Survey of Information Searching and Retrieving Challenges in Databases in Connection with Persian Language Writing Features

Hoda Homavandi
PhD Candidate in Knowledge and Information Science;
Tehran University;
Corresponding Author h.homavandi@ut.ac.ir

Yaghub Norouzi
PhD in Knowledge and Information Science; Associate Professor;
Qom University ynorouzi@gmail.com

Moluk S. Hoseini Beheshti
PhD in General Linguistics; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc);
Tehran, Iran beheshti@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 24, Jun. 2016 **Accepted: 30, Jan. 2017**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 33 | No. 3 | pp. 1087-1110
Spring 2018
https://doi.org/10.35050/JIPM010.2018.042

Abstract: The present research was carried out with the aim of explicating the major writing and semantic problems of Persian language when using data environments and determining the degree of compatibility and attention to these features in Persian databases. This research is of survey analytical type being conducted through direct observation. Having reviewed the related literature, we kept a checklist of search keywords. Each of these keywords was searched in the databases under study, such as Iranian Research Institute for Information Science and Technology, Regional Centre for Information Science and Technology, NoorMagaz, and Scientific Information Database affiliated with Jahad Daneshgahi, and the number of retrieved findings was recorded.

Some of the writing and semantic features of Persian language contribute to problems associated with retrieving information from the selected databases. Some of these features include connected and disconnected forms of writing of derivative, compound, and derivative-compound words, diversity of plural forms, loanwords and their equivalents in writing as well as polysemy, homonymy, etc., in semantics. For instance, retrieving different results for various writing forms of the keywords «فناوری و فن آوری» as derivative-compound words or «پتاسیوم و پتاسیم» as various forms of recording words, or retrieving different findings for keywords «دریای خزر، دریای مازندران و دریای کاسپین» as well as lack of their

appropriate coverage as synonymous words and giving the user information about it in order to improve the exploration process, for it has negative effects on search and retrieval process. Findings indicated that Persian databases do not pay adequate attention to writing and semantic features of Persian language, and disregard many of its features in searching and retrieving information. In connection with the impact of these features on the interaction of users with databases, Persian-speaking users' need for native exploration tools and databases designed in accordance with the features of their own language have become more and more urgent. The present research has examined the ability of Persian databases in covering some of the features of this language, which have a noticeable impact on the process of searching and retrieval, pinpointing the weak points and strengths of these databases. The results of the present research could be utilized to improve the performance of the above-mentioned databases.

Keywords: Information Retrieval, Databases, Persian Language, Writing Features

بررسی مشکلات جست‌وجو و بازیابی اطلاعات در پایگاه‌های اطلاعاتی از جنبهٔ ویژگی‌های نگارشی زبان فارسی

هدی هماوندی
دانشجوی دکتری علم اطلاعات و دانش‌شناسی؛ دانشگاه تهران؛
پدیدآور رابط h.homavandi@ut.ac.ir

یعقوب نوروزی
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ دانشگاه قم ynorouzi@gmail.com

ملوک‌السادات حسینی بهشتی
دکتری تخصصی زبان‌شناسی‌عمومی؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
beheshti@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۵/۰۴/۰۴ پذیرش: ۱۳۹۵/۱۱/۱۱ مقاله برای اصلاح به مدت ۵۸ روز نزد پدیدآوران بوده است.

چکیده: این پژوهش با هدف تشریح مشکلات عمدهٔ نوشتاری و معنایی زبان فارسی در استفاده از محیط‌های اطلاعاتی و تعیین میزان انطباق و توجه به این ویژگی‌ها هنگام جست‌وجو و بازیابی در پایگاه‌های اطلاعاتی فارسی و به روش پیمایشی-تحلیلی و با استفاده از شیوهٔ مشاهدهٔ مستقیم انجام گرفت. پس از مرور پژوهش‌های مرتبط، کلیدواژه‌های کاوش در قالب یک سیاهه شکل گرفت. هر یک از این کلیدواژه‌ها در پایگاه‌های اطلاعاتی مورد مطالعه شامل «پژوهشگاه علوم و فناوری اطلاعات ایران»، «پایگاه استنادی علوم جهان اسلام»، «پایگاه مجلات تخصصی نور» و «پایگاه اطلاعات علمی جهاد دانشگاهی» جست‌وجو و تعداد نتایج بازیابی‌شده ثبت گردید. سپس، به بررسی میزان انطباق پایگاه‌های اطلاعاتی با این ویژگی‌ها پرداخته شد.

برخی ویژگی‌های نوشتاری و معنایی زبان فارسی سبب بروز مشکلاتی در بازیابی اطلاعات از پایگاه‌های اطلاعاتی منتخب می‌شوند. مواردی مانند پیوسته‌نویسی و جدانویسی واژگان مشتق، مرکب و مشتق- مرکب، گوناگونی جمع‌ها، واژگان دخیل و معادل آن‌ها در بخش نوشتاری و چندمعنایی، همنامی و ... در بخش معنایی از این دست ویژگی‌ها هستند. فقدان پوشش مناسب ویژگی‌های یادشده در مراحل ذخیره‌سازی و پردازش و عدم آگاه‌نمودن کاربر از آن جهت اصلاح فرایند کاوش در مرحلهٔ بازیابی

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۳ | شمارهٔ ۳ | ص‌ص ۱۰۸۷-۱۱۱۰
بهار ۱۳۹۷
https://doi.org/10.35050/JIPM010.2018.042

اطلاعات در پایگاه‌های اطلاعاتی مورد پژوهش، اثرات نامطلوبی بر فرایند کاوش و بازیابی دارد. یافته‌ها نشان داد که پایگاه‌های اطلاعاتی فارسی نسبت به ویژگی‌های نوشتاری و معنایی زبان فارسی توجه کافی نداشته و بسیاری از ویژگی‌های آن را در مراحل ذخیره‌سازی و پردازش اطلاعات نادیده می‌گیرند. با توجه به تأثیر این ویژگی‌ها در تعامل کاربران با پایگاه‌های اطلاعاتی، احتیاج کاربران فارسی‌زبان به ابزارهای کاوش بومی و پایگاه‌های اطلاعاتی که مبتنی بر ویژگی‌های زبانیِ خودشان طراحی شده باشد، بیش از پیش احساس می‌شود. پژوهش حاضر با بررسی میزان توانایی پایگاه‌های اطلاعاتی فارسی‌زبان در پوشش برخی ویژگی‌های این زبان که در فرایند جست‌وجو و بازیابی تأثیر قابل توجهی دارند، نقاط ضعف و قوت این پایگاه‌ها را مشخص نموده است. نتایج آن می‌تواند در جهت بهبود و اصلاح عملکرد پایگاه‌های مذکور مورد استفاده قرار گیرد.

کلیدواژه‌ها: بازیابی اطلاعات، پایگاه‌های اطلاعاتی، زبان فارسی، ویژگی‌های نگارشی

۱. مقدمه

رایانه‌ها و بستر شبکهٔ جهانی وب بدون اغراق از عمده‌ترین و پراستفاده‌ترین ابزارهای دسترسی به اطلاعات در دنیای امروز هستند. این موضوع چالش‌ها و فرصت‌هایی را موجب شده است. یکی از این چالش‌ها، تنوع زبان‌های مورد استفاده و نحوهٔ پشتیبانی ابزارهای دسترسی به اطلاعات از آن‌هاست. مسئلهٔ استفاده از زبان طبیعی در محیط‌های رایانه‌ای از دیدگاه Ruiter (2006) ابعاد گوناگونی دارد که بحث بازیابی اطلاعات معمولاً به‌عنوان دغدغه‌ای همیشگی در آن طرح می‌شود. تعامل کاربران با اطلاعات و شبکه‌های ارتباطی و اطلاعاتی نظیر اینترنت، وب، پایگاه‌های اطلاعاتی و موتورهای کاوش همواره با میانجی زبان است. آن‌ها از این طریق مفاهیم مورد نظر خود را بیان و اطلاعات را کاوش و بازیابی می‌کنند. «مارچیونینی» عقیده دارد که امروزه یکی از اساسی‌ترین دغدغه‌ها در حوزهٔ علم اطلاعات عبارت است از توضیح چگونگی تعامل انسان با آنچه که خود ساخته است و این ابعاد و مباحث گوناگونی از جمله زبان را دربرمی‌گیرد. همین تنوع و تکثر است که آن را به یک حوزهٔ میان‌رشته‌ای تبدیل کرده است (Marchionini 2008). «زره‌ساز و فتاحی» (۱۳۸۵) نیز اشاره می‌کنند که در سال‌های اخیر، این حوزه با تأثیرپذیری از حوزه‌هایی مانند روان‌شناسی، علوم رایانه، علم اطلاعات، جامعه‌شناسی، و سایر حوزه‌های مشابه گسترش بسیاری یافته و متون و تحقیقات فراوانی نیز در این زمینه پدید آمده است. با توجه به این رویکرد، دور از ذهن نیست اگر یکی از موارد مؤثر ذکرشده را، که رایانه‌ها

به‌عنوان ابزارهای اطلاعاتی همواره در مواجهه با پیچیدگی‌های آن مشکلاتی دارند، زبان طبیعی دانست. البته، در این بیان منظور از زبان طبیعی، زبانی است که انسان‌ها معمولاً در محاورات یا نوشته‌های خود از آن استفاده می‌کنند و به ناچار ممکن است چالش‌هایی را در این زمینه ایجاد کند.

آمارهای مربوط به سال ۲۰۱۵ در مورد استفاده از اینترنت بر اساس زبان حاکی از آن است که حدود ۶۲/۴ درصد کاربران، انگلیسی‌زبان و ۳۷/۶ درصد غیرانگلیسی‌زبان هستند (Internet world stats: Usage and population statistics 2015). این امر از رشد روزافزون کاربران غیرانگلیسی‌زبان نشان داشته و لزوم توجه به نیازهای زبانی و حل مسائل مربوط به آن‌ها را (با توجه به این‌که زبان غالب در شبکهٔ وب و ابزارهای کاوش انگلیسی است) یادآور می‌شود. کاربران فارسی‌زبان نیز با ویژگی‌های خاص زبانی خود در زمرهٔ دومین گروه هستند. از همین‌روست که در سال‌های اخیر، بسیاری از تحقیقات بین‌رشته‌ای به بررسی تأثیر مسائل زبانی در تعامل کاربران و محیط وب معطوف شده‌اند؛ تحقیقاتی که هر یک از دیدگاهی بحث زبان در کاوش و بازیابی اطلاعات را در انواع رسانه‌های اطلاعاتی اعم از پایگاه‌های تحت وب، موتورهای کاوش، وب‌سایت‌های تجاری، کتابخانه‌ای و ... مورد کنکاش قرار داده و مشکلات ناشی از آن را بررسی نموده‌اند. عدم توجه به ویژگی‌های تأثیرگذار زبان فارسی از جمله ویژگی‌های نگارشی در مراحل ذخیره‌سازی و پردازش، جست‌وجو و بازیابی اطلاعات از پایگاه‌های اطلاعاتی موجب ایجاد چالش‌ها و موانعی پیشِ روی کاربران فارسی‌زبان در دستیابی به اطلاعات مورد نیازشان شده است. به‌عنوان مثال، بازیابی نتایج متفاوت برای صورت‌های مختلف نوشتاری کلیدواژه‌های «گلابگیری»، «گلاب‌گیری» به‌عنوان واژگان مشتق-مرکب و یا «انفولانزا»، «آنفلوآنزا»، «آنفولانزا» به‌عنوان صورت‌های مختلف ضبط واژگان یا بازیابی نتایج متفاوت برای کلیدواژه‌های مترادف «دریای خزر»، «دریای مازندران» و «دریای کاسپین» از پیامدهای عدم پوشش مناسب آن‌ها در مرحلهٔ نمایه‌سازی و پردازش است. این است که در مطالعهٔ پیشِ رو، به تبیین مشکلات عمدهٔ نگارشی زبان فارسی، اعم از نوشتاری و معنایی، در تعامل انسان و محیط‌های کاوش وب مانند پایگاه‌های اطلاعاتی پرداخته می‌شود. همچنین، در پژوهش حاضر تلاش بر این است که میزان انطباق و سازگاری نمونه‌هایی از پایگاه‌های اطلاعاتی برگزیدهٔ داخلی، شامل پایگاه‌های اطلاعاتیِ «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)»، «پایگاه استنادی علوم جهان اسلام (آی‌اس‌سی)»، «پایگاه مجلات تخصصی نور (نورمگز)» و «پایگاه

اطلاعـات علمـی جهـاد دانشـگاهی (سـید)» بـا مؤلفه‌هـای زبـان فارسـی از حیـث جسـت‌وجو و بازیابـی اطلاعـات بررسـی شـده و ضمـن شناسـایی چالش‌هـا در بخـش نوشـتاری و معنایـی زبـان، بیـن پایگاه‌هـای یادشـده نیـز مقایسـه‌ای انجـام گیـرد تـا بدین‌ترتیـب، راهکارهایـی بـرای پیشـگیری و حـل مشـکلات یادشـده ارائـه گـردد.

۲. مروری بر پژوهش‌های مرتبط

تاکنـون پژوهش‌هـای گوناگونـی انجـام گرفتـه کـه هـر یـک بـه نوعـی مسـائل زبـان فارسـی را در تعامـل انسـان بـا رایانـه، موتورهـای کاوش، پایگاه‌هـای اطلاعاتـی مختلـف بررسـی نموده‌انـد. مطالعـات یادشـده در اکثـر مـوارد معطـوف بـه ویژگی‌هـای نوشـتاری زبـان فارسـی در بازیابـی اطلاعـات متنـی بوده‌انـد. برخـی پژوهش‌هـای صورت‌گرفتـه در ایـن زمینـه در جـدول ۱، ارائـه شـده اسـت.

جدول ۱. پیشینه‌های داخلی مربوط به موضوع پژوهش

نویسنده/ سال	عنوان	هدف	روش	نتیجه
حری (۱۳۷۲)	کامپیوتر و رسم‌الخط فارسی	تبیین مشکلات رسم‌الخط فارسی در ذخیره و بازیابی اطلاعات در محیط رایانه‌ای	مطالعهٔ موردی	استفاده از سیاست‌گذاری واحد برای یکسان‌سازی رسم‌الخط
نشاط (۱۳۷۹)	مسائل رسم‌الخط فارسی در رویارویی با فناوری نوین اطلاعاتی	توصیف ناهماهنگی‌های میان زبان فارسی و نظام‌های رایانه‌ای	توصیفی	ارزیابی راهکارهای ممکن جهت تطبیق رسم‌الخط فارسی با محیط‌های رایانه‌ای
اکبری‌نژاد (۱۳۷۶)	فاصلهٔ خالی میان واژه‌ها در ذخیره و بازیابی رایانه‌ای اطلاعات	تبیین و تشریح تأثیر فاصلهٔ خالی میان واژه‌ها در ذخیره و بازیابی اطلاعات	پیمایشی- توصیفی	هماهنگی و یکدستی در فاصله‌گذاری
مرتضائی (۱۳۸۰)	مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات	ارائهٔ نمونه‌هایی از تجربه‌های واژه‌گزینی در ذخیرهٔ اطلاعات به‌منظور تسریع و تسهیل ذخیره و بازیابی اطلاعات به زبان فارسی	کتابخانه‌ای- توصیفی	مسائل زبان و خط فارسی سبب کندی مراحل ذخیره و بازیابی اطلاعات، کاهش نسبت بازیافت اطلاعات و تأثیر منفی بر جامعیت نتیجهٔ یک جست‌وجو می‌شوند.

نویسنده/ سال	عنوان	هدف	روش	نتیجه
اسلامی (۱۳۸۱)	دشواری‌های پردازش رایانه‌ای برای پردازش رایانه‌ای خط فارسی	تبیین مشکلات خط فارسی کتابخانه‌ای		تفکیک و دسته‌بندی دشواری‌های خط فارسی در مقولاتی مثل نحو و ایهام در نقش‌ها هنگام همنشینی واژگان و ...
راثی ساربانقلی (۱۳۸۵)	مشکلات جست‌وجو و بازیابی اطلاعات اینترنت از دیدگاه کاربران به زبان فارسی در مرکز اینترنت دانشگاه آزاد واحد شبستر موردی: کاربران مرکز اینترنت دانشگاه آزاد اسلامی واحد شبستر	تبیین مشکلات جست‌وجو و بازیابی اطلاعات در اینترنت، مطالعهٔ	پیمایشی- توصیفی	استفادهٔ اغلب کاربران از موتور کاوش گوگل و وجود مشکلات به‌دلیل عدم توجه به شکل‌های مختلف نوشتاری واژه‌ها
روح‌پرور و بی‌جن‌خان (۱۳۸۶)	به‌کارگیری یک نظام برچسب‌دهی برای تعبیر و تفسیر یک پیکرهٔ متنی زبان فارسی	به‌کارگیری استانداردی برای حل مسئلهٔ هم‌نگاره‌ها و مشخص‌کردن مرز گروه‌های نحوی به‌عنوان مشکلاتی بر سر راه پردازش متون فارسی	توصیفی	تأثیر زیاد استفاده از برچسب‌دهی سلسله‌مراتبی بر حل مسئلهٔ هم‌نگاره‌ها و مشخص‌کردن مرز گروه‌های نحوی
رسولی و بیدگلی (۱۳۸۷)	روشی جدید در خطایابی املایی در زبان فارسی	ارائهٔ روشی برای یافتن خطاهای املایی واژگان با استفاده از الگوریتم‌های هوشمند و یادگیرنده	پیمایشی- توصیفی	بررسی مشکلات موجود در رسم‌الخط رایانه‌ای زبان فارسی در مورد حروفی که در رسم‌الخط رایانه‌ای دارای چند نوع حرف هستند و رفع مشکلات رسم‌الخط با ارائهٔ پیشنهادهای صحیح به کاربران از طریق الگوریتم خطایاب
عبدالهی نورعلی و جوکار (۱۳۸۸)	چالش‌های شیوهٔ نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش‌هستند. وب	بررسی مسائلی که کاوشگران فارسی در کاوش ریخت‌های مختلف‌اسنادی یک واژه با آن روبه‌رو	پیمایش مقایسه‌ای و	عدم توجه موتورهای کاوش وب به شیوه‌های نگارش زبان فارسی به‌منظور بهبود کاوش و وجود رابطهٔ معنادار بین شکل واژه و نوع ابزار جست‌وجو
گل‌تاجی و بذرگر (۱۳۸۹)	بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی «آی‌اس‌اسی»، «سید»، و «ایرانداک»	بررسی مشکلات ریخت‌شناسی زبان فارسی در پایگاه‌های اطلاعاتی فارسی	پیمایش مقایسه‌ای	تأثیر زیاد چالش‌های ریختی شناخته‌شدهٔ زبان فارسی بر بازیابی اطلاعَات در هر یک از سه پایگاه مورد نظر و عدم پرداختن به حل مسائل ریخت‌شناسی واژگان فارسی به شیوه‌ای جامع و قابل ملاحظه از سوی پایگاه‌های مورد پژوهش

نویسنده/ سال	عنوان	هدف	روش	نتیجه
آخشیک و فتاحی (۱۳۹۱)	تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی	بررسی وضعیت توجه نویسندگان و همچنین پایگاه‌های مورد مطالعه به ویژگی پیوسته‌نویسی و جدانویسی واژگان فارسی و ارائهٔ راهکارهایی برای حل این مشکلات	تحلیل محتوا	عملکرد بهتر پایگاه اطلاعاتی «ایرانداک» نسبت به پایگاه «آی‌اس‌سی» در بازیابی عنوان پایان‌نامه‌ها در حالت‌های مختلف پیوسته و جدا نوشته‌شده و لزوم تأکید به نگارندگان پایان‌نامه‌ها به استفاده از قواعد یکدست ملی به‌ویژه در نگارش کلمات دو جزیی و مشتق
ستوده و هنرجویان (۱۳۹۱)	مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آن‌ها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات	بررسی متون و پیشینه‌های موجود به‌منظور تبیین چالش‌های نگارش فارسی،محتوا تأثیر آن‌ها بر اثربخشی بازیابی اطلاعات، و ارائهٔ پیشنهادهایی در جهت رفع این دشواری‌ها	متن‌پژوهی با رویکرد تحلیل	ذکر بیش از ۴۰ دشواری نگارشی مؤثر بر بازیابی در رابطه با جست‌وجو و بازیابی اطلاعات فارسی در آثار مورد بررسی و ضرورت توجه به هنجارسازی چنددستی‌های نگارشی و دستوری در طراحی الگوریتم‌های سامانه‌های جست‌وجو و بازیابی فارسی، تدوین استاندارد نگارش فارسی و تجهیز پایگاه‌های‌اطلاعاتی به اصطلاحنامه و فرهنگ‌های املایی و ...

همان‌طور که در جدول ۱، مشاهده می‌شود، مرور پیشینه‌های مورد مطالعه که روش غالب آن‌ها پیمایش است، در مجموع، تبیین و شناسایی بسیاری از ویژگی‌ها و مشکلات نگارشی زبان فارسی حاکی از آن است که رسم‌الخط فارسی یکی از متغیرهای عمده در ذخیره و بازیابی اطلاعات خصوصاً از موتورهای کاوش و پایگاه‌های اطلاعاتی زبان فارسی است.

در خارج از ایران و پیرامون سایر زبان‌ها نیز مطالعات مشابه گوناگونی انجام شده است که برخی از آن‌ها در جدول ۲، ارائه شده است.

جدول۲. پیشینه‌های خارجی مربوط به موضوع پژوهش

نویسنده/سال	عنوان	هدف	روش	نتیجه
Lazarinis (2008)	بهبود مبتنی بر مفهوم بازیابی تصاویر در وب از طریق ترکیب کلیدواژه‌های مشابهِ معنایی در زبان یونانی	بررسی عملکرد موتورهای کاوش (گوگل، یاهو و اماس‌ان) در ارتباط با سؤالات تک‌زبانهٔ غیرانگلیسی (یونانی) و ارائهٔ ابزار فراکاوش برای اصلاح نقایص و در جهت ایجاد موتورهای جست‌وجوی کارآمد بومی	پیمایشی	تأثیر قابل توجه ریخت‌شناسی کلمات (استفاده از حروف کوچک یا بزرگ، استفاده یا عدم استفاده از علائم تلفظ زبانی، و حتی نقش واژه‌ها) بر بازیابی نتایج (تعداد و ربط تصاویر) و تکیهٔ موتورهای جست‌وجو بر شکل کلیدواژه‌ها به جای تمرکز بر نیاز واقعی کاربران و عدم توجه موتورهای کاوش مورد پژوهش به ویژگی‌های زبان‌شناختی دیگر زبان‌ها
Lazarinis (2007)	قابلیت‌های جست‌وجوی وب‌سایت‌های الکترونیکی تجاری در مورد زبان‌های غیرانگلیسی (مطالعهٔ موردی یونانی)	بررسی قابلیت‌های جست‌وجوی وب‌سایت‌های الکترونیکی تجاری در مورد زبان یونانی	پیمایشی (مطالعهٔ موردی)	عدم توجه موتورهای جست‌وجوی محلی به ریخت‌شناسی سؤالات که امری مهم و قابل توجه در مورد زبان‌های غیرانگلیسی و غیرلاتین از جمله زبان یونانی است و در نتیجه، شکست جست‌وجوی کاربر
Zhang& Lin (2007)	پشتیبانی چندزبانه به‌وسیلهٔ موتورهای جست‌وجو	بررسی ویژگی‌های پشتیبانی چندزبانه به‌وسیلهٔ موتورهای جست‌وجوی شبکهٔ اینترنت	پیمایش مقایسه‌ای	پشتیبانی چندزبانه بهتر موتورهای جست‌وجوی گوگل، EZ۲Find و Onlinelink در بین بسیاری از موتورهای جست‌وجوی مجهز به ویژگی پشتیبانی چندزبانه
Lewandowski (2008)	مشکلات استفاده از موتورهای جست‌وجوی وب برای یافتن نتایج به زبان‌های خارجی	بررسی توانایی موتورهای جست‌وجوی پُراستفاده و اصلی از جمله گوگل، یاهو، اماس‌ان و اسک در تشخیص و تمایز میان مدارک به‌زبان آلمانی از پیشینه‌هایی با زبان انگیسی	پیمایشی	روبه‌روشدن کاربر با مشکلاتی در موتورهای کاوش گوگل و اماس‌ان وقتی نتایج به زبان خاصی محدود می‌شوند، در حالی که هیچ‌یک از موتورهای کاوش در بازیابی نتایج به زبان صفحهٔ رابط کاربر استفاده‌شده (زبان انتخابی) با مشکل مواجه نمی‌شوند، عدم تأثیر راهبُرد محدودیت زبانی در بهبود جست‌وجو در همهٔ موارد و اثرگذاری بهتر استفاده از صفحهٔ میانجی به زبان بومی در جست‌وجو و بازیابی اطلاعات در برخی از مواقع

نویسنده/سال	عنوان	هدف	روش	نتیجه
Hammo (2009)	افزایش کارایی موتورهای جست‌وجو برای پیشینه‌های نشانه‌گذاری‌شده به زبان عربی	ارائهٔ یک قالب کاری برای افزایش کارایی موتورهای کاوش برای متون عربی دارا و فاقد اِعراب‌گذاری از طریق روش‌های گسترش سؤال (کلیدواژه)	پیمایشی	تأثیر گسترش سؤال بر بهبود جست‌وجو و بازیابی متون عربی و افزایش کارایی موتورهای کاوش با استفاده از ابزارهای پیشرفتهٔ پردازش زبان طبیعی

بر اساس مندرجات جدول ۲، مروری بر پیشینه‌های خارجی که بیشتر با روش پیمایشی انجام شده‌اند، نشان می‌دهد که این پژوهش‌ها اغلب به بررسی قابلیت‌ها و کاستی‌های موتورهای کاوش در پوشش زبان‌های غیرانگلیسی و با هدف شناخت و ارائهٔ راهکارهایی جهت اصلاح چالش‌های زبانی پرداخته‌اند. یافته‌های حاصل از آن‌ها نشان می‌دهد که ریخت‌شناسی واژه‌ها و عبارت‌های جست‌وجوشده بر بازیابی نتایج اثر دارد و ابزارهای جست‌وجو به جای تمرکز بر نیاز واقعی کاربران در جهت بهبود فرایند کاوش، بیشتر بر شکل کلیدواژه‌ها تکیه می‌کنند. حتی بعضی ابزارهای جست‌وجوی محلی نیز ریخت‌شناسی سؤالات را در نظر نمی‌گیرند. بنابراین جست‌وجوی کاربر با شکست مواجه می‌شود.

بررسی پیشینه‌های پژوهش که شناسایی چالش‌های زبانی در ارتباط با جست‌وجو و بازیابی اطلاعات از موتورهای کاوش و پایگاه‌های اطلاعاتی از اهداف عمدهٔ آن‌ها بوده و اغلب با روش پیمایشی انجام شده‌اند، در مجموع حاکی از نکاتی است که در ادامه ذکر خواهد شد. با وجود مقبولیت و استفادهٔ زیاد از فضای وب و موتورهای کاوش برای یافتن انواع اطلاعات توسط طیف گسترده‌ای از کاربران و فراهم‌شدن ابزارهای متنوع با امکانات متفاوت برای جست‌وجو در محیط وب، ویژگی‌ها و مشکلات زبانی همچنان به‌عنوان عاملی مهم در بحث ذخیره و بازیابی اطلاعات مطرح هستند و مطالعات بسیاری به آن‌ها پرداخته و راهکارهایی نیز ارائه شده است. اما در مورد زبان فارسی به‌دلیل ماهیت و خصوصیات منحصربه‌فرد آن، نحوهٔ تعامل فارسی‌زبانان با محیط وب و تولید روزافزون صفحات وب و وبلاگ‌های فارسی‌زبان که تا حدودی به آن‌ها اشاره شد، هنوز جای تحقیق و کار بسیار است. همچنین، با توجه به آنچه بیان شد، پیشینه‌های داخلی اغلب به ویژگی‌های ریخت‌شناسی زبان فارسی تأکید داشته و ویژگی‌های معنایی آن را کمتر مورد بررسی قرار داده‌اند. علاوه بر این، تاکنون مطالعهٔ جامعی که پایگاه‌های بررسی‌شده

در پژوهش حاضر را با استفادهٔ طیف گسترده‌ای از کاربران از آن‌ها به‌طور همزمان مطالعه کند، یافت نشد. در این میان تنها تعدادی از پژوهش‌ها به توصیف این ویژگی‌ها پرداخته و مشکلات احتمالی رسم‌الخط فارسی را در محیط‌های رایانه‌ای مورد بحث قرار داده‌اند. این در حالی است که مطالعهٔ حاضر ضمن گردآوری مجموعه‌ای جامع از ویژگی‌هایی مانند ریخت‌شناسی و معنایی زبان فارسی قصد دارد با جست‌وجو در محیط پایگاه‌های اطلاعاتی فارسی‌زبان و بر مبنای نتایج بازیابی‌شده، مسائل زبانی عمده و اثرگذار در این حوزه را که کاربر با آن‌ها مواجه است، به‌صورت عینی شناسایی کند. به‌علاوه، انجام مطالعاتی از این دست می‌تواند از طریق تشخیص نقاط ضعف و قوت پایگاه‌ها (به لحاظ پوشش ویژگی‌های زبانی) زمینه‌ساز بهبود طراحی و اصلاح پایگاه‌های اطلاعاتی و حتی موتورهای جست‌وجوی بومی و ملی باشد.

پرسش پژوهش

میزان انطباق و توجه به ویژگی‌های نگارشی زبان فارسی در پایگاه‌های اطلاعاتی فارسی چگونه است؟

۳. روش پژوهش

این پژوهش به‌روش پیمایشی-تحلیلی و با استفاده از مشاهدهٔ مستقیم انجام گرفت. بدین منظور، پس از بررسی منابع مرتبط و بر مبنای پیشینه‌های فارسی پژوهش، آن دست از ویژگی‌های زبان فارسی که باعث ایجاد مسائلی در کاوش و بازیابی اطلاعات هستند دسته‌بندی و نسبت به تهیهٔ جداول ۳ و ۴ اقدام شد. جدول ۳، حاوی دسته‌بندی مشکلات نوشتاری است که بر اساس پیشینهٔ پژوهش باید در ذخیره‌سازی، پردازش و بازیابی منابع مورد توجه قرار گیرند. جدول ۴، نیز طبقه‌بندی ویژگی‌های معنایی زبان فارسی را نشان می‌دهد که بر بازیابی اطلاعات با این زبان اثرگذار هستند. هر یک از ویژگی‌های ذکرشده در جداول نیز شامل مصادیقی (کلیدواژه‌هایی) است که هر یک بازنمون چالش‌های نوشتاری یا معنایی زبان فارسی در جست‌وجو و بازیابی اطلاعات در محیط‌های رایانه‌ای است. از این کلیدواژه‌ها به منزلهٔ سیاهه‌ای برای بررسی و شناسایی مشکلات عمدهٔ ناشی از ویژگی‌های نوشتاری و معنایی زبان فارسی در تعامل کاربران فارسی‌زبان با پایگاه‌های اطلاعاتی و در ارتباط با کاوش و بازیابی اطلاعات و نیز چگونگی

ایجاد این مسائل استفاده شده است؛ بدین‌گونه که برای هر یک از ویژگی‌های نوشتاری و معنایی زبان فارسی واژه‌ای انتخاب شد تا به‌عنوان کلیدواژهٔ کاوش مبنا قرار گیرد و توسط پژوهشگران در پایگاه‌های مورد مطالعه جست‌وجو و نتایج بازیابی‌شدهٔ حاصل از آن بررسی و ثبت شود. روایی سیاههٔ مذکور که شامل کلیدواژه‌های موجود در جداول ۵ و ۶ است نیز با مشورت شش نفر از اساتید علم‌اطلاعات و دانش‌شناسی و زبان و ادبیات فارسی مورد تأیید قرار گرفت. به‌عنوان نمونه، برای ویژگی «همنامی» از واژهٔ «توپ» جهت آزمون مشکلات پیش‌آمده در کاوش و بازیابی اطلاعات در پایگاه اطلاعاتی مورد نظر، در رابطه با این ویژگی استفاده شد. به‌منظور گردآوری داده‌ها، هر یک از کلیدواژه‌های موجود در سیاهه توسط پژوهشگران به تفکیک وارد بخش جست‌وجوی پایگاه‌های اطلاعاتیِ مورد مطالعه شد و نتایج حاصل ثبت گردید[1]. این پایگاه‌ها عبارت‌اند از: «ایرانداک» (که کاوش در آن بالاخص در زمینهٔ جمع‌آوری پیشینه در موضوعات مختلف پژوهشی بسیار مورد توجه محققان و به‌ویژه دانشجویان و اساتید است)، «آی‌اس‌سی» (پایگاه دیگری که مورد توجه پژوهشگران و دانشجویان است)، «نورمگز» (که این پایگاه نیز به‌ویژه در حوزهٔ علوم انسانی و اسلامی مراجعان زیادی دارد) و «سید» (این پایگاه نیز به‌دلیل محتوایی که دارد، مورد مراجعهٔ بسیاری از پژوهشگران و دانشجویان است). در بخش مسائل نوشتاری صورت‌های مختلف متصور برای هر واژه درج و تعداد کل نتایج بازیابی‌شده ثبت شد. در بخش مسائل معنایی نیز به‌همین ترتیب عمل شد با این تفاوت که در مورد کلمات فاقد صورت‌های مختلف نوشتاری و دارای معانی گوناگون، مانند واژه‌های «توپ» و« شور» در میان ۲۰ نتیجهٔ نخست بازیابی‌شده، تعداد نتایجی که حاوی معانی مختلفِ کلیدواژهٔ جست‌وجوشده بودند، شمارش و ثبت گردید.

٤. یافته‌های پژوهش

برای پاسخگویی به پرسش پژوهش، همان‌طور که ذکر شد، ابتدا با مرور متون و با توجه به پژوهش‌های انجام‌شده و بر اساس مطالعهٔ «هماوندی» (۱۳۹۲) ویژگی‌های نوشتاری و معنایی عمدهٔ زبان فارسی که در روند بازیابی اطلاعات توسط کاربران، به‌ویژه در پایگاه‌های اطلاعاتی چالش ایجاد می‌کنند، استخراج و پس از تغییراتی متناسب با

---

۱. ۱۶ بهمن ماه ۱۳۹۴ (لازم است تاریخ جست‌وجو و بازیابی نتایج به‌دلیل تغییرات مداوم محتوای پایگاه‌های اطلاعاتی مد نظر قرار گیرد).

محیـط پایگاه‌هـا به‌طـور خلاصـه در جـدول شـماره ۳ و ۴ ارائـه شـد.

جدول ۳. ویژگی‌های نوشتاری زبان فارسی مؤثر در بازیابی اطلاعات در محیط‌های اطلاعاتی

مسئلهٔ نوشتاری	مثال
نحوهٔ ضبط واژگان لاتین	انفولانزا، آنفلوآنزا، آنفولانزا/ پتاسیم، پتاسیوم/ امریکا، آمریکا
واژگان دخیل و انواع معادل آن‌ها	ایمیل، رایانامه/ سیستم، نظام، سامانه
پیوسته‌نویسی و جدانویسی انواع واژگان	مشتق: پرستش گاه، پرستشگاه
	مرکب: جوانمرد، جوان‌مرد/ بزرگسال، بزرگ سال
	مشتق‌مرکب: گلابگیری، گلاب‌گیری/ فن‌آوری، فناوری
	علائم جمع: دانشگاهها، دانشگاه‌ها
انواع جمع‌ها (جمع‌های فارسی و مکسر)	حوادث، حادثه‌ها/ اساتید، استادان، استادها
طریقهٔ نگارش الف مقصوره	موسی، موسا/ مصلی، مصلا
استفاده یا عدم استفاده از اِعراب‌گذاری و سایر علائم (واژه‌هایی با شکل نوشتاری یکسان و تلفظ متفاوت)	مسکن (مُسَکّن، مَسکَن)/ ملک (مِلک، مُلک، مَلَک، مَلِک)
نحوهٔ نگارش همزهٔ میانی و پایانی کلمات با کرسی واو، دندانه، الف و بدون کرسی	جبرئیل، جبرییل/ مؤذن، موذن/ املا، املاء/ مسأله، مساله، مسئله
استفاده و عدم استفاده از علامت تشدید	محمد، محمّد/ معلم، معلّم
کسرهٔ اضافه و بدل‌های آن	اهدا کتاب، اهداء کتاب، اهدای کتاب
واژه‌های دو املائی (واژه‌هایی با واج یا آوای مشترک و شکل نوشتاری متفاوت)	آزوقه، آذوقه/ تهران، طهران
جابه‌جایی ی و همزه در کلمات فارسی	پایین، پائین/ آیین، آئین
نحوهٔ نگارش ه غیر ملفوظ و ی میانجی	زلزله بم، زلزله‌ی بم، زلزلهٔ بم
استفاده از زبان محاوره (شکل عامیانه)	خونه، خانه/ زمونه، زمانه
کاربرد و حذف مد در کلمات فارسی	فناوری، فنآوری/ دستاورد، دستآورد

بـا دقـت در مـوارد ذکرشـده در جـدول ۳، می‌تـوان دریافـت کـه بسـیاری از ویژگی‌هـای ریخت‌شناسـی و نوشـتاری زبـان فارسـی ریشـه در دو یـا چندگانه‌نویسـیِ یـک واژه دارنـد کـه همیـن امـر به‌دلیـل ایجـاد ناهماهنگـی سـبب بـروز مسـائلی در بازیابـی اطلاعـات می‌شـود. به‌عنـوان مثـال، جسـت‌وجوی صورت‌هـای مختلـف نوشـتاری واژهٔ «فنـاوری» سـبب بازیابـی نتایجی متفـاوت می‌شـود. بعضـی از ویژگی‌هـای معنایـی زبـان فارسـی نیـز در ایـن زمـره هسـتند.

مـرور پیشـینه‌ها نیـز حاکـی از ایـن اسـت کـه مـواردی ماننـد همنامـی یـا واژه‌هـای یکسـان بـا معانـی متفـاوت و واژگان هم‌نویسـه بـا معانـی متفـاوت، چندمعنایـی[1]، هم‌معنایـی و تـرادف از ایـن قبیـل هسـتند. همان‌گونـه کـه «حسـینی بهشـتی» (۱۳۸۲) نیـز عقیـده دارد، در ایـن بیـن چندمعنایـی و تـرادف دو مشـکل عمـده در ارتباطـات معنایـی واژگان هسـتند.

جدول ٤. ویژگی‌های معنایی زبان فارسی مؤثر در بازیابی اطلاعات در محیط‌های اطلاعاتی

مسئلهٔ معنایی	مثال
همنامی یا واژه‌های یکسان با معانی متفاوت (واژگان مشترکِ لفظی/ هم‌آوا و هم‌نویسه)	شیر (نام یک حیوان، مادهٔ لبنی، شیر آب)/ توپ (توپِ بازی، توپِ جنگی، واحد شمارش پارچه، معنای عوامانه)
واژگان هم‌نویسه با معانی متفاوت	شور (شور بودن طعم، شور و شوق)
چندمعنایی	روان (روح و روان، جاری)/ قلب (عضو بدن، ضمیر و خاطر، وارونه‌بودن، مرکز)
هم‌معنایی و ترادف	دریای خزر، دریای کاسپین، دریای مازندران

ویژگی‌هـای معنایـی ذکرشـده در جـدول شـماره ۴، کـه در پژوهش‌هـای ایـن حـوزه کمتـر مـورد توجـه قـرار گرفته‌انـد، اغلـب بـا ایجـاد تعـدد معنایـی سـهم عمـده‌ای از مشـکلات یادشـده در حـوزهٔ جسـت‌وجو و بازیابـی اطلاعـات را بـه خـود اختصـاص می‌دهنـد. به‌عنـوان مثـال، کاربـر بـا جسـت‌وجوی کلیـدواژهٔ «تـوپ» کـه دارای ویژگـی همنامـی اسـت، ممکـن اسـت بـا نتایجـی حـاوی چندیـن معنـا مواجـه شـده و ایـن کاوش بـه بازیابـی نتایـج نامربـوط منجـر شـود و یـا این‌کـه بـه نتایجـی حـاوی چنـد مـورد از معانـی دسـت یابـد، امـا در بیـن آن‌هـا نتایجـی بـا معنـای مـورد نظـر وی بازیابـی نشـود.

پرسـش پژوهـش: میـزان انطبـاق و توجـه بـه ویژگی‌هـای نگارشـی زبـان فارسـی در پایگاه‌هـای اطلاعاتـی فارسـی چگونـه اسـت؟

به‌منظـور پاسـخ بـه پرسـش پژوهـش، شـکل‌های گوناگـون ایـن کلیدواژه‌هـا توسـط پژوهشـگران در قسـمت جسـت‌وجوی پایگاه‌هـای «ایرانـداک»، «آی‌اس‌سـی»، «نورمگـز»، و «سـید» وارد و سـپس، تعـداد نتایـج بازیابی‌شـده بـرای هـر کلیـدواژه ثبـت و در جـدول شـماره ۵

۱. چندمعنایـی مختـص زبـان فارسـی نیسـت، بلکـه از ویژگی‌هـای سـایر زبان‌هـا نیـز هسـت کـه در ایـن جـا اختصاصاً در مـورد زبـان فارسـی بررسـی شـده اسـت.

ارائـه شـده اسـت.

جدول ۵. نتایج بازیابی‌شده برای کلیدواژه‌های منتخب با صورت‌های گوناگون نوشتاری در پایگاه‌های مورد مطالعه

کلیدواژه‌های مورد جست‌وجو	مجموع نتایج بازیابی‌شده در پایگاه‌های اطلاعاتی			
	ایرانداک	آی‌اس‌سی	نورمگز	سید
پتاسیوم	۷	۱	۲۱	۱
پتاسیم	۵۷۳۶	۱۶۶۳	۶۵۹	۴۲۱
سامانه	۴۵۲۸	۳۶۷۳	۳۹۳۱	۵۵۳
سیستم	۷۳۳۱۶	۲۱۶۳۲	۵۶۳۸۶	۳۱۳۲
نظام	۲۲۲۴۱	۱۴۳۱۶	۱۴۱۷۲۶	۱۷۲۴
پرستشگاه	۱۲	۶	۱۰۷۸	۰
پرستش گاه	۲۷	۱۰	۸۰۰۳۰	۰
بزرگسال	۸۱۰	۳۳۰	۱۱۹۸۱	۹۳
بزرگ سال	۳۷۸۱	۱۱۷۵	۳۰۱۳۵۵	۱۹
فناوری	۱۸۹۳۰	۱۰۷۴۹	۱۸۹۶۰	۱۰۲۶
فن آوری	۳۷۷۲	۱۹۱۸	۱۱۰۱۷۱	۲۰۳
دانشگاهها	۱۴۸۶	۵۰۹	۱۴۳۲۷۵	۲۴
دانشگاه ها	۵۸۸۸۰	۹۱۹۹	۱۸۲۱۹۹	۲۹۵
اساتید	۳۷۳۶	۱۳۶۵	۱۰۳۰۷۷	۲۲۱
استادان	۸۶۳	۹۰۸	۱۰۳۰۷۷	۹۰
استاد ها	۳	۱	۱۰۳۰۷۷	۰
موسا	۵	۳	۱۴۸	۰
موسی	۳۷۰۲	۶۰۹	۴۳۹۳۰	۵۴
مسکن	۵۵۹۵	۲۱۳۷	۲۵۵۱۵	۳۲۴
مُسَکِّن	۱۷۲	۰	۲۵۵۱۵	۳۲۴
مَسکن	۴	۰	۲۵۵۱۵	۳۲۴
مسأله	۵۸۵۶	۴۶۷۷	۸۲۳۶۸	۸۵۶
مساله	۱۱۱۳۴	۴۶۷۷	۸۲۳۶۸	۸۵۶

کلیدواژه‌های مورد جست‌وجو	مجموع نتایج بازیابی‌شده در پایگاه‌های اطلاعاتی			
	ایرانداک	آی‌اس‌سی	نورمگز	سید
مسئله	۲۱۳۵۴	۱۳۷۹	۱۷۷۶۲۲	۱۹
معلم	۷۵۸۱	۱۴۹۰	۶۱۲۷۲	۱۳۷
معلّم	۷۲	۲۲	۶۱۲۷۲	۱۳۷
اهدا کتاب	۷	۵	۲۶۹۸۲۲	۰
اهداء کتاب	۵	۱	۲۶۸۲۰۴	۰
اهدای کتاب	۶	۵	۲۶۹۸۲۲	۰
تهران	۸۱۸۳۴	۱۰	۱۴۲۸۳۷	۶۵۱۵
طهران	۶۷	۷۹	۱۱۲۳۳	۱۰
زلزله بم	۳۰۳	۲۲۷	۲۲۸۹۵	۸۸
زلزله‌ی بم	۳۲	۱	۲۲۸۹۷	۸۹
زلزلهٔ بم	۰	۰	۲۲۸۹۵	۸۸
خانه	۶۵۷۱	۴۰۷۲	۹۲۹۷۹	۴۴۲
خونه	۱۰	۸	۲۰۹۲	۰
دستاورد	۸۹۷	۴۸۴	۲۸۸۶۲	۸
دستآورد	۱۴	۴۸۴	۲۸۸۶۲	۸
آیین	۴۵۰۲	۲۲۲۵	۴۸۷۰۸	۱۷۵
آئین	۱۷۵۵	۱۱۴	۴۸۷۰۸	۱۷

همان‌طور که در نتایج جدول ۵ دیده می‌شود، تعداد کل نتایجی که پایگاه «ایرانداک» برای کاوش صورت‌های مختلف نوشتاری واژه‌های برگزیده بازیابی می‌کند، در مورد همهٔ کلیدواژه‌ها (ویژگی‌های نوشتاری) متفاوت است. همچنین، یافته‌های جدول ۵، در مورد پایگاه «آی‌اس‌سی» نیز حکایت از فقدان توجه کافی به ویژگی‌های نوشتاری فارسی دارد که اختلاف نتایج بازیابی‌شده گواه آن است. به‌عنوان نمونه، موارد بارزی از این اختلاف را می‌توان در کلیدواژه‌هایی مانند «پتاسیم» و «پتاسیوم» مشاهده نمود.

از سوی دیگر، نتایج حاصل از جست‌وجو در پایگاه «نورمگز» نیز نشان می‌دهد که تعداد نتایج بازیابی‌شده در مورد اغلب کلیدواژه‌ها (ویژگی‌های نوشتاری) متفاوت است؛ مانند آنچه که در جست‌وجوی کلیدواژهٔ «فناوری» و «فن‌آوری» اتفاق می‌افتد که

تعداد نتایج بازیابی‌شده برای این دو اختلاف زیادی با هم دارند. موارد استثنا نیز شامل کلیدواژه‌های «اساتید»، «استادها»، «استادان»، «زلزله بم»، «زلزله‌ی بم»، «زلزلهٔ بم» و «آیین»، «آئین» هستند که صورت‌های نوشتاری گوناگون در نتایج بازیابی‌شده دیده می‌شود.

در رابطه با پایگاه «سید» نیز اختلاف نتایج بازیابی‌شده برای صُوَر گوناگون نگارشی واژه‌ها در بسیاری از موارد باعث ایجاد اختلال و ارائهٔ نتایج ناقص به کاربر می‌شود. به‌عنوان نمونه، شاهد این اختلاف در مورد کلیدواژه‌های «پتاسیوم»، «پتاسیم» و یا «دانشگاهها»، «دانشگاه‌ها» هستیم. همچنین «سید» در مورد کلیدواژه‌هایی مانند «مسکن»، «مُسَکِّن» و مَسکن» و «دستاورد» و«دستآورد» قادر به شناسایی اعراب‌گذاری و علائم به‌کاررفته نبوده و نتایجی با محتوای محل سکونت بازیابی می‌نماید. در دیگر پایگاه‌های بررسی‌شده نیز کاربر یا با اتفاقی مشابه آنچه گفته شد، روبه‌رو می‌شود و یا در صورت اعراب‌گذاری با نتایجی بدون ارتباط مواجه خواهد شد.

در مورد ویژگی‌های معنایی زبان فارسی نیز مشکلاتی وجود دارد که برخی مصادیق آن در ارتباط با پایگاه‌های اطلاعاتی مورد پژوهش در جدول شماره ۶، مشاهده می‌شود.

جدول ٦. نتایج بازیابی‌شده برای کلیدواژه های منتخب با معانی گوناگون[1] در پایگاه‌های مورد مطالعه

کلیدواژهٔ مورد جست‌وجو		مجموع نتایج بازیابی در پایگاه‌های اطلاعاتی			
		ایرانداک	آی‌اس‌سی	نورمگز	سید
توپ	توپِ بازی	۱۰	۱۴	۷	۶
	توپِ جنگی	۵	۳	۱۲	۴
	واحد شمارش پارچه	۰	۰	۰	۰
	قسمتی از یک آنتی‌ژن (علم ژنتیک)	۵	۳	۰	۶
	معنای عوامانه	۰	۰	۱	۰
شور	شوق	۱	۱	۱۴	۲
	شور بودن	۱۸	۱۷	۳	۱۶
	مشورت	۰	۰	۳	۲
	غیرمرتبط	۱	۲	۰	۰

۱. در مورد ویژگی‌های معنایی به‌دلیل نبودن حالت‌های مختلف نوشتاری کلمات، (به جز کلیدواژه‌های مربوط به دریای مازندران) معانی مختلف بازیابی‌شده برای هر کلیدواژه، بین ۲۰ نتیجهٔ نخست بازیابی‌شده شمارش و ثبت شدند.

کلیدواژهٔ مورد جست‌وجو	مجموع نتایج بازیابی در پایگاه‌های اطلاعاتی			
	ایرانداک	آی‌اس‌سی	نورمگز	سید
روان روح	۱۶	۱۸	۲۰	۲۰
جاری	۴	۲	۰	۰
دریای خزر	۲۳۶۰	۱۶۵۴	۵۸۰۳۲	۶۵۱
دریای کاسپین	۲۷	۲۵	۵۶۸۹۷	۳
دریای مازندران	۴۳۹	۱۷۰	۶۳۸۸۲	۵۴

همان‌طور که در جدول ۶ مشهود است، با توجه به نتایج حاصل از کاوش کلیدواژه‌ها در پایگاه «ایرانداک» شاهد عدم توجه پایگاه‌های اطلاعاتی به واژگانی هستیم که دارای معانی مختلف و صورت نوشتاری یکسانی هستند که موجب می‌شود گاه معانی مختلفی از یک واژه بازیابی شود، اما معنای مورد نظر کاربر در میان آن‌ها نباشد. همچنین، نتایج نشان می‌دهد که پایگاه «آی‌اس‌سی» نیز توجه کافی به ویژگی‌های یادشده جهت تأمین نیازهای کاربران فارسی‌زبان در رابطه با توجه به ویژگی‌های معناشناسی زبان فارسی ندارد. به‌عنوان مثال، در مورد کلیدواژهٔ «دریای مازندران» اختلاف قابل توجهی میان تعداد نتایج و مدارک بازیابی‌شده برای کلیدواژه‌های مترادف آن وجود دارد که جامعیت کاوش را تحت تأثیر قرار می‌دهد. همچنین، مانند نتایج دیگر پایگاه‌ها در بخش کاوش معنایی، در مورد واژه‌هایی مانند «توپ» و «شور» کاربر با بازیابی نتایج نامرتبط مواجه شده و یا ممکن است هرگز بدون استفاده از کلیدواژهٔ کامل‌کننده به نتایج مورد نظر دست پیدا نکند.

علاوه بر موارد یادشده، با دقت در نتایج حاصل درمی‌یابیم که در پایگاه اطلاعاتی «نورمگز» نیز مشکلاتی وجود دارد، چرا که گاهی کاربران در جست‌وجوی معانی مختلف یک واژه دچار مشکل شده و از بازیابی نتایج مطلوب بازمی‌مانند و یا با نتایج نامربوط مواجه می‌شوند. (مثال: بازیابی نتایجی با معنای «مشورت و هم‌اندیشی» برای جست‌وجوی کلیدواژهٔ «شور» که در دیگر پایگاه‌ها نتایجی با این مضمون بازیابی نشد و بسته به نیاز کاربر می‌تواند مطلوب یا نامطلوب باشد). البته، ذکر این نکته ضروری است که در قسمت واژگان مترادف از بخش معنایی، این پایگاه ضمن بازیابی نتایجی نزدیک به هم برای کلیدواژه‌های «دریای خزر، دریای مازندران و دریای کاسپین»، پس از بازیابی نتایج

کاوش برای هر یک از کلیدواژه‌های «دریای خزر و مازندران» دیگری را نیز برای کاوش به کاربر پیشنهاد می‌دهد که خود گامی مؤثر در بهبود بازیابی کلیدواژه‌های مترادف است.

همچنین، بر اساس نتایج بازیابی‌شده، پایگاه «سید» نیز توجه کافی به ویژگی‌های معنایی زبان فارسی نداشته است، به‌طوری که در مورد ویژگی ترادف در واژه‌هایی که به «دریای مازندران» اشاره دارند، اختلاف زیادی در تعداد نتایج بازیابی‌شده وجود دارد.

۵. بحث و نتیجه‌گیری

هر یک از ویژگی‌های نوشتاری و معنایی زبان فارسی که در این پژوهش ذکر شد، هنگام کاوش و بازیابی اطلاعات از منابع مختلف، مانند موتورهای کاوش وب و پایگاه‌های اطلاعاتی، چالش‌هایی را به‌نحوی فراروی کاربران فارسی‌زبان قرار می‌دهد. علاوه بر نتایج و مصادیق ذکرشده که تأثیر قابل توجه خصوصیات یادشده را نشان می‌دهند، شرحی از این مشکلات به همراه مصادیق آن‌ها بر پایهٔ یافته‌های پژوهش‌های انجام‌شده نیز می‌تواند تأییدی بر نتایج به‌دست‌آمده در این پژوهش باشد. برای این منظور و برای به‌دست آوردن تصویری بهتر از مسائل و چالش‌های زبان فارسی در زمینهٔ جست‌وجو از پایگاه‌های اطلاعاتی، ابتدا شرحی از ویژگی‌های نوشتاری و سپس، ویژگی‌های معنایی بیان می‌شود.

ویژگی‌های نوشتاری

در بخش مسائل نوشتاری، نتایج پژوهش‌های «آخشیک و فتاحی» (۱۳۹۱) و «گل‌تاجی و بذرگر» (۱۳۸۹) نشان داد که چالش‌های ریختیِ زبان فارسی، از جمله پیوسته‌نویسی‌و جدانویسی واژگانی مانند واژگان مشتق و مشتق-مرکب و برخی ویژگی‌های مندرج در جدول ۳، تأثیر زیادی بر بازیابی اطلاعات از پایگاه‌های اطلاعاتی داخلی از جمله از پایگاه‌های «آی‌اس‌سی»، «ایرانداک» و «سید» دارد که نتایج جدول شماره ۵ نیز مؤید آن است. بدین‌ترتیب، می‌توان به‌وضوح دریافت که حتی پایگاه‌های داخلیِ فارسی‌زبان نیز توجه کافی به ملزومات و خصیصه‌های زبان فارسی ندارند. کاربران این قبیل پایگاه‌ها که در بسیاری موارد در پی جست‌وجوی اطلاعات جهت انجام کارهای پژوهشی و مطالعاتی هستند، به‌واسطهٔ این مشکل از دستیابی جامع به اطلاعات مورد نیازشان بازمی‌مانند؛ مانند

آنچه در جست‌وجوی واژهٔ «فناوری» و «فن‌آوری» به‌دست آمد. این، در حالی است که کاربران در سطوح مختلف و با عادت‌های نوشتاری گوناگون ممکن است هر یک از این صورت‌ها را برای جست‌وجو استفاده کنند و در بسیاری موارد ناخودآگاه از دستیابی به نتایجی که حاصل کاوش شکل دیگر واژه هستند، محروم بمانند و به همان نتایج اولیه بسنده کنند. البته، موارد استثنایی هم وجود دارد. چنان‌که به نظر می‌رسد پایگاه «نورمگز» با توجه به بازیابی تعداد مساوی و یا نزدیک نتایج برای صورت‌های مختلف برخی کلمات و همچنین، دیده‌شدن صورت‌های مختلف هر یک از واژه‌ها در پیِ کاوش تنها یک صورت از آن در بین نتایج بازیابی‌شده، در بعضی موارد فارغ از صور نوشتاری گوناگون و بر مبنای مفهوم واژه‌ها عمل نموده و همهٔ منابع حاوی آن‌ها را در یک دستهٔ معنایی قرار داده است.

در مجموع، شناسایی آن دسته از ویژگی‌های زبانی که نقش پررنگ‌تر و عمده‌ای در بروز چالش‌های پیشِ روی کاربران دارند نیز نکته‌ای است که نباید از آن غافل شد. چنان‌که «هماوندی» (۱۳۹۲) در پژوهش خود مواردی مانند پیوسته‌نویسی و جدانویسیِ واژگانِ مشتق، مشتق- مرکب، انواع جمع‌های فارسی و مکسرِ عربی، نگارش همزهٔ بدونِ کرسی و استفاده از زبان محاوره را از مشکلات عمدهٔ زبان فارسی در جست‌وجوو بازیابی تصاویر خاصه در موتورهای کاوش می‌داند. همچنین، یافته‌های پژوهش «عبدالهی نورعلی و جوکار» (۱۳۸۸) نشان می‌دهد که بین شکل نوشتاری واژه و ابزار جست‌وجو (موتورهای کاوش) رابطه معناداری وجود دارد. بنابراین، می‌توان نتیجه گرفت که به‌کار بردن یک شکل خاص از کلیدواژه و نیز استفاده از یک ابزار جست‌وجوی خاص در بازیابی اطلاعات اثرگذار است. به‌عنوان نمونه، چنانچه کاربری کلیدواژه «باغها» را با شکل پیوسته انتخاب کند، بیشتر اطلاعات موجود را که با کلیدواژهٔ «باغ‌ها» نمایه‌سازی و ذخیره شده‌اند، از دست می‌دهد و از طرف دیگر، انتخاب کلمه «باغ‌ها» نیز منجر به بازیافت نتایج نامربوط می‌شود. همچنین است تفاوت بین تعداد نتایج بازیابی‌شدهٔ حاصل از کاوش کلیدواژه‌های «پرستش گاه و پرستشگاه»، «فناوری و فن‌آوری» و «بزرگسال و بزرگ سال» که در جدول شماره ۴، مشاهده شد. در این بین، فقدان توجه و آگاهی کاربر از این مسئله می‌تواند باعث عدم دسترسی وی به نتایج حاوی صورت دوم واژه شود. همان‌طور که مشهود است، وجود مسئلهٔ پیوسته‌نویسی و جدانویسی در نتایج همهٔ پژوهش‌های یادشده، گواهی بر تأثیر زیاد این خصیصه بر کاوش و بازیابی اطلاعات به زبان فارسی است. این در حالی است

کـه پیوسته‌نویسـی و جدانویسـی واژگان در بسـیاری از مـوارد به‌صـورت سـلیقه‌ای انجـام شـده و خیلـی از نویسـندگان از قواعـد مربـوط بـه زبـان در مـورد آن پیـروی نمی‌کننـد. وجـود مسـائلی ماننـد اسـتفاده از فاصلـه و نیم‌فاصلـه و جـدا در نظـر گرفتـه شـدن اجـزای یـک واژهٔ مرکـب ماننـد «کتاب‌خانـه» و بازیابـی نتایـج مسـتقل بـرای هـر یـک از ایـن اجـزاء نیـز موجـب تشـدید ایـن مـوارد می‌شـود. در مجمـوع، در بخـش نوشـتاری، بـا دقـت در نتایـج حاصـل از پژوهـش می‌تـوان دریافـت کـه عـلاوه بـر مشـکلات مربـوط بـه پیوسته‌نویسـی و جدانویسـی واژگان، کلمـات عربـیِ موجـود در زبـانِ فارسـی ماننـد انـواع جمع‌هـا و انـواع شـیوه‌های نـگارش همـزه، واژگان دخیـل و معـادل آن‌هـا سـهم زیـادی از مشـکلات پیـشِ روی کاربـران فارسـی‌زبان را در تعامـل بـا محیط‌هـای اطلاعاتـی و خاصـه پایگاه‌هـای اطلاعاتـی فارسـی، بـه خـود اختصـاص داده‌انـد. به‌طـور کلـی، اختـلاف نتایـج بازیابی‌شـده بـرای حـالات مختلـف نوشـتاری کـه در اکثـر مـوارد در همـهٔ پایگاه‌هـا مشـابه اسـت، به‌نوعـی حاکـی از الگـوی نوشـتاری غالـب در بیـن نویسـندگان منابـع هـر پایـگاه اسـت و می‌تـوان از آن بـرای اسـتخراج الگـو به‌منظـور نمایه‌سـازی یـا پیشـنهاد کلیـدواژهٔ کاوش اسـتفاده نمـود؛ مثـل واژه‌هـای مرکـب «بزرگسـال و بزرگ‌سـال» کـه در هـر چهـار پایـگاه مـورد بررسـی، جسـت‌وجوی صـورتِ جـدای آن منجـر بـه بازیابـی نتایـج بیشـتری از فـرم پیوسـتهٔ آن شـد. همچنیـن، در بسـیاری مـوارد کـه صورت‌هـای نوشـتاری مختلـف واژه‌هـا دارای گونه‌هـای بـا اِعـراب و بـدون اِعـراب بوده‌انـد، بـه نظـر می‌رسـد کـه ترجیـح نویسـندگان بـه اسـتفاده از صورت‌هـای بـدون اِعـراب اسـت کـه می‌توانـد ریشـه در عـادات نوشـتاری فارسـی‌زبانان و عـدم تمایـل آن‌هـا بـرای اسـتفاده از اِعراب‌گـذاری و همچنیـن، پیچیده‌بـودن اِعراب‌گـذاری در صفحـه کلیدهـا اشـاره نمـود؛ ماننـد نتایـج بازیابی‌شـده بـرای واژه‌هـای «معلـم و مسـکن».

ویژگی‌های معنایی

بـا دقـت در پژوهش‌هـای یادشـده می‌تـوان دریافـت کـه ابعـاد معنایـی زبـان در آن‌هـا مهجـور مانـده، در حالـی کـه ویژگی‌هـای خـاص معنایـیِ زبـان فارسـی نقـش قابـل توجهـی در مسـئلهٔ کاوش و بازیابـی اطلاعـات داشـته و سـبب ایجـاد مشـکلاتی در تعامـل بیـن کاربـران و ابزارهـای اطلاع‌رسـانی هسـتند. در ایـن مـورد نتایـج پژوهـش «هماونـدی» (۱۳۹۲) بـا نتایـج حاضـر همسوسـت. در ایـن صـورت، جسـت‌وجوی کاربـر بـا مسـائلی ماننـد بازیافـت نتایـج نامربـوط مواجـه می‌شـود و در مـواردی حصـولِ نتیجـهٔ مطلـوب مسـتلزم تکـرار جسـت‌وجو بـا شـیوه‌های

مختلف و واژگان تکمیلی است؛ مانند آنچه که در مورد کاوش واژهٔ «توپ» و یا واژه‌های «دریای خزر، دریای مازندران و دریای کاسپین» (در مورد اخیر پایگاه «نورمگز» نسبت به دیگر پایگاه‌ها، عملکرد قابل قبولی داشت) اتفاق افتاد (جدول شماره ۶). در مجموع، در بخش معنایی، برخی ویژگی‌ها مانند ترادف و همنامی و هم‌نویسگی، علی‌رغم کم‌توجهی به آن‌ها در بسیاری از پژوهش‌های مربوطه، از جملهٔ مشکلات جدی کاربران در استفاده از پایگاه‌های اطلاعاتی هستند. چه‌بسا اگر در فرایند نمایه‌سازی و کاوش، انتخاب کلیدواژه بر اساس یک اصطلاح‌نامه و با ارجاع به فرم صحیح و منتخب واژه انجام می‌شد، تا حد زیادی به بهبود کاوش این دست واژه‌ها مؤثر بود.

آنچه ذکر شد، خلاصه‌ای از پیامدهای کم‌توجهی و یا نادیده انگاشتن خصوصیات زبان فارسی توسط کاربران و طراحان موتورهای کاوش و پایگاه‌های اطلاعاتی است که در نهایت، منجر به اختلال در کاوش و بازیابی انواع گوناگون اطلاعات می‌شود. در مجموع، می‌توان گفت که پایگاه‌های اطلاعاتی (حتی انواع بومی آن‌ها) نسبت به ویژگی‌های نوشتاری و معنایی زبان فارسی توجه کافی ندارند، در حالی‌که بخش بزرگی از این مسائل را می‌توان به مرحلهٔ نمایه‌سازی مربوط دانست که باید طی آن ویژگی‌های زبان فارسی را مد نظر قرار داد و با تمهیداتی از جمله یکدست‌سازی واژگان و استفاده از اصطلاح‌نامه‌ها نسبت به حل آن اقدام نمایند. این مسئله موجب می‌شود که احتیاج فارسی‌زبانان، به‌ویژه به ابزارهای کاوش بومی و پایگاه‌های اطلاعاتی که مبتنی بر ویژگی‌های زبانیِ خودشان طراحی شده و در تعامل با کاربران فارسی‌زبان به ظرائف و باید و نبایدهای زبانی آن‌ها بیشتر توجه شود، بیش از پیش احساس گردد. از سوی دیگر، پایگاه‌های اطلاعاتی نیز باید نسبت به برآورده‌ساختن نیازهای زبانی کاربران و اصلاح تعامل با کاربران بیشتر تلاش کنند، چرا که نادیده‌انگاشتن و یا کم‌توجهی به شاخصه‌ها و ویژگی‌های زبانی کاربران موجب بروز مسائلی در امر جست‌وجو و بازیابی اطلاعات می‌شود که در نهایت، از دست رفتن اطلاعات مفید و یا بازیابی اطلاعات ناخواسته را به همراه خواهد داشت.

با توجه به یافته‌های پژوهش، پیشنهادهای زیر می‌تواند در پیشگیری و رفع این چالش‌ها و بهبود تعامل میان کاربران و ابزارهای کاوش و بازیابی اطلاعات مؤثر باشد:

- بر اساس یافته‌های پژوهش و با توجه به تأثیر قابل توجه صورت‌های مختلف نوشتاری واژگان بر جست‌وجو و بازیابی نتایج در پایگاه‌های اطلاعاتی، لازم است از ابتدا شیوه‌های نگارشی در محیط دیجیتال و حداقل در مورد متون علمی و تخصصی

تا حد امکان یکپارچه و هماهنگ باشند. وجود یک نرم‌افزار واژه‌پردازِ منطبق با ویژگی‌های زبان فارسی[1]، می‌تواند یکی از راه‌های نیل به این مقصود باشد و در صورت بروز خطاهای نوشتاری و املایی، برای تصحیح به کاربر هشدار دهد. طرح سامانهٔ «استانداردساز و خطایاب متون فارسی» که توسط «پژوهشکدهٔ مدیریت دانش ایرانداک» به شورای پژوهش ارائه و تصویب شده و در دست اجراست (که حاصل آن نرم‌افزاری برای خطایابی متون فارسی و برای ویرایش پایگاه «گنج» این پژوهشگاه خواهد بود) نمونه‌ای از این تلاش‌هاست.

- در مواردی مانند واژگان دخیل و معادل آن‌ها، که در پایگاه‌های مورد بررسی توجه لازم به آن صورت نگرفته است، استفاده از واژگان معادل و تلاش برای رواج آن‌ها از طریق رسانه‌های مکتوب و غیرمکتوب، خصوصاً از طریق رسانه‌های رسمی از ابتدای ورود و استفاده از یک فناوری یا مفهوم (مثل واژه‌های یارانه و سوبسید[2] و یا پیامک و اس‌ام‌اس[3])، می‌تواند در پذیرش و کاربرد گسترده و هماهنگ آن توسط افراد مؤثر باشد.
- در مورد مسئلهٔ تنوع صورت‌های نوشتاری واژگان، پایگاه‌های اطلاعاتی می‌توانند از طریق فراهم‌ساختن نمایه‌های مناسب و استفاده از اصطلاح‌نامه‌ها و واژه‌نامه‌ها، کاربران را از وجود صورت‌های مختلف نوشتاری یک واژه آگاه کنند و در صورت لزوم به آن‌ها ارجاع دهند. نمونه‌ای از این‌گونه تلاش‌ها را می‌توان در «بخش تحقیق و توسعهٔ ایرانداک» مشاهده نمود. این پژوهشگاه حدود یک سال است که نرم‌افزاری طراحی نموده و به‌وسیلهٔ آن نمایه‌سازان پژوهشگاه با دسترسی به مجموعهٔ اصطلاح‌نامه‌های تدوین و ترجمه‌شدهٔ گروه اصطلاح‌شناسی این پژوهشگاه و بر اساس اصطلاح‌نامه‌ها و مجموعهٔ واژگان‌های موجود نمایه‌سازی و سازماندهی مدارک را انجام می‌دهند. این تلاش در نهایت، پشتوانهٔ مناسبی برای کاوش کاربران و حل مسائل و چالش‌های زبانی آن‌ها خواهد بود.
- با توجه به یافته‌های پژوهش که بخش عمده‌ای از مشکلات ناشی از ناهماهنگی در نوشتار صورت‌های مختلف واژگان است، در نظر گرفتن سازوکارهایی برای

---

۱. نرم‌افزار «ویراستیار» از جملهٔ این نوع تلاش‌هاست.

2. subsidy

3. Short Message Service (SMS)

اطلاع‌رسانی و آموزش به کاربران فارسی‌زبان، چه در مقام نویسنده در وب در جهت ایجاد یکدستی و هماهنگی در متون و چه در جایگاه کاوشگران اطلاعات در جهت آموزش ویژگی‌های خاص زبانی و انواع روش‌های بهبود کاوش، می‌تواند در جلوگیری از بروز مشکلات یادشده در پژوهش مؤثر باشد.

- ◇ برای پوشش صورت‌های گوناگون نوشتاری و معنایی واژه‌ها، کاربرد نمایه‌سازی مشارکتی[1] و توجه به آنچه که کاربران مختلف از طیف‌های گوناگون و با عادات نوشتاری متفاوت پیشنهاد و جست‌وجو می‌کنند، می‌تواند راهکاری جهت بهبود جست‌وجو و بازیابی باشد.
- ◇ هوشمندسازی نمایه‌سازی و واژگان در پایگاه‌های اطلاعاتی در مرحلهٔ ذخیره و بازیابی و نیز هوشمندسازی الگوریتم‌ها و مدل‌های بازیابی اطلاعات (مانند آنچه که گوگل در زمینهٔ همانندها، خطاهای املایی، و ... انجام می‌دهد).

فهرست منابع

آخشیک، سمیه‌سادات و رحمت‌الله فتاحی. ۱۳۹۱. تحلیل چالش‌های پیوسته‌نویسی و جدانویسی واژگان فارسی در ذخیره و بازیابی اطلاعات در پایگاه‌های اطلاعاتی. *فصلنامه کتابداری و اطلاع‌رسانی* ۱۶ (۳): ۹- ۳۰.

اسلامی، محرم. ۱۳۸۱. دشواری‌های پردازش رایانه‌ای خط فارسی. *فصلنامه نشر دانش* ۱۹ (۳): ۳۲-۲۸. http://www.noormags.com/view/fa/articleage/47746 (دسترسی در ۱۳۹۳/۲/۱۳).

اکبری‌نژاد، سعید. ۱۳۷۶. فاصلهٔ خالی میان واژه‌ها در ذخیره و بازیابی رایانه‌ای اطلاعات. *فصلنامه کتاب* ۸ (۱ و ۲): ۵۶-۴۹. http://www.noormags.com/view/fa/articlepage/87024 (دسترسی در ۱۳۹۳/۲/۱۳).

حری، عباس. ۱۳۷۲. کامپیوتر و رسم‌الخط فارسی. *فصلنامه تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی* ۳ (۱): ۱۱-۶. http://www.noormags.com/view/fa/articlepage/396231 (دسترسی در ۱۳۹۳/۲/۱۳).

حسینی بهشتی، ملوک‌السادات. ۱۳۸۶. معنی‌شناسی واژگانی فرااصطلاحنامه و بازیابی اطلاعات. *کتاب ماه کلیات مجموعه اطلاع‌رسانی و کتابداری* ۱۰ (۱۰): ۳۰- ۳۷.

رائی ساربانقلی، محمد. ۱۳۸۵. مشکلات جست‌وجو و بازیابی اطلاعات به‌زبان فارسی در اینترنت، مطالعهٔ موردی: کاربران مرکز اینترنت دانشگاه اسلامی واحد شبستر. *فصلنامه کتاب* ۱۷ (۳): ۱۹۶-۱۷۹. http://www.noormags.com/view/fa/articlepage/159553. (دسترسی در ۹۱/۶/۱).

رسولی، محمدصادق و بهروز مینایی بیدگلی. ۱۳۸۷. روشی جدید در خطایابی املایی در زبان فارسی. دومین کنفرانس داده‌کاوی ایران، تهران. ۲۲- ۲۱ آبان ۱۳۸۷.

1. folksonomy

روح‌پرور، رحیمه و محمود بی‌جن‌خان. ۱۳۸۶. به‌کارگیری یک نظام برچسب‌دهی برای تعبیر و تفسیر یک پیکر متنی زبان فارسی. هفتمین همایش زبان‌شناسی ایران، تهران. ۲۰-۲۱ آذر ۱۳۸۶.

زره‌ساز، محمد و رحمت‌الله فتاحی. ۱۳۸۵. ملاحظات اساسی در طراحی رابط کاربر نظام‌های رایانه‌ای و پایگاه‌های اطلاعاتی. *مطالعات ملی کتابداری و سازماندهی اطلاعات* ۱۷ (۲): ۲۵۱- ۲۶۸.

ستوده، هاجر و زهره هنرجویان. ۱۳۹۱. مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آن‌ها بر اثر بخشی پردازش خودکار متن و بازیابی اطلاعات. *فصلنامه کتابداری و اطلاع‌رسانی* ۱۵ (۴): ۵۹- ۹۲.

عبدالهی نورعلی، محمدصادق و عبدالرسول جوکار. ۱۳۸۸. چالش‌های شیوهٔ نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب. *مطالعات تربیتی و روان‌شناسی* ۱۰ (۲): ۶۷- ۹۰.

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. *فصلنامه کتابداری و اطلاع‌رسانی* ۱۳ (۲): ۱۹۹- ۲۲۲.

لیلا مرتضائی. ۱۳۸۰. مسائل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. *پژوهشنامه پردازش و مدیریت اطلاعات* ۱۷ (۱ و ۲): ۱۹-۲۶.

نشاط، نرگس. ۱۳۷۹. مسائل رسم‌الخط فارسی در رویارویی با فناوری نوین اطلاعاتی. *در فهرست‌های رایانه‌ای، کاربرد و توسعه*. مجموعه‌مقالات همایش کاربرد و توسعهٔ فهرست‌های رایانه‌ای در کتابخانه‌های ایران. آبان ۲۷-۲۸، (۴۰۱-۴۰۸). مشهد: دانشگاه فردوسی مشهد.

هماوندی، هدی. ۱۳۹۲. بررسی مشکلات جست‌وجو و بازیابی تصاویر در موتورهای کاوش برگزیده مبتنی بر ویژگی‌های نگارشی زبان فارسی، پایان‌نامهٔ کارشناسی ارشد، دانشگاه قم.

Hammo, B. H. 2009. Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents. *Information Retrieval* 12 (3): 300-323. http://link.springer.com/article/10.1007/s10791-008-9081-9. (Accessed July 19, 2012)

Internet world stats: Usage and population statistics. 2015. http://www.internetworldstats.com/stats7.htm. (accessed January 20, 2016)

Lazarinis, F. 2007. At the sharp END evaluating the searching capabilities of commerce websites in a non-English language A Greek case study. *Online Information Review* 31 (6): 881-891. http://www.emeraldinsight.com/journals.htm?articleid=1640585. (accessed July 17, 2012).

_____. 2008. Improving concept-based web image retrieval by mixing semantically similar Greek queries. *Program: electronic library and information systems* 42 (1): 56-67. http://www.emeraldinsight.com/journals.htm?articleid=1674242. (accessed July 17, 2012).

Lewandowski, D. 2008. Problems with the use of Web search engines to find results in foreign languages. *Online Information Review* 32 (4): 668-672. http://www.emeraldinsight.com/journals.htm?articleid=1747662. (accessed June 15, 2012).

Marchionini, G. 2008. Human information interaction research and development. *Library and information science research* 30 (4): 165-174. http://ils.unc.edu/~march/Marchionini_Inf_interact_LISR_2008.pdf. (accessed May 4, 2014)

Ruiter, De. J. 2006. Natural Language Interaction - the understanding computer. Essay for the course:

Human Computer Interaction. http://www.jdruiter.nl/published_work/Natural_language_interaction.pdf. (accessed June 10, 2014)

Zhang, J and L. Suyu. 2007. Multiple language supports in search engines. *Online Information Review* 31 (4): 516-532. http://www.emeraldinsight.com/journals.htm?articleid=1621798. (Accessed July 13, 2012.

هدی هماوندی

دانشجوی دکتری علم اطلاعات و دانش‌شناسی (گرایش بازیابی اطلاعات) دانشگاه تهران است.

ذخیره و بازیابی تصاویر، نمایه‌سازی و سازماندهی اطلاعات با تمرکز بر مسائل زبان‌شناختی از علایق پژوهشی وی است.

یعقوب نوروزی

متولد سال ۱۳۵۱، دارای مدرک تحصیلی دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی از دانشگاه آزاد اسلامی واحد علوم تحقیقات تهران است. ایشان هم‌اکنون دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه قم است. کتابخانه‌های دیجیتالی، سازماندهی اطلاعات، طراحی رابط کاربری، فناوری اطلاعات، نرم‌افزارهای کتابخانه‌ای و اطلاع‌رسانی از جمله علایق پژوهشی وی است.

ملوک‌السادات حسینی بهشتی

دانش‌آموخته دکتری تخصصی زبان‌شناسی از دانشگاه تهران است. ایشان هم‌اکنون استادیار پژوهشگاه علوم و فناوری اطلاعات، گروه پژوهشی اصطلاح‌شناسی و هستان‌شناسی است.

مدیریت دانش، مطالعه اصطلاح‌شناسی، سازماندهی اطلاعات، مدیریت اطلاعات و پردازش زبان طبیعی از جمله علایق پژوهشی وی است.