

Persian Writing in GANJ: Investigating the Impact of Morphology, Semantics, and Writing Style on Iran's Treasure of Scientific and Technical Information¹

Somaye Sadat Akhshik

Assistante Professor, Department of Library and Information Studies, Faculty of Psychology and Education, Kharazmi University, Tehran, Iran (**Corresponding Author**).
akhshik@khu.ac.ir

Khadije Negahdari

M.A., Information Science and Knowledge, Kharazmi University, Tehran, Iran.
negahdari1371@gmail.com

Alireza Emami

Assistante Professor, Department of Persian Language and Literature, University of Tehran, Iran.
aremami@ut.ac.ir

Abstract

Objectives: Persian writing has some difficulties that neglecting can affect information retrieval. This study was conducted to investigate the effect of Persian writing problems on the retrieval of documents on the (GANJ) database based on morphological, semantical, and orthographical aspects.

Methods: This research is a practical-purpose study that was conducted in a qualitative manner using content analysis techniques. The used method was a researcher-made checklist. The research population was all the records recorded in the GANJ until the time of the survey. Sampling was done by standard sampling. In the category of conjugational problems, the impact of these problems on the retrieval of documents in the database was examined.

Results: In the semantic problem category, the effects of semantic ambiguity or semantic differences on the information retrieval of documents were examined. This case was done in the category of orthography problems to determine difficulties arising from morphological and written features of Persian writing.

Conclusions: According to the data obtained from searching for keywords related to each difficulty in the Ganj, it was observed that inconsistencies in the text affect the retrieval result.

1. **The article is taken from:** the master's thesis in the field of Knowledge and Information Science entitled: **Analysis of the effect of morphological, semantic and Persian writing problems on information retrieval in Iranian scientific database (Ganj)**. Student: Khadijeh Negahdari, Supervisor Professor: Dr. Somayeh Sadat Akhshik, Faculty of Psychology and Educational Sciences, Department of Knowledge and Information Science, Kharazmi University, Defense Date: 2020.

Cite: Akhshik, Somaye Sadat; Negahdari, Khadije & Emami, Alireza (2022). Persian Writing in GANJ: Investigating the Impact of Morphology, Semantics, and Writing Style on Iran's Treasure of Scientific and Technical Information. *Sciences and Techniques of Information Management*, 8(1): 193-220. DOI: 10.22091/stim.2021.6418.1505

Received: 2020-12-28 ; **Revision:** 2021-04-12 ; **Accepted:** 2021-04-19

© the authors / **Publisher:** University of Qom



The results of the present study showed that morphological, semantic, and orthographical problems affect the information retrieval results in the database. It was also found that in the morphological group only for the "adjective morphological morphemes", and in the orthography group, for the problems of "accent mark", "ی and ء transposition in Persian words", "writing of middle and end ء (tittle) with الف seat", "writing of middle and end ء (tittle) with واو seat", "remove or writing tilde Mark" and "writing of consonantal ی after inarticulate ء" have been deliberated and difficulties in the semantic group were completely neglected. Since Ganj is the basis of other Irandoc databases, regardless of morphological, semantic, and calligraphic problems can affect the work of other systems as well.

Keywords: Persian Writing, Information Retrieval, Ganj Database, Semantic problems, Writing style, Morphology problems, Iranian Research Institute for Information Science and Technology (IranDoc).

خط فارسی در گنج: واکاوی تأثیر مسائل صرف، معنا و رسم الخط بر بازیابی اطلاعات در پایگاه اطلاعات علمی ایران^۱

سمیه سادات آخشیک

استادیار، گروه علم اطلاعات و دانش‌شناسی، دانشکده روانشناسی و علوم تربیتی، دانشگاه خوارزمی، تهران، ایران
(نویسنده مسئول): akhshik@khu.ac.ir

خدیجه نگهداری

کارشناسی ارشد، گروه علم اطلاعات و دانش‌شناسی، دانشگاه خوارزمی، تهران، ایران. negahdari1371@gmail.com

علیرضا امامی

استادیار، گروه زبان و ادبیات فارسی، دانشگاه تهران، تهران، ایران. aremami@ut.ac.ir

چکیده

هدف: خط فارسی دشواری‌هایی دارد که بی‌توجهی به آنها، بازیابی اطلاعات را تحت تأثیر قرار می‌دهد. پژوهش حاضر با هدف بررسی تأثیر مشکلات خط فارسی در ابعاد صرفی، معنایی و رسم الخط بر بازیابی مدارک پایگاه گنج انجام شده است.

روش: این پژوهش از نظر هدف کاربردی بوده که به شیوه کیفی و با استفاده از تکنیک تحلیل محتوا انجام شد. ابزار مورد استفاده، سیاهه واری محقق ساخته بود. جامعه پژوهش نیز تمامی مدارک ثبت شده در گنج تا زمان اجرای پژوهش می‌باشد. نمونه‌گیری به صورت نمونه‌گیری معیاری انجام شد. با توجه به داده‌های به دست آمده، توصیف و تحلیل انجام گرفت.

نتایج: در دسته مشکلات صرفی، تأثیر مسائل و دشواری‌های مربوط به صرف یا ساخت‌واژه در بازیابی مدارک و در دسته مشکلات معنایی تأثیر دشواری‌هایی که از نظر معنا در واژه ابهام یا تفاوت معنایی ایجاد می‌کنند، بر بازیابی مدارک بررسی شد. در مشکلات رسم الخط، این بررسی برای دشواری‌های ناشی از ویژگی‌های ریخت‌شناسی و نوشتاری خط فارسی انجام شد.

۱. **پژوهش حاضر برگرفته از:** پایان‌نامه کارشناسی ارشد رشته علم اطلاعات و دانش‌شناسی، با عنوان **واکاوی تأثیر مشکلات صرفی، معنایی و رسم الخط فارسی بر بازیابی اطلاعات در پایگاه اطلاعات علمی ایرانیان (گنج)**، دانشجو: خدیجه نگهداری، استاد راهنما: سمیه سادات آخشیک، گروه علم اطلاعات و دانش‌شناسی، دانشکده روانشناسی و علوم تربیتی، دانشگاه خوارزمی تهران، ارائه شده در سال ۱۳۹۸ است.

استاد به این مقاله: آخشیک، سمیه سادات؛ نگهداری، خدیجه؛ امامی، علیرضا (۱۴۰۱). خط فارسی در گنج: واکاوی تأثیر مسائل صرف، معنا و رسم الخط بر بازیابی اطلاعات در پایگاه اطلاعات علمی ایران. *علوم و فنون مدیریت اطلاعات*، ۸(۱): ۱۹۳-۲۲۰. DOI: 10.22091/stim.2021.6418.1505

تاریخ دریافت: ۱۳۹۹/۱۰/۰۸ ؛ تاریخ اصلاح: ۱۴۰۰/۰۱/۲۳ ؛ تاریخ پذیرش: ۱۴۰۰/۰۱/۳۰

ناشر: دانشگاه قم

نتیجه‌گیری: پژوهش حاضر نشان داد مشکلات صرفی، معنایی و رسم‌الخط، نتایج بازیابی اطلاعات در گنج را تحت تأثیر قرار می‌دهند. همچنین در این پایگاه در گروه صرفی تنها برای تک‌واژه‌های صرفی صفت، و در گروه رسم‌الخط، برای مشکلات علامت تشدید، جابه‌جایی «ی» و همزه در کلمات فارسی، نگارش همزه میانی و پایانی با کرسی «الف»، نگارش همزه میانی و پایانی با کرسی «واو»، «حذف یا نوشتن علامت مد» و «نوشتن صامت «ی» بعد از «ه» غیرملفوظ، چاره‌اندیشی شده و در گروه معنایی به هیچ یک از دشواری‌ها توجهی نشده است.

کلیدواژه‌ها: مسائل صرفی، مسائل معنایی، رسم‌الخط، خط فارسی، بازیابی اطلاعات، پایگاه گنج، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک).

۱. مقدمه

خط، چهره مکتوب زبان است و همان‌گونه که زبان از مجموعه اصول و قواعدی به نام «دستور زبان» پیروی می‌کند، خط نیز باید پیرو اصول و ضوابطی باشد. تدوین ضوابط و قواعد خط فارسی خصوصاً در سال‌های اخیر که استفاده از رایانه در عرصه خط و زبان روزافزون شده و حروف‌نگاری و صفحه‌آرایی و ویرایش و نمونه‌خوانی و غیره در حوزه نگارش بر عهده رایانه قرار گرفته، ضرورت و اهمیت بیشتری پیدا کرده است (فرهنگستان زبان و ادب فارسی، ۱۳۹۴).

یکی از متغیرهای عمده در ذخیره و بازیابی اطلاعات فارسی، رسم الخط یا شیوه نگارش فارسی است و ورود ناگهانی رایانه‌ها، به صاحب‌نظران این حوزه مجال این را نداده که راه‌حلی بنیانی برای مقابله با این دشواری‌ها طراحی و عرضه کنند، چرا که این پیچیدگی در نظام دستی چندان مشکل‌آفرین نبوده است (حری، ۱۳۷۲). از آنجایی که زبان فارسی، در مواجهه با محیط الکترونیکی، از جهت شیوه نگارشی، دارای مشکلاتی است که بر کیفیت کاوش در محتویات آن تأثیر می‌گذارد، تأثیر برطرف‌سازی این موانع در طراحی هر پایگاه اطلاعاتی فارسی زبان بر میزان بازیابی رکوردهای مرتبط، چه از لحاظ کمیت و چه از لحاظ محتوای رکوردهای بازیابی شده، بر کسی پوشیده نیست (گل تاجی و بذرگر، ۱۳۸۹). البته این مسأله خاص زبان فارسی نیست و اغلب زبان‌ها با این مسأله مواجه هستند که این دسته مشکلات ناشی از ویژگی‌های خاص هر زبان و همین‌طور تفاوت در صورت نوشتاری و گفتاری واژه‌ها می‌تواند باشد.

پایگاه اطلاعات علمی ایران یا همان گنج^۱، دستاورد کاربرد فناوری اطلاعات در خدمات مدیریت اطلاعات علم و فناوری در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. این پایگاه هم‌اکنون مرجع بسیاری از پژوهشگران ایران و جهان بوده و روزانه بیش از ده هزار کاربر، ده‌ها هزار جست‌وجو در آن انجام می‌دهند. این سامانه بنیان سامانه‌های دیگری در ایرانداک همچون سامانه همانندجو، سامانه پیشینه پژوهش، و برخی از داشبوردهای رصدخانه پژوهش و فناوری نیز است. مدارک منتشر شده این پایگاه شامل مقالات مجله‌های علمی، گزارش دولتی، طرح پژوهشی، پایان‌نامه ایرانیان داخل و خارج از کشور، مقاله‌های همایش و پیشنهاد است. کاربران می‌توانند تمام متن برخی از پایان‌نامه‌ها یا رساله‌ها را دریافت کنند، به همین دلیل، این

پایگاه جزو پایگاه‌های مهم برای پژوهشگران به حساب می‌آید. بنابراین، بی‌توجهی به دشواری‌ها یا مشکلات خط فارسی در این پایگاه می‌تواند کاربر را به مدرک مورد نظر خود نرساند و یا زمان زیادی از کاربر تا رسیدن به نتایج مرتبط بگیرد. بنابراین، پژوهش حاضر به بررسی مشکلات خط فارسی در ابعاد صرفی، معنایی و رسم الخط بر بازیابی عنوان مدارک پایگاه اطلاعات علمی ایران (گنج) می‌پردازد تا تأثیر آنها را بر بازیابی مدارک این پایگاه نشان دهد. در واقع پژوهش حاضر به دنبال یافتن پاسخ پرسش‌های زیر است:

۱. مشکلات صرفی خط فارسی چه تأثیری بر بازیابی اطلاعات در گنج دارد؟

منظور از مشکلات صرفی در این پژوهش، دشواری‌های ناشی از صرف (مطالعه ساخت درونی کلمات و روابط نظام‌مند صورت و معنا در کلمات) است که بر بازیابی اطلاعات تأثیر می‌گذارند. موارد بررسی شده در این پژوهش عبارتند از: وام‌گیری واژگانی، کلمات مرکب و ترکیبات نحوی، جمع مکسر اسم‌ها، تک‌واژه‌های صرفی صفت و علامت‌های جمع.

۲. مشکلات معنایی خط فارسی چه تأثیری بر بازیابی اطلاعات در گنج دارد؟

منظور از مشکلات معنایی در این پژوهش شش دسته از دشواری‌هاست که نتایج بازیابی را از حیث معنا تحت تأثیر قرار می‌دهند و موارد زیر را شامل می‌شوند: واژه‌های دخیل و برابره‌های فارسی آنها، هم‌معنایی، نوشتن یا ننوشتن مصوت‌های کوتاه (ـَ، ـِ، ـُ)، تمایز آ و الف، هم‌نامی و هم‌نویسی.

۳. مشکلات رسم الخط فارسی چه تأثیری بر بازیابی اطلاعات در گنج دارد؟

دشواری‌های ناشی از ویژگی‌های ریخت‌شناسی یا رسم الخط، باعث بروز مسائلی در بازیابی اطلاعات می‌شوند که در اینجا شامل دشواری‌های زیر می‌شوند: علامت تشدید، نحوه نوشتن «الف» مقصوره، واژه‌های دارای دو هجای میانی «وو»، واژه‌های دو املایی، جابه‌جایی «ی» و «همزه» در کلمات فارسی، نوشتن اعداد به حروف و عدد، نگارش همزه میانی با کرسی‌های «الف»، «واو»، کرسی دندان، نگارش همزه بدون کرسی، جدانویسی و پیوسته‌نویسی، نحوه نوشتن صامت «ی» بعد از «ه» غیر ملفوظ، برخی واژگان دخیل عربی، نزدیک‌نویسی و جایگاه علائم نگارشی (گیومه، ویرگول، دونقطه و نقطه).

۲. پیشینه پژوهش

تاکنون پژوهش‌های گوناگونی پیرامون تأثیر مسائل زبانی بر بازیابی اطلاعات انجام شده است

که مرور پیشینه‌های موجود نشان می‌دهد این مسأله از ابتدا مورد توجه پژوهشگران علم اطلاعات و دانش‌شناسی و همین‌طور محققان حوزه رایانه و زبان‌شناسی بوده است که در ادامه به برخی از این پژوهش‌ها در داخل و خارج اشاره می‌شود.

برای نمونه، پژوهش راثی ساربانقلی (۱۳۸۵) نشان می‌دهد، بیشتر مشکل کاربران در بازیابی اطلاعات، عدم توجه به شکل‌های مختلف نوشتاری واژه بوده و نیمی از موارد عدم بازیابی مطالب مورد نظر، به مسائل مربوط به زبان و خط فارسی است.

آخشیک و فتاحی (۱۳۹۱)، در پژوهشی پیوسته‌نویسی و جدانویسی واژگان فارسی را بر روی پایگاه‌های اطلاعاتی علوم و فناوری اطلاعات ایران و مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری بررسی کردند. نتایج نشان داد تقریباً دو سوم از کلماتی که در عنوان به کار رفته‌اند، درست نوشته شده‌اند و این کلمات اغلب دو جزئی و سه جزئی بوده و اغلب نویسندگان پایان‌نامه‌ها در مورد نحوه نگارش کلماتی که دو جزئی و مشتق هستند، دچار خطا شده‌اند که در پایگاه منطقه‌ای اطلاع‌رسانی علوم و فناوری، جستجو با همان حالت ثبت شده عنوان‌ها، منجر به بازیابی شده و در پژوهشگاه علوم و فناوری اطلاعات ایران تقریباً نیمی از عنوان‌ها با تغییر رسم الخط همچنان بازیابی شده‌اند.

همچنین گل تاجی و بذرگر (۱۳۸۹)، در پژوهشی با عنوان بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارك علمی ایران و جهاد دانشگاهی، نشان دادند که هیچ یک از این سه پایگاه به شیوه‌ای جامع به حل مشکلات ریخت‌شناسی واژگان فارسی نپرداخته‌اند و هر پایگاه به صورت جداگانه از میان ۱۷ چالش پیش رو، تنها به رفع تعداد محدودی از آنها اقدام کرده است.

در پژوهشی دیگر، ارسطوپور و احمدی نسب (۱۳۹۱)، چالش‌های رسم الخط، مسائل صرفی و مسائل معنایی را بر روی بانک اطلاعات نشریات کشور، پایگاه مقالات فارسی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پایگاه مجلات نور و موتور جستجوی گوگل بررسی کرده و پیشنهادهایی که برخی از پژوهشگران برای مقابله با این چالش‌ها بیان کرده‌اند را با نتایج پژوهش خود تطبیق داده و پیشنهادهایی در این زمینه ارائه دادند.

ستوده و هرنجویان (۱۳۹۱)، نیز در پژوهشی، مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آنها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات داشته‌اند و معتقدند گوناگونی نگارشی، بازیابی را از منظر کاهش دقت و ریزش کاذب تحت تأثیر قرار می‌دهد.

آخشیکی (۱۳۹۴)، در پژوهشی، عنوان مقالات موجود در بانک اطلاعات نشریات کشور را از نظر مسائل صرفی مورد بررسی قرار داده و بیان می‌کند که پربسامدترین اشتباه‌های رایج در عنوان مقاله‌های علمی-پژوهشی حوزه‌های علوم انسانی به دو دسته خطاهای مربوط به فاصله‌گذاری نادرست میان کلمه‌ها و عدم استفاده از «الف» همزه و «واو» همزه مربوط هستند.

هماوندی، نوروزی و بهشتی (۱۳۹۷)، نیز در پژوهشی به بررسی مشکلات جستجو و بازیابی اطلاعات از جنبه ویژگی‌های نگارشی زبان فارسی روی پایگاه‌های اطلاعاتی پژوهشگاه علوم و فناوری اطلاعات، پایگاه استنادی علوم جهان اسلام، پایگاه مجلات نور و پایگاه اطلاعات علمی جهاد دانشگاهی پرداختند. آنها با تهیه سیاهه‌ای از کلیدواژه‌های کاوش، پژوهش خود را انجام داده‌اند. نتایج این بررسی نشان می‌دهد مواردی مانند، پیوسته‌نویسی و جدانویسی واژگان مشتق، مرکب و مشتق- مرکب، گوناگونی جمع‌ها، واژگان دخیل و معادل آنها در بخش نوشتاری و چندمعنایی و هم‌معنایی و... در بخش معنایی، از جمله ویژگی‌های نوشتاری و معنایی زبان فارسی است که بازیابی اطلاعات از این پایگاه‌های اطلاعاتی را با مشکل مواجه می‌کند. آنها همچنین اشاره می‌کنند که عدم پوشش مناسب ویژگی‌های یاد شده در مراحل ذخیره‌سازی و پردازش و عدم آگاه نمودن کاربر از آن، جهت اصلاح فرایند کاوش در مرحله بازیابی اطلاعات، اثرات نامطلوبی بر فرایند کاوش و بازیابی دارد.

رنجبر و همکاران (۱۳۹۷)، نیز میزان انطباق رفتار نگارشی نگارندگان و کاربران پایگاه‌های اطلاعات علمی را با دستورالعمل‌های مصوب فرهنگستان زبان و ادب فارسی در ارتباط با پیوسته‌نویسی، نزدیک‌نویسی و جدانویسی مورد بررسی قرار داده و پژوهش خود را بر مدارک موجود در سه پایگاه بانک نشریات ایران (مگیران)، مرکز منطقه‌ای علوم و فناوری و پایگاه مرکز جهاد دانشگاهی انجام دادند و رفتار نگارشی دانشجویان تحصیلات تکمیلی دانشگاه شیراز (سال تحصیلی ۹۳-۹۴) به عنوان کاربران این پایگاه‌ها مورد ارزیابی قرار گرفت. نتایج نشان می‌دهد که رفتار نگارشی درصد کمی از کاربران، مطابق با دستورالعمل‌های مصوب فرهنگستان زبان و ادب است و کم‌تر از نیمی - چهل درصد - نگارندگان به موارد مغایر با قواعد این فرهنگستان عمل کرده‌اند. کاربران علی‌رغم آشنایی با نحوه درج نیم‌فاصله از طریق صفحه کلید رایانه، نزدیک‌نویسی را رعایت نکرده‌اند.

رنجبر و عباسپور (۱۳۹۷)، پژوهشی را در مورد تأثیر پیوسته‌نویسی و جدانویسی، روی گسترش‌پذیری جستجو و بازیابی مدارک، در پایگاه‌های اطلاعات علمی فارسی انجام داده و نشان

دادند که میان پایگاه‌های اطلاعات علمی فارسی از نظر قابلیت بازیابی مدارک با هر یک از شکل‌های نگارشی، اعم از پیوسته‌نویسی و جدانویسی، تفاوت وجود دارد و تنها درصد خیلی کمی از مدارک انتخابی در هنگام جستجو در پایگاه‌های اطلاعاتی مورد مطالعه، با همه شکل‌های نگارشی قابل بازیابی است.

از جمله پژوهش‌های خارجی مرتبط با پژوهش حاضر می‌توان به پژوهش مونز و ریکه^۱ (۲۰۰۲) و مقدار^۲ (۲۰۰۵) اشاره کرد که مونز و ریکه (۲۰۰۲) با تمرکز بر اثرات تحلیل‌های ریخت‌شناسی همچون ریشه‌سازی و جداسازی کلمات مرکب، کارایی بازیابی اطلاعات را بررسی کردند. این مطالعه بر روی زبان‌های هلندی، آلمانی و ایتالیایی انجام شده است. نتایج نشان داد بازیابی اطلاعات حدود ۲۵٪ برای زبان آلمانی، ۶۹٪ برای زبان هلندی و ۲۵٪ برای زبان ایتالیایی بهبود یافت. مقدار (۲۰۰۵) نیز در پژوهش خود، عملکرد سه ابزار جستجوی عمومی را با سه جستجوگر عربی (که اختصاصاً مسائل زبان‌شناختی عربی را لحاظ می‌کنند) مقایسه کرد. نتایج نشان داد جستجوگرهای عمومی، نظیر آلدوب، آلتاویستا و گوگل در بازیابی مدارک عربی، ناقص عمل می‌کنند.

تاث^۳ (۲۰۰۶)، نیز در پژوهشی به بررسی قابلیت‌های زبان‌شناختی جستجوگرهای انگلیسی و مجاری پرداخت. محقق سه ابزار جستجوی انگلیسی به نام‌های گوگل، آلتاویستا و آلدوب را با پنج جستجوگر محلی مقایسه نمود. تحلیل داده‌ها بر پایه چند شاخص انجام شد که عبارت بودند از: ریشه‌سازی، بازیابی لهجه‌های مختلف، کوتاه‌سازی و جستجوی مترادف‌ها. نتایج حاکی از آن بود که جستجوگرهای محلی، مسائل زبان مجاری را بهتر از جستجوگرهای انگلیسی مورد توجه قرار داده‌اند. ابزارهای انگلیسی زبان، لهجه‌های مختلف زبان مجاری را به خوبی پشتیبانی نمی‌کردند، که این امر به بازیابی ضعیف اطلاعات منجر می‌شد.

مرور پژوهش‌ها نشان می‌دهد بررسی تأثیر خط فارسی و نگارش بر بازیابی اطلاعات، از همان سال‌های اولیه گسترش وب مورد توجه پژوهشگران بوده است. برخی، تأثیر دشواری‌های مختلف را بر نتیجه بازیابی موتورهای جستجو بررسی کرده بودند، برخی در پایگاه‌های اطلاعاتی داخلی

1. Monz & Rijke

2. Moukddad

3. Tóth

مثل نورمگز و مگ ایران و... این مسأله را مورد مطالعه قرار داده‌اند، عده‌ای نیز به بررسی یک دشواری خاص پرداخته بودند، مانند جدانویسی و پیوسته‌نویسی که از شایع‌ترین مشکلات مربوط به رسم‌الخط است. بررسی پیشنهادها نشان می‌دهد که اغلب به مشکلات نگارشی و معنایی پرداخته شده و به مشکلات صرفی فارسی توجه زیادی نشده است. البته وام‌گیری واژگانی که جزو این دسته مشکلات است، مد نظر پژوهشگران بوده است که برخی، آن را در دسته مسائل مربوط به رسم‌الخط بررسی کرده‌اند. بررسی دشواری‌هایی مثل تک‌واژه‌های صرفی صفت، جمع مکسر اسم‌ها و کلمات مرکب و ترکیبات نحوی که در این پژوهش مورد بررسی قرار گرفته است، جزو جنبه‌ها جدید این پژوهش است.

۳. روش پژوهش

پژوهش حاضر از نظر هدف، کاربردی و از نوع کیفی است. با توجه به اینکه داده‌های زبانی را نمی‌توان بر پایه روش‌های آماری مورد تجزیه و تحلیل قرار داد (طاهری اسکویی و همکاران، ۱۳۹۵)، در این پژوهش از روش تحلیل محتوا استفاده شد. جامعه این پژوهش، مدارک ثبت شده در گنج است. تعداد این مدارک ۷۴۴۵۹۶ عنوان تا زمان اجرای این پژوهش، یعنی پایان دی‌ماه ۱۳۹۸ بود. نمونه این پژوهش به شیوه نمونه‌گیری معیاری^۱ انتخاب شد. در این شیوه نمونه‌گیری، مواردی از جامعه انتخاب می‌شوند که معیارهای موردنظر پژوهش را داشته باشند (پالینکاس و همکاران^۲، ۲۰۱۵). معیار مورد نظر برای انتخاب نمونه در این پژوهش، دارا بودن حداقل یکی از دشواری‌های مربوط به سه دسته رسم‌الخط، معنایی یا صرفی در یکی از واژه‌های به‌کار رفته در عنوان ثبت شده در گنج بود. ابزاری که برای گردآوری داده‌ها استفاده شد، سیاهه واری محقق ساخته است. این سیاهه بر پایه مشکلات خط فارسی در سه دسته رسم‌الخط، معنایی و صرفی طراحی شد. در این مطالعه برای تأیید روایی، سیاهه تهیه شده که زیر نظر اساتید راهنما و مشاور تدوین شده بود، در اختیار ۶ تن از متخصصان علم اطلاعات و دانش‌شناسی و متخصصان ادبیات و زبان فارسی قرار داده شد تا از روایی آن اطمینان حاصل شود. براساس نظر این افراد در سیاهه واری تغییرهایی اعمال و سیاهه واری نهایی شد. برای اطمینان از پایایی ابزار، مصداق‌های

مربوط به هر دشواری بعد از زمان اجرای پژوهش در پایگاه گنج جستجو شد که نتایج حاصل از آن با نتیجه به دست آمده از پژوهش مطابقت داشت.

۴. یافته‌های پژوهش

تجزیه و تحلیل داده‌ها در ادامه به تفکیک سه دسته مشکلات صرفی، معنایی و رسم‌الخط بیان شده است.

مصادیق مربوط به دشواری‌های صرفی در ۵ دسته، همان‌طور که در جدول ۱ آمده است، در پایگاه گنج جستجو شدند و فراوانی رکوردهای جستجو شده برای هر کدام در این جدول آمده است.

جدول ۱ - نتایج حاصل از جستجوی مصادیق مربوط به مشکلات صرفی در گنج

ردیف	دشواری	مصادیق ۱	فراوانی رکوردها	مصادیق ۲	فراوانی رکوردها	یکسانی رکوردهای بازیابی شده	
						بله	خیر
۱	وام‌گیری واژگانی	اورانیم اورانیوم	۲۴۰ ۴۶۰	مرفین مورفین	۳۸۸ ۵۷۱		*
۲	کلمات مرکب و ترکیبات نحوی	محیط زیست محیط‌زیست	۲۱۸۵۸ ۱۷۳۱۳	آب سیاه آب‌سیاه	۷۴۰ ۸۳		*
۳	جمع مکسر اسم‌ها	کتابها کتاب	۲۱۵۶۱ ۷۴۳۰	میدان میادین	۱۸۲۸۷ ۱۱۲۳		*
۴	تک‌واژه‌های صرفی صفت	"کودکان بزرگتر از"	۲۰	"بیشترین مساحت"	۱۴۲	*	
		"کودکان بزرگ تر از"	۲۰	"بیش ترین مساحت"	۱۴۲		
۵	علامت‌های جمع	کتاب‌ها	۲۵۶	داستان‌ها	۸۳۸	*	
		معلمان	۸۲۰۶	مهندسان	۱۶۰۴	*	
		معلمین	۷۳۱۵	مهندسین	۲۲۴۱	*	
		حیوانات	۲۷۴۵	انتقادات	۵۸۳	*	
		میوه‌جات	۶۴۴۳	سبزیجات	۷۵۰	*	
		انقلابی انقلابیون	۱۶۴۳ ۱۶۴۳	روحانی روحانیون	۳۱۰۵ ۴۶۶	*	

http://stlm.gom.ac.ir

نتیجه جستجوی مصادیق مربوط به دشواری‌های صرفی در ۵ دسته وام‌گیری واژگانی، کلمات مرکب و ترکیبات نحوی، جمع مکسر اسم‌ها، تک‌واژه‌های صرفی و علامت‌های جمع که در جدول ۱ آمده، نشان داد در وام‌گیری واژگانی، فراوانی رکوردهای بازیابی شده برای هر دو مصداق، متفاوت است. برای نمونه، تعداد مدارک برای واژه اورانیم و اورانیوم به ترتیب ۲۴۰ و ۴۶۰

بود.^۱ جستجوی مصداق‌های کلمات مرکب و ترکیبات نحوی نشان داد که پایگاه گنج هنگام جستجو، این کلمات را جدا از هم جستجو می‌کند و زمانی که این کلمه‌ها با نیم‌فاصله^۲ جستجو شدند، فراوانی رکوردهای بازیابی شده، تغییر کرد، یعنی جستجوی کلمات مرکب به دلیل ویژگی این ترکیبات (تشکیل شده از دو یا چند بخش)، بر نتیجه بازیابی اثرگذار است. برای نمونه وقتی ترکیب «آب‌سیاه» با یک فاصله جستجو شد، تعداد رکوردهای بازیابی شده ۷۴۰ مورد و در جستجو با نیم‌فاصله، ۸۳ بود. در مورد جمع مکسر، ثفات زیادی در فراوانی رکوردهای بازیابی شده وجود داشت و واژه‌ها عیناً با همان املاء در نتایج بازیابی شده مشاهده شدند. برای نمونه تعداد رکوردهای بازیابی شده برای واژه کتاب‌ها و کتب به ترتیب ۲۱۵۶۱ و ۷۴۳۰ بود.

نتایج جستجوی مربوط به مصداق‌های گروه تک‌واژه‌های صرفی صفت نشان داد، در پایگاه گنج با جستجوی این واژگان در حالت‌های مختلف، صفت تفضیلی و عالی به همان صورت نوشته شده بازیابی می‌شود و اگر تک‌واژه صرفی صفت (تر-ترین)، به صورت جدا نوشته شود، تعداد رکوردهای بازیابی شده، یکسان است. برای نمونه رکوردهای جستجو برای «کودکان بزرگ تر از» و «کودکان بزرگتر از»، یکسان و برابر با ۲۰ رکورد بود.

در مورد علامت‌های جمع، براساس نتایج حاصل از بازیابی رکوردهای مربوط به دو مصداق جستجو شده می‌توان گفت پایگاه گنج بین حالت مفرد و حالت جمع آن با پسوند «ها»، تفاوتی قائل نیست. علامت جمع «ها» در راهنمای جستجوی گنج جزو «ایست واژه‌ها»^۳ آمده است؛ یعنی واژگان پرسامدی که به دلیل اخلال، در فرایند جست‌وجو شرکت داده نخواهند شد. شمار این واژگان در گنج محدود است و بیشتر دربرگیرنده برخی نشانه‌های جمع مانند «ها»، و واژگان ربط یا فعل‌های پرتکرار (مانند یا، که، و) می‌باشد.^۴

برعکس علامت جمع «ها»، برای دیگر علامت‌های جمع، رکوردهای بازیابی شده در هر مورد یکسان نبود، به این معنی که این مصادیق با همان املاء در نتایج جستجو ظاهر شدند. برای

۱. هنگام جستجوی واژه «اورانیوم»، عنوان‌هایی با املاهای دیگر هم مشاهده شد، اما برای «مورفین»، واژه عیناً با همان املاء در جستجو‌ها دیده شد.

۲. نیم‌فاصله در گنج به این صورت تعریف شده است که بین اجزای کلمه، ویرگول بدون فاصله با آنها نوشته شود.

3. Stop-words

4. <https://ganj.irandoc.ac.ir/#/help>

نمونه تعداد رکوردهای بازیابی شده برای هر دو واژه «کتاب» و «کتابها»، ۲۵۶۱ بود. بنابراین، می‌توان گفت به احتمال زیاد در پایگاه گنج برای علامت جمع «ها»، چاره‌ای اندیشیده نشده است.

به طور کلی نتایج حاصل از جستجوی مشکلات صرفی در پایگاه گنج نشان داد این مشکلات به طور کلی بر نتیجه بازیابی تأثیرگذار هستند.

همان‌طور که در مورد کلمات مرکب و ترکیبات نحوی دیده شد، تعداد رکوردهای بازیابی‌شده، تفاوت زیادی دارند که در این مورد، باعث ریزش کاذب می‌شد و این یعنی هدر رفتن وقت کاربر. اما از طرفی مشاهده شد که پایگاه گنج به تک‌واژه‌های صرفی صفت توجه دارد. در مورد علامت‌های جمع، و جمع مکسر، چون کاربران در سطوح مختلف ممکن است هر یک از صورت‌ها را برای جستجو انتخاب کنند، در بسیاری موارد امکان دارد ناخودآگاه از دستیابی به نتایجی که حاصل جستجوی شکل دیگر واژه هستند، محروم بمانند، و به همان نتایج اولیه بسنده کنند.

مصادیق مربوط به دشواری‌های معنایی در ۶ دسته، همان‌طور که در جدول ۲ آمده است، در پایگاه گنج جستجو شدند و فراوانی رکوردهای جستجو شده برای هر کدام در این جدول آمده است. اما برای دشواری‌های «هم‌نامی» و «هم‌نویسی»، برای اینکه مثال‌ها از عنوان‌های بازیابی شده نشان داده شود، به صورت جداگانه در جدول شماره ۳ ارائه شده‌اند.

جدول ۲- نتایج حاصل از جستجوی
مصادیق مربوط به مشکلات معنایی در گنج

ردیف	دشواری	مصادیق ۱	فراوانی رکوردها	مصادیق ۲	فراوانی رکوردها	یکسانی رکوردهای بازیابی شده	
						بله	خیر
۱	واژه‌های دخیل و برابره‌های فارسی آنها	ایمیل رایانامه	۳۰۸ ۱۱۷	آنلاین برخط	۲۵۲۱ ۹۷۹		*
۲	هم‌معنایی	دریای خزر دریای مازندران	۱۹۴۶ ۱۵۵۷	قرآن کریم قرآن مجید	۶۷۲۵ ۴۷۲		*
۳	نوشتن مصوت‌های کوتاه (اَ، اِ، اُ)	گرم گِرم	۲۷۵۶۳ ۲۷۵۶۳	خرف جِرَف	۱۵۹۸ ۱۵۹۸	*	
۴	تمایز آ و الف	آسم اسم	۲۲۰۲ ۲۲۰۲	آبله ابله	۱۱۷ ۱۱۷	*	

جدول ۳- نتایج حاصل از جستجوی مصادیق مربوط به
دشواری‌های هم‌نامی و هم‌نویسی در پایگاه گنج

ردیف	دشواری	مصادیق	مثال‌هایی از عنوان‌های بازیابی شده	توجه به دشواری	
				بلی	خیر
۱	هم‌نامی	شانه	امکان‌سنجی استفاده از سیستم انتقال قدرت پنوماتیکی در دروگرهای شانه‌ای	*	
			ارزیابی عملکرد شانه بازیکنان دختر نخبه بدمینتون با تکنیک CONSTANT SCORE		
			بررسی ریخت‌شناسی و رژیم تغذیه‌ای شانه داران سواحل جنوبی دریای خزر		
		شیر	بررسی تغییرات میزان فلزات سنگین شیر در مراحل مختلف تولید شیر خشک صنعتی		
			بررسی استفاده از شیر پروانه‌ای در خطوط انتقال سیالات جهت کنترل جریان		
			بازتاب نمادین شیر در هنر اسلامی ایرانی		
	علی		توصیف بعضی از شرایط علی روی پیوستگی لورنتسی		
			تربیت سیاسی از دیگاه امام علی(ع)		
			ترجمه و نقد بخشی از کتاب: اضا علی متشابهات القرآن/ به قلم خلیل یاسین		

نتیجه جستجوی مصداق‌های مربوط به دشواری‌های معنایی در ۶ دسته واژه‌های دخیل و برابره‌های فارسی آنها، هم‌معنایی، نوشتن مصوت‌های کوتاه (ـ، ـِ، ـُ)، تمایز «آ و الف»، هم‌نامی و هم‌نویسی که در جدول ۲ و ۳ آمده‌اند، نشان داد در خصوص واژه‌های دخیل و برابره‌های فارسی آنها، فراوانی رکوردهای بازیابی شده، یکسان نیست. برای نمونه تعداد رکوردهای بازیابی شده برای واژه «آنلاین» ۲۵۲۱ و برای برابرواژه آن «برخط»، ۹۷۹ مورد بود. از طرفی مرور نتایج و مدارک بازیابی شده، نشان داد برای واژه «ایمیل» در برخی مدارک، این مسأله مورد توجه بوده است، به این معنی که اگر در عنوان مدرک، واژه «ایمیل» ذکر شده است، در کلیدواژه‌های آن مدرک از معادل فارسی آن هم استفاده شده است و از این رو حین جستجو بازیابی شده است. اما برای واژه «آنلاین» و برابر فارسی آن عیناً با همان املاء در مدارک دیده شدند.

تعداد نتایج بازیابی شده برای هر دو مصداق در هم‌معنایی، متفاوت بود. برای نمونه تعداد رکوردهای بازیابی شده برای دریای خزر ۱۹۴۶ و برای دریای مازندران ۱۵۵۷ رکورد بود، در صورتی که هر دو به یک مفهوم (دریایی در شمال ایران) اطلاق می‌شوند. پس، احتمالاً جستجوی

یکی باعث از دست رفتن مدارکی می‌شود که در عنوان آنها از واژه دیگر آن استفاده شده است. لازم به توضیح است از آنجایی که امکان شمارش رکوردهای قسمت معنایی وجود نداشت، تحلیل‌های این بخش براساس مرور تعدادی از نتایج ارائه شده است؛ چرا که گستره نتیجه جستجو شده بسیار وسیع است و در عمل امکان بررسی تک تک آنها وجود ندارد. در مورد نوشتن مصوت‌های کوتاه، نتایج بازیابی، تفاوتی را در تعداد رکوردهای بازیابی شده نشان نداد، به این معنا که حرکت‌گذاری در نتیجه بازیابی تأثیری نداشت و در عنوان‌های بازیابی شده، واژه‌ها با معانی مختلف‌شان مشاهده شدند. لازم به ذکر است که حین جستجوی واژه «کره»، پایگاه گنج چهار واژه «کره توخالی»، «کره جنوبی»، «کره کاکائو» و «کره ریمان» را به صورت پیش‌فرض پیشنهاد می‌داد.

موارد بازیابی شده برای مسأله تمایز «آ و الف» نشان داد که پایگاه گنج بین «الف و آ» تمایزی قائل نیست که این مسأله در مشکلات رسم الخطی هم بررسی شده است. تمایز «آ» و «الف» تا زمانی که معنی واژه مورد نظر با علامت مد و بدون علامت مد، فرقی نکند، مشکل ساز نیست؛ اما اگر حذف مد، معنا را تغییر دهد، در بازیابی اطلاعات، بهتر است به آن توجه شود. همان طور که در جدول شماره ۲، آمده است، فراوانی رکوردهای بازیابی شده برای هر دو املاء در مصداق «آسم» و «اسم» یکسان بوده و برابر با ۲۲۰۲ رکورد است. در صورتی که اگر واژه «آسم» بدون علامت مد نوشته شود، معنی دیگری می‌دهد، چون می‌تواند به شکل دیگری تلفظ شود.

نتایج بازیابی در مورد هم‌نامی، همان‌طور که مثال‌هایی از عنوان مدارک بازیابی شده در جدول ۳ آمده است، نشان داد زمانی که واژه‌شانه را جستجو کنیم، نتایجی که دلالت بر معانی متفاوت آن دارند، بازیابی می‌شوند. عنوان‌های بازیابی شده برای واژه «شیر» نیز همین را نشان داد. ذکر این نکته در اینجا ضروری است که پایگاه گنج مثلاً هنگام جستجو برای «شیر»، شیر (شیر خوراکی) را پیشنهاد می‌داد.

در مورد هم‌نویسی نیز نتیجه به دست آمده مطابق با دشواری هم‌نامی بود و واژه‌های کره و علی با معانی مختلف در عنوان مدارک بازیابی شده مشاهده شدند و همان‌طور که در مورد نوشتن مصوت‌های کوتاه هم گفته شد، راهی برای تمایز واژه‌های هم‌نویسه حین جستجو توسط کاربر وجود ندارد.

مصادیق مربوط به دشواری‌های رسم الخط در ۱۵ دسته به صورتی که در جدول‌های شماره ۴، ۵ و ۶ آمده، در پایگاه گنج جستجو شدند و فراوانی رکوردهای جستجو شده برای هر کدام در این جدول‌ها ارائه شده است. جدول ۵ تنها به دشواری نیم‌فاصله (نزدیک‌نویسی) مربوط بوده و داده‌های مربوط به جایگاه علائم نگارشی نیز جداگانه در جدول ۶ آمده است.

جدول ۴- نتایج حاصل از جستجوی مصادیق مربوط به مشکلات رسم الخط در پایگاه گنج

ردیف	دشواری‌ها	مصدق ۱	فراوانی رکوردها	مصدق ۲	فراوانی رکوردها	یکسانی فراوانی رکوردها	بلی	خبر
۱	علامت تشدید	معلم	۷۲۸۶	مدرس	۲۸۳۳	*		
		معلم	۷۲۸۶	مدرس	۲۸۳۳			
۲	نحوه نوشتن «الف» مقصوره	عیسی	۴۴۵۱	اسمعیل	۱۴۴۳	*		
		عیسا	۲	اسماعیل	۱۱۹۵۹			
۳	واژه‌های دارای هجای میانی «وو»	طاووس	۸۳	داوود	۲۲۶۱	*		
		طاوس	۲۲	داود	۶۰۰۴			
۴	واژه‌های دو املایی	زغال	۱۱۹۲	تاقدیس	۵۹۲	*		
		ذغال	۲۶۵	طاقدیس	۱۱۴			
۵	جابه‌جایی «ی» و همزه در کلمات فارسی	ساییدگی	۱۵۲	پاییز	۲۳۷۶	*		
		سایتیدگی	۱۵۲	پائیز	۲۳۷۶			
۶	نوشتن اعداد به عدد و حروف	دو بعدی	۳۹۷۵	چهار بعدی	۷۳	*		
		۲ بعدی	۲۵	۴ بعدی	۱۳			
۷	نگارش همزه میانی و پایانی با کرسی «الف»	تالیف	۲۲۸۳	منشا	۶۱۵۸	*		
		تالیف	۲۲۸۳	منشأ	۶۱۵۸			
۸	نگارش همزه میانی و پایانی با کرسی «واو»	موسسه	۴۹۷۵	تاللو	۲۴	*		
		مؤسسه	۴۹۷۵	تلاؤ	۲۴			
۹	نگارش همزه با کرسی دندانه	مساله	۳۹۲۸۲	جرات	۷۶۱	*		
		مسأله	۳۹۰۱۳	جرئت	۳۴			
		مسئله	۲۴۹۸۱					
۱۰	نگارش همزه بدون کرسی	شی	۱۳۶۳	املا	۱۶۷	*		
		شیء	۹۸۶	املاء	۸۶			
۱۱	جدانویسی و پیوسته‌نویسی	آموزشگاه	۱۰۳۸	خاکبرداری	۱۸۱	*		
		آموزش گاه	۴۲۳	خاک برداری	۴۳۶			
۱۲	نوشتن یا ننوشتن علامت مد	اسیا	۲۴۲۱	گردآوری	۳۲۴۳۲	*		
		آسیا	۲۴۲۱	گردآوری	۳۲۴۳۲			
۱۳	نگارش صامت میانجی «ی» بعد از «ه» غیر ملفوظ	خانه روستایی	۴۱۶	شیوه زندگی	۷۶۹۵	*		
		خانه روستایی	۴۱۶	شیوة زندگی	۷۶۹۵			
		خانه،ی روستایی	۴۱۶	شیوه،ی زندگی	۷۶۹۵			
		خانه ی روستایی	۴۱۶	شیوه ی زندگی	۶۹۵			
۱۴	نگارش برخی واژه‌های دخیل عربی	جعفر بن محمد	۶۳	علی بن موسی	۱۶۹	*		
		جعفر ابن محمد	۲۱	علی ابن موسی	۳۹			
		زکوه	۳	حیوه	۴	*		
		زکات	۳۲۵	حیات	۱۱۵۷۹			
		دایره المعارف	۲۳۴	حیوه	۴	*		
		دایرةالمعارف	۲۳۴	حیوة	۴			

جدول ۵- نتایج حاصل از جستجوی مصادیق مربوط به
دشواری نزدیک‌نویسی (نیم‌فاصله) در پایگاه گنج

دشواری	مصدق با نیم فاصله	فراوانی رکوردها	مصدق با یک فاصله	فراوانی رکوردها	یکسانی رکوردهای بازیابی شده	
					بله	خیر
نزدیک‌نویسی (نیم‌فاصله) و جدانویسی (یک فاصله)	آب، شکستگی	۷	آب شکستگی	۵۵۲		*
	عقب، افتاده	۸۵	عقب افتاده	۱۱۷		*
	زبان، آموزی	۳۶۶	زبان آموزی	۴۸۴		*
	قفسه، سینه	۴۱۵	قفسه سینه	۴۳۴		*
	هشت، بهشت	۴۷	هشت بهشت	۶۲		*
	بی، حسی	۲۶۵	بی حسی	۵۶۲		*
	تنظیم، کننده	۲۳۹۶۴	تنظیم کننده	۲۳۹۶۴	*	
	بازیابی شده	۷۶۱۴	بازیابی شده	۷۶۱۴	*	
	نظام، مند	۱۶۸۳	نظام مند	۲۳۵۵		*

جدول ۶- نتایج حاصل از جستجوی مصادیق مربوط به جایگاه علائم نگارشی در پایگاه گنج

دشواری	مصدق	فراوانی	مصدق	فراوانی	یکسانی رکوردهای بازیابی شده	
					بله	خیر
گیومه	" بزرگ سال "	۶۷	" بزرگ سال "	۶۷	*	
ویرگول	بزرگ، سال بزرگ ، سال	۷۷۰۴ ۷۷۰۴	بزرگ، سال	۶۷		*
دو نقطه	بزرگ: سال بزرگ: سال	۷۷۰۴ ۰	بزرگ : سال	۷۷۰۴		*
نقطه	بزرگ. سال بزرگ. سال	۷۷۰۴ ۰	بزرگ . سال	۷۷۰۴		*

تأثیر مشکلات رسم الخط اغلب با تفاوت در تعداد نتایج بازیابی شده خود را نشان داد. نتایج به دست آمده گویای آن است که پایگاه گنج به برخی از دشواری‌های رسم الخط توجه داشته است، برای مثال در مورد نگارش همزه میانی و پایانی با کرسی «الف» و نگارش همزه میانی و پایانی با کرسی «واو»، تعداد رکوردهای بازیابی شده یکسان بود، همچنین در جابه‌جایی «ی» و همزه در کلمات فارسی یا نحوه نوشتن صامت «ی» بعد از «ه» غیر ملفوظ. در مورد علامت تشدید و حذف یا نوشتن علامت «مد» نیز نتایج بازیابی شده یکسان بود و تأثیری بر بازیابی نداشت که در بخش مشکلات معنایی نشان داده شد که بر نتیجه بازیابی تأثیر می‌گذاشتند، چرا که تمایز «آ و الف» از این‌رو مهم است که در برخی موارد تفاوت معنایی ایجاد می‌کند. همچنین تأثیر نادیده گرفتن

علامت تشدید که در هم‌نویسی مشاهده شد، بر نتیجه بازیابی تأثیر دارد. اما در مورد واژه‌های دو املائی و واژه‌هایی با دو هجای میانی «وو»، نحوه نوشتن «الف» مقصوره و نوشتن اعداد به حروف و عدد و نیز برخی واژگان دخیل عربی و نگارش همزه بدون کرسی، نتایج بازیابی متفاوت بود؛ به این معنی که اگر کاربر از این مشکلات آگاهی نداشته باشد و یا مسأله را فراموش کند، مدارکی را از دست خواهد داد. در مورد جدانویسی و پیوسته‌نویسی، حین جدانویسی موجب ریزش کاذب شده و باعث هدر رفتن زمان کاربر می‌شود و نزدیک‌نوسی نیز در همه موارد پیشنهاد نمی‌شود، چون ممکن است کاربر از دستیابی به برخی مدارک محروم بماند. در خصوص علائم نگارشی درست است که برای «نقطه» و «دو نقطه» بدون فاصله از واژه‌های قبل و بعد، جستجو نتیجه‌ای در بر نداشت، اما در سایر حالت‌ها، تأثیری بر بازیابی نمی‌گذاشت. برای «گیومه» نیز به همین صورت بود. برای ویرگول هم اگر در راهنمای جستجوی پایگاه گنج در مورد آن توضیح داده شود، مفید واقع خواهد شد.

در اینجا در مورد یافته‌های پژوهش بحث و اظهارنظر می‌شود که به ترتیب به سه پرسش پاسخ داده شده است.

پرسش اول: مشکلات صرفی خط فارسی چه تأثیری بر بازیابی اطلاعات در گنج دارد؟

صرف به مطالعه ساختمان درونی واژه‌ها می‌پردازد (طباطبائی، ۱۳۸۶). دشواری‌های صرفی مورد بررسی در گنج پنج دسته وام‌گیری واژگانی، کلمات مرکب و ترکیبات نحوی، جمع مکسر اسم‌ها، تک‌واژه‌های صرفی صفت و علامت‌های جمع را شامل می‌شد. نتایج حاصل از جستجوی مشکلات صرفی در پایگاه گنج نشان داد که این دشواری‌ها بر نتیجه بازیابی تأثیرگذار هستند. همان‌طور که در مورد کلمات مرکب و ترکیبات نحوی مشاهده شد، تعداد رکوردهای بازیابی شده تفاوت زیادی دارند که در این مورد، باعث ریزش کاذب می‌شد. اما از طرفی مشاهده شد که پایگاه گنج به تک‌واژه‌های صرفی صفت توجه دارد و در مورد علامت‌های جمع دیده شد که رکوردهای بازیابی شده، متفاوت است، جز در علامت جمع «ها». در مورد «علامت‌های جمع» و «جمع مکسر»، چون کاربران در سطوح مختلف ممکن است هر یک از صورت‌ها را برای جستجو انتخاب کنند، در بسیاری موارد امکان دارد ناخودآگاه از دستیابی به نتایجی که حاصل جستجوی شکل دیگر واژه هستند، محروم بمانند و به همان نتایج اولیه بسنده کنند.

در پژوهش همانندی، نوروزی و بهشتی (۱۳۹۷)، نتیجه جستجوی وام‌گیری واژگانی در پایگاه‌های آی. اس. سی، نورمگز، اس. آی. دی. و ایرانداک با آنچه که در پایگاه گنج برای آن

مشاهده شد، همسو است و تأثیر آنها بر بازیابی به این صورت است که جستجوی یکی از واژه‌ها باعث از دست رفتن مدارکی می‌شود که املاهای دیگر واژه در عنوان آمده است. پس، کاربران ناگزیرند چندین فرایند جستجو را انجام دهند یا فرمول‌های جستجوی پیچیده‌تری را به کار گیرند. آشکار است که جامعیت چنین جستجویی در گرو آگاهی کاربر از همه تنوع‌های نگارشی و ظرایف جستجوی بولی است و در عین حال برخورداری از وقت و حوصله کافی است که با توجه به اصل کم‌ترین کوشش و همچنین آسان‌گیری کاربران در رفتار جستجوی خود، احتمال نادیده گرفتن چنین راهکارهایی وجود دارد (ستوده و هنرجویان، ۱۳۹۳).

برای دشواری وام‌گیری واژگانی، پیشنهاد می‌شود در نمایه‌سازی مدارک به این مسأله توجه شود. به این معنا که املاءهای متفاوت واژه‌های قرضی در مرحله ذخیره و نمایه‌سازی، مورد توجه باشد. در مورد علامت جمع «ها» مشاهده شد که پایگاه گنج آن را حین بازیابی در نظر نمی‌گرفت و رکورد‌های بازیابی شده برای صورت‌های نوشتاری هر مصداق یکسان بود، اما در پژوهش هم‌آوندی، نوروزی و بهشتی (۱۳۹۷)، فراوانی رکورد‌های بازیابی شده برای علامت جمع «ها»، یکسان نبوده و در این مورد با پژوهش حاضر هم‌سو نیست. نتایج حاصل از جستجوی دیگر علامت‌های جمع (ان، ین، ون، ات، جات) در پایگاه گنج یکسان نیست. برای مثال جستجوی واژه معلمین، می‌تواند باعث از دست رفتن مدارکی با املاهای معلمان یا معلم‌ها شود که این‌ها همه ناشی از صرف فارسی است که در نوشتار ظاهر می‌شود. همچنین در این مورد نیز پیشنهاد می‌شود که حین ذخیره‌سازی مدارک در پایگاه، به این دشواری‌ها توجه شود. به این معنا که در نظر گرفتن صورت جمع اسامی با علامت‌های جمع متفاوت و جمع مکسر می‌تواند به بازیابی مدارک مرتبط کمک کند.^۱

پرسش دوم: مشکلات معنایی خط فارسی چه تأثیری بر بازیابی اطلاعات در گنج دارد؟

زبان طبیعی از بُعد معنایی آکنده از ابهام است و در نظام‌های بازیابی مبتنی بر جستجوی متنی، ابهام‌های معنایی، مشکلاتی برای جستجوگران اطلاعات به همراه دارد (ارسطوپور و احمدی‌نسب، ۱۳۹۱). دشواری‌های شناسایی شده در بخش مشکلات معنایی مربوط به بازیابی اطلاعات بیانگر این است که اغلب ابهام‌های معنایی حاصل از هم‌نگاشتی و هم‌آوایی، واژه‌ها است و اغلب کاربران زبان با استفاده از دانش زبانی و زمینه «گفتمان» در برخورد با این مسائل

۱. منظور از مدارک مرتبط در اینجا، مدارکی است که حاوی کلیدواژه مورد جستجو (در عنوان، کلیدواژه، پدیدآور یا چکیده‌شان) با املاءهای متفاوت آن هستند.

می‌توانند به ابهام‌زدایی بپردازند، اما این ابهام در نوشتار به قوت خود باقی است، مثل چیزی که در هم‌نویسی و تمایز «آ» و «الف» در بازیابی پایگاه گنج دیده شد.

نتایج پژوهش حاضر در بخش مشکلات معنایی با پژوهش هماوندی، نوروزی و بهشتی (۱۳۹۷)، همسو است و شاهد عدم توجه پایگاه‌های اطلاعاتی به واژگانی هستیم که دارای معانی مختلف و صورت نوشتاری یکسان هستند که موجب می‌شود گاه معانی مختلفی از یک واژه بازیابی شود، اما معانی مورد نظر کاربر در میان آنها نباشد. در خصوص برخی دشواری‌ها مثل هم‌معنایی، اگر یکی از واژه‌های هم‌معنی در پایگاه جستجو شود، مدارک زیادی از دید کاربر پنهان می‌ماند.

ویژه‌نامه فرهنگستان زبان و ادب فارسی که در بهار سال ۱۳۹۸ با عنوان مطالعات واژه‌گزینی منتشر شده است، اهمیت مسأله واژه‌گزینی در خصوص واژگان دخیل و برابری آنها را می‌رساند. همان‌طور که در پایگاه گنج نیز مشاهده شد، این دشواری جزو مسائلی بوده که به آن توجه نشده است، این دیدگاه نیز با پژوهش هماوندی، نوروزی و بهشتی (۱۳۹۷) همسو است. لذا، کاربر حین جستجو به دلایل مختلف ممکن است از رسیدن به مدارک مناسب با نیاز خود باز بماند.

یکی از راه‌هایی که برای ابهام‌زدایی در نوشتار می‌تواند مؤثر باشد، نوشتن مصوت‌های کوتاه است. با توجه به نتایج به‌دست آمده برای مشکلات معنایی مشاهده شد که نوشتن مصوت‌های کوتاه، تأثیری بر نتیجه بازیابی نمی‌گذارد. در خصوص هم‌نامی، ارائه راهکار، به سادگی نوشتن مصوت‌های کوتاه نیست، چرا که در بافت و جمله است که می‌توان معنای آنها را دریافت. برای مثال اگر کسی بگوید «من شیر خوردم»، کسی از او نمی‌پرسد شیر جنگل یا شیر پاستوریزه؟ اما در جستجوها که عموماً مبتنی بر کلیدواژه هستند، این مسأله به سادگی قابل حل نیست و بر نتایج بازیابی تأثیرگذار خواهد بود. در پایگاه گنج مشاهده شد که پایگاه حین جستجوی واژه‌های هم‌نویسه و هم‌نام، به صورت پیش‌فرض کلیدواژه‌هایی را برای جستجو پیشنهاد می‌کند، اما اینکه تا چه میزان دشواری‌های نام‌برده با این راهکار قابل حل هستند، نیازمند تحقیق و بررسی بیشتر در این زمینه است.

به عقیده جلالی (۱۳۸۷)، وجود نداشتن واژگان مشابه میان دو متن، دلیل بر نامربوط بودن آنها به یکدیگر نیست. حالت‌های متعددی ممکن است وجود داشته باشد که در آنها کاربر نیاز اطلاعاتی خود را با لغات و واژگانی متفاوت از آنچه در مستندات سیستم وجود دارد، بیان کند. از این‌رو استفاده از هستان‌شناسی‌ها می‌تواند به بازیابی معنایی، به ویژه در مورد دشواری هم‌معنایی کمک کند.

پرسش سوم: مشکلات رسم الخط فارسی چه تأثیری بر بازیابی اطلاعات در گنج دارد؟

بسیاری از ویژگی‌های ریخت‌شناسی و نوشتاری فارسی ریشه در دو یا چندگانه‌نویسی یک واژه دارد که همین امر به دلیل ایجاد ناهماهنگی، سبب بروز مسائلی در بازیابی اطلاعات می‌شود (هماوندی، نوروزی و بهشتی، ۱۳۹۷). تحلیل داده‌های مربوط به مشکلات صرفی در پایگاه گنج نشان داد که تأثیر مشکلات صرفی اغلب با تفاوت در فراوانی رکوردهای بازیابی شده خود را نشان می‌دهد. برای مثال در مورد نگارش همزه میانی و پایانی با کرسی «الف» و نگارش همزه میانی و پایانی با کرسی «واو»، تعداد رکوردهای بازیابی شده یکسان بود. همچنین در جابه‌جایی «ی» و همزه در کلمات فارسی، یا نحوه نوشتن صامت «ی» بعد از «ه» غیر ملفوظ به همین شکل است. در مورد علامت تشدید و حذف یا نوشتن علامت «مد» نیز نتایج بازیابی شده یکسان بود و تأثیری بر نتیجه بازیابی نداشت. اما مثلاً در مورد جدانویسی و پیوسته‌نویسی، واژه‌های دو املائی، نحوه نوشتن «الف» مقصوره و نگارش همزه بدون کرسی، تعداد رکوردهای بازیابی شده متفاوت بود.

نتایج پژوهش حاضر با پژوهش رنجبر و عباسپور (۱۳۹۷) در مورد دشواری جدانویسی و پیوسته‌نویسی و نزدیک‌نویسی، هم‌راستا است. پژوهش آنها نشان داد که در پایگاه‌های اطلاعات علمی مرکز اطلاعات علمی جهاد دانشگاهی، علاوه بر بازیابی مدارک دارای شکل نگارشی جدانویسی، مدارکی دارای یکی از اجزای مصداق‌ها هم بازیابی می‌شدند و این امر موجب ریزش کاذب در مدارک بازیابی شده می‌گردید. همچنین در خصوص برخی دشواری‌ها نیز یافته‌های پژوهش حاضر با نتایج پژوهش گل‌تاجی و بذرگر (۱۳۸۹) هم‌سو است، مانند نتیجه به دست آمده از آن برای واژه‌های دو املائی و واژه‌های با دو هجای میانی «وو» و نیز نحوه نوشتن «الف» مقصوره، نحوه نوشتن همزه بدون کرسی و علامت‌های جمع؛ اما در خصوص برخی دشواری‌ها، نتایج پژوهش حاضر، با نتایج پژوهش آنها هم‌سو نیست. مثلاً در پژوهش حاضر فراوانی رکوردهای بازیابی شده در مورد نوشتن یا ننوشتن علامت «مد» یکسان بود، اما در پژوهش آنها نتایج در هر سه پایگاه مورد بررسی (مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، ایرانداک و جهاد دانشگاهی) متفاوت است.

نتایج این پژوهش در خصوص دشواری علامت تشدید، با نتایج پژوهش هماوندی، نوروزی و بهشتی (۱۳۹۷)، کاملاً هم‌سو نیست؛ چون در پژوهش آنها رکوردهای بازیابی شده برای علامت تشدید در پایگاه‌های «آی‌اس‌سی» و «ایرانداک» یکسان نیست، اما در پایگاه «اس. آی. دی.» و «نورمگز» تعداد رکوردها یکسان است. در خصوص دشواری‌هایی که نتایج بازیابی یکسان نبود،

مشکل ریزش کاذب وجود دارد، یا برخی مدارک بازیابی نشده است. برای اولی مانند چیزی که در جدانویسی و پیوسته‌نویسی دیده شد و برای دومی مانند چیزی که در واژه‌های دو املائی، نگارش همزه بدون کرسی، واژگان دخیل و برابره‌های فارسی آنها و... مشاهد شد و این یعنی جستجو با یک صورت املائی واژه باعث می‌شود مدارکی که با صورت دیگر واژه نوشته می‌شوند، بازیابی نشوند. پس به دلیل وجود چنین دشواری‌هایی در رسم الخط فارسی، اگر شکل‌های مختلف نگارشی توسط طراحان و نمایه‌سازان پایگاه‌های اطلاعات علمی فارسی بهنجارسازی نشود، کاربران با نگارش هر یک از شکل‌های نگارشی واژه، تنها به مدارکی که بدان شکل نگارش یافته است، دسترسی خواهند یافت و سایر مدارک مرتبط نگارش یافته با دیگر شکل‌ها را از دست خواهند داد (رنجبر و عباسپور، ۱۳۹۷).

۵. نتیجه‌گیری

با توجه به داده‌های حاصل از جستجوی کلیدواژه‌های مربوط به هر دشواری در پایگاه گنج، مشاهده شد که عدم یکدستی در نوشتار، بر نتیجه بازیابی تأثیر می‌گذارد. به جرأت می‌توان گفت، دشواری‌های صرفی و معنایی را نیز می‌توان ناشی از ویژگی‌های رسم الخط فارسی دانست. در خصوص مشکلات معنایی، به این دلیل مشکلات معنایی خواننده می‌شوند که تفاوت رسم الخط، در معنای واژه‌ها تغییر ایجاد می‌کند و برای صرف هم که ناظر بر ساخت واژه است و وام‌گیری واژگانی، کلمات مرکب و ترکیبات نحوی و اسامی جمع نتیجه صرف در زبان است. مثلاً در وام‌گیری واژگانی با واژه‌های جدید روبرو می‌شویم که تفاوت در نوشتار و رسم الخط وام‌واژه‌های مختلف می‌تواند بر نتیجه بازیابی تأثیر بگذارد. پس، در بحث رسم الخط به مواردی پرداخته شد که تفاوت در نگارش مصداق‌های مربوط به هر دشواری در برخی موارد بر نتیجه بازیابی تأثیرگذار بود. «دستور خط فارسی» که به عنوان معیاری برای یکدستی در نوشتار از سوی فرهنگستان زبان و ادب فارسی تدوین شده است، هنوز نتوانسته به دشواری‌های رسم الخط فارسی فائق آید. علاوه بر دستور خط فارسی، کتاب «فرهنگ املائی خط فارسی» نیز به اهتمام این سازمان تهیه شده است، اما چیزی که وجود دارد این است که استفاده از چنین معیارهایی برای نگارش، تا چه حد می‌تواند بر نتیجه بازیابی، اثر مطلوب بگذارد؟ چون در دستور خط و فرهنگ املائی در برخی قواعد دست نویسنده را باز گذاشته‌اند. اگر فرض بر این گذاشته شود که استفاده از یک دستور خط مشترک برای نگارش و همین‌طور برای ذخیره‌سازی و نمایه‌سازی پایگاه‌های اطلاعاتی تا حدی می‌تواند

مشکلات را حل کند، اما مدارک مربوط به سال‌های قبل از تدوین این دستورها، با این مشکلات دست به گریبان خواهند بود، مگر اینکه هر پایگاه هزینه کند و به ذخیره مدارک قبلی مطابق با دستور خط مشترک بپردازد و به نظر می‌رسد چنین کاری مقرون به صرفه نخواهد بود و این روش برای بسیاری از دشواری‌ها کارساز نیست. مثلاً در خصوص وام‌گیری واژگانی که در عصر جدید نمی‌توان وام‌گیری واژگانی را به کلی نفی کرد (طباطبایی، ۱۳۸۶)، دادوستد زبانی از قدیم بوده، از این پس نیز خواهد بود و در حقیقت بده‌بستان کالایی و فرهنگی چنین مبادله‌ای را به همراه یا در پی دارد (سمعی، ۱۳۷۵) یا در مورد هم‌نامی که واژه‌ها هم‌نویسه و هم‌آوا هستند، البته مسأله خط و ارتباط آن بازیابی اطلاعات پیچیده‌تر از این حرف‌هاست که بتوان چنین نسخه ساده‌ای را برای آن پیچید. از طرفی پیشرفت فناوری در حال حرکت به سمت پردازش زبان طبیعی است. توجه به مسائل معنایی، به ویژه از این جهت، نیاز به توجه جدی دارد. شاید در خصوص دشواری‌های رسم‌الخط، با آگاهی از دو یا چندگانه‌نویسی واژه‌ها و یا با پیشنهاد استفاده از یک دستور خط مشترک و رعایت یکدستی در نوشتار و مواردی اینچنینی بتوان آن را حل کرد، اما در مورد مشکلات معنایی این‌گونه نیست. همان‌طور که گفته شد، یکی از روش‌های رفع ابهام برای واژه‌های هم‌نویسه استفاده از مصوت‌های کوتاه است که نتایج به دست آمده در پایگاه گنج نشان داد که این پایگاه مصوت‌ها را مانند دیگر علامت‌ها مثل تشدید و علامت مد نادیده می‌گیرد که با توجه به پژوهش هماوندی، نوروزی و بهشتی (۱۳۹۷) و همین‌طور پژوهش‌های دیگر، عدم توجه به مشکلات معنایی در پایگاه‌های دیگر نیز مشاهده شده است. از این رو، توجه به این دسته از مشکلات بسیار مهم است که متخصصان علم اطلاعات و سازمان‌های اداره‌کننده این پایگاه‌ها، باید به آن توجه ویژه‌ای داشته باشند؛ چرا که احتیاج فارسی‌زبانان به ویژه به ابزارهای کاوش بومی و پایگاه‌های اطلاعاتی که مبتنی بر ویژگی‌های زبانی خودشان طراحی شده و در تعامل با کاربران فارسی‌زبان به ظرائف و باید و نبایدهای زبانی آنها بیشتر توجه شود، بیش از پیش احساس می‌گردد. از سوی دیگر، پایگاه‌های اطلاعاتی نیز باید نسبت به برآورده ساختن نیازهای زبانی کاربران و اصلاح تعامل با کاربران بیشتر تلاش کنند (هماوندی، نوروزی و بهشتی، ۱۳۹۷).

با توجه به اینکه یکی از مهم‌ترین چالش‌های حوزه ذخیره و بازیابی اطلاعات، نزدیک‌تر کردن زبان موضوعی جویندگان و ارائه‌دهندگان اطلاعات است (رضایی شریف‌آبادی و همکاران، ۱۳۸۹)، از این رو پژوهش‌های انجام شده در زمینه کارآمدی کلیدواژه‌ها بر این عقیده‌اند که یک پایگاه اطلاعاتی در حوزه موضوعی خاص وقتی می‌تواند معتبر باشد که درصد زیادی از

کلیدواژه‌های موضوعی آن حوزه را تحت پوشش قرار دهد و به تبع آن برای آن حوزه موضوعی بیشترین منابع را در اختیار بگذارد (شهبازی و شاهینی، ۱۳۹۴). با توجه به این مهم می‌توان گفت که یکی از عامل‌هایی که باید برای بررسی کارآمدی کلیدواژه‌ها در نظر داشت، توجه به مسأله خط و نگارش کلیدواژه‌ها است که باید در ابعاد مختلف صرفی، معنایی و رسم الخط مورد توجه ارائه‌دهندگان اطلاعات باشد.

همان‌طور که پیش‌تر اشاره شد، پایگاه گنج بنیان سامانه‌های دیگر ایرانداک نیز است، لذا، بی‌توجهی به مشکلات صرفی، معنایی و رسم الخط می‌تواند بر کار دیگر سامانه‌ها نیز اثر بگذارد. برای مثال در سامانه پیشینه پژوهش، دانشجویان کارشناسی ارشد و دکتری باید عنوان پژوهش خود را در آن ثبت کنند تا اطمینان یابند کار مشابهی با کار آنها انجام نشده یا نمی‌شود. برای این کار پژوهشگر باید عنوان و کلیدواژه‌های کار خود را در سامانه ثبت کند و منتظر پاسخ باشد. لذا، احتمال می‌رود این بی‌توجهی به دشواری‌های گفته شده که در سه دسته مشکلات صرفی، معنایی و رسم الخط شرح آن رفت، بر درستی و دقت عملکرد این سامانه و حتی دیگر پیشینه‌ها تأثیر بگذارد. برای مثال اگر در عنوان پژوهشی واژه‌ای دو املائی آمده باشد و در پژوهش دیگری نیز همان واژه با املائی دیگر آن ثبت شده باشد، این احتمال وجود دارد که سامانه به درستی این مسأله را تشخیص ندهد و در واقع پیشینه‌ای را نشان ندهد.

۶. پیشنهادهای پژوهش

۶-۱. پیشنهادهای کاربردی

با توجه به داده‌های حاصل از جستجوی کلیدواژه‌های مربوط به هر دشواری که در پایگاه گنج مشاهده شد، عدم یک‌دستی در نوشتار، بر نتیجه‌بازایی تأثیر می‌گذارد. بنابراین، پیشنهاد می‌شود:

- پایگاه گنج حین ذخیره اطلاعات مربوط به مدارک در پایگاه، بهتر است با توجه به یک دستور خط مشترک (دستور خط فارسی توصیه می‌شود)، اطلاعات را در پایگاه ذخیره کنند؛ چرا که احتمال استفاده از شیوه‌های نگارشی از سوی نویسندگان بیشتر است. البته باید از پژوهشگران هم بخواهند مدارک خود را براساس دستور خط مشترک به نگارش درآورند و بر این مسأله تأکید زیادی داشته باشند تا پژوهشگران این مسأله را جدی بگیرند.

- همچنین پیشنهاد می‌شود در مرحله نمایه‌سازی، دشواری‌هایی مثل وام‌گیری واژگانی، واژه‌های دخیل و برابرای فارسی آنها و هم معنایی مد نظر قرار گیرد. بسیاری از مشکلات مربوط

به رسم الخط مانند «الف» مقصوره، واژه‌های دو املائی، واژه‌هایی با هجای میانی «وو» را نیز در مرحله نمایه‌سازی می‌توان مرتفع ساخت.

● سامانه تحلیل‌گر صرفی زبان فارسی با عنوان اختصاری «پارس مورف» که توسط محققان دانشگاه صنعتی شریف و دانشگاه زنجان ساخته شده است، از پیشرفت‌های حوزه پردازش زبان طبیعی می‌باشد. در مشکلات صرفی برای دشواری کلمات مرکب ترکیبات نحوی مشاهده شد، جدانویسی واژه‌ها بر نتیجه بازایی تأثیرگذار است و این به دلیل معتبر نبودن فاصله به عنوان مرز کلمه است. یکی از قابلیت‌های سامانه «پارس مورف» در تجزیه ستاک‌های پیچیده این است که ترکیب‌های بالقوه را در اختیار می‌گذارد. از این‌رو، استفاده از چنین سامانه‌هایی در پایگاه‌های اطلاعات علمی پیشنهاد می‌شود.

۶-۲. پیشنهادهای پژوهشی

با توجه به نتایج به دست آمده، پیشنهادهای زیر برای پژوهش‌های آینده پیشنهاد می‌شود: برای بررسی رفتارهای نگارشی افراد در پایگاه‌های اطلاعات علمی پیشنهاد می‌شود که پژوهشی انجام شود تا مشخص گردد که افراد در برخورد با دشواری‌های نگارشی چگونه برخورد می‌کنند. آیا با شیوه‌های نگارشی واژه‌ها آشنایی دارند؟ و اگر دارند، حین برخورد با آنها از شکل‌های مختلف نگارشی استفاده می‌کنند؟ همچنین پیشنهاد می‌شود رفتار نگارشی افراد در پایگاه‌های اطلاعات علمی با رفتار نگارشی افراد در موتورهای جستجو جداگانه بررسی شود، چرا که به دلیل ماهیت علمی پایگاه‌های اطلاعات علمی، مخاطبان آنها متفاوت هستند و در موتورهای جستجو بهتر است عموم کاربران در نظر گرفته شوند.

در پایگاه گنج حین جستجوی واژه‌های هم‌نام و هم‌نویسه، به صورت پیش‌فرض، کلیدواژه‌هایی پیشنهاد می‌شود. لذا، می‌تواند این مسأله در پایگاه‌های اطلاعاتی بررسی شود تا مشخص گردد که این راه‌حل برای دشواری‌های هم‌نامی و هم‌نویسی تا چه میزان مفید فایده است؟

از طرفی به دلیل اهمیت پرداختن به مسائل معنایی، خصوصاً هم‌نامی و هم‌نویسی، پیشنهاد می‌شود پژوهشگران حوزه علم اطلاعات و دانش‌شناسی با پژوهشگران حوزه زبان و ادبیات فارسی و نیز متخصصان مهندسی رایانه و هوش مصنوعی در این زمینه به انجام پژوهش‌های بنیادی دست بزنند.

منابع

- آخشیک، س. (۱۳۹۴). بازتاب دشواری‌های نگارش کلمه در بازیابی اطلاعات بانک نشریات کشور (مگ ایران). در: اولین کنفرانس بین‌المللی بازیابی تعاملی اطلاعات.
- آخشیک، س.س.، فتاحی، ر. (۱۳۹۱). تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی. کتابداری و اطلاع‌رسانی، ۱۵(۳): ۹-۳۰.
- ارسطوپور، ش.، احمدی‌نسب، ف. (۱۳۹۱). آسیب‌شناسی زبان و خط فارسی در بازیابی اطلاعات: نگاهی به موتورهای کاوش و پایگاه‌های برخط. در: نخستین کنفرانس ملی مدیریت منابع اطلاعاتی وب، سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران.
- جلالی، و. (۱۳۸۷). بازیابی معنایی اطلاعات با استفاده از بسط مفاهیم حاصل از جستجوی مبتنی بر کلیدواژه. پایان‌نامه کارشناسی ارشد. دانشگاه صنعتی امیرکبیر.
- حری، ع. (۱۳۷۲). کامپیوتر و رسم الخط. پیام کتابخانه، ۳(۱).
- رائی ساربانقلی، م.ص. (۱۳۸۵). مشکلات جستجو و بازیابی اطلاعات به زبان فارسی در اینترنت، مطالعه موردی: کاربران مرکز اینترنت دانشگاه آزاد اسلامی واحد شبستر. فصلنامه کتاب، ۱۷(۳): ۱۷۹-۱۹۶.
- رضایی‌شریف‌آبادی، س.، خسروی، ع.، حاجی‌زین‌العابدینی، م. (۱۳۸۹). امکان‌سنجی مهار مستندات موضوعی پزشکی در پایگاه‌های اطلاعاتی فارسی موجود بر روی وب. مطالعات تربیتی و روان‌شناسی، ۸(۳): ۱۸۳-۲۰۱.
- DOI: 10.22067/RIIS.V8I3.5725
- رنجبر، ا.، عباسپور، ج. (۱۳۹۷). گسترش‌پذیری جستجو و بازیابی مدارک در پایگاه‌های اطلاعات علمی فارسی: موردپژوهی پیوسته‌نویسی و جدانویسی. کتابداری و اطلاع‌رسانی، ۲۱(۳): ۵۷-۹۰.
- DOI: 10.30481/LIS.2018.67485
- ستوده، ه.، هنرجویان، ز. (۱۳۹۱). مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آنها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات. کتابداری و اطلاع‌رسانی، ۱۵(۴): ۵۹-۹۲.
- ستوده، ه.، هنرجویان، ز. (۱۳۹۳). بررسی تنوع الگوهای نگارشی فارسی و تأثیر آن بر جامعیت بازیابی اطلاعات (مطالعه موردی: پیکره همشهری). کتابداری و اطلاع‌رسانی، ۱۷(۲): ۳۱-۴۹.
- سمیعی، ا. (۱۳۷۵). یادداشت سردبیر. نامه فرهنگستان، ۵: ۱-۳.
- شهبازی، م.، شاهینی، ش. (۱۳۹۴). بررسی میزان کارایی پایگاه‌های اطلاعاتی مگ ایران، نورمگز و اس.آی.دی در بازیابی و ربط مباحث علم اطلاعات و دانش‌شناسی با استفاده از کلیدواژه‌های آزاد و مقایسه آنها از نظر میزان استفاده از کلیدواژه‌های مهار شده. پژوهش‌نامه پردازش و مدیریت اطلاعات، ۳۱(۲): ۴۵۴-۴۳۱.
- طاهری اسکویی، م.، پروینی‌راد، ز.، تباری، پ. (۱۳۹۵). بررسی روابط معنایی در اسامی مرکب برون مرکز زبان فارسی. زبان و ادب فارسی، ۸(۲۷-۲۶): ۲۶۱-۲۷۸.
- طباطبائی، ع. (۱۳۸۶). صرف زبان فارسی. بخارا، ۱۰(۶۳): ۲۱۲-۲۴۲.
- فرهنگستان زبان و ادب فارسی (۱۳۹۴). دستور خط فارسی. تهران: فرهنگستان زبان و ادب فارسی.
- فرهنگستان زبان و ادب فارسی (۱۳۹۸). مطالعات واژه‌گزینی. ویژه‌نامه فرهنگستان، ۳.
- گل‌تاجی، م.، بذگر، س. (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).

هماوندی، ه.، نوروزی، ی.، بهشتی، م. (۱۳۹۷). بررسی مشکلات جستجو و بازیابی اطلاعات در پایگاه‌های اطلاعاتی از جنبه ویژگی‌های نگارشی زبان فارسی. پژوهش‌نامه پردازش و مدیریت اطلاعات، ۳۳(۳).

References

- Academy of Persian Language and Literature (2015). **Persian calligraphy**. Tehran: Academy of Persian Language and Literature. [in persian]
- Academy of Persian Language and Literature (2019). Vocabulary studies. *Special Issue of the Academy*, 3. [in persian]
- Akhshik, S.S. & Fattahi, R. (2012). **Analysis of the challenges of spelling and separating Persian words in information storage and retrieval in databases**. *Library and Information Quarterly*, 15(3): 9-30. [in persian]
- Akhshik, S.S. (20). **Reflection of difficulty of writing the word in retrieving the information of the country's publications bank (Magiran)**. In: The first International Interactive Information Retrieval Conference. [in persian]
- Arastoopour, Sh. & Ahmadi Nasab, F. (2012). **Pathology of Persian language and script in information retrieval: A look at search engines and online databases**. In: The First National Conference on Web Information Resources Management, National Library and archives of Iran. [in persian]
- Goltagi, M. & Bazrgar, S. (2010). Investigation of Persian language morphology problems in three databases of Regional Information Center for Science and Technology, Iranian Research Institute for Information Science and Technology, and Jihad Daneshgahi. *Quarterly Journal of Library and Information*, 13(2). [in persian]
- Homavandi, H., Norozi, Y. & Beheshti, M. (2018). Survey of Information Searching and Retrieving Challenges in Databases in Connection with Persian Language Writing Features. *Journal of Information Processing and Management*, 33(3): 1087-110. [in persian]
- Horri, A. (1993). Computer and calligraphy. *Payame Payame Ketabkhaneh*, 3(1). [in persian]
- Jalali, V. (2008). **Semantic retrieval of information by expanding present concepts from keyword-based search**. Master Thesis. Amirkabir University of Technology. [in persian]
- Monz, C. & De Rijke, M. (2002). **Shallow Morphological Analysis in Monolingual Information Retrieval for Dutch, German, and Italian**. Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross Language Evaluation Forum, CLEF 2001, Darmstadt, Germany.
- Moukdad, H. (2005). Lost in cyberspace: How Do Search Engines Handle Arabic Queries? *The international information & library review*, 37(4): 237-394.
- Palinkas, L.A., Horwitz, S.M., Green, C.A., Wisdom, J.P., Duan, N. & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research*, 42(5): 533-544.
- Ranjbar, A. & Abaspor, J. (2018). Extension of search and retrieval of documents in Persian science databases: A case study of Connected Spelling and Separation. *Quarterly Journal of Library and Information*, 21(3): 57-90. [in persian] DOI: 10.30481/LIS.2018.67485

- Rasi Sarbangholi, M.S. (2006). Problems of searching and retrieving information in Persian on the Internet, a case study: Users of the Internet Center of Islamic Azad University, Shabestar Branch. *Faslname Ketab*, 17(3): 179-196. [in persian]
- Rezaei Sharifabadi, S., Khosravi, A. & Haji Zeinolabedini, M. (2010). Feasibility study of controlling medical subject documents in Persian databases available on the web. *Educational and Psychological Studies*, 8(3): 183-201. [in persian] DOI: 10.22067/RIIS.V8I3.5725
- Samiei, A. (1996). Editor in chief lecture. *Academy letter*, 5: 1-3. [in persian]
- Shahbazi, M. & Shahini, Sh. (2015). Evaluation of the efficiency of Magiran, Noormagz, and SID databases in retrieval and relevance of information science and knowledge issues using free keywords and their comparison in terms of use of Restrained keyword. *Journal of Information Processing and Management*, 31(2): 431-454. [in persian]
- Sotodeh, H. & Honarjoyan, Z. (2012). A review of the difficulty of the Persian language in the digital environment and their effects on the effectiveness of automated text processing and data retrieval. *Quarterly Journal of Library and Information*, 15(4): 59-92. [in persian]
- Sotodeh, H. & Honarjoyan, Z. (2014). Investigating the Diversity of Pattern of Persian writing and its Effect on Information Retrieval Comprehensiveness (Case Study: Hamshahri Figure). *Quarterly Journal of Library and Information*, 17(2): 31-49. [in persian]
- Tabatabaei, A. (2007). Morphology of Persian language. *Bukhara*, 10(63): 212-242. [in persian]
- Taheri Oskouei, M., Parvini Rad, Z. & Tabari, P. (2016). Investigating Semantic Relationships in Compound nouns outside of the Persian language center. *Journal of Persian Language and Literature, Islamic Azad University of Sanandaj*, 8(26-27): 261-278. [in persian]
- Toth, E. (2006). Exploring the Capabilities of English and Hungarian Search Engine for Various Queries. *Libri*, 56: 38-47.