بررسی تنوع الگوهای نگارش فارسی و تأثیر آن بر جامعیت بازیابی اطلاعات (مطالعهٔ موردی: پیکرهٔ همشهری)

هاجر ستوده[1]
زهره هنرجویان[2]

چکیده

هدف: در سبک نگارش فارسی، برخی واژه‌ها را می‌توان با درج، حذف یا جایگزینی نویسه‌ای خاص نوشت و به این ترتیب، برای واژه‌ای واحد دو یا چند الگوی نگارشی متفاوت پدید آورد. این نایکدستی در الگوهای نگارش می‌تواند بر اثربخشی بازیابی اطلاعات فارسی تأثیر منفی داشته باشد. آشکار است که تلاش در جهت لحاظ کردن همهٔ الگوها در الگوریتم‌های بازیابی اطلاعات فارسی، بر پیچیدگی آنها می‌افزاید و کارآیی سامانه‌های بازیابی اطلاعات را کاهش می‌دهد. از این رو، ضروری است با بررسی رفتارهای نگارندگان فارسی، میزان چندگانگی الگوها و تأثیر آن بر بازیابی اطلاعات در عمل و در نتیجه، ضرورت لحاظ کردن آنها در الگوریتم‌های فارسی آشکار گردد.

روش پژوهش: در پژوهش حاضر که به روش تحلیل محتوای مفهومی انجام گرفته است، ۷ چالش از میان ۴۳ چالش نگارشی که با مرور ادبیات مربوط، شناسایی شده بود انتخاب گردید و پس از محاسبهٔ تنوع و فراوانی رویداد آنها در متون پیکرهٔ همشهری، میزان انطباق شیوه نگارش آنها با دستور خط مصوب فرهنگستان زبان و ادب فارسی بررسی شد.

یافته‌ها: نتایج پژوهش نشان داد نگارندگان متون پیکره به طور کلی تمایل به حذف یا جایگزینی نویسه‌های چالشی دارند. بنابراین، به نظر می‌رسد دست‌کم دربارهٔ هفت چالش مورد بررسی در این پژوهش، با نادیده گرفتن این چالش‌ها در سامانه‌های بازیابی اطلاعات، اثربخشی بازیابی چندان متأثر نمی‌شود. مقدار کلی شاخص«ضریب درگیری» برابر با ۰/۰۳۳ به دست آمد که بیانگر انطباق نداشتن گستردهٔ الگوی نگارشی نگارندگان با دستور خط مصوب فرهنگستان است. از دلایل این امر

---

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

می‌توان به تمایل نگارندگان به ساده‌نگاری در اثر «اصل کمترین کوشش» و عدم احساس ضرورت رعایت رسم‌الخط مرسوم عربی توصیه شده در دستور خط مصوب فرهنگستان، اشاره کرد.

کلیدواژه‌ها: بازیابی اطلاعات، نگارش فارسی، چالش، پیکرهٔ همشهری.

مقدمه

در نگارش فارسی، برخی واژه‌ها را می‌توان با درج، حذف یا جایگزینی نویسه‌ای[1] خاص نوشت و به این ترتیب، برای واژه‌ای واحد، دو یا چند الگوی نگارشی متفاوت پدید آورد. الگوهای نگارشی متنوع می‌تواند سبب ناهماهنگی در متون شود و بر بازیابی اطلاعات به لحاظ جامعیت نتایج بازیابی شده، تأثیر منفی بگذارد. شمار بسیار بالای چالش‌های شناسایی شده در الگوهای نگارش فارسی (ستوده و هنرجویان، ۱۳۹۱؛ محقق‌زاده و زارعیان، ۱۳۸۳؛ مرتضایی، ۱۳۸۱؛ حری، ۱۳۷۲) این ضرورت را پیش می‌آورد که هنگام طراحی الگوریتم‌های سامانه‌های فارسی، فنونی برای بهنجارسازی[2] چندگانگی املایی واژگان نمایه یا واژگان جستجو اندیشیده شود. با این حال، در بسیاری از سامانه‌های بازیابی اطلاعات فارسی، هنوز تأثیر صورت‌های مختلف نگارشی یک واژه بهنجار نمی‌شود (شهیدی، صدیقی و زمانی‌فر، ۱۳۸۳). از این رو، کاربران ناگزیرند چندین فرایند جستجو یا فرمول‌های جستجوی پیچیده‌تری را به کار گیرند. آشکار است که جامعیت چنین جستجویی در گروِ آگاهی کاربر از همهٔ تنوع‌های نگارشی و ظرایف جستجوی بولی و در عین حال، برخورداری وی از وقت و حوصله کافی است. با توجه به اصل کم‌ترین کوشش و همچنین آسان‌گیری کاربران در رفتار جستجوی خود (مانینگ و همکاران، ۲۰۰۸)، احتمال نادیده گرفتن چنین راهکارهایی وجود دارد. بنابراین، بهنجارسازی الگوهای نگارشی در الگوریتم‌ها ضروری می‌نماید. آشکار است که هر چه تنوع الگوهای نگارش یک زبان بیشتر باشد، الگوریتم حاصل پیچیده‌تر و احتمال تأثیر منفی آن بر کارآیی سامانه بیشتر خواهد بود. از این رو، این پرسش فراروی مدیران و برنامه‌نویسان سامانه‌های بازیابی اطلاعات فارسی خواهد بود

1. Character.
2. Normalize.

که چه میزان بهنجارسازی نگارشی در الگوریتم‌های بازیابی ضروری است.

بروز چالش‌های ریختی در پایگاه‌های فارسی، موتورهای کاوش عمومی، نشریات و در میان کاربران، و همچنین تأثیر آنها بر بازیابی اطلاعات، تأیید شده است (گل‌تاجی و بذرگر، ۱۳۸۹؛ عبدالهی نورعلی و جوکار، ۱۳۸۸؛ راثی ساربانقلی، ۱۳۸٤الف و ب). همچنین، در مطبوعات فارسی‌زبان وقوع اشتباهات آوایی، دستوری و واژگانی و در عین حال رعایت نکردن دستور خط زبان فارسی مشاهده شده است (ذوالفقاری و همکاران، ۱۳۸۵). با این حال، در رابطه با چالش پیوسته یا جدانویسی، یکدستی در الگوی نگارش و انطباق بالا با دستور خط فرهنگستان زبان و ادب فارسی گزارش شده است (فتاحی و آخشیک، ۱۳۹۱). آشکار است که صرف رویداد هر چالش، ضرورت لحاظ کردن آن در الگوریتم‌های بازیابی یا راهبرد جستجوی کاربران را توجیه نمی‌کند، زیرا ممکن است نگارندگان در رفتار نگارشی معمول خود، یک صورت نوشتاری را به صورت دیگر اولویت دهند و از صورت(های) دیگر صرف نظر نمایند یا به‌ندرت از آنها استفاده کنند. در این صورت، آشکار است که هزینه ـ سودمندی سامانه ایجاب می‌کند که الگوریتم را به ازای دستاوردی بسیار اندک، پیچیده نسازیم. از این رو، ضروری است رفتار عملی نگارندگان متن بررسی شود، تا میزان تنوع الگوهای نگارشی و اثرگذاری آنها بر جامعیت بازیابی اطلاعات، آشکار گردد.

بدین منظور، تحقیق حاضر می‌کوشد با بررسی تنوع الگوهای نگارشی در متون پیکرهٔ همشهری، رفتار نگارشی کلی نگارندگان فارسی را روشن سازد و لزوم در نظر گرفتن تنوع نگارشی در الگوریتم‌های نمایه‌سازی، الگوریتم‌های بازیابی یا راهبرد جستجوی کاربران را به بوتهٔ آزمون گذارد. همچنین، با مقایسهٔ این رفتار با دستور خط رسمی فارسی مصوب فرهنگستان زبان و ادب فارسی[۱]، میزان انطباق بین رفتار نگارشی نگارندگان و دستور خط رسمی، آشکار خواهد گردید.

همهٔ چالش‌های شناسایی شده (ستوده و هنرجویان، ۱۳۹۱) به دلایل گوناگون قابلیت بررسی در این پژوهش را نداشت. از جمله، ارتباط با ابعاد و ویژگی‌های

---

۱. مصوب ۱۳۸۰/٤/۳۰ (http://www.persianacademy.ir/fa/first.aspx)

غیرنگارشی (مانند معناشناسی، دستور زبان یا حروف پیش‌گزیدهٔ سامانه)، نیاز به ابزارها و روش‌های متفاوت جهت بررسی جامع، بروز ریزش کاذب به دلیل رویداد بسیار بالا در پیکره. به این‌ترتیب، هفت نویسهٔ چالشی شامل همزه بر پایه الف (أ)، همزه بر پایه و (ؤ)، تنوین نصب (اً)، همزه پایانی (ء)، همزه مختوم به یا (ئی)، تای گرد (ة) وتشدید، جهت بررسی انتخاب شد.

پرسش‌های پژوهش

۱. فراوانی هریک از الگوهای چندنگارشی در متون فارسی چه اندازه است؟

۲. درهرالگوی چندنگارشی، کدام صورت از فراوانی بیشتری برخوردار است؟

۳. درصد انطباق فراوانی الگوهای چندنگارشی با دستورخط رسمی فارسی چقدر است؟

تعاریف مفهومی

پیکره: مجموعه‌ای از متون نوشتاری یا گفتاری آوانویسی شده است که می‌توان آن را به عنوان مبنایی برای تحلیل و توصیف زبانی به کار برد (کندی، ۱۹۹۸). پیکره می‌تواند ویژهٔ بررسی خاصی فراهم شود و یا دربرگیرندهٔ مجموعه عظیم و بی‌ساختاری از متون گوناگون باشد که برای منظورهای گوناگون به کار رود (عاصی، ۱۳۸۵). پیکره‌های ویژهٔ بازیابی اطلاعات، به هدف آزمایش اثربخشی فنون یا روش‌های خاص در بازیابی اطلاعات طراحی می‌شود و مشتمل بر مجموعه‌ای مشخص از مدارک است که ویژگی‌های متون، برای مثال ربط موضوعی آنها، از قبل مشخص شده است (مانینگ، راگاوان و شوتس، ۲۰۰۸).

ضریب درگیری: این شاخص نخستین بار در حوزهٔ تعلیم و تربیت، جهت تجزیه و تحلیل محتوای کتاب‌های درسی و میزان درگیری فعالانهٔ دانش‌آموزان با آموزش و محتوای یادگیری، به کار گرفته شد. نسبت بین مقوله‌هایی که دانش‌آموزان در آنها فعالانه به آموزش و یادگیری می‌پردازند، به مقوله‌هایی که در آنها دانش‌آموزان به معنای

واقعی به فعالیت علمی نمی‌پردازند، ضریب درگیری دانش‌آموز با محتوا را تشکیل می‌دهد (فضل‌اللهی و ملکی توانا، ۱۳۸۹).

تعاریف عملیاتی

الگوی چندنگارشی فارسی: نگارش یک حرف فارسی به بیش از یک صورت که سبب شکل‌گیری سبک‌های نگارشی متفاوت در میان نگارندگان متون فارسی می‌شود. وضعیت نگارش در دو الگو بررسی شده است: ۱) سره‌نگاری که منظور از آن پایبندی به شکل مرسوم سنتی رسم‌الخط عربی است که در آن عین نویسه چالشی به نگارش در می‌آید. ۲) فارسی‌نگاری، که منظور از آن حذف نویسه چالشی یا جایگزینی آن با نویسه‌ای دیگر است. آشکار است که در این میان، برخی مصداق‌ها ممکن است به طور محض سره‌نگاری شوند؛ یعنی تنها و تنها به شکل مرسوم عربی آن در پیکره پدیدار شوند و برخی دیگر با نگارش ترکیبی ظاهر شوند؛ یعنی یک مصداق گاه سره‌نگاری و گاه فارسی‌نگاری شود. صورت محتمل دیگر، فارسی‌نگاری محض است. با توجه به نبود ویژگی شاخصی برای تمایز، جستجو و بازیابی صورت نگارشی اخیر، بررسی آن در پژوهش حاضر ممکن نبود.

نگارنده: فردی که متن روزنامهٔ همشهری را حروف‌چینی نموده یا متن حروف‌چینی شده را ویرایش کرده است، اعم از حروف‌چین، خبرنگار، ویراستار، اعضای هیئت تحریریه روزنامه و جز آن.

مصداق: واژگانی که در نگارش آنها یکی از حروف چالشی نگارشی به کار رفته باشد. برای نمونه، «تأیید» و «تأکید» دو مصداق برای چالش «همزه بر پایه الف» به شمار می‌آیند.

ضریب درگیری انطباق با فرهنگستان: یا به اختصار ضریب درگیری در پژوهش حاضر، عبارت است از نسبت بین فراوانی واژه‌های منطبق با دستور خط رسمی فارسی و فراوانی واژه‌هایی که با این دستور خط انطباق ندارند.

روش‌شناسی پژوهش

پژوهش حاضر به روش تحلیل محتوای مفهومی[1] صورت گرفت. متون موجود در ویرایش دوم پیکرهٔ همشهری[2] که توسط آزمایشگاه پایگاه داده دانشگاه تهران تهیه و به صورت دستی برچسب‌گذاری شده، برای بررسی انتخاب گردید. پس از اخذ مجوز دسترسی از آزمایشگاه پایگاه داده، کل محتوای پیکره بارگذاری شد.

انتخاب این پیکره برای بررسی، به چند دلیل صورت گرفت: نخست، این مجموعه بزرگترین پیکرهٔ متنی فارسی و مشتمل بر بیش از ۱۶۰،۰۰۰ مقاله خبری در موضوعات مختلف در یک بازه زمانی ۱۲ ساله (۱۳۷۵–۱۳۸۶)است (آل احمد[3] و همکاران، ۲۰۰۹). همچنین، در تهیه متن یک روزنامه، طیف نسبتاً متنوعی از افراد با سطوح متفاوتی از تحصیلات و آگاهی زبانی دخالت دارند، مانند حروف‌چینان که متن مخابره شده را حروف‌چینی می‌کنند، گزارشگران، خبرنگاران و اعضای هیئت تحریریه روزنامه و دیگر نگارندگان متون الکترونیکی که شخصاً به حروف‌چینی متن خود می‌پردازند و متون آنها به طور مستقیم روبرداری می‌شود. تنوع طیف نگارندگان متن روزنامه، سبب بازتاب عادت‌های نگارشی مختلف در متن روزنامه خواهد شد. از آنجا که بسیاری از مردم از میان مطالب مکتوب مختلف، تنها به خواندن نشریات (و آن هم اغلب نشریه‌ای خاص) اکتفا می‌کنند، نثر مطبوعات، خواه ناخواه بر چگونگی به کارگیری زبان یا مهارت نوشتن آنان تأثیر می‌گذارد (ذوالفقاری و همکاران، ۱۳۸۵). بنابراین، روزنامه‌ها نمونه مناسبی برای ملاحظهٔ الگوهای نگارشی غالب در میان مردم جامعه به نظر می‌رسند. نکته آخر و بسیار مهم آن است که در تهیهٔ روزنامه‌ها، به دلیل سرشت روزنگاشت آنها، سرعت مخابره خبر یا تهیهٔ مقالات به روز، اهمیت بسیار دارد. سرعت در نگارش، باعث می‌شود نگارندگان به طور ناخودآگاه و غیرفعالانه به نگارش متن بپردازند. از این رو، در متن حروفچینی شده، عادات ناخودآگاه آنان بازتاب می‌یابد و رفتار نگارشیِ اندیشیده و آگاهانهٔ آنان مشهود نیست.

---

1. Conceptual Content Analysis.
2. http://ece.ut.ac.ir/dbrg/hamshahri.
3. Ale Ahmad.

روش و ابزار گردآوری داده‌ها

به منظور جستجو در محتوای متنی پیکره، نرم‌افزارهایی بررسی شدند[1] که هیچ‌یک برای جستجوی نویسه‌ها و واژه‌ها و همچنین گزارش نتایج در قالب مورد نیاز این پژوهش، مناسب تشخیص داده نشد. از این رو، نرم‌افزاری ویژهٔ جستجو در پیکرهٔ همشهری، توسط یک متخصص رایانه طراحی شد.

روایی ابزار پژوهش

از آنجا که پیکره‌ها عموماً با هدف تحقیقات بازیابی موضوعی طراحی می‌شوند، ویژگی‌های املایی متون را مشخص نمی‌سازند. بنابراین، پیکرهٔ همشهری مختصاتی را در اختیار نمی‌گذارد که بر پایهٔ آن بتوان دربارهٔ صحت و دقت عملکرد نرم‌افزار قضاوت کرد. از این رو، به منظور آزمایش قابلیت اطمینان نتایج به دست آمده از نرم‌افزار، یکی از فایل‌های پیکره به صورت تصادفی انتخاب و فراوانی رویداد ۵ نویسهٔ چالشی (شامل تشدید، تنوین نصب، «أ»، «ئ» و «ء») در آن به صورت دستی محاسبه شد. سپس فراوانی این نویسه‌ها با استفاده از نرم‌افزار به دست آمد. در نهایت، به کمک نرم‌افزار SPSS، میزان همبستگی میان این دو دسته فراوانی با استفاده از آزمون ضریب همبستگی پیرسون محاسبه شد. نتیجه، همبستگی بسیار قوی را نشان داد (N=5, r=1, sig.=0.01) که دقت و صحت عملکرد نرم‌افزار را تأیید می‌کند.

بازیابی نویسه‌های چالشی و مصداق‌های آنها

به منظور اطمینان از یافتن همهٔ مصداق‌های دارای حروف چالشی در پیکره، از تعیین مصداق‌ها از قبل خودداری و تلاش شد تا جستجو در ریزترین سطح ممکن، یعنی تک‌نویسه صورت گیرد. به این ترتیب، واژه‌های دربردارندهٔ آن نویسه یعنی مصداق‌های واژه‌ای آن، بازیابی گردید. سپس، هر مصداق با نگارش‌های مختلف آن جستجو و فراوانی هریک ثبت شد. برای مثال، با جستجوی نویسه «ّ»،واژه‌های حاوی

1.http://lucene.apache.org,http://terrier.org,
http://lemurproject.org/lemur/retrieval.php.

این نویسه شناسایی و در گام بعد هر یک از واژه‌ها، یک بار با علامت تشدید و بار دیگر بدون آن، مورد جستجو قرار گرفت. ذکر چند نکته در این باره ضروری می‌نماید:

۱) آن دسته از واژه‌های عربی که بخشی از یک آیهٔ قرآن کریم و یا حدیثی از معصومین یا یک جملهٔ عربی بوده‌اند، در محاسبهٔ فراوانی آن واژه لحاظ نشده‌اند، زیرا این واژگان از سبک نگارشی زبان عربی پیروی می‌کنند که ممکن است چالش‌های آن با چالش‌های زبان فارسی متفاوت باشد.

۲) واژه‌های هم‌نگاشت[۱] از پژوهش حذف شدند، زیرا نرم‌افزار قادر به تمایز آنها نبود و در عین حال، به دلیل رخداد فراوان آنها، امکان وارسی متن پیکره برای درک معنا و در نتیجه تعیین شکل نگارشی درست آنها وجود نداشت.

۳) به دلیل بروز پاره‌ای اختلال‌های نویسه‌ای در پیکره، تعیین مرز واژه‌ها بر اساس علایم سجاوندی یا فاصله، با خطای زیاد همراه بود. از این رو، از مرزبندی کلمات در طراحی نرم‌افزار خودداری شد. این امر، مرحله جستجوی مصداق‌ها را با ریزش کاذب همراه کرد. برای مثال «سید»، هم به شکل واژه و هم پاره‌واژه (مانند «اسید»، «رسید»، و «پرسید») بازیابی شد. بنابراین، در گزارش نتایج، همهٔ واژه‌ها به صورت دستی بررسی و پس از حذف موارد ریزش کاذب، فراوانی مصداق‌ها محاسبه گردید.

روش تجزیه و تحلیل

به منظور بررسی رفتار نگارشی نگارندگان، از آمار توصیفی (شامل فراوانی و درصد) استفاده شد. ضریب درگیری، بر پایهٔ نسبت فراوانی واژگان منطبق با دستور خط رسمی فارسی به واژگان نامنطبق با این دستور محاسبه شد. چنانچه رفتار نگارندگان در دو گروه (منطبق و نامنطبق با دستور خط) با هم یکسان باشند، ضریب درگیری به سمت یک میل می‌کند.

---

۱. مانند «مسکن» به معنی محل سکونت یا آرام‌بخش، و «رویت» به معنی دیدن ، یا روی تو.

یافته‌ها

بر اساس یافته‌ها که بخشی از آنها در جدول ۱ آمده است، در تمام نویسه‌های چالشی، فراوانی الگوی ترکیبی بیش از الگوی سره‌نگاری محض است؛ یعنی مصداق‌ها در غالب موارد، گاه سره‌نگاری و گاه فارسی‌نگاری شده‌اند. این امر نشانگر آن است که نادیده گرفتن نویسه‌های چالشی در راهبرد جستجوی کاربر یا در الگوریتم‌های سامانه‌های بازیابی، سبب از دست رفتن بخشی از مدارک می‌شود. با این حال، در این الگو، در همهٔ نویسه‌ها اکثریت با فارسی‌نگاری است، به نحوی که در بیشتر موارد، جستجو با کلماتی که در آنها نویسهٔ چالشی حذف یا با نویسه‌ای دیگر جایگزین شده است، به بازیابی بیش از ۹۰٪ واژگان منجر می‌شود. تنها استثنا در این باره، دو چالش همزه بر پایه «و» و تنوین است که این مقدار در آنها به حدود ۷۰٪ می‌رسد.

جدول ۱: فراوانی صورت‌های نگارشی نویسه‌های چالشی در پیکرهٔ همشهری

چالش	صورت‌های نگارشی		مصداق‌ها		رویداد واژه‌ها	
			فراوانی	درصد	فراوانی	درصد
تشدید (ّ)	سره‌نگاری محض		۱۲	۳/۰۵	۱۵	۰/۰۰
	ترکیبی	سره‌نگاری	۳۸۲	۹۶/۹۵	۱۰۲۸	۰/۰۳
		فارسی‌نگاری			۳,۳۶۷,۵۷۳	۹۹/۹۷
	جمع		۳۹٤	۱۰۰	۳,۳٦۸,٦۱٦	۱۰۰
تای گرد (ۀ)	سره‌نگاری محض		۱۸	۲۹/۰۳	۱۹	۰/۳۵
	ترکیبی	سره‌نگاری	٤٤	۷۰/۹۷	۳۱٦	۵/۷۳
		فارسی‌نگاری			۵,۱۷۷	۹۳/۹۲
	جمع		٦۲	۱۰۰	۵,۵۱۲	۱۰۰
همزه پایانی (ء)	سره‌نگاری محض		۳	۲/٤۸	۵	۰/۰۰
	ترکیبی	سره‌نگاری	۱۱۸	۹۷/۵۲	٤۰,۱۰۲	۸/٤۹
		فارسی‌نگاری			٤۳۲,۱۳٤	۹۱/۵۱
	جمع		۱۲۱	۱۰۰	٤۷۲,۲٤۱	۱۰۰

همزه بر پایه «و»	سره‌نگاری محض		۰	۰/۰۰	۰	۰/۰۰
	ترکیبی	سره‌نگاری	۲۸	۱۰۰	۳۶,۶۲۶	۲۸/۵۵
		فارسی‌نگاری			۹۱,۶۶۸	۷۱/۴۵
	جمع		۲۸	۱۰۰	۱۲۸,۲۹۴	۱۰۰
همزه مختوم به یا (ئی)	سره‌نگاری محض		۹	۹/۳۸	۶۳۵	۰/۴۶
	ترکیبی	سره‌نگاری	۸۷	۹۰/۶۲	۷۶۶۲	۵/۵۵
		فارسی‌نگاری			۱۲۹,۶۸۲	۹۳/۹۹
	جمع		۹۶	۱۰۰	۱۳۷,۹۷۹	۱۰۰
همزه بر پایه الف (أ)	سره‌نگاری محض		۰	۰/۰۰	۰	۰/۰۰
	ترکیبی	سره‌نگاری	۵۴	۱۰۰	۱۳۶,۹۰۷	۲۷/۸۳
		فارسی‌نگاری			۳۵۵,۰۰۷	۷۲/۱۷
	جمع		۵۴	۱۰۰	۴۹۱,۹۱۴	۱۰۰
تنوین نصب (اً)	سره‌نگاری محض		۱۲	۶/۳۸	۱۳	۰/۰۰
	ترکیبی	سره‌نگاری	۱۷۶	۹۳/۶۲	۱۰۹,۱۵۵	۲۶/۶۹
		فارسی‌نگاری			۲۹۹,۸۵۰	۷۳/۳۱
	جمع		۱۸۸	۱۰۰	۴۰۹,۰۱۸	۱۰۰

در جدول ۲ نمونه‌هایی از واژه‌هایی که سره‌نگاری محض شده‌اند، معرفی شده است. چنان که مشاهده می‌شود، این مصداق‌ها عمدتاً از واژه‌های عربی هستند که در زبان فارسی رواج کمتری دارند. فراوانی رویداد هر مصداق (جدول ۱) نیز مؤید بروز بسیار اندک این واژه‌ها در پیکرهٔ مورد بررسی است. بر این اساس، احتمال این که در زبان فارسی عمومی، واژه‌ای که به طور بالقوه دارای نویسهٔ چالشی است، لزوماً با درج این نویسه نگاشته شود، بسیارضعیف است.

جدول ۲: نمونه‌هایی از واژه‌های سره‌نگاری شده

نویسه چالشی	واژه‌ها
تشدید	«انیّه»، «متطّب»، «ادسّر»، «مهنّد»، «باهر النّور»، «ظلیّه»، «ملکوتییّن»، «احدیّات»، «تمطّق»، «تدنّی»، «جرّه» و «علی السّوا»

تای گرد	«القدوة العارفین»، «المرقاة»، «تکملة الاصناف»، «سفینة البحار»، «روضة الشهدا»، «مادة المواد»، «معرفة النفس»، «علیه الصلوة»، «زبدة التواریخ»، «صلواة الله»
همزه پایانی	«رشاء»، «نصحاء»، «استحصاء»
ئی	«الظوائی»، «المسائی»،«بطئی»، «مرجئی»
تنوین	«عنفاً»، «متبرکاً»، «مزیداً»، «مترسلاً»، «متمرداً»، «غریزتاً»

وضعیت نگارش حروف چالشی بر اساس دستور خط رسمی

با توجه به آنچه از دستور خط فرهنگستان بر می‌آید[1]، فرهنگستان جز در مورد چالش‌های همزهٔ پایانی و تشدید، نگارش حروف چالشی را به همان شکل مرسوم سنتی خود توصیه می‌کند. برای نمونه، درباره حرف همزه میانی ذکر شده است: اگر حرف پیش از آن مفتوح باشد، علامت همزه روی کرسی«ا» نوشته می‌شود، مگر آن که پیش از آن، مصوت «ای»، «او» و یا «ـِـ» باشد، که در این صورت روی کرسی «ی» نوشته می‌شود؛ مانند «رأفت»، «تأسف»، «مأنوس». اگر حرف پیش از آن مضموم باشد، روی کرسی «و» نوشته می‌شود، مگر آن که پیش از آن، مصوت «او» باشد، که در این صورت روی کرسی «ی» نوشته می‌شود؛ مانند «رؤیا»، «رؤسا»، «مؤسسه». چنانچه حرف پیش از آن مفتوح یا ساکن و پس از آن حرف «آ» باشد، به صورت ـ آ/ آ نوشته می‌شود، مانند «مآخذ»، «لآلی»، «قرآن». از این دستور بر می‌آید که فرهنگستان به هیچ روی، جایگزینی حرف همزهٔ میانی را با حروف دیگر مانند «ا»، «و»، یا «ی» مد نظر نداشته است.

دستور فرهنگستان دربارهٔ چالش تشدید چنین است: « گذاشتن تشدید همیشه ضرورت ندارد، مگر در جایی که موجب ابهام و التباس شود، که یکی از مصداق‌های آن، هم‌نگاشت‌ها است، مانند معین/ معیّن؛ علی/علّی» تنها موردی که فرهنگستان درج تشدید را ضروری دانسته، در متون آموزشی برای نوآموزان و غیر فارسی‌زبانان و نیز در اسناد دولتی است.

در خصوص چالش همزهٔ پایانی، دستور این گونه بیان شده است: اگر حرف پیش

1. http://www.persianacademy.ir/fa/first.aspx.

از آن، مصوت «آ»، «او» و یا «ای» باشد، بدون کرسی نوشته می‌شود. با این حال، یک تبصره این دستور را متفاوت می‌سازد: کلماتی مانند «انشاء»، «املاء»، «اعضاء» در فارسی بدون همزه پایانی هم نوشته می‌شوند که صحیح است.

یک استثنا نیز در دستور نگارش «ة» وجود دارد. طبق تصریح فرهنگستان، گاهی در بعضی ترکیبات عربی رایج در فارسی، «ة» با «ت» جایگزین می‌گردد که آن هم صحیح است. مانند «حجت الاسلام»و «آیت الله». با این حال، این دستور شامل جایگزینی «ة» با «ه» نمی‌شود.

بنابراین، دستور فرهنگستان هر شکل نگارشی مرسوم در دو چالش تشدید، و همزه پایانی را مجاز و در نتیجه منطبق با شیوهٔ نگارش رسمی فارسی می‌داند. به این ترتیب، می‌توان رفتار نگارشی نویسندگان را در مورد ۵ چالش دیگر در دو گروه منطبق و نامنطبق با دستور خط رسمی زبان فارسی دسته‌بندی کرد و مورد مطالعه قرار داد. همچنین، برخی از مصداق‌های یافت شده مانند «پارسائی»، «ایتالیایی» و «لاشائی» به غلط با «ئی» نوشته شده‌اند. این کلمات در نگارش صحیح خود واجد همزه نیستند و بنابراین، مصداق‌های واقعی چالش همزه مختوم به یا (ئی) به شمار نمی‌آیند. این واژگان نیز به دلیل این که تنها حکایت از اشتباهات مصطلح نگارشی دارند، در این پژوهش بررسی قرار نشدند.

فرهنگستان، نگارش حروف چالشی مورد بحث را به همان شکل سره‌نگاری محض عربی تجویز می‌کند. تنها استثنا، دو نویسه همزه پایانی و تشدید است که نگارش‌های متفاوتی برای آنها بسته به شرایط، مجاز دانسته شده است. به این ترتیب، انطباق رفتار نگارشی نگارندگان با دستور خط رسمی، تنها در مورد ۵ چالش دیگر قابل مطالعه است.

جدول۳. وضعیت انطباق نگارش واژه‌های پیکره همشهری با دستور خط رسمی فارسی

چالش	واژه‌های نامنطبق با دستور خط رسمی فارسی		ضریب درگیری
	فراوانی	درصد	
همزه مختوم به یا (ئی)	۱۲۹,۶۸۲	۹۳/۹۹	۰/۰۶۴
تای گرد (ة)	۵۱۷۷	۹۳/۹۳	۰/۰۶۵
تنوین نصب (اً)	۲۹۹,۸۵۰	۷۳/۳	۰/۳۶۴

همزه با پایه الف (أ)	۳۵۵,۰۰۷	۷۲/۱	۰/۳۸٦
همزه با پایه و (ؤ)	۹۱,٦٦۸	۷۱/۵	۰/٤
جمع	۸۸۱,۳۸٤	۹٦/۷	۰/۰۳۳

ضریب درگیری کلی، ۰/۰۳۳ به دست آمده است. به این ترتیب، در بخش بسیار ناچیزی از رویدادهای چالش‌ها (۳/۳٪)، دستور خط فارسی رسمی رعایت شده است و در مقابل، در اکثریت قریب به اتفاق (۹٦/۷٪) گرایش نگارندگان متون پیکره همشهری به نادیده گرفتن این دستور بوده است (جدول ۳).

چنانچه هریک از چالش‌ها را به تفکیک مد نظر قرار دهیم، چالش همزه با پایه و (ؤ) با ضریب درگیری ۰/٤ بیشترین میزان انطباق و چالش همزه مختوم به یا (ئی) با ۰/۰٦٤ کمترین میزان انطباق را با دستورخط فرهنگستان نشان می‌دهند. به این ترتیب، در همهٔ چالش‌های مورد بررسی، فراوانی واژگان نگاشته شده نامنطبق با دستور خط فرهنگستان زبان و ادب فارسی، با تفاوت بسیار چشمگیری از واژگان منطبق با این دستور پیشی گرفته است.

بحث

مصداق‌های هفت چالش مورد بررسی، گاه سره‌نگاری و گاه فارسی‌نگاری شده اند. به جز دو نویسه «ؤ» و «أ»، درپنج چالش دیگر، تعدادی از مصداق‌ها تنها سره‌نگاری شده‌اند. این امر برخورد احتیاط‌آمیز با این نویسه‌ها را به هنگام طراحی الگوریتم‌ها یا تدوین راهبرد جستجو ایجاب می‌کند، زیرا در غیر این صورت میزان بازیافت این مصداق‌ها به صفر خواهد رسید و کاربر هیچ‌گونه پاسخی را از سامانه دریافت نخواهد کرد. این امر به‌ویژه در مورد نویسه «ة» صدق می‌کند که شمار مصداق‌های تک‌شکل آن (۲۹/۰۳٪) نسبتاً قابل توجه است. به این ترتیب، درصد مصداق‌هایی که در صورت بهنجار نشدن این نویسه در الگوریتم‌های سامانه بازیابی به طور کامل از دسترس کاربران دور خواهند ماند، قابل توجه خواهد بود. با این حال، شمار این مصداق‌ها در دیگر نویسه‌ها بسیار اندک است و نشان از آن دارد که نگارندگان متون پیکره همشهری به طور کلی تمایل به فارسی‌نگاری دارند. به این ترتیب، با وجود نگرانی‌ها و

هشدارهای پژوهشگران درباره تأثیر گونه‌گونی نگارش بر اثربخشی بازیابی اطلاعات (حری، ۱۳۷۲؛ گل‌تاجی و بذرگر، ۱۳۸۹؛ عبدالهی نورعلی و جوکار، ۱۳۸۸) به نظر می‌رسد دست کم در مورد هفت نویسهٔ مورد بررسی در این پژوهش، نادیده انگاشتن این چالش‌ها، اثربخشی بازیابی اطلاعات را به لحاظ میزان بازیافت چندان متأثر نسازد. چنان که پیشتر بیان شد، احتمال دارد برخی مصداق‌ها، فارسی‌نگاری محض شده باشند؛ یعنی صرفاً با حذف یا جایگزینی نویسهٔ چالشی نوشته شده باشند که در این پژوهش امکان بررسی آنها نبود. چنانچه این احتمال را نیز در نظر آوریم، وزنه به نفع فارسی‌نگاری سنگین‌تر خواهد شد.

قضاوت دقیق‌تر دربارهٔ میزان تأثیر چالش‌ها بر اثربخشی بازیابی اطلاعات، به میزان رواج این مصداق‌ها در متون بستگی دارد. در نویسه‌های تشدید، «تای گرد»، همزه پایانی، در بیش از ۹۰٪ موارد، فارسی‌نگاری روی داده است. بنابراین، رفتار کلی نگارندگان متون مورد بررسی، تمایل به فارسی‌نگاری را نشان می‌دهد و به نظر می‌رسد لحاظ نکردن این نویسه‌ها در الگوریتم‌های سامانه‌ها یا راهبردهای جستجوی کاربران، آسیب چندانی به جامعیت بازیابی وارد نمی‌آورد.

چالش تشدید به درج نویسهٔ افزوده‌ای نیاز دارد که نه تنها کلید آن چندان شناخته شده نیست، بلکه گویش‌ور فارسی، ضرورت درج آن را به جهت کمک به شناخت کلمه چندان احساس نمی‌کند. حرف «تای گرد» نیز در زبان فارسی مهجور است و رواج چندانی ندارد. بنابراین، کلید مربوط به این حرف نیز در اثر کاربرد کم، در میان نگارندگان کم و بیش ناشناخته می‌ماند، به‌ویژه این که درج این حروف، نیاز به استفاده از کلید مبدله[1] دارد که یک گام فعالانهٔ اضافی را به هنگام حروفچینی بر نگارنده تحمیل می‌کند. مسئله مهم دیگر این که بسیاری از فونت‌های رایانه‌ای این حرف را به شکل «ۀ» درج می‌کنند و نه «ۀ». در واقع این فونت‌ها، شکل صحیح این حرف را دارا نیستند. این نکات، انگیزه نگارندگان را برای سره‌نگاری این حرف کاهش می‌دهند.

همزه پایانی، نه تنها مستلزم درج نویسه‌ای اضافی است، بلکه، از حروف عربی

1. Shift.

محض به شمار می‌آید که در الفبای زبان فارسی وجود ندارد. نگارنده فارسی‌زبان، عموماً – آگاهانه یا ناخودآگاه– سعی در نگارش کلمات به صورت هرچه شبیه‌تر به شیوهٔ نگارش فارسی خواهد داشت، مگر در جایی که ضرورت رعایت رسم‌الخط زبان عربی را احساس کند. برای مثال، به هنگام تلاش برای درج عین کلمه، یا نقل عبارتی از متون عربی یا قرآن کریم و جز آن. مسئله مهم دیگر آن که در زبان فارسی، نویسه همزه پایانی اغلب تلفظ نمی‌شود، جز در مواردی که علامت جمع، «یا»ی نکره، نسبت، مضاف‌الیه یا صفت به کلمات واجد این نویسه اضافه می‌شود و در آن صورت این حرف یا با آوای اصلی خود خوانده می‌شود یا با حرف «ی» جایگزین می‌شود (مثل سمائی یا سمایی). این مسئله قطعاً در پرهیز از درج این نویسه تأثیر بسزایی دارد. فرهنگستان زبان و ادب فارسی نیز نگارش برخی کلمات واجد همزه پایانی را بدون درج این نویسه مجاز دانسته است. بنابراین، نگارندگان بنا بر اصل کمترین کوشش، دلیلی برای درج آن در نگارش کلمات نیافته، اغلب تمایل به حذف آن خواهند داشت.

تمایل به فارسی‌نگاری در چالش «ؤ»ـ در مقایسه با نویسه‌های پیش‌گفته ـ کمتر است. شاید دلیل کاربرد بیشتر این نویسه را بتوان در این نکته جستجو کرد که حرف «و» در برخی کلمات چون «تو»، «خود»، «خورد»، مصوت کوتاه «ـُ» و در اکثر کلمات، آوای بلند «او» را ایجاد می‌کند، مانند «دوست»، «گوش»، «زود». در حالی که نویسه «ؤ» در واقع «همزه» است که بر کرسی «و» قرار گرفته است و آوایی کاملاً متفاوت را تولید می‌کند. نگارش کلمات واجد نویسه «ؤ» به شکل «و» به خوانش بد این کلمات، حداقل در نگاه اول، یا برای نوآموزان، کم‌سوادان و کسانی که به زبان و نگارش فارسی تسلط ندارند، منجر خواهد شد. مثالی از این مطلب، نحوهٔ خوانش کلماتی چون «مؤثر»، «موذن»، «روسا»، با صدای «او» یا «ـُ» می‌باشد. ممکن است به‌کارگیری بیشتر این نویسه، ریشه در تلاش نگارندهٔ فارسی زبان برای کمک به خوانش بهتر متن داشته باشد.

در دو چالش تنوین نصب و همزه بر پایهٔ الف نیز، گرایش کلی نگارندگان به فارسی‌نگاری است. با این حال، سره‌نگاری این دو نویسهٔ فراوانی قابل ملاحظه‌ای را نشان می‌دهد که قابل چشم‌پوشی نیست. تمایل به فارسی‌نگاری و وقتگیر بودن استفاده از کلید مبدله برای درج نویسه «أ»، نگارندگان را به جایگزینی این نویسه با حرف و

کلید ساده‌تر و شناخته شده‌تر «ا» سوق داده است. همچنین، کلماتی که با تنوین نصب نوشته شده‌اند، کاملاً عربی هستند و کاربرد آنها نه در زبان فارسی عمومی، بلکه عمدتا در متون فقهی، حقوقی و مذهبی است که از کلمات عربی صرف استفاده بیشتری می‌شود. شاید این امر را بتوان دلیلی بر پایبندی نگارندگان به استفاده از تنوین به منظور هرچه نزدیک‌تر ساختن آنها به صورت عربی دانست. به طور کلی، التزام نداشتن نگارندگان به رعایت رسم‌الخط عربی، عدم احساس ضرورت سره‌نگاری عربی، نامأنوس بودن و یا رواج کم برخی کلمات، تمایل به نزدیک ساختن نگارش به محاوره، تلاش برای افزایش سرعت حروفچینی و – به گفته مانینگ و همکاران (۲۰۰۸) ـ عادات بازمانده از قدیم[۱]، از دلایل احتمالی این گرایش در نگارش است.

شایان ذکر است، این پژوهش بر بررسی رویداد واژگان تمرکز داشته و فراوانی مدارک محاسبه نشده است. آشکار است که به دلیل احتمال رویداد چندبارهٔ مصداق‌ها و واژگان در مدرکی واحد، فراوانی مدارک مربوط می‌تواند برابر یا کمتر از فراوانی واژه‌ها باشد. به این ترتیب، احتمال این که درصد مدارک بازیابی نشده از این هم کمتر باشد، وجود خواهد داشت.

با وجود این واقعیت، میزان واژگانی که به سبب توجه نکردن به نویسه‌های چالشی تنوین نصب، همزه بر پایهٔ الف و همزه با پایه واو بازیابی نخواهند شد، حدود یک‌چهارم از کل واژگان را تشکیل می‌دهد که برخورد احتیاط‌آمیز به هنگام سیاست‌گذاری به منظور طراحی سامانه و تدوین الگوریتم‌های نمایه‌سازی و بازیابی یا تدوین راهبرد جستجو را می‌طلبد.

رفتار نگارشی رایج در میان نگارندگان مورد بررسی، به شدت از آنچه فرهنگستان به عنوان دستور خط رسمی تجویز کرده است، فاصله دارد. دلایل چندی را می‌توان در این راستا برشمرد. از یک سو، اعمال دستور خط رسمی مصوب فرهنگستان به هیچ روی برای افراد، نهادها یا سازمان‌ها الزام‌آور نیست. از سوی دیگر، فرهنگستان بیشتر به سره‌نگاری عربی پایبند است. این در حالی است که رسم‌الخط عربی به دلیل

۱. منظور قبل از پیدایش کدهای اسکی پیشرفته و امکان درج نویسه‌های چالشی است.

دشواری و نأمانوس‌تر بودن، با اصل کمترین کوشش مغایرت دارد، زیرا نیازمند درج نویسه‌ای افزوده از سوی نویسنده (مانند نویسه تشدید و تای گرد)، آشنایی با برخی نویسه‌ها و کلیدهای مربوط وجستجوی فعالانه برای یافتن آنها در صفحه کلید رایانه (مانند تای گرد) است. همچنین، به نظر می‌رسد دستور خط فارسی نیازهای زبانی اساسی گویش‌وران فارسی را نادیده گرفته است. برخی نویسه‌ها (مانند همزه پایانی) در فارسی به تلفظ در نمی‌آیند یا به همان سختی ادا نمی‌شوند.

نتیجه‌گیری

بر پایهٔ یافته‌های این پژوهش، رفتار نگارشی نگارندگان متون پیکره همشهری، گرایش به ساده‌نگاری، تلاش در جهت هر چه «فارسی‌تر» کردن نگارش و عدم انطباق کلی با دستور خط رسمی فارسی را نشان می‌دهد. مسئله توانش خط فارسی در تولید آواهای فارسی و همچنین نیاز نداشتن به برخی حروف (مانند ص، ض، ذ، ع) که در عربی، برخلاف فارسی، برای تولید آواهای متفاوت به کار می‌روند، از دیرباز مورد توجه بوده است. فراوانی چشمگیر غلط‌های املایی در محیط‌های عمومی وب که کاربران در آنها آزادانه و به دور از نظارت ویراستاران می‌نویسند، خود مؤید احساس نیاز گویش‌وران فارسی به ساده‌سازی املای فارسی بر اساس نیازهای عملی خود است. آشکار است که بحث و بررسی دربارهٔ پیامدهای رعایت یا عدم رعایت دستور خط فارسی، در صلاحیت متخصصان زبان و ادب فارسی است. به هر حال، دستاوردی که پژوهش حاضر در بر دارد، آن است که این گرایش به نادیده گرفتن رسم الخط عربی و تلاش برای «فارسی‌تر» کردن آن باید در سامانه‌های بازیابی اطلاعات، خواه به هنگام نمایه‌سازی و خواه به هنگام بازیابی اطلاعات، مد نظر قرار گیرد.

در عین حال، درصدی از واژگان سره‌نگاری شده، به‌ویژه آنهایی را که به طور محض سره‌نگاری شده‌اند را نیز نباید از نظر دور داشت. میزان تخصصی بودن سامانهٔ بازیابی اطلاعات و جامعه هدف آن، در تصمیم‌گیری برای بهنجار کردن یا نکردن چالش‌های نگارشی برخوردار، تعیین کننده است. بر این اساس، بسته به اهداف سامانهٔ بازیابی اطلاعات، ویژگی‌های جامعه هدف، پوشش موضوعی و میزان جامعیت مورد نظر، می‌توان سامانه‌هایی با سطوح مختلف حساسیت را طراحی نمود. در سامانه‌های

عمومی و بسیار بزرگ مانند محیط‌های وبی که جامعیت صددر صدی مد نظر نیست، می‌توان نویسه‌های چالشی را در الگوریتم‌های نمایه‌سازی یا بازیابی نادیده گرفت. این امر، به سادگی الگوریتم‌ها و در نتیجه افزایش کارآیی سامانه منجر خواهد شد. با این حال، در سامانه‌های تخصصی مانند پایگاه‌های مجلات و همچنین فهرست‌های کتابخانه‌ای که شمار مدارک به طور نسبی کمتر است و حوزه‌های موضوعی تخصصی وابسته به زبان عربی- مانند حقوق، ادبیات عرب، فلسفه اسلامی، الهیات و معارف اسلامی- را پوشش می‌دهند، به هنجارسازی این چالش‌ها در الگوریتم‌ها چندان از کارآیی سامانه نخواهد کاست و در عین حال به افزایش قدرت جستجوی واژگان چالشی نیز منجر خواهد شد.

با آن که احتمال می‌رود گرایش به ساده‌نگاری که نزد نگارندگان متون مشاهده شده است، در میان کاربران پایگاه‌های اطلاعاتی و شبکه‌ها نیز رواج داشته باشد، به منظور اطمینان یافتن از انطباق رفتار نگارشی این دو گروه - که موفقیت بازیابی اطلاعات را تضمین خواهد نمود- ضروری است پژوهشی در جهت مقایسهٔ این دو رفتار صورت گیرد. همچنین، در پژوهش حاضر، متون یک روزنامه در یک بازهٔ زمانی گسترده برای بررسی انتخاب گردید. آشکار است که طیف نگارندگان یک روزنامه، هر چند متنوع باشد، در مقایسه با تنوع طیف نگارندگان رسانه‌های متنی گوناگون، محدود است. توصیه می‌شود در پژوهش‌های دیگر، با انتخاب یک بازهٔ زمانی کوتاه‌تر، انواعی از رسانه‌ها (مانند کتاب، روزنامه، مجله، وبلاگ‌ها و وب‌سایت‌ها) و در نتیجه، طیف متنوع‌تری از نگارندگان متون، مورد بررسی قرار گیرد.

منابع

- آخشیک، سمیه سادات و رحمت‌الله فتاحی(۱۳۹۱). «تحلیل چالشهای پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاههای اطلاعاتی». کتابداری و اطلاع‌رسانی، ۱۶ (۳): ۳۰-۹.
- حری، عباس(۱۳۷۲). «کامپیوتر و رسم‌الخط فارسی». پیام کتابخانه، ۳ (۱) : ۱۱-۶.
- ذوالفقاری، حسن و همکاران (۱۳۸۵). «الگوهای غیر معیار در زبان مطبوعات». طرح پژوهشی ،به سفارش دفتر مطالعات و توسعه رسانه‌ها. معاونت مطبوعاتی و اطلاع‌رسانی وزارت ارشاد بازیابی به تاریخ ۲۵ شهریور ۱۳۹۲ از:

http://www.rasaneh.org/NSite/FullStory/News/?Id=832

- راثی ساربانقلی، محمدصابر(۱۳۸٤ الف). «بررسی مشکلات جستجو و بازیابی اطلاعات به زبان فارسی از اینترنت با مطالعه موردی بر روی کاربران مرکز اینترنت دانشگاه آزاد اسلامی واحد شبستر». پایان‌نامه کارشناسی ارشد. دانشگاه آزاد اسلامی، واحد تهران شمال.

ـــــــــــــــــــــ (۱۳۸٤ ب). «مهارت در جستجوی اطلاعات فارسی از اینترنت». ارتباط علمی، ۵ (۱): ۲۸-۱٦.

- ستوده، هاجر و زهره هنرجویان(۱۳۹۱). «مروری بر دشواری‌های نگارش فارسی در محیط دیجیتال و تأثیرات آنها بر اثربخشی بازیابی اطلاعات». کتابداری و اطلاع‌رسانی، ۱۵ (٤): ۹۲-۵۸.

- شهیدی، مجتبی، محسن صدیقی و کامران زمانی‌فر(۱۳۸۳). «روشی برای رفع چالش‌های محتواکاوی وب‌های فارسی». علوم و فناوری اطلاعات، ۲۱ (۲):٦۹-٤۷. بازیابی به تاریخ ۲۰ اسفند ۱۳۹۰ از:

http://jist.iranoc.ac.ir/browse.php?a_id=97&slc_lng=fa&sid=1&ftxt=1

- فضل‌الهی، سیف‌الله و منصوره ملکی توانا (۱۳۸۹). «روش‌شناسی تحلیل با تأکید بر تکنیک‌های خوانایی سنجی و تعیین ضریب درگیری متون». پژوهش، ۳: ۹٤-۷۱.

- عاصی، مصطفی (۱۳۸۵). «از پیکره زبانی تا زبان‌شناسی پیکره‌ای». پژوهشگران، ۸ و ۹. بازیابی به تاریخ ۲۵ مرداد ۱۳۹۳ از:

http://www.hawzah.net/fa/magazine/magart/6280/6293/69809

- عبداللهی نورعلی، محمدصادق و عبدالرسول جوکار(۱۳۸۸). «چالشهای شیوه نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب». مطالعات تربیتی و روانشناسی، ۳٦: ۹۰-٦۷.

- گل‌تاجی،مرضیه و سعیده بذرگر(۱۳۸۹). «بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی». کتابداری و اطلاع‌رسانی، ۱۳ (۲) :۲۱٤-۱۹۱ .

- محقق‌زاده، محمدصادق و کاظم زارعیان (۱۳۸۳). «ارائه راه حل برای برخی مسائل اتوماسیون نگارش فارسی». اطلاع‌رسانی، ۱۹ (۳- ٤) : ۱۰-۱.

- مرتضایی، لیلا (۱۳۸۱). «مسائل زبان و خط فارسی در ذخیره و بازیابی اطلاعات». اطلاع‌رسانی، ۱۷ (۱-۲) : ۷-۱.

- AleAhmad, A., H.Amiri, E.Darrudi, M.Rahgozar, &F.Oroumchian (2009). "Hamshahri: A standard Persian text collection".Knowledge Based Systems, 22 (5): 382-387, DOI: 10.1016/j.knosys.2009.05.002.
- Kennedy, Graeme. An Introduction to Corpus Linguistics. London : Longman,1998.
- Maning, CD., p.Raghavan, &H. Schutze.Introduction to Information Retrieval. Cambridge:Cambridge University Press, 2008.