف یا خطای املایی تایپیست یا کاربر می‌تواند اثرگذار باشد. چنان‌که از فراوانی متون مورد بررسی برمی‌آید، آنها بیش از همه به مسئله «پیوسته‌نویسی، یا جدانویسی» پرداخته‌اند. پس از آن، «تنوع نشانه‌های جمع» (، «تفاوت در آوا / اعراب‌گذاری»، «تنوع دگرنوشته‌ها»، «الف کوتاه»، «فاصلهٔ بین حروف واژه»، و «نگارش از راست به چپ» فراوانی بالایی دارند [برای نمونه نگاه کنید به حری، ۱۳۷۲؛ راثی ساربانقلی، ۱۳۸۴الف؛ ۱۳۸۴ب؛ عبداللهی نورعلی، ۱۳۸۶؛ گل تاجی و بذرگر، ۱۳۸۹، محقق زاده و زارعیان، ۱۳۸۳؛ اسلامی، ۱۳۸۱؛ مرتضایی، ۱۳۸۱؛ جرات و سمایی، ۱۳۸۳؛ معصومی همدانی، ۱۳۸۱؛ صدیق بهزادی، ۱۳۷۷؛ حسینی بهشتی، ۱۳۸۲؛ مرعشی، ۱۳۸۳).

نوع و خاستگاه چالشها

چالشهای برشمرده در متون را می‌توان به سه سطح معنایی، نحوی، و ریخت‌شناختی تقسیم کرد. برخی مشکلات، بیش از آنکه به ریخت‌شناسی فارسی بازگردند، به دستور زبان فارسی مربوط می‌شوند. برای نمونه، اختیاری بودن فاعل در جمله‌های فارسی یا آرایش آزاد سازه‌های جمله. چالشهای معنایی را می‌توان به تنوع واژگان و غنای زبان و همچنین وابستگی به زبانهای بیگانه نسبت داد. تنوع در کاربرد واژه می‌تواند بر اثربخشی بازیابی اطلاعات تأثیر گذارد. برای نمونه، واژه‌های وام گرفته، مترادفها، آوانویسی واژه‌های خارجی به جای ترجمهٔ آنها و چنددستی در نگارش یا تنوع واژگان به کار گرفته برای تبیین اسامی‌مشهور یا علمی، می‌تواند جامعیت جستجو را بویژه در محیطهای وبی که امکان استفاده از اصطلاحنامه یا دیگر ابزارهای مهار واژگان وجود ندارد، تحت تأثیر قرار دهد. آشکار است که این مشکلات، مختص زبان فارسی نیست، با این حال، به دلیل وابستگی زیاد زبان فارسی به زبانهای خارجی و نبود استانداردی برای آوانویسی واژگان خارجی، به نظر می‌رسد این مسئله

1. Speech recognition.

بازیابی فارسی را به شدت با دشواری روبرو سازد. به منظور کاهش تأثیر این عوامل، کاربر باید به هنگام جستجو، واژگان را با همهٔ تنوع آنها مد نظر داشته باشد تا بتواند در پیوندی انفصالی، آنها را در یک راهبرد جستجو کند و بدین ترتیب، تا جایی که ممکن است به جامعیت بیشتر نزدیک شود.

دسته‌ای دیگر از چالشها به تنوع ریخت‌شناختی نگارش فارسی باز می‌گردد که به نگارش، عدم نگارش یا تنوع در نگارش حروف، علایم یا اعراب منجر می‌شود (مانند همزهٔ پایانی یا میانی، «های» غیرملفوظ، «ی» ك پیش از «یای» وحدت، الف (کوتاه یا بلند)، تای نقطه‌دار، «ی» صامت میانجی، خط تیره، نقطه، فاصله یا نیم فاصله). به نظر می‌رسد حدس زدن و اعمال تمامی‌این جزئیات برای کاربر در راهبرد جستجو دشوار باشد. با این حال، با توجه به قاعده‌مندی بسیاری از این ریختها، می‌توان در الگوریتم جستجو، واژه‌ها را به نحوی بهنجار[1] کرد که واژه صرف نظر از ریختهای مختلف آن، بازیابی شود. تنوع فونتها بویژه تفاوت بین فونتهای قدیمی‌و جدید (با نگارش فارسی و عربی) به دسته‌ای دیگر از مشکلات دامن می‌زند که به ظاهر به ریخت‌شناسی کلمه باز می‌گردد، اما در واقع به تفاوت نویسه‌های فارسی و عربی مربوط می‌شود (مثل عربی (بی‌سرکش) و ک فارسی (با سرکش) یا ی فارسی و عربی).

بدین ترتیب، مشاهده می‌شود که برخی چالشها، مانند تنوع مترادفها و املاهای واژگان، ذاتی هر زبانی است، اما برخی مانند حذف یا درج حرف همزه یا «ی»به سرشت زبان فارسی یا استاندارد نبودن نگارش آن باز می‌گردد. همچنین، ریشهٔ بروز این چالشها را می‌توان در مراحل مختلف چرخهٔ حیات یک مدرک علمی‌از مرحلهٔ تایپ متن به هنگام تولید مدرک، تا آخرین مرحله که دروندهی عبارت جستجوست، یافت. نبود استاندارد نگارش فارسی و در نتیجه سلیقه‌ای عمل کردن نویسندگان یا تایپیستها، نبود صفحه‌کلید و کدهای استاندارد، عادت به آسان‌نویسی و رعایت نکردن پیچیدگیهای نگارش به هنگام تایپ می‌تواند به چنددستی در نگارش واژگان

1. Normalize.

نویسندگان، تایپیستها، نمایه‌سازان و کاربران منجر شود (عبداللهی نورعلی، ۱۳۸۶؛ محقق‌زاده و زارعیان، ۱۳۸۳).

نوع تأثیر و فعالیت متأثر از چالشها

آشکار است که بدون بهنجارسازی چالشهای نگارشی و دستوری در الگوریتمهای سامانه‌های جستجو و بازیابی فارسی، اثربخشی بازیابی مطلوب نخواهد بود. در بسیاری از موارد، انتخاب یکی از صورتهای نگارشی و نادیده گرفتن دیگری، سبب کاهش بازیافت می‌شود. افزون بر این، گاهی چنددستی در شیوهٔ نگارش، به ریزش کاذب نیز منجر می‌شود. برای مثال، جستجو به دنبال واژهٔ «معین» بدون تشدید، نه تنها به از دست رفتن مدارکی حاوی این واژه با نگارش تشدیددار، بلکه به بازیابی مدارک حاوی واژه «معین» (به معنی «کمکی») منجر می‌شود. به عنوان نمونه‌ای دیگر، بی‌دقتی در فاصله‌گذاری بین کلمات مرکب حاوی حروف ناچسبان (مثل کدگذاری)، می‌تواند به انفصال یا اتصال کاذب، تغییر معنی (حری،۱۳۷۲) و در نهایت ریزش کاذب بینجامد. از این گذشته، اصل چسبیده‌نویسی حروف در فارسی، که بر خلاف لاتین جدا جدا نوشته نمی‌شوند، تشخیص مرز بین حروف را دشوار می‌سازد. این خود می‌تواند دقت تایپیست یا جستجوگر به هنگام ورود داده را کاهش دهد یا به بروز خطاهای مکرر در نرم‌افزارهای تشخیص نوری نویسه، منجر شود.

به همین ترتیب، شیوهٔ اعراب‌گذاری می‌تواند به بازیابی واژه‌هایی با املای مشابه اما آوای متفاوت و در نتیجه ریزش کاذب منجر شود. این امر می‌تواند امکان بهنجارسازی اعراب و علایم در الگوریتمهای جستجو را نیز محدود سازد. علاوه بر این، ناتوانی خط فارسی در نشان دادن تلفظ واژه‌های ایران باستان و میانه و نیز گویشها و لهجه‌ها، حتی با نشانه‌ها، کاهش بازیافت اطلاعات را در پی‌خواهد داشت. همچنین، ممکن نبودن تمایز بین اسم خاص و عام در زبان فارسی، برای مثال نبود حروف دوگانهٔ بزرگ و کوچک، می‌تواند به‌ریزش کاذب بینجامد. برای نمونه، در جستجوی «حافظ» (شاعر قرن هشتم هجری) که یک اسم خاص است، همه مدارکی که واژهٔ «حافظ» به معنای

عام در آنها وجود دارد نیز بازیابی می‌شوند. آشکار است که با شیوهٔ کنونی نگارش متن، نمی‌توان الگوریتمهای جستجو را به نحوی طراحی کرد که با تمایز خودکار بین اسامی‌خاص و عام، دقت جستجو را افزایش دهند.

هر یک از این چالشها، بسته به فراوانی رویداد آنها ـ در متن یا در عبارت جستجو ـ نتایج بازیابی را با درجات متفاوتی متأثر می‌سازند. برای نمونه، همان‌گونه که «مانینگ و همکاران»[1] می‌نویسند، بسیاری از کاربران پرسشها را بدون علایم آوایی می‌نویسند. این کار برای بالا بردن سرعت، از روی تنبلی یا محدودیت نرم‌افزاری، یا به دلیل عادتهایی بازمانده از روزگار گذشته که استفاده از متن غیر اسکی در بسیاری از نظامهای رایانه‌ای دشوار بود، صورت می‌گیرد (مانینگ، راگاوان و شوتس، ۲۰۰۸). از این رو، احتمال می‌رود علایمی‌مانند اعراب‌گذاری، همزهٔ پایانی و تشدید، در هر دو دستهٔ کاربران و تایپیستها، به یک اندازه نادیده گرفته شود. در نتیجه، این موارد در مقایسه با تنوع در املا، همزهٔ میانی، پیوسته یا جدانویسی واژه‌های مرکب، گوناگونی برابرنهادهای علمی، و دگرنویسی مشکل کمتری را به لحاظ جامعیت بازیابی پیش می‌آورند، با این حال، در مواردی ریزش کاذب را افزایش می‌دهند.

همچنین، سطح تأثیر این چالشها به لحاظ عملیات و فعالیتهای مختلف، متفاوت است. برای نمونه، رعایت نکردن اعراب‌گذاری نه تنها در مرحلهٔ درونداد اطلاعات (به هنگام تولید مدرک یا جستجو) رخ می‌نماید و نتایج جستجو را متأثر می‌سازد، بلکه به هنگام پردازش خودکار نوشتار، بویژه در زمینهٔ بازسازی گفتار و ترجمهٔ ماشینی مشکلاتی را به همراه دارد. معلوم نیست برای یک صورت نوشتاری واحد، کدام زنجیرهٔ واجی را باید در نظر گرفت. همچنین، تشخیص تلفظ صحیح واژه برای برنامه‌های گویا دشوار خواهد بود. یا به عنوان نمونه‌ای دیگر، یکسانی تلفظ برخی حروف مانند «س»، «ث»، و «ص» باعث کُندی و پیچیدگی کار پردازش نوشتار می‌گردد، زیرا برنامهٔ پردازشگر نوشتار ناچار است دائم به واژگان مراجعه و برای هر کدام از واحدهای

1. Manning, Raghavan & Schuts.

نوشتار، یک صورت واجی از واژگان اخذ کند (اسلامی، ۱۳۸۱). یا بـه عنـوان نمونـه‌ای دیگر، آشکار است که وجود دندانه‌ها و نقطه‌های متعـدد، چسـبیدگی و شـباهت شـکل برخی حروف، ورود داده‌ها را به شکل دستی و خودکار دچار مشـکل مـی‌کنـد؛ بـدین ترتیب که دقت تایپیست یا کـاربر در ورود صـحیح امـلای واژه را کـاهش مـی‌دهـد و تشخیص نوری نویسه‌ها[1] را هم دشوار می‌سازد. این امر در مورد اعداد نیز صادق اسـت (مانند شباهت صفر و نقطه و همچنین ۱، ۲ و ۳) (راثی ساربانقلی، ۱۳۸۴الف).

افزون بر این، پردازش خودکار متن می‌تواند در اثر وجود چندین چیـدمان نویسـه‌ای در متن با دشواریهایی روبه رو شود. برای نمونه، بر خلاف متن فارسی کـه از راسـت بـه چـپ چیده می‌شود، متون ریاضی، شیمی، نت‌های موسـیقی، و دسـتورهای شـطرنج از چـپ بـه راست نوشته می‌شوند. از این رو، گاه در یک متن چندین بار جهت چیدمان نویسه‌هـا تغییـر می‌کند. نرم‌افزار پردازش خودکار ناچار است بارها جهت خوانـدن را از راسـت بـه چـپ و بالعکس تغییر دهد. آشکار است که در این میان امکان بروز خطـا بسـیار افـزایش مـی‌یابـد. علاوه بر این، یکسانی علامت نکره و اسم ساز و صفت‌ساز، یکسانی نشـانهٔ واژه‌بسـت‌هـای ربطی فعل«بودن» و «م» مالکیت، اختیاری بودن فاعل، نبود نشانهٔ نوشتاری برای کسرهٔ اضـافه و آرایش آزاد سازه‌های جمله باعث می‌شود تشخیص مرز و نقش گروه‌هـای نحـوی بـرای پردازش خودکار متن یا ترجمهٔ ماشینی با چـالش روبـه رو شـود (اسـلامی، ۱۳۸۱). آشـکار است، وقتی چند مورد از این چالشها در یک اصطلاح یا عبارت واحد روی دهد، اثربخشـی بازیابی کمتر شده و ضرورت تدوین راهبرد پیچیده‌ای برای جستجو بیشـتر و در عـین حـال انجام آن دشوارتر می‌شود. برای نمونه، در جستجو به دنبـال واژهٔ «دایـرةالمعـارف»، مسـتلزم پیوند انفصالی چندین املا در یک راهبرد واحد است تا جامعیت جستجو تضمین گـردد: ۱– سه شکل مختلف حرف «ی» عربی (با دو نقطه زیرین)، فارسی و ئ؛ ۲– دو شـکل مختلـف «ه» (تای گرد نقطه‌دار و بدون نقطه) ۳– گسسته نویسی و پیوسته‌نویسـی «ه» (بـی‌فاصـله، بـا فاصله یا نیم فاصله).

1. Optical Character Recognition (OCR).

جامعیت چالشهای معرفی شده در آثار

با نگاهی به آنچه تاکنون بیان شد، روشن می‌شود شمار بسیاری از چالشها در آثار مورد بررسی معرفی شده‌اند. با این حال، نمی‌توان نسبت به جامعیت آنها مطمئن بود، زیرا برخی چالشها در این متون نادیده گرفته شده یا به‌طور گذرا به آن پرداخته شده است. احتمال می‌رود با پژوهشهای زبانشناختی بیشتر بتوان به نمونه‌های دیگری نیز دست یافت. برای مثال، مسائلی چون«یکسانی علامت نکره و اسم ساز و صفت ساز» و یا «یکسانی نشانهٔ واژه بستهای ربطی فعل «بودن» و «م» مالکیت»، با وجود تأثیر بسزایی که می‌توانند در میزان موفقیت و ثمربخشی جستجو داشته باشند، کمتر مورد توجه بوده‌اند. همچنین، به برخی موارد در متون هیچ‌گونه اشاره‌ای نشده است:

۱. استفاده از مصوتهای کوتاه به جای مصوت بلند «و» یا «ا» (مانند کوه/که؛ گوهر/گهر؛ کاه/که)
۲. کاربرد دو مصوت کوتاه و بلند «ُ» و «و» به جای هم (مانند خرسند و خورسند؛ خرجین/ خورجین)
۳. یکسانی واژه‌بستهای ربطی فعل «بودن»و «ی» وحدت یا نکره (مانند «خانه‌ای»، که در آن «ای» می‌تواند نقش فعلی (در خانه هستی) یا نشانهٔ نکره (یک خانه) داشته باشد)
۴. تأثیر به‌کارگیری فونتهای قدیمی‌و جدید که ذاتی زبان فارسی نبوده، بلکه از پویایی و تنوع فناوری سرچشمه می‌گیرد، چندان مد نظر قرار نگرفته است. این چالش در بخش بعد به اختصار شرح داده خواهد شد.

نقش نوع فونت

نقش کدگذاری و نوع فونت، تنها در (طرح جامع) به طور گذرا مورد اشاره قرار گرفته است. این امر بویژه از آن رو اهمیت دارد که کاربر به دلیل شباهت نمایش این فونتها، متوجه تفاوت نویسه‌ای آنها با هم نیست. از این رو، احتمال این که به هنگام جستجو در پی لحاظ کردن هر دو نوع فونت باشد، بسیار اندک و در نتیجه احتمال از دست دادن منابع بسیار زیاد است. مثال بارزی در این باره، حرف «ی» است که به دو شیوه کدگذاری می‌شود. بسته به این که در صفحه کلید، کدام نوع فونت به عنوان

پیش‌گزیده به کار رفته باشد، دسته‌ای از منابع با فونتِ دیگر بازیابی نخواهند شد. دو چالش « ک در شکلهای مختلف»، و نیز «تای نقطه‌دار» که در متون به آنها اشاره شده است، می‌تواند ناشی از تنوع در فونتهای مورد استفاده در رایانه‌های مختلف باشد. تأثیر تفاوت فونت بر جامعیت نتایج را با جستجو در اینترنت می‌توان آشکارا دید. برای مثال، جستجو با حرف کاف (بدون سرکش) در گوگل به دنبال واژه «کودکان» به بازیابی ۵۴ میلیون و ۹۰۰ هزار پیشینه منجر شد. اما حاصل جستجو به دنبال همین واژه با کاف سرکش‌دار ۳۲ میلیون و ۷۰۰ هزار پیشینه بود که تفاوت چشمگیری را نشان می‌دهد. همچنین، جستجو به دنبال کلیدواژه «روانشناسی» با یای عربی (با دو نقطه در زیر) به بازیابی ۳۲۵ میلیون و با یای فارسی (بدون نقطه) به ۶۱۸ هزار پیشینه انجامید (جستجو به تاریخ ۲۵ بهمن ماه ۱۳۹۰). اگر کاربر این دو نوع حرف را با پیوند انفصالی جستجو نکند، بخش عمده‌ای از نتایج را از دست خواهد داد. البته، تدوین راهبرد جامع جستجو در چنین شرایطی بسیار دشوار خواهد بود، زیرا ممکن است فرد راهکار دسترسی به هر دو نوع فونت را نداند. نکتهٔ دیگر در مورد تفاوت صفحه کلیدها یا برنامه‌ها به لحاظ شیوهٔ تعریف یک نویسه است. برای نمونه، شیوهٔ اعمال نیم‌فاصله که برای پیشگیری از چسبیدن دو جزء یک واژه مرکب به هم اعمال می‌شود، در محیطهای مختلف با هم متفاوت است. در واژه‌پرداز وُرد[1]، نیم فاصله را می‌توان به دو شیوهٔ Shift+ Space و نیز (_) + Ctrl درج کرد. حال آنکه در رابط کاربر گوگل تنها شیوهٔ نخست اعمال می‌شود و شیوهٔ دوم با «فاصله» یکسان تلقی می‌شود. اگر کاربر از این تفاوتها آگاه نباشد، به سادگی می‌تواند بخشی از منابع را از دست بدهد.

راهکارهای ارائه شده در متون

هریک از پژوهشهای مورد بررسی برای رفع یا تقلیل این مشکلات نگارش فارسی در محیط دیجیتالی، راهکارهایی را ارائه نموده‌اند (جدول۲). برخی، راهکارهایی بنیانی

1. Word.

برای حلّ ریشه‌ای این مشکلات هستند و برخی ناظر بر یک یا چند مشکل نگارشی محدود. هر راهکار را می‌توان به یک یا چند مرحلهٔ خاص از چرخهٔ حیات مدرک یعنی پیش از بازیابی، و به هنگام بازیابی نسبت داد. دستهٔ اول، راهکارهایی است برای نویسندگان و تایپیستها به هنگام تولید مدرک یا ذخیره‌سازی آن. همچنین، این راهکار می‌تواند به هنگام نمایه‌سازی به منظور تولید بازنمونهای مدرک نیز به کار گرفته شود. بنابراین، مخاطب این راهکارها، گاه کاربران، گاه نمایه سازان، و گاه هر دو قشر می‌باشند. راهکارهای دستهٔ دوم، متوجه تمام افرادی است که در محیطهای دیجیتالی به جستجوی اطلاعات می‌پردازند.

این راهکارها ناظر به دو روش کلیِ ایجاد ابزارها و قواعد برای استانداردسازی نگارش متن (مدرک، اصطلاحات نمایه و اصطلاحات پرسش) است. در راهکار «هماهنگی رسم الخط» تأکید بر آن است که مرجعی قابل اطمینان، استانداردی را برای شیوهٔ نگارش تصویب و عرضه کند و اجرای آن نیز الزام آور باشد تا بتوان مرز و شیوهٔ نگارش کلمات را تابع قاعدهٔ واحدی کرد. برای تحقق چنین امری، پیشنهاد شده است فرهنگستان زبان کمیته‌ای را مأمور تدوین راهکاری برای شیوه خط فارسی کند. راهکار دیگر، استفاده از سیاههٔ آماده است. در این شیوه، به کمک سیاهه‌ای از پیش تعیین شده، احتمالات گوناگون شیوهٔ نگارش از طریق ارجاعات با یکدیگر مرتبط می‌شود. پیشنهادی دیگر، تدوین فرهنگ جامع املایی است که در آن فهرستی جامع از واژه‌های دارای گوناگونی املایی گردآوری و برای ایجاد یکدستی و هماهنگی، به همهٔ سازمانها ابلاغ شود و در کتابهای آموزشی و رسمی‌اعمال گردد. راهکار دیگر، تدوین اصطلاحنامه‌های تخصصی در زبان فارسی است که حاوی اصطلاحات معیار در هر رشته و شیوه نوشتاری مورد قبول باشد. این راهکار نیازمند اقدامهایی مؤثر، هماهنگ و حساب شده از طرف سازمانهای ذیربط است (حری، ۱۳۷۲؛ عبداللهی نورعلی، ۱۳۸۶؛ مرتضایی، ۱۳۸۱).

دسته‌ای دیگر از راهکارها قواعدی را برای یکدستی نگارش فارسی پیشنهاد می‌کنند. برای نمونه، در روش هماهنگ کردن حروف، همهٔ حروف به شکل مستقل، بزرگ و در کنار هم نوشته می‌شوند (مثلاً « م ا س ت» به جای « ماست»). پیشنهادی

دیگر، ناظر بر نگارش تکواژها به طور مستقل است. پیشنهاد تکمیلی برای بهبود این کار آن است که تکواژها با فاصله‌ای تعریف شده نسبت به یکدیگر، متفاوت با فاصلهٔ معمول میان کلمات نوشته شوند (برای مثال، «من زبان شناس‌ی نه می‌دان م». یعنی نخست، تکواژهای تشکیل دهنده هر کلمه شناسایی و از هم جدا می‌شوند، با این حال، بی‌فاصله نوشته می‌شوند (حری،۱۳۷۲). برخلاف برخی که فراهم کردن امکان اعراب گذاری را در واژه‌پردازهای فارسی پیشنهاد می‌کنند، برخی حذف تمامی نشانه‌های اعراب گذاری در نگارش را پیشنهاد می‌کنند، برخی نیز آوانگاری حروف (یعنی تکرار حرف مشدد به جای علامت تشدید، نوشتن نون خیشومی‌از روی زبر زنجیره به روی زنجیره نوشتار در مورد تنوین (محقق‌زاده و زارعیان، ۱۳۸۳).

راهکار دیگر، استفاده از هر دو شکل مفرد و جمع در نمایه‌سازی است. با این حال، معنای صورت جمع و مفرد برخی کلمات در زبان تخصصی متفاوت است. برای نمونه، «آثار باستانی» رایج‌تر از «اثر باستانی» است، «منسوجات نظامی» را نمی‌توان به شکل مفرد «منسوج» به کار برد. در واژهٔ «مهمات» ارتباط معنایی صورت مفرد و جمع ضعیف شده است (سمایی، ۱۳۷۹). نگاشت یکسان حروفی مانند «ا» و «آ» از دیگر پیشنهادهاست. از آنجا که بین نگارش این دو مصوت کوتاه و بلند تمایزی وجود ندارد، با حذف علامت مد روی الف، املای کلماتی چون آرام، آن، انار، و ابر یکسان خواهد شد و تمایز بین این دو مصوت کوتاه و بلند در نمایش گرافیکی از میان می‌رود. همچنین، چیدمان از چپ به منظور یکدستی چیدمان انواع درونداده‌ای متنی، عددی و علایم پیشنهاد شده است. بدین ترتیب، یکدستی چیدمان از چپ نه تنها باعث هماهنگی زبان و متون ریاضی و شیمی، نتهای موسیقی، خط تصویری یا علائم گرافیکی مورد استفاده در سراسر جهان می‌شود، بلکه نگارش و مطالعه را هم برای انسان و هم برای ماشین ساده می‌سازد (محقق‌زاده و زارعیان، ۱۳۸۳). همچنین، تجهیز پایگاه اطلاعاتی به اصطلاحنامه می‌تواند کاربران را از ریختهای مختلف واژه به اصطلاح پذیرفته شده راهنمایی کند. ایجاد تمهیداتی برای آموزش و راهنمایی کاربران دربارهٔ استفاده از پایگاه، راهکار دیگری برای بهبود راهبردهای جستجوست (گل تاجی و بذرگر، ۱۳۸۹).

تحلیل راهکارهای ارائه شده در متون

گرچه راهکارهای ارائه شده در مجموع بهترین راهکارهای ممکن را تشکیل می‌دهند، با این حال، همان‌گونه که برخی نویسندگان خود نیز اذعان داشته‌اند هر راهکار به گونه‌ای قابل انتظار از جامعیت به دور است و در عین حال دارای کاستیهای خاص خود است. برای نمونه، در راهکار هماهنگ نوشتن حروف (حری، ۱۳۷۲)، احتمال خطا بسیار کاهش می‌یابد، با این حال، احتمال اقبال به این شیوه نگارش اندک است. زیرا مستلزم تغییر رفتار و نگرش کاربران است. بویژه، احتمال مقاومت در برابر آن، به دلیل دوری از شیوه سنتی نگارش فارسی، بیم گسستن پیوند با گذشته و دشواری خواندن متون کهن فارسی وجود دارد. البته می‌توان نمایش و ذخیره‌سازی متن به شیوه‌های متفاوت صورت گیرد، به نحوی که اولی به روش متعارف و دومی به روش «هماهنگ شدهٔ پیشنهادی» روی دهد. اما حتی در این صورت نیز این راهکار تنها بخشی از دشواریهای نگارش را رفع می‌کند و چالشهایی چون کلمات مرکب، اعراب‌گذاری، تفاوت در املا، عدم تمایز بین اسامی‌خاص و عام همچنان به قوّت خود باقی خواهد ماند. از سوی دیگر، در این روش به دلیل نیاز به تقطیع حروف، زمان زیادی به هنگام ذخیره‌سازی، کاوش و همچنین نمایش متن صرف می‌شود که کارآیی سامانه را کاهش می‌دهد. در راهکار استفاده از تکواژها نیز همان‌گونه که حرّی خود تأکید می‌کند، تعیین تکواژها نیازمند دانشی است که تنها نزد متخصصان یا پژوهندگان زبان‌شناسی است. از این رو، عملیاتی کردن این راهکار به سادگی ممکن نیست (حری،۱۳۷۲).

کاستی راهکار استفاده از سیاههٔ آماده، به پویایی زبان باز می‌گردد. در بهترین حالت، سیاههٔ آماده تنها در نقطه‌ای از زمان کامل است و هیچ‌گاه به نقطهٔ کمال خود نخواهد رسید. از این رو، به بازنگری مستمر نیاز دارد. همچنین، بیم آن می‌رود که در دراز مدت، به دلیل بی‌دقتی یا سلیقه‌ای عمل کردن، سیاهه دچار ناهماهنگی شود. از این گذشته، کارآیی سامانه به لحاظ فضا و زمان کاهش می‌یابد، زیرا به ناچار حجمی رو به رشد از واژگان و صورتهای مختلف آن در سامانه ذخیره می‌شود و از آنجا که هر فقره اطلاعات هنگام بازیابی ناگزیر باید از غربال سیاههٔ مورد نظر بگذرد، زمان کاوش اطلاعات افزایش یافته،

کار بازیابی کُند می‌شود. اما این شیوه را می‌توان در نبود مرجعی واحد و موثّق برای یکسان‌سازی شیوهٔ نگارش، جایگزینی مناسب تلقی کرد(حری،۱۳۷۲).

در راهکار پیوند ساختگی میان کلمات، که بر تعریف فاصله‌های درونی اجزای کلمه استوار است، این اشکال عمده وجود دارد که قبل از درونداد اطلاعات، متخصصان باید کلماتی را که احتمال جدا یا پیوسته نوشتن اجزای آنها می‌رود، شناسایی و با کد مربوط مجهز کنند. در این روش، امکان پردازش خودکار متن نیست، زیرا عملیات مقدماتی باید قبل از ورود صورت گیرد و از طریق صفحه کلید به نظام خورانده شود. اما این روش، همان‌گونه که حرّی بیان می‌دارد، برای حلّ مسائل مقطعی برنامه‌های فارسی موجود مطلوب است(حری،۱۳۷۲).

در روش هماهنگی رسم الخط، تأکید بر تدوین و تصویب رسم الخط واحد و الزامی‌کردن اجرای آن است(حری،۱۳۷۲). آشکار است که این راهکار، نه تنها از منظر بازیابی اطلاعات که به لحاظ رفع آشفتگی و چندگونگی نگارش و در نتیجه بقا و اعتلای زبان فارسی، بسیار ارزشمند است. با این حال، وابستگی آن به تغییر رفتار و عادات کاربران اثربخشی آن را در کوتاه مدت زیر سؤال می‌برد. حتی اگر با ابلاغ قوانین و مقررات استاندارد نگارش، افراد را به رعایت نگارش تجویز شده وادار کنیم، باز هم نهادینه شدن آن بسیار به طول خواهد انجامید. به طور کلی، پیشنهادهایی از این دست، به ایجاد تغییراتی زیربنایی و گسترده در بافتاری نزدیک به بیش از یک هزار ساله نیاز دارند. بویژه، این گونه راهکارها نیازمند هم‌رأیی و همراهی تودهٔ مردم ـ خواه عوام یا خواص ـ است که چه بسا لزوم این تغییرات اساسی را درک نکنند. از سوی دیگر، از آنجا که ابتکار فردی جای خود را به نگارش دستوری خواهد داد، با پراگماتیک زبان مغایر خواهد بود، چه، زبان در بستر عملی و در جریان طبیعی خود، راه بقای خود را می‌یابد و چندان با روشهای دستوری سازگار نیست. از این رو، این راهکارها بیشتر متناسب هدفهای راهبردی و بلندمدت است، که آن نیز مستلزم نقش‌آفرینی بنیادین‌ترین نهاد یعنی نظامهای آموزش و پرورش است. از سوی دیگر، به نظر می‌رسد جمع میان این ۵ راهکار به دلیل به کارگیری مبناهای متفاوت برای تقطیع عناصر زبان‌شناختی

دشوار است. در پیشنهاد نگارش هماهنگ حروف، حرف به عنوان عنصر اصلی مبنا قرار گرفته است و در پیشنهاد دوم، تکواژ. اما در پیشنهادهای بعدی، مبنای تقطیع کلمه است. اگر بپذیریم که این تدابیر برای یافتن راهی روشن به منظور برقراری ارتباط انسان ـ ماشین ـ انسان است و کاربران رایانه عمدتاً افرادی با تخصصهای گوناگون هستند که قصد ارتباط کلامی با یکدیگر دارند، بنابراین، کوچکترین واحد معنادار برای آنها کلمه است، نه حرف یا تکواژ. از این گذشته، راهکار تهیهٔ سیاههٔ آماده و همچنین پیوند ساختگی میان کلمات به دلیل وابستگی به مداخلهٔ نیروی انسانی، با روند خودکارسازی پردازش متن در تعارض است. از این لحاظ، هماهنگ کردن رسم‌الخط فارسی، معقول‌تر از سایر پیشنهادها به نظر می‌رسد(حری،۱۳۷۲). با این حال، این راهکار نیز همان‌گونه که گفته شد، به دلیل دستوری بودن و نیاز به نهادینه شدن در میان کاربران، در کوتاه مدت به بار نخواهد نشست.

جدول۲. راهکارهای پیشنهادی در مراحل مختلف چرخه زندگی مدرک

مرحله	راهکار پیشنهادی	
مرحلهٔ پیش از بازیابی	ایجاد ابزارهایی برای استانداردسازی نگارش به هنگام تولید مدرک	تدوین فرهنگ جامع املایی
		تدوین اصطلاحنامه‌های تخصصی در زبان فارسی
		تدوین اصول برگردان کلمات خارجی
		هماهنگ کردن رسم الخط
	ایجاد ابزارهایی برای ارتقای نگارش به هنگام تولید بازنمونهای مدرک	استفاده از سیاههٔ آماده
		ایجاد نظامهای ریشه‌یابی در فارسی
		تجهیز واژه‌پردازهای فارسی به غلط‌یاب املایی
		هماهنگ کردن حروف
		استفاده از تکواژها
		حذف اعراب‌گذاری
		چپ‌نویسی
		استفاده از هر دو صورت مفرد و جمع در نمایه‌سازی

		کاهش شمار نویسگان با قایل شدن دو حالت بزرگ و کوچک برای حروف
		نگارش واژه محور و قرار دادن فاصله بین کلمات برای تعیین مرز بین آنها
		نشانه‌گذاری اسامی‌خاص از طریق تفکیک حالت بزرگ و کوچک حروف
		قرار دادن نشانهٔ یکسان برای حروف دارای چند تلفظ مانند س، ث، ص
		قرار دادن نشانهٔ نوشتاری خاص برای کسرهٔ اضافه در همه شرایط
		قرار دادن نشانهٔ جداگانه برای «ی» نکره و «ی» تکیه بر اسم ساز و صفت ساز
		قرار دادن نشانهٔ جداگانه برای واژه بستهای ربطی فعل«بودن»
		قرار دادن نشانه«ـ» در بین کلمات ترکیبی
مرحلهٔ ذخیره‌سازی	ایجاد ابزارهایی برای ارتقای نگارش / نمایه‌سازی	درج حروفی که خوانده ولی نوشته نمی‌شوند
		عدم تمایز بین «ا» و «آ»
		پیوند ساختگی میان کلمات
	قواعد یکدستی نگارش	واگذاری حل مشکل کلمات ترکیبی به رایانه
		بی‌فاصله‌نویسی کلمات مرکب
		درج نکردن فاصله میان مقلوب عبارتهای اسمی‌مانند«زردکوه»
		درج نکردن فاصله میان عبارتها و واژه‌های لاتین که دقیقا منعکس کننده لفظ خارجی است، مانند «سوپرساب» و نه «سوپر ساب»
		درج فاصله قبل و بعد از حرف ربط، مانند «مواد دیداری و شنیداری»
		درج فاصله قبل و بعد از حرف ربط، مانند «مواد دیداری و شنیداری»
مرحلهٔ بازیابی	تجهیز پایگاه اطلاعاتی به اصطلاحنامه	
	آموزش و راهنمایی کاربران	
	استفاده از واسط کاوش فارسی برای رفع چالشهای رسم‌الخط و مفهومی	

به همین ترتیب، روش چیدمان چپ‌نویس (محققزاده و زارعیان، ۱۳۸۳) از همین کاستیِ نیاز به تغییر عادتها و نهادینه شدن در طول زمان رنج می‌برد. با این حال، این روش را می‌توان بر ذخیره‌سازی متن و نه لزوما نمایش آن پیاده کرد. بدین ترتیب، خواندن متن برای رایانه ساده‌تر می‌شود و کاربر نیز با روش مألوف خود به خواندن متن نمایش داده شده می‌پردازد. البته، این تمایز بین سبک ذخیره‌سازی و نمایش، به الگوریتمی پیچیده نیاز دارد که خواه ناخواه کارآیی سامانه را متأثر خواهد ساخت.

روش کاهش شمار نویسگان پیشنهاد می‌کند که از میان شکلهای متعدد برای یک حرف، تنها دو حالت بزرگ و کوچک را برای هر حرف بپذیریم. هر چند این پیشنهاد در جهت کاهش شمار نویسگان و حل مشکل کمبود کلید بر صفحه کلید بسیار مفید به نظر می‌رسد، حالت کوچک و بزرگ پیشنهادی برای این حروف، تفاوتی چشمگیر ندارند (نگاه کنید به محققزاده و زارعیان، ۱۳۸۳). علاوه بر این، شکل بزرگ و کوچک حروفی چون «د»، «ذ»، «ر»، «ز»، «ژ»، «و» و «ء» هم برای انسان و هم برای رایانه (به هنگام تشخیص نوری نویسه‌ها) تقریباً قابل تشخیص نیست. همچنین، موفقیت این روش نیز در گروِ تغییر در رفتار و نگرش کاربران است.

روش دیگر، پیشنهاد یکسان‌سازی نگارش حروفی مانند «س»، «ث» و «ص» است که در زبان فارسی تلفظ یکسان دارند. به نظر می‌رسد این راهکار و دیگر راهکارهایی از این دست مانند نوشتن حروفی که خوانده اما نوشته نمی‌شوند، با گرایشهای نگارشی نسل جدید نیز انطباق داشته باشد. نگاهی گذرا به نوشته‌های فارسی در جای جای اینترنت روشن می‌سازد کاربر جوان بیش از آن‌که به املای کلمه توجه داشته باشد، آن را با آوانویسی ساده می‌کند. برای مثال، فراوانی املای «راجب» به جای «راجع به» نمونه‌ای از این گرایش است که یا ناشی از املای ضعیف است یا تمایل به ساده‌سازی و ساده‌نویسی املای فارسی. به نظر می‌رسد کاربر امروز با این رفتار ـ آگاه یا ناخودآگاه ـ نشان می‌دهد که ضرورتی برای رعایت نگارش عربی نمی‌شناسد و مایل است پیچیدگی نگارش تنها بر حسب ضرورت زبان فارسی روی دهد و نه ضرورتهای برخاسته از زبان مبدأ. با این حال، این گونه راهکارها هدف اصلی زبان را به چالش

می‌کشد. زیرا، نه تنها رسالت اصلی زبان را که برقراری ارتباط است محقق نمی‌کند و باعث گسست در درک خواننده می‌گردد، بلکه به دوگانگی متون چاپی و رایانه‌ای نیز منجر می‌شود، که این امر آسیب شدیدی به ارتباطات و نیز فرهنگ نوشتاری وارد می‌سازد.

از طرفی، با توجه به آمیختگی شدید زبان عربی و فارسی، تغییر املایی این واژه‌ها به منظور هماهنگی با رسم الخط فارسی، سبب از بین رفتن و یا دگرگونی معنای آنها و در نتیجه ابهام، بدفهمی‌و حتی گاهی درک نشدن واژه توسط خواننده می‌گردد و درصد ریزش کاذب را در نتایج بازیابی نیز افزایش می‌دهد. برای مثال، اگر واژه «قالب» به معنای «شکل» به صورت «غالب» نگارش شود، معنی «پیروز» از آن برداشت می‌شود، یا نگارش واژه «صبور» به صورت «سبور»، برای خواننده کاملاً نامأنوس بوده، ممکن است سبب درک نشدن آن شود. علاوه بر این، روی آوردن به چنین راهکاری، موجب گسستی عمیق بین حال و گذشتهٔ ادبی، فرهنگی و تاریخی می‌شود و تردید بسیاری را بر جای می‌گذارد.

حرکت به سوی خودکارسازی پردازش متن فارسی

چنان‌که گفته شد، به طور کلی دو دسته راهکار ایجاد ابزار و استانداردسازی تولید متن را می‌توان در جهت کاهش دشواریهای بازیابی فارسی به کار گرفت. ایجاد و تدوین ابزارهایی چون اصطلاحنامه‌ها، فرهنگهای املایی و قواعد نگارش استاندارد، گامی‌مؤثر در افزایش اثربخشی بازیابی به شمار می‌آید. این ابزارها، ضمن توسعهٔ معنایی اصطلاحات جستجو و نمایه، می‌توانند با هدف یکسان‌سازی نگارش و از بین بردن گوناگونی نحوی و ریخت‌شناختی نیز به کار روند. یکسان‌سازی نگارش می‌تواند متن مدرک، اصطلاحات نمایه یا اصطلاحات پرسش را در برگیرد. از این رو، این روش را می‌توان در هر مرحله‌ای از چرخهٔ زندگی اطلاعات، از تولید، ذخیره‌سازی، نمایه‌سازی گرفته تا جستجو و بازیابی، اعمال کرد. اما این راهکار زمانی بیشترین بازده را خواهد داشت که بیش از آنکه به قضاوت و تصمیم کاربر یا تغییر عادتها و رفتار وی وابسته

باشد، بر خودکارسازی پردازش متن، نمایه‌سازی، یا ترجمهٔ ماشینی استوار باشد. چه در روش خودکار، می‌توان صورتهای متغیر کلمه را صرف نظر از عادتهای نگارشی افراد، یکدست و بهنجار کرد.

همان‌گونه که در متن اشاره شد، با توجه به قاعده‌مندی بسیاری از چالشها مانند اعراب، علایم جمع، همزهٔ پایانی و برخی وندهای اسم‌ساز و صفت‌ساز، می‌توان در الگوریتمهای جستجو، این واژه‌ها را به نحوی بهنجار کرد که واژه صرف نظر از ریختهای مختلف آن، بازیابی شود. آشکار است که به سادگی نمی‌توان به الگوریتمی‌تمام‌عیار با اثربخشی مطلق دست یافت. برای نمونه، در مورد علامت جمع، شاید بتوان واژه‌های جمع و مفرد را با حذف «ها» و «ان» یکسان کرد. با این حال، زمانی که این علایم بخشی از واژه باشند، مانند «تنها»، «رها»، «زمان»، «نان» یا «انسان» احتمال بروز خطا می‌رود. البته در برخی از این موارد، این امکان وجود دارد که با فنون سنجش در الگوریتم، در صورتی که تعداد نویسه‌ها کمتر از دو نویسه باشد، بهنجارسازی را اعمال نکرد. برای نمونه‌ای دیگر، همان‌گونه که پیشتر ذکر شد، در برخی موارد بهنجارسازی صورت جمع با صورت مفرد کلمه باعث تغییر معنا می‌شود (مانند مصالح / مصلحت). همچنین، همیشه نمی‌توان شکلهای بلند یک واژه را به شکل کوتاه آن یا برعکس بهنجار کرد، زیرا در پاره‌ای موارد شکل اختصاری با واژه‌ای دیگر هم‌املا می‌شود یا معنای آن به کلی تغییر می‌کند. (مانند کوه/که؛ آگاهی/آگهی). با این حال، باید توجه داشت که بروز درصدی از خطا ذاتیِ هر گونه روش «اکتشافی»[1] است و حتی در الگوریتمهای موفق و رایجی مانند پرتر[2] نیز ممکن است روی دهد. از این‌رو، پیش از طراحی این گونه الگوریتمها، بررسی قاعده‌مندیهای نگارش زبان فارسی و درصد واژه‌هایی که این قاعده‌مندیها را نقض می‌کنند، می‌تواند ما را نسبت به میزان رواداری[3] این الگوریتمها آگاه سازد.

ایجاد الگوریتمهای ریشه‌یابی کلمات فارسی که در متون نیز آمده بود، به بخشی از

1. Heuristic.
2. Porter.
3. Tolerance.

راهکارهای خودکارسازی پردازش متن اشاره دارد. چنانچه منظور از ریشه‌یابی حذف وندهای کلمه باشد[1] می‌تواند بسیار راهگشا باشد، زیرا در زبان فارسی، واژه‌سازی بیشتر به کمک پیشوندها و پسوندها صورت می‌گیرد که ریخت واژه را چندان دستخوش تغییرات بنیادین نمی‌کند. با این حال، چنانچه منظور از ریشه‌یابی طراحی الگوریتمی‌برای یافتن بن‌واژه[2] باشد، کار یافتن قاعده‌مندیها دشوارتر خواهد شد، زیرا تغییر ریخت واژگان در فارسی، بیشتر بر واژگان وام گرفته عربی روی می‌دهد. برای مثال، جمع مکسر، یا صرف کلمه در بابهای مختلف (مانند تعمیر یا استعمار). آشکار است که تقلیل این صورتهای صرف شده به ریشهٔ آنها نه به سادگی ممکن است و نه مطلوب، زیرا در بسیاری از موارد جمع مکسر یا صرف کلمه در بابی دیگر به تغییری بنیادین در معنا می‌انجامد. از این گذشته، به‌کارگیری فنون بازیابی روادار[3] بویژه فنون تصحیح املا که نسبت به گونه‌گونی ریختی یا صرفی واژه نیرومند باشد، از دیگر راهکارهای ممکن است. در این فنون، املاهای مختلف، خواه ناشی از اشتباه کاربر باشد یا تنوع املایی واژه، به یک ریخت واحد تقلیل می‌یابد و در نتیجه همهٔ احتمالات ممکن مورد جستجو قرار می‌گیرد (مانینگ، راگاوان و شوتس، ۲۰۰۸). فنون تصحیح املا بر بازیابی فارسی در گوگل به کار گرفته شده است. برای نمونه، جستجو به دنبال «یگتا» یا «اسربخشی» ضمن ارائه نتایج حاصل از جستجوی این دو املای غلط، نتایج مربوط به واژهٔ «یکتا» یا «اثربخشی» را نیز پیشنهاد می‌دهد.

نتیجه‌گیری

به طور کلی، ۴۳ گروه چالش نگارشی در متون معرفی شده است. آنچه بیش از همه مد نظر پژوهشگران بوده مسئله «پیوسته یا جدانویسی»، «تنوع نشانه‌های جمع»، «تفاوت در آوا / اعراب‌گذاری»، «تنوع دگرنوشته‌ها»، «الف کوتاه»، «فاصلهٔ بین حروف

1. Stemming.
2. Lemmatization.
3. Tolerant Retrieval Techniques.

واژه»، و «نگارش از راست به چپ» بوده است. برخی از چالشها نیز کمتر مورد توجه قرار گرفته یا به طور کلی نادیده گرفته شده است. با توجه به اینکه در هر گروه ممکن است بیش از دو شکل املایی روی دهد، آشکار خواهد شد نگارش فارسی اصولا به شیوه‌ای بسیار متنوع صورت می‌گیرد. آشکار است که این گونه‌گونی نگارشی به نایکدستی و دگرگونی بسیار در نگارش فارسی می‌انجامد که می‌تواند اثربخشی بازیابی را بویژه از منظر کاهش دقت یا ریزش کاذب و نیز کاهش جامعیت بازیابی، متأثر سازد.

اگرچه راهکارهای ارائه شده در متون از کاستیهایی بویژه نداشتن جامعیت رنج می‌برند، کم و بیش اثربخش به نظر می‌رسند. با این حال، با توجه به اینکه راهکارهای انسانی نیازمند مشارکت فعالانهٔ نویسندگان متون (تایپیستها و کاربران) است و از روندی کُند، بلندمدت و هزینه‌بر برخوردار است، ضروری است راهکارهای خودکارسازی پردازش متن و نمایه‌سازی بیش از پیش مورد تأکید قرار گیرد. مرور آثار پژوهشی در بخش پیشینهٔ پژوهش نشان داد شمار پژوهشها در حوزه طراحی و آزمایش تکنیکها، ابزارها و الگوریتمهای خودکارسازی بازیابی زیاد است که نشان از پیشرفتها و دستاوردهای روزافزون در این حوزه دارد. با این حال، دانش اندکی در مورد میزان به‌کارگیری این فنون در سامانه‌های اطلاعاتی مختلف و میزان اثربخشی آنها در بافتار عملی در دست است. از این رو، ضروری است ضمن آنکه در طراحی سامانه‌های فارسی به این چالشها توجه می‌شود، مطالعات مقدماتی به منظور سنجش میزان اثربخشی و همچنین هزینه ـ سودمندی راهکارها انجام شود. چه، طراحی الگوریتمی‌که تنها به ازای درمان یک چالش نادر یا ناچیز، پیچیدگی زیادی را بر سامانه تحمیل کند، به کاهش کارآیی آن و افزایش هزینه ـ سودمندی منجر خواهد شد. از این رو، یکی از گامهای بنیادین در پژوهشهای بازیابی فارسی، بررسی میزان رویداد هر یک از چالشها و میزان تأثیر آنها بر اثربخشی بازیابی است.

گام بنیادین دیگر در این راستا، تدوین شیوه‌نامهٔ نگارش فارسی، اصطلاحنامه‌ها و فرهنگهای املایی در محیط دیجیتالی است. مشارکت متخصصان موضوعی، زبان و ادب فارسی، رایانه و کتابداری در این امر ضروری است. کتابخانه ملی یکی از سازمانهای

مهم و تأثیرگذار است که می‌تواند در تدوین استانداردها با طراحان پایگاه‌های اطلاعاتی و نرم‌افزارها مشارکت کند. با توجه به آنکه این راهکار در بلندمدت به بار می‌نشیند، پیشنهاد می‌شود هم‌زمان با اقدامهای پژوهشی و زیربنایی، اقدامهای عملی نیز از سوی کتابخانه‌ها و مراکز اطلاع‌رسانی به منظور افزایش بهره‌وری پایگاه‌ها و سامانه‌های اطلاعاتی صورت گیرد. برای نمونه، تدوین دستنامه یا راهنمای جستجو می‌تواند کاربران را در رابطه با تدوین راهبردهای جستجوی موفق آموزش دهد. لازم است در این راهنما، در کنار شرح فنون و تسهیلات جستجو مانند امکانات جبر بولی و جز آن، نکات مهمّ نگارش فارسی مؤثر بر اثربخشی بازیابی اطلاعات آموزش داده شود. همچنین، در طراحی پایگاه‌های اطلاعاتی، الگوریتمهای متفاوت بسته به نوع پایگاه و پوشش موضوعی آن به کار گرفته شود. برای نمونه، در برخی رشته‌های علمی‌مانند شیمی‌و ریاضی، فرمول‌نویسی مشکل غالب است، حال آنکه در متون مذهبی یا متون فارسی– عربی، احتمالاً اعراب‌گذاری تأثیر بسزایی بر بازیابی اطلاعات خواهد داشت.

منابع

- اسلامی، م (۱۳۸۱). *دشواریهای پردازش رایانه‌ای خط فارسی*. نشر دانش. ۱۹ (۳): ۲۸–۳۲.
- اشرف‌زاده، ب (۱۳۸۱). ایرادهای «قواعد کلی» فرهنگستان (۲). بازیابی شده به تاریخ ۱۵ آذر ماه ۱۳۹۰ از نشانی http://ashrafzaade.persianblog.ir/post/48
- باطنی، م (۱۳۷۲). *رابطه خط و زبان*. آدینه. ۷۸و۷۹: ۷۰–۷۱.
- جرأت، ع. و س. سمایی (۱۳۸۳). «چرا جهانی نمی‌شویم؟ نگاهی به مشکلات خط فارسی زمینه با فناوری اطلاعات»، *ابتکار* (۴ بهمن).
- حری، ع (۱۳۷۲). کامپیوتر و رسم الخط فارسی. *پیام کتابخانه*. ۳ (۱): ۶–۱۱.
- حسینی بهشتی، م (۱۳۸۲). کاربرد اصطلاح شناسی و واژه گزینی در نمایه‌سازی ماشینی و بازیابی اطلاعات. *علوم اطلاع‌رسانی*. ۱۸ (۳): ۳۱–۴۴.
- خیام، م (۱۳۸۶). خط فارسی کارایی علمی‌و فنی ندارد. *حیات نو* (۱۷ تیر).
- دشتی، الف (۱۳۸۳). «بررسی سه پیشنهاد در شیوه نگارش خط فارسی: حتا، مثلن، خاهر»، *شرق* (۱۰ تیر).
- راثی ساربانقلی، م.ص (۱۳۸۴ب). مهارت در جستجوی اطلاعات فارسی از اینترنت. *ارتباط*

علمی. ۵ (۱): ۱۶–۲۸.

- ـــــــــــــ (۱۳۸۴الف). بررسی مشکلات جستجو و بازیابی اطلاعات به زبان فارسی از اینترنت با مطالعه موردی بر روی کاربران مرکز اینترنت دانشگاه آزاد اسلامی‌واحد شبستر. پایان نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی، دانشگاه آزاد اسلامی، واحد تهران شمال.
- رجایی، م (۱۳۷۲). خط فارسی به اصلاحات بنیادی نیاز دارد. *آدینه*. ۸۰: ۶۹–۷۱.
- زندی مقدم، ز. (۱۳۸۸). نکاتی دربارهٔ شیوه نگارش چند تکواژ دستوری زبان فارسی. *نامهٔ فرهنگستان*. ۳: ۲۰۳–۲۲۰
- سرمستانی، ج (۱۳۸۸). کاستیها و نادرستیهای دستور خط فارسی فرهنگستان. بازیابی شده به تاریخ ۱۵ آذر ماه ۱۳۹۰ از نشانی http://vista.ir/?view=context&id=309089
- سمایی، م (۱۳۷۹). مفرد و جمع در نمایه سازی. *علوم اطلاع‌رسانی*.۱۶ (۱و۲): ۲۷–۳۱.
- شهیدی، م، صدیقی، م، زمانی فر،ک (۱۳۸۴) روشی برای رفع چالش‌های محتواکاوی در وب‌های فارسی زبان فصلنامه علوم و فناوری اطلاعات ۲۱ (۲): ۴۷– ۶۹
- صدیق بهزادی، م (۱۳۷۱). ناهماهنگی ضبط نامهای بیگانه در فارسی. *فرهنگ*.کتاب سیزدهم:۱۰۳–۱۱۶.
- ـــــــــــــ (۱۳۷۵). زبان فارسی در تبادل اطلاعات کتابشناختی. *نامهٔ فرهنگستان*. ۸: ۷۰–۸۱.
- صنعتی، م (۱۳۷۱). دشواری‌های زبان فارسی با کامپیوتر. *آدینه*. ۷۲: ۵۶– ۵۷.
- ضیاء، الف (۱۳۸۱). شیوهٔ خط فارسی، دیدگاهی دیگر. *چیستا*. ۱۹۰: ۷۹۴–۷۹۷.
- طباطبایی، ع (۱۳۸۱). در دشواری‌های رایانه‌ای زبان فارسی. *نشر دانش*. ۳۶:۱۰۶–۳۸.
- طرح جامع پیکره زبان فارسی با موضوع فاز اول مطالعاتی ایجاد پیکره متنی زبان فارسی: ارایه مشاوره در پروژه‌های ذخیره و بازیابی اطلاعات متون زبان فارسی (۱۳۸۸). ویرایش ۱. طرح مصوب شورای عالی اطلاع‌رسانی و دانشگاه علم و صنعت، بازیابی شده به تاریخ ۲ تیر ماه از نشانی http://arooz.net/attachments/059_20-IR.pdf
- عبداللهی نورعلی، م.ص (۱۳۸۶). کند و کاو مسائل ریخت شناسی زبان فارسی در بازیابی اطلاعات از جستجوگرهای وب. پایان نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی، دانشگاه شیر از، شیراز.
- فرهنگستان زبان و ادب فارسی (۱۳۸۳). *دستور خط فارسی*. تهران: فرهنگستان زبان و ادب

فارسی

- کابلی، الف (۱۳۷۲). احیای نشانهٔ اضافه. *آدینه*. ۸۰: ۶۸-۶۹.
- گزنی، ع (۱۳۸۵) استخراج خودکار عبارتهای کلیدی از متون مقاله‌های فارسی. *کتابداری و اطلاع‌رسانی*، ۹(۳): ۹۷- ۱۰۸
- گل تاجی، م. و س بذرگر (۱۳۸۹). بررسی مشکلات ریخت شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی‌ایران و جهاد دانشگاهی. *کتابداری و اطلاع‌رسانی*. ۱۳ (۲):۱۹۱-۲۱۴
- محققزاده، م.ص. و زارعیان، ک (۱۳۸۳). ارائه راه حل برای برخی مسائل اتوماسیون نگارش فارسی. *اطلاع‌رسانی*، ۱۹ (۳- ۴): ۱-۱۰.
- مرتضایی، ل (۱۳۸۱). مسائل زبان و خط فارسی در ذخیره و بازیابی اطلاعات. *اطلاع‌رسانی*، ۱۷ (۱-۲): ۱-۷.
- مرعشی، ع (۱۳۸۳). خط فارسی چگونه با اینترنت کنار علمی‌آید؟. *رشد تکنولوژی آموزشی*. ۱۶۱: ۳۳-۳۵.
- معصومی‌همدانی، ح (۱۳۸۱). خط فارسی و رایانه. **نشر دانش**. ۱۰۱: ۲.

- AleAhmad, A, Hakimian, P, Mahdikhani, F and Oroumchian, F, (2007) N-gram and local context analysis for Persian text retrieval, 9th International Symposium on Signal Processing and Its Applications - ISSPA 2007, Sharjah, United Arab Emirates, 12-15 February 2007.
- AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F. (2009). Hamshahri: A standard Persian text collection.Knowledge Based Systems, 22(5): 382-387, DOI: 10.1016/j.knosys.2009.05.002
- Alizadeh, H., Fattahi, R. (2010) International Journal of Information Science and Management. 8 (2): 89-98.
- Bar-Ilan,J. & Gutman,T(2002). How do Search engines Handle Non-English Queries? A Case Study. Retrieved 10/5/2011 from Science Direct database.
- Berenjkoob, M., Mehri, M., Khosravi, H., Nematbakhsh, Ma. (2009). A Method for Stemming and Eliminating Common Words for Persian Text Summarization. In the proceedings of the International Conference on Computer Engineering and Technology.978-1-4244-4538-7/09/$25.00 ©2009IEEE. Retrieved 7 June 2012 from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5313836

- Danesh, M., Minaei, B. and Kashefi, O. (2011) Challenging Massive Information Retrieval in Persian. International Journal of Information and Education Technology (1) 3: 212-220.
- Farhoodi,M., Mahmoudi, M., Zare Bidoki, AM., Yari, AR., and Azadnia, M. (2009). Query Expansion Using Persian Ontology Derived from Wikipedia. World Applied Sciences Journal 7 (4): 410-417.
- Hedlund, T. et al (2000). Aspects of Swedish Morphology and Semantics from the Perspective of Mono- and Cross-Language Information Retrieval. Retrieved 10/12/2011 from Elsevier database.
- Iranpour Mobarakeh, M., and Minaei-Bidgoli, B., (2009). Verb Detection in Persian Corpus. International Journal of Digital Content Technology and its Applications 3 (1): 58-65. Retrieved 7 June 2012 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.8313&rep=rep1&type=pdf
- Karimpour, R., Ghorbani, A., Pishdad, A., Mohtarami, M., Aleahmad, A., Amiri, H. & Oroumchian, F. (2009), Improving Persian information retrieval system using stemming and part of speech tagging, in Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, 17-19 Sept. 2008, Aarhus, Denmark, Lecture Notes in Computer Science, vol. 5706, pp. 89-96.
- Keyvan, F. Borjian, H., Kasheff, M. and Fellbaum, C. (2006) Developing PersiaNet: The Persian Wordnet. In Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 315–318. Retrieved 7 June 2012 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.148.7473&rep=rep1&type=pdf
- Khalifehsoltani, SN., Cholmaghani, A., Vahdani, A., and Moallemi, R. (2010a)Towards Acquisition of a Thematic Persian Corpus from the Tebyan Portal: TebCorp. In the proceedings of 2nd International Conference on Computer Engineering and Technology ,978-1-4244-6349-7/10/$26.00 _c 2010 IEEE Pp: v7-682-686
- Khalifehsoltani, SN., Cholmaghani, A., Vahdani, A., and Moallemi, R. (2010b) Building a Large Per sian Verb Collection: A Generative Approach. In the proceedings of 2nd International Conference on Computer Engineering and Technology, 978-1-4244-6349-7/10/$26.00 _2010 IEEE Pp: v7-687-691.
- Khosravi, F. and Vazifedoost, A. (1386) Creating a Persian ontology through thesaurus reengineering for organizing the Digital Library of the National Library of Iran. Faslnameh – ye- Ketaab, 70 (Summer

1386): 19-36.
- Lazarinis, F., Vilares, J., Tait, J. Efthimiadis, EN. (2009). Current research issues and trends in non-English Web searching.Information Retrieval, 12:230–250. DOI 10.1007/s10791-009-9093-0.
- Manning CD.; Raghavan, P.; Schutze, H. (2008)Introduction to Information Retrieval. Cambridge: Cambridge University Press
- Mehrad, J. and Berenjiann S. R. (2011). Providing a Persian Language Singular-Stemmer System (RICeST Stemmer). International Journal of Information Science and Management, 1(2): 13-22.
- Mohtarami, M., Amiri, Hadi., Oroumchian, F. & Rahgozar, M. 2008, 'Using heuristic rules to improve Persian part of speech tagging accuracy', International Conference on Information and Knowledge Engineering, Universal Conference Management Systems and Support, California, USA.
- Monz,C. & De Rijke,M (2002).Shallow Morphological Analysis in Monolingual Information Retrieval for Deutch, German, and Italian. In proceedings of Evaluation of Cross- Language Evaluation forum, CLEF 2001. September 3-4 2001. Darmstadt, Germany.
- Mosavi Miangah, T. (2006): Automatic lemmatization of Persian words, Journal of Quantitative Linguistics, 13:01, 1-15. DOI: 10.1080/09296170500500884.
- Mosavi Miangah, T. (2007) Solving the Polysemy Problem of Persian Words. In Proceedings of the Corpus Linguistics Conference (CL2007). Retrieved 7 June 2012 from ucrel.lancs.ac.uk/publications/CL2007/paper/8_Paper.pdf
- Moukdad, H (2005). Lost in Cyberspace: How do Search Engines Handle Arabic Queries? The International Information & Library Review, 37(4): 237-394.
- Moukdad, H. and Cui, H. (2005). How Do Search Engines Handle Chinese Queries?. Webology, 2 (3). Retrieved 10 June 2012 from http://www.webology.org/2005/v2n3/a17.html
- Moukdad, H. and Large, A. (2001). Information Retrieval from Full-Text Arabic Databases: Can Search Engines Designed for English Do the Job? Libri, 51: 63–74.
- Taghva, K., Beckley, R. and Sadeh, M. (2005) "A Stemming Algorithm for the Farsi Language," in Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - I (01).

- Tashakori,M. Meybodi, MR., and Oroumchian, F. (2002). Bon: The Persian Stemmer EurAsia-ICT 2002: Information and Communication Technology. Lecture Notes in Computer Science, 2002, Volume 2510/2002, 487-494, DOI: 10.1007/3-540-36087-5_57.
- Teymoorian, F., Mohsenzadeh, M., and Seyyedi, MA. (2009) English-Persian Text Retrieval Using Concept Graph. In the proceedings of the International Conference on Computer Engineering and Technology 978-1-4244-4520-2/09/$25.00 ©2009 IEEE: 447-451.
- Tóth, E. (2006). Exploring the Capabilities of English and Hungarian Search Engines for Various Queries. Libri, 56: 38–47
- Yoosofan, A., Rahimi, A., Rastgoo, M. and Mojiri, MM. (2010). Automatic Stemming of Some Arabic Words Used in Persian through Morphological Analysis without a Dictionary. World Applied Sciences Journal, 8 (9): 1078-1085. Retrieved 7 June 2012 from www.idosi.org/wasj/wasj8(9)/7.pdf