شناسایی واژه‌های غیرمفهومی (رایج) در نمایه سازی خودکار مدارک فارسی[1]

مجیده سنجی[2]

دکتر محمدرضا داورپناه[3]

چکیده:

پژوهش حاضر با هدف شناسایی واژه‌های غیرمفهومی در زبان فارسی و تهیه سیاهه‌ای از این واژه‌ها برای نمایه‌سازی خودکار متنهای فارسی در رشته‌های روانشناسی، علوم‌تربیتی و کتابداری و اطلاع‌رسانی انجام شده است. این پژوهش با روش تحلیل محتوا صورت گرفته‌است. جامعه آماری این پژوهش را مقاله‌های مندرج در آخرین شماره منتشرشده در مجله‌های علمی و پژوهشیِ رشته‌های علوم‌تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی در سال ۱۳۸۵ تشکیل می‌دهد. نمونه شامل ٦٣ مقاله است. گردآوری داده‌ها با استفاده از تفکیک واژگان به صورت ماشینی و دستی صورت گرفت. نتایج پژوهش نشان داد: ۱- افعال (معین و همراه‌شونده)، قیدها، ضمایر، حروف، اصوات، اعداد و علائم سجاوندی به‌عنوان واژه نمایه‌ها ظاهر نمی‌شوند، بنابراین، آنها را می‌توان واژه‌های غیرمفهومی یا به اصطلاح واژه‌های بازدارنده تلقی کرد. ۲- بدون احتساب علائم سجاوندی، در رشتهٔ علوم تربیتی ۳۹/۹٦٪، در رشته روانشناسی ۳۸/۵۷٪ و در رشته کتابداری ۳۸/۱۲٪ از حجم متون را واژه‌های غیرمفهومی تشکیل می‌دهد. ۳- واژه‌های بازدارندهٔ پربسامد در هرسه حوزه تقریبا مشابه است. ٤- از تعداد ۲٤۸۵۵۲ واژه (بدون احتساب علائم سجاوندی) که پیکرهٔ زبانی مورد مطالعه را تشکیل می‌داد، ۹۷۲۸۰ واژه که ۳۸/۹٤٪ کل واژه‌ها را

---

۱. برگرفته از پایان‌نامه کارشناسی ارشد با همین عنوان در رشته علوم کتابداری و اطلاع‌رسانی.

۲. کارشناس ارشد علوم کتابداری و اطلاع‌رسانی و مسئول بخش فهرستنویسی کتابخانه مرکزی دانشگاه امام رضا (ع).

۳. دانشیار گروه کتابداری و اطلاع رسانی دانشگاه فردوسی مشهد.

شامل می‌شود، جزء واژه‌های بازدارنده هستند.۵- نتیجهٔ مقایسه فهرست فارسی حاصل از پژوهش با فهرست واژه‌های بازدارنده انگلیسی نشان داد بین این دو فهرست به میزان ۲۸/۵٪ همپوشانی وجود دارد. ٦. همچنین ۲۰/۳۸٪ از واژه‌ها فاقد توزیع بسامدی یکسان در سه رشتهٔ مورد مطالعه می‌باشند.

کلیدواژه‌ها: واژه‌های غیرمفهومی، واژه‌های بازدارنده، نمایه‌سازی خودکار، ساخت واژگانی زبان فارسی.

مقدمه

نمایه‌سازی فرایند تحلیل محتوای اطلاعاتی، پیشینه‌ای از دانش و بیان‌کردن محتوای اطلاعاتی در زبان نمایه‌سازی از طریق اصطلاحات نمایه‌ای است. به این ترتیب، نمایه‌سازی حداقل سه مرحله دارد:

۱. انتخاب مفاهیم قابل نمایه در یک مدرک
۲. بیان‌کردن این مفاهیم در زبان نمایه‌سازی (به صورت واژه نمایه‌ها[1])
۳. تهیه یک فهرست مرتب‌شده از این شناسه‌ها (لاتینن[2]، ۲۰۰۰:ص. ۲۹)

نمایه‌سازی یا به‌صورت دستی و یا به‌صورت ماشینی[3] است. هنگامی‌که اصطلاحات نمایه‌ای توسط نمایه‌ساز انتخاب شود، نمایه‌سازی دستی است و چنانچه همه امور انتخاب کلیدواژه، ترتیب‌بندی و ... توسط رایانه انجام شود، نمایه‌سازی ماشینی است. اکثر روشهای نمایه‌سازی خودکار موجود، برای انتخاب اصطلاحات نمایه‌ای از زبان طبیعی[4] استفاده می‌کنند. در این شیوه تکواژه‌ها و عبارتهای چندواژه‌ای برای انعکاس محتوای متن، مستقیماً از عنوان، چکیده و متن کامل یک مدرک انتخاب می‌شوند(موئنز[5]،۲۰۰۳: ص. ۲۴).

در مجموع، در نظامهای نمایه‌سازی رایانه‌ای سعی نشده‌است تا عملکرد ذهنی شخص نمایه‌ساز تقلید شود. برنامه‌ریزی رایانه‌ای به منظور گزینش اصطلاحات حایز اهمیت از متنی با زبان طبیعی، مستلزم این است که برنامه درک خوبی از زبانشناسی و

. . . . . .

1. Index Term.
2. Timo Lahtinen.
3. Automatic Indexing.
4. Natural language.
5. Marie- Francine Moens.

دانش لازم در مورد موضوعی که تحت بررسی است، داشته‌باشد. البته، این امر در حال حاضر برای تمام و یا اکثر نظامهای بازیابی کار بسیار بزرگی است. در عوض، نمایه‌سازی خودکار به روشهایی که بر فراوانی نسبی کلمات در متن مبتنی است، تکیه دارد (ویکری و یکری[1]، ۱۳۸۰:ص ۱۸۰).

می‌توان مراحل زیر را در فرایند نمایه‌سازی خودکار درنظر گرفت:

۱. شناسایی واژه‌های انفرادی از متن که تحلیل واژگان[2] نامیده‌می‌شود (موئنز، ۲۰۰۳: ص. ۷۷)

۲. برداشتن واژه‌های کاربردی و واژه‌های با بسامد تکرار بالا که در ارائه محتوای متن بی‌تأثیرند، با استفاده از فهرست واژه‌های غیرمجاز.

۳. تبدیل واژه‌های باقی‌مانده به شکل ریشه[3] آنها؛ یعنی حذف پسوندها یا پیشوندها تا هر کلمه تا حد ریشه‌اش کوتاه شود.

۴. محاسبهٔ رایانه‌ای بسامد رخدادهای ریشه‌هایی که در متن تحلیل شده‌اند، به منظور تعیین تابع ارزشگذاری هر ریشه.

۵. ریشه‌هایی که نسبت به بعضی ارزشهای قراردادی آستانه‌ای[4]، ارزشگذاری بزرگتری دارند، برای متنی که در آن ظاهر شده، به عنوان کلیدواژه تعیین می‌شود. در بعضی نظامها، کلیدواژه ممکن است ارزشی متناسب با ارزش تابع ارزشگذاری داشته باشد (ویکری و ویکری[5]، ۱۳۸۰: ص. ۱۸۳).

به هر حال، تعیین واحدهای متنی و مشخص کردن حد و حدود واژه برای ماشین از مسائل اساسی در گزینش اصطلاحات نمایه‌ای در نمایه‌سازی خودکار است (گیلوری، ۱۳۷۹). به علاوه، امکان تشخیص واژه‌های مفهومی از واژه‌های غیرمفهومی، در فرایند انتخاب اصطلاحات نمایه تأثیر بسزایی دارد. آنچه مسلم است، ماشین امکان تشخیص را

. . . . . .

1. Brian C. Vickery and Elian Vickery.
2. Lexical Analysis.
3. Stemming.
4. Threshold Value.
5. Brian C. Vickery and Elian Vickery.

تنها از طریق تطبیق واژه‌های استخراج شده از متن یا منتسب شده به متن با فهرستی که واژه‌های غیرمجاز نامیده می‌شود، به دست می‌آورد. در اختیار داشتن فهرستی از این واژه‌ها و ارائه آنها به برنامه رایانه‌ای برای ممانعت از ورود آنها به فهرست واژه‌های مفهومی مطلوب برای نمایه‌شدن، یکی از اقدامهای سودمند در نمایه‌سازی خودکار مبتنی بر کلیدواژه‌هاست.

تحلیل کلمات یک متن نشان می‌دهد گروهی از کلمات بی‌اهمیت وجود دارد که به‌فراوانی در متن ظاهر می‌شود (مانند یک، به، نه، برای، با، چه کسی، چه موقع، است، آن). گروهی نیز وجود دارد که بندرت در متن می‌آیند و ممکن است نشان‌دهنده محتوای اطلاعاتی متن نباشند (ویکری و ویکری، ۱۳۸۰:ص ۱۸۰). این دسته از واژه‌ها به تنهایی بارمعنایی ندارند (حاوی معنا و مفهوم نیست) بلکه در ارتباط با واژه‌های دیگر معنا می‌یابند. به این ترتیب، مفهوم و محتوای متن را نشان نمی‌دهند. از طرف دیگر، بود یا نبود آنها نه تنها در پرسش کاربر تأثیری ندارد، بلکه در میزان ربط یا عدم ربط مدارک بازیابی شده نیز تأثیری نخواهدداشت. این واژه‌ها با عنوان واژه‌های غیرمجاز برای ورود به نمایه معرفی می‌شوند. درصورتی‌که واژه‌های غیرمجاز قبل از فرایند نمایه‌سازی مدارک مشخص و فهرست آنها برای کنترل به رایانه داده‌شود، علاوه‌بر صرفه‌جویی در زمان و حجم بایگانیهای نمایه، به میزان زیادی از بازیابی مدارک نامرتبط و ریزش کاذب در جستجو جلوگیری خواهدشد (زو و دیگران[1]، ۲۰۰۶).

با گسترش مدارک الکترونیکی فارسی و به تبع آن کاربران فارسی زبان و همچنین ویژگیهای خاص زبان و خط فارسی، نیاز به توجه و بهبود روشهای نمایه‌سازی بیش از پیش احساس می‌شود. زبان فارسی مانند هر زبان دیگری واژه‌هایی دارد که هیچ گونه سهمی در ارائه بار معنایی مدرک ندارند، ولی فهرستی از پیش آماده از این‌گونه واژه‌ها در دسترس نبوده و مشخص نیست چگونه باید انتخاب شوند. بنابراین، مسئله اساسی این پژوهش آن است که: معیارهای شناسایی واژه‌های غیرمفهومی در زبان فارسی کدام است؟

1. Feng Zou and Others.

و آیا می‌توان سیاهه‌ای از این واژه‌ها را برای نمایه‌سازی خودکار متنهای فارسی در حوزه روانشناسی، علوم‌تربیتی و کتابداری و اطلاع‌رسانی تهیه کرد؟

هدفهای پژوهش

این پژوهش با بهره‌گیری از اصول ساختاری زبان فارسی، درپی دستیابی به هدفهای زیر است:

- شناسایی معیارهای نحوی و معنایی زبان فارسی در تشخیص واژه‌های غیرمفهومی
- تهیهٔ سیاهه‌ای از واژه‌های غیرمفهومی در زبان فارسی
- مقایسه واژه‌های غیرمفهومی فارسی و انگلیسی
- بررسی چگونگی توزیع بسامد واژه‌های غیرمفهومی فارسی در سه رشتهٔ علوم تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی.

سؤالهای پژوهش

۱. معیارهای نحوی و معنایی شناسایی واژه‌های غیرمفهومی در زبان فارسی کدام است؟

۲. واژه‌های غیرمفهومی در زبان فارسی که بتواند در نمایه‌سازی خودکار مدارک فارسی هر یک از حوزه‌های مورد مطالعه نادیده گرفته‌شود، کدام است؟

۳. واژه‌های غیرمفهومی چه حجمی از متون زبان فارسی هر یک از حوزه‌های مورد مطالعه را تشکیل می‌دهد؟

۴. بین سیاههٔ واژه‌های غیرمفهومی انگلیسی و واژه‌های غیرمفهومی به دست آمده در این پژوهش، چه وجوه تشابه و تفاوتی وجود دارد؟

۵. توزیع بسامد واژه‌های غیرمفهومی در رشته‌های علوم‌تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی چگونه است؟

پیشینهٔ پژوهش

تاریخچهٔ نمایه‌سازی خودکار بر مبنای بسامد تکرار واژه، به دههٔ ۱۹۵۰ و تحقیقات «لوان»[1](۱۹۷۵) و «باکسن دال»[2] (۱۹۵۸) بر می‌گردد. پیش فرض کلی در بازیابی اطلاعات این است که اصطلاحات با بسامد متوسط، مناسب‌ترین اصطلاحات برای نمایه‌سازی هستند. این فرض اساس کار «زیف»[3] (۱۹۴۹) قرار گرفت.

«فاکس»[4] (۱۹۹۰) از یک شیوهٔ رایج، یعنی واژه‌های پربسامد گروهی بزرگ از آثار نوشته شده در همان زبان، استفاده کرد. بسامد واژه‌ها در یک مجموعه مواد انگلیسی با عنوان پیکره زبانی براون در حوزه‌های موضوعی متنوع تعیین شده‌است. مراحل کار فاکس را می توان در مراحل زیر خلاصه کرد:

- تعیین بسامد واژه‌ها در یک مجموعه از مدارک نوشته شده
- محدودکردن فهرست به اندازهٔ مناسب برای استفاده (تعیین نقطه برش. فاکس ۳۰۰ بار تکرار واژه را انتخاب نمود)
- حذف واژه های بامعنی اما با بسامد بالا در سیاههٔ واژه‌های بازدارندهٔ اولیه.
- اضافه کردن واژه‌های پربسامد و کم‌اهمیتی که نقطه برش را ازدست‌داده‌اند.

«فرانسیس و کوسرا»[5] (نقل شده در : گیلوری ۱۳۷۹) در مرحله اول ده کلمه پر بسامد را تعیین و در مرحله دوم فهرستی از ۴۲۵ کلمه ممنوعه را تهیه کردند . «رایجزبرگن»[6] (نقل شده در: گیلوری ۱۳۷۹) نیز فهرستی ارائه داد که از ۱۵۰ کلمه بازدارنده تشکیل شده بود. فهرست برگمن در سال ۱۹۷۵ منتشر شد.

در سال ۱۹۸۷، در کتابخانهٔ بازمن[7] دانشگاه ایالت مونتانا فهرستی از واژه‌های بازدارنده نرم و سخت برای فهرست پیوسته آن با هدف جلوگیری از نمایه‌سازی آنها و

. . . . . .

1. H. P. Luhn.
2. Boxendal.
3. George Zipf.
4. Christopher Fox.
5. Francis and Kucera.
6. Van Rijsbergen.
7. Bazemen.

درنتیجه بازیابی آنها تهیه شد. در سال ۱۹۹۲ به دلیل تغییر در نرم افزار پایگاه اطلاعاتی و نیاز به نمایه‌سازی مجدد مدخلهای آن، بهترین فرصت جهت ارزیابی و بهبود فهرست واژه‌های بازدارنده به دست آمد.

«سیروتکین و ویلبور»[1] (۱۹۹۲) با استفاده از این پیش فرض که واژه‌های بازدارنده به عنوان واژه‌هایی که احتمال رخداد یکسانی در مدارک مرتبط و نامرتبط با درخواست اطلاعاتی دارند، شناخته می‌شوند، پژوهشی را انجام دادند. آنها بیان می‌کنند که این احتمال وجود دارد واژه‌های بازدارنده در یک مجموعه با استفاده از یک روش بازیابی برداری مبتنی بر ضریب تأثیر کسینوس مدارک و تشابه مدارک انجام گیرد. نتیجه بررسی سیروتکین و ویلبور در مجموعه‌ای از مدارک مدلاین (۷۱۳۱۱ مدرک) در حوزهٔ بیوتکنولوژی کاهش ۲۰۳۰۴۰ تک‌واژه در این مدارک به ۵۰۵۰۸ (کاهش ۷۸ درصدی) تک‌واژه است.

«یمین و ویلبور»[2] (۱۹۹۶) با استفاده از روش ویلبور و سیروتکین (۱۹۹۲) و دو روش طبقه‌بندی آماری (Expert Network و LLSF) برای بازیابی و دسته‌بندی مدارک و یک روش تطابق واژه‌محور برای جستجو در پایگاه‌ها (WBM[3]) استفاده نمودند. در این بررسی، به عنوان مثال روش Expert Network روی مدارک کتابخانه ملی پزشکی آمریکا، با حذف ۸۷٪ واژه‌های بازدارنده، ۸۰۰۲ واژه به ۱۰۴۲ واژه کاهش یافت و در نتیجه باعث صرفه‌جویی ۶۳ درصدی زمان، ۷۴ درصدی در حجم فایل مقلوب و در نهایت بهبود دقت بازیابی تا ۱۰٪ گردید.

«برگ»[4] (۱۹۹۷) با استفاده از روش «ادمونسون و وایلز»[5] (۱۹۵۹) درمورد تعیین اهمیت واژه، پژوهشی را در زمینه تعیین واژه‌های بازدارنده در یک حوزه موضوعی مشخص بر اساس مجموعه‌ای از مدارک نیروی هوایی آمریکا انجام داد.

. . . . . .

1. W. John Wilbur and Karl Sirotkin.
2. Yang Yiming an W. John Wilbur.
3. Word - based Matching.
4. Craig N. Berg.
5. H. P. Edmondson and R.E. Wyllys.

«ساووی»[1] (۱۹۹۹) برای شناسایی واژه‌های بازدارنده در زبان فرانسوی از پیکره‌ای شامل دو مجموعه آزمایشی استفاده کرد. وی با پیروی از خط مشی فاکس ابتدا ۲۰۰ واژه پربسامد را استخراج کرد؛ سپس با بازبینی فهرست به دست آمده، تمام اسمها و صفتهایی که بسامد بالایی داشتند ولی با موضوعات اصلی پیکره مرتبط بودند، از فهرست حذف شدند. ضمن اینکه بعضی واژه‌های غیر اطلاعی که در ۲۰۰ واژه پربسامد ظاهر نشده بودند مانند ضمایر ملکی و شخصی، حروف اضافه، ربط و تعریف و بعضی از اشکال فعل Be به این فهرست ابتدائی اضافه شد. فهرست نهایی شامل ۲۱۵ واژه است و هنگامی که از چنین فهرستی استفاده می شود، اندازه فایل مقلوب برای مجموعه آزمایشی اول تا حدود ۲۱٪ و برای مجموعه آزمایشی دوم حدود ۳۵٪ کاهش می یابد.

«هو»[2] (۱۹۹۹) پژوهشی را با استفاده از این حقیقت زبانشناسی که بیش از نیمی از واژه‌های یک نمونه صفحه انگلیسی از میان ۱۵۰ واژه‌بازدارنده رایج هستند، یک شیوهٔ سریع برای تعیین محل واژه‌های بازدارنده ارائه می‌دهد. این روش از سنجش دامنه واژه‌های انفرادی و واژه‌های همجوار آنها استفاده می‌کند. در یک آزمایش با استفاده از ۴۰۰ تصویر از صفحات، این روش ۶۳٪ از واژه‌های زاید متن را حذف کرد. وی بیان می‌کند تمایز بین واژهٔ بازدارنده و غیربازدارنده اغلب به تشخیص کلیدواژه کمک می‌کند.

«ساووی» ۲۰۰۱، ۲۰۰۲ و ۲۰۰۷) براساس کار قبلی خود فهرستی از واژه‌های بازدارنده را برای زبانهای فنلاندی، ایتالیایی، آلمانی، اسپانیایی و بلغاری شناسایی نمود. با استفاده از همین شیوه، «ساووی و راسولوفو»[3] (۲۰۰۳) فهرست واژه‌های بازدارنده در زبان عربی را نیز ایجاد کردند. فهرست ایجاد شده مبتنی بر پیکره زبان عربی است که توسط دیوید کرافت و کوین والکرد در کنسرسیوم اطلاعات زبانشناسی در فیلادلفیا شکل گرفته و دربرگیرندهٔ ۸۷۲/۳۸۳ مدرک (۷۶ میلیون علامت) حاوی بیش از ۶۶۶/۰۹۴ تک واژه است.

. . . . . .

1. Jacques Savoy.
2. Tin Kam Ho.
3. Jacques Savoy and Rasolofo.

«ابوالخیر» در پایان‌نامه دکتری خود با عنوان «اثربخشی روشهای پردازش متن برای بازیابی زبان عربی» به فهرستی از واژه‌های بازدارنده نیازداشت. به این منظور، با استفاده از سه شیوهٔ متفاوت: ۱- مبتنی بر ساختار و ویژگیهای زبان عربی ۲- مبتنی بر آمارهای پیکره زبانی و ۳- ترکیبی از این دو روش، فهرستی از واژه‌های بازدارنده زبان عربی را تهیه نمود.

در پژوهشی دیگر، «زو و دیگران»[1] (۲۰۰۶) برای استخراج واژه‌های بازدارنده در زبان چینی از یک روش خودکار یکپارچه مبتنی بر الگوهای اطلاعاتی و آماری استفاده کردند. الگوی آماری بر اساس احتمال و توزیع واژه و الگوی اطلاعاتی بر اساس اهمیت واژه با استفاده از نظریه اطلاعات عمل می‌کند. در انتها نتایج به دست‌آمده از این دو الگو یکپارچه می‌شوند.

«لازارینیس»[2] (۲۰۰۷) مقاله ای را با هدف پردازش ساختمان فهرستی از واژه‌های بازدارنده برای زبانهای غیر لاتین و ارزیابی تأثیر حذف این واژه ها از پرسش کاربران ارائه داد. برای انجام این پژوهش، مجموعه ای از ۳۲ پرسش موثق و صحیح که توسط کاربران پیشنهادشده‌بود، یک نوبت با درنظرگرفتن واژه‌های بازدارنده و نوبت بعد بدون این واژه‌ها به موتور کاوش گوگل داده شد. به این ترتیب، اهمیت حذف واژه‌های بازدارنده از عبارت پرسش بر حسب ربط در ۱۰ نتیجه برتر گوگل ارزیابی شده‌است.

اما بررسی متون و مرور نوشتار نشان می‌دهد در مورد واژه‌های بازدارنده در زبان فارسی کارهای معدود و پراکنده‌ای صورت پذیرفته است که در ادامه به آن پرداخته خواهد شد.

«تقوا، بکلی و سده»[3] (۲۰۰۳) مجموعه‌ای متشکل از ۱۸۵۰ مدرک در یک دورهٔ ۶ ماهه از میان وب‌سایتهای فارسی با حوزه‌های موضوعی متنوع که عمدتاً نسخه الکترونیکی روزنامه‌ها و مجله های مشهور ایرانی بودند، و یا وب‌سایتهای فارسی طراحی‌شده در آمریکا را جمع‌آوری کردند. آنها فهرست خود را بر اساس پراکندگی

. . . . . .

1. Feng Zou And Others.
2. Fotis Lazarinis.
3. Kazem Taghva, Russel Bechley and Mohammad Sadeh.

واژه‌ها تهیه نمودند. در فهرست نهایی ۱۲ فعل وجود داشت که «تقوا و همکارانش» به جای فهرست‌کردن همه وجوه و زمانهای این افعال، تنها زمان گذشته و حال بن فعل را در فهرست خود وارد کردند.

«پور اسماعیل و رستمی» (۱۳۸۴) ابتدا فهرست تقوا، بکلی و سده(۲۰۰۳) را برای صیغه‌ها و زمانهای مختلف کامل کردند. این فهرست شامل ۲۰۴ فعل فارسی است. سپس با کمک مجموعه آزمون استاندارد محک که بالغ بر ۳۰۰۰ مستند را شامل می‌شود، بسامد کلمات مختلف را محاسبه نمودند و از میان آنها در مرحله مقدماتی کلماتی را که بسامدی بیشتر از ۱۰۰۰ بار داشته اند انتخاب و سپس فهرست نهایی را که شامل ۳۴ واژه می‌شود، استخراج کردند.

«داورپناه و بلندیان» (۱۳۸۶) پژوهشی را با موضوع نمایه‌سازی ماشینی متون فارسی براساس قانون زیف انجام دادند. نتایج نشان داد توزیع فراوانی واژگان در متون فارسی دارای الگوی پیش‌بینی‌پذیر است. کاربرد واژه‌های با بسامد بالا و بسامد پایین درمقاله های فارسی، از قانون زیف پیروی می کند. همچنین، مشخص شد بسامد واژگانی می‌تواند به عنوان معیاری برای نمایه‌سازی ماشینی متون فارسی در نظرگرفته شود. وضعیت همخوانی کامل بین بسامد واژگانی و کلیدواژه‌های موضوعی در شیوهٔ تفکیک صرفا ماشینی بدون دخالت عامل انسانی به طور متوسط در کل مقاله های مورد بررسی به میزان ۲۱/۵۰٪ است. در شیوهٔ تفکیک ماشینی با دخالت عامل انسانی، میزان همخوانی به ۵۲٪ می‌رسد. وضعیت همخوانی کامل بسامد واژگانی با کلیدواژه‌های عنوانی در شیوهٔ صرفاً ماشینی بدون دخالت عامل انسانی، به طور متوسط در کل، مقاله های مورد بررسی ۹/۲۰٪ است که در شیوهٔ ماشینی با دخالت عامل انسانی این میزان بیشتر از ۵ برابر شده و به ۵۴/۱۴٪ می‌رسد.

روش پژوهش، جامعه آماری و حجم نمونه

این پژوهش با استفاده از روش تحلیل محتوا انجام پذیرفت. جامعه آماری این پژوهش، مقاله های مندرج در آخرین شماره منتشر شده در مجله های علمی و پژوهشی در

رشته‌های علـــوم‌تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی در سال ۱۳۸۵ است. این مجله ها از فهرست مجله های مورد تأیید وزارت علوم، تحقیقات و فناوری در همین سال شناسایی شد که به شرح ذیل است:

- پژوهش در مسائل تعلیم و تربیت/ انجمن ایرانی تعلیم و تربیت
- نوآوریهای آموزشی/ وزارت آموزش و پرورش
- آموزش عالی ایران/ انجمن آموزش عالی ایران
- پژوهش و برنامه‌ریزی در آموزش عالی/ مؤسسه پژوهش و برنامه‌ریزی آموزش عالی
- روانشناسی و علوم‌تربیتی/ دانشگاه تهران
- روانشناسی و علوم‌تربیتی/ دانشگاه تبریز
- علوم تربیتی و روانشناسی/ دانشگاه شهید چمران
- مطالعات تربیتی و روانشناسی/ دانشگاه فردوسی مشهد
- پژوهشهای روانشناختی/ رضا زمانی (بخش خصوصی)
- تازه‌ها و پژوهشهای مشاوره/ انجمن مشاوره ایران
- روانشناسی/ انجمن ایرانی روانشناسی
- کتابداری و اطلاع‌رسانی/ کتابخانه مرکزی و مرکز اسناد آستان قدس رضوی

با استفاده از جدول تعیین حجم نمونه مورگان، برای ۷۳ عنوان مقاله (کتابداری ۲۰ عنوان، روانشناسی ۲۲ عنوان و علوم‌تربیتی ۳۱ عنوان مقاله) حجم نمونه پژوهش ۶۳ عنوان مقاله است؛ و از آنجا که تعداد مقاله های سه رشته با هم برابر نبود، تعداد مقاله‌های هر رشته براساس فرمول زیر محاسبه و تعیین شد:

$$\text{تعداد مقاله های هر رشته در نمونه} = \text{حجم نمونه} \times \frac{\text{حجم جامعه آن طبقه}}{\text{حجم کل جامعه}}$$

با روش نمونه‌گیری تصادفی طبقه‌ای، در رشته کتابداری ۱۷ عنوان مقاله، رشته روانشناسی ۱۹عنوان و در رشته علوم‌تربیتی ۲۷ عنوان مقاله انتخاب شد.

گردآوری داده‌ها

برای تهیه سیاههٔ واژه‌های غیرمجاز از متون مورد مطالعه، اولین گام، تفکیک واژگان این متون بود. برای رسیدن به این هدف، مراحل زیر انجام پذیرفت:

۱. تهیه متن الکترونیکی

در اولین گام نسخه الکترونیکی مقاله های منتشرشده در آخرین شماره مجله‌های علمی و پژوهشی در سه رشته مورد مطالعه در سال ۱۳۸۵ در محیط نرم‌افزاری Word که امکان تفکیک واژگان متن در آن وجود دارد، تهیه شد.

۲. تفکیک واژگان

برای استفاده از اصول به دست آمده و استخراج واژه‌های غیرمفهومی از میان دیگر واژه‌ها، واژگان هر یک از مقاله های حجم نمونه (۶۳ مقاله) تفکیک شد. تفکیک واژگان متن هر یک از مقاله ها به صورت ماشینی و با استفاد از فرامین موجود در نرم افزار Word صورت پذیرفت. سپس واژگان تفکیک شده هر متن براساس معیارهای زبانشناسی، قواعد دستوری و آیین نگارش فارسی ازلحاظ نوع و بار معنایی به صورت دستی بررسی و ویرایش گردید. به این ترتیب، در تایپ مجدد متن مقاله ها و در تفکیک واژگان، معیارهای زیر مورد استفاده قرارگرفت:

- همه صیغه‌ها، وجوه و انواع فعل، به‌صورت یک واژه ← فراهم شده‌است، رفته بودم
- افعال مرکب به‌صورت یک واژه ← بناکرد
- افعال پیشوندی به‌صورت یک واژه ← ازدست‌داده‌است
- مصدرهای مرکب به‌صورت یک واژه ← پیش رفتن
- اسامی مرکب به‌صورت یک واژه ← آیین نامه
- اسامی پیشوندی به‌صورت یک واژه ← بی‌گناه ، به سرعت

- اسامی میانوندی به‌صورت یک واژه ← خودبه‌خود
- پاره‌های غیرمستقل و واژه‌های ترکیبی به‌صورت یک واژه ← همکار
- عبارتهایی که به‌عنوان گروه اسمی، گروه قیدی، گروه حرف اضافه شناخته می‌شوند، چون براساس مفهومی که حامل آن هستند غیرقابل تفکیک می‌باشند، به صورت یک واژه ← محمدحسین دیانی، شنای صدمتر
- نام‌آواها به‌صورت یک واژه ← جیک‌جیک ، وزوز
- فاصله درمورد افعال حذف می‌شوند ← می‌گیرد
- فاصله درمورد علامت جمع (ها، های، هایی) حذف شده و شمارش نمی‌شوند ← ماشین‌ها
- فاصله درمورد تکواژهای صرفی (تر، ترین) حذف شده و شمارش نمی‌شوند ← خوشبخت ترین
- آیه‌های قرآن و واژه‌های انگلیسی (خارجی) درصورت وجود به دلیل غیرفارسی بودنشان حذف می‌شوند.
- اختصارات حذف‌شده و شمارش نمی‌شوند ← ج. Cm (داورپناه و بلندیان، ۱۳۸۶)

۳.شمارش بسامد واژگان

برای شمارش واژگان مرتب‌شده هرمتن، از دستور Word Count استفاده‌شد.

یافته‌های پژوهش

با استفاده از داده‌های گردآوری شده به سؤالهای پژوهش پاسخ داده شد که توضیح آن به شرح زیر است:

۱. معیارهای نحوی و معنایی شناسایی واژه‌های غیرمفهومی در زبان فارسی کدام است؟

با مطالعه متون و کتابهای مربوط به حوزه زبانشناسی[۱]، دستور زبان فارسی[۲] و متون مربوط به تهیه و تدوین اصطلاحنامه‌ها ساخت واژگانی زبان فارسی مورد مطالعه قرارگرفت؛ سپس اصول و قواعدی مشخص و مستدل استخراج شد که با استناد به آنها، شناسایی و استخراج واژه‌های کم‌معنا یا بدون معنا در زبان فارسی امکان پذیر خواهد بود. این اصول و قواعد عبارتند از:

. . . . . .

۱. باقری، مهری(۱۳٦۷). «مقدمات زبانشناسی». تبریز: دانشگاه تبریز.
صفوی، کورش(۱۳۶۰). «درآمدی بر زبانشناسی». تهران: بنگاه ترجمه و نشر.
نجفی، ابوالحسن(۱۳۸۰). «مبانی زبانشناسی و کاربرد آن در زبان فارسی». تهران: نیلوفر.
هادسن، گرور(۱۳۸۳). «مباحث ضروری و بنیادین زبانشناسی مقدماتی (ضرورت زبانشناسی مقدماتی)». ترجمه علی بهرامی. تهران: رهنما.
پالمر، فرانک(۱۳۶۶). «نگاهی تازه به معنی شناسی». ترجمه کورش صفوی. تهران: مرکز.
مشکوة الدینی، مهدی(۱۳۸۲). *دستور زبان فارسی بر پایه نظریه گشتاری* (ویرایش ۲). مشهد: فاطمی.
۲. شفائی، احمد(۱۳۶۳). «مبانی علمی دستور زبان فارسی». تهران: نوین.
بابک، علی(۱۳۸۳). «دستور زبان فارسی پژوهشی معاصر». تهران: دانشگاه آزاد اسلامی مشهد: سخن.
مشکوةالدینی، مهدی(۱۳۸۴). «دستور زبان فارسی (واژگان و پیوندهای ساختی)». تهران: سمت.
وحیدیان کامکار، تقی؛ عمران، غلامرضا، (۱۳۸۵). *دستور زبان فارسی (۱)*. تهران: سازمان مطالعه و تدوین کتب علوم انسانی(سمت).
ناتل خانلری، پرویز (۱۳۵۹). *دستور زبان فارسی* (با تجدیدنظر). تهران: توس.
مشکوة الدینی، مهدی (۱۳۸۴). *دستور زبان فارسی. واژگان و پیوندهای ساختی*. تهران: سازمان مطالعه و تدوین کتب علوم انسانی(سمت).
معین، محمد(۱۳۷۸). *فرهنگ فارسی (متوسط): شامل یک مقدمه و سه بخش لغات، ترکیبات خارجی، اعلام ....* تهران: امیرکبیر.
مرزبان راد، علی(۱۳۷۸). *دستور سودمند*. تهران: دانشگاه صنعتی امیرکبیر.
محتشمی، بهمن(۱۳۷۰). *دستور کامل زبان فارسی*. تهران: اشراقی.
صهبا، عبدالرشید (۱۳۷۱). *حرفهای ربط، اضافه، نشانه در دستور زبان فارسی برای استفاده دانش آموزان، دانشجویان و پژوهندگان*. تهران: غزل.
غلامعلی زاده، خسرو(۱۳۷۴). *ساخت زبان فارسی*. تهران: احیاء الکتاب.
فرشیدورد، خسرو(۱۳۸۲). *دستور مفصل امروز*. تهران: سخن.
فرشیدورد، خسرو(۱۳۸۶). *دستور برای لغت سازی: فرهنگ پیشوندها و پسوندهای فارسی به همراه گفتارهایی درباره دستور زبان فارسی*. تهران: زوار.
کلباسی، ایران(۱۳۸۰). *ساخت اشتقاقی در فارسی امروز*. تهران: پژوهشکده علوم انسانی و مطالعات فرهنگی.
دهخدا، علی اکبر(۱۳۸۳). *لغتنامه*. (با همکاری محمد معین، جعفر شهیدی). تهران: موسسه لغتنامه دهخدا.
خطیب رهبر، خلیل(۱۳۷۹). *دستور زبان فارسی: کتاب حرف اضافه و ربط مشتمل بر تعریف و تقسیم و شرح اصطلاحات و معانی و کاربرد حروف*. تهران: مهتاب.
خطیب رهبر، خلیل(۱۳۸۱). *دستور زبان فارسی: برای پژوهش دانشجویان و ادب دوستان در آثار شاعران و نویسندگان بزرگ ایران*. تهران: مهتاب.
باطنی، محمدرضا(۱۳۸۲). *توصیف ساختاری دستوری زبان فارسی بر بنیاد یک نظریه عمومی زبان*. تهران: امیرکبیر
انوری، حسن(۱۳۸۱). *فرهنگ بزرگ سخن*. تهران: سخن.
انوری، حسن؛ احمدی گیوی، حسن (۱۳۷۷). *دستور زبان فارسی* ۲ (ویرایش ۲). تهران: فاطمی
احمدی گیوی، حسن(۱۳۸۰). *دستور تاریخی فعل*. تهران: قطره.

■ به کوچکترین واحد معنادار که در ساخت واژه مشخص می گردد، تکواژ گفته می‌شود.

■ تکواژها از دید کم و بیشی در تعداد بسامد (کاربرد) به دو گروه محدود یا بسته و نامحدود یا باز تقسیم می‌شوند.

■ تکواژهای زبان از نظر ایفای نقش به دو گروه تکواژهای قاموسی و تکواژهای دستوری تقسیم می‌شوند.

■ تکواژهای قاموسی معنای مستقلی داشته و بر اشیا، اعمال و کیفیات خاص که قابل حس و لمس و درک هستند دلالت دارند. تعداد اجزا و آحاد این گروه از واژه‌ها، ثابت، معیّن و محدود نیست و فهرست آنها در زبان باز است.

■ تکواژهای دستوری اغلب به تنهایی به‌کار نمی‌روند (کارکرد دستوری دارند) و معنای آنها با پیوستن به تکواژهای دیگر آشکار می‌شود. این گروه دارای شمار معیّن و ثابتی از اعضا و اجزا هستند. فهرست این تکواژها بسته و محدود است.

■ هرچه تعداد آحاد و تکواژها بیشتر باشد، بسامد آنها کمتر می‌شود.

■ هرچه تعداد آحادِ تکواژها کمتر باشد، بسامد آنها (یعنی میزان کاربرد آنها در جمله های مختلف) بیشتر می‌شود.

■ تکواژی که متعلق به گروه محدود و بسامد آن بالا باشد، تکواژ قاموسی است.

■ تکواژهای دستوری شامل ضمایر، قیود، حروف، اصوات، اعداد و افعال معین می‌باشند.

در پایان، بر اساس معیارهای ذکر شده در بالا، می‌توان چنین استنباط نمود که، افعال (معین و همراه شونده)، قیدها، ضمایر، حروف، اصوات، اعداد و علایم سجاوندی به عنوان واژه نمایه ها ظاهر نمی شوند. این قواعد مبنایی را برای شناسایی و تهیه فهرست واژه‌های بازدارنده در زبان فارسی فراهم‌می‌کند.

۲. واژه‌های غیرمفهومی در زبان فارسی که بتواند در نمایه‌سازی خودکار مدارک فارسی هر یک از حوزه های مورد مطالعه نادیده گرفته‌شود، کدام است؟

به منظور شناسایی واژه های غیرمفهومی با توجه به معیارهای استخراج شده ذیل سؤال اول پژوهش، ابتدا نوع دستوری واژگان بررسی شد. برای تعیین نوع دستوری واژه ها از فرهنگهای لغت فارسی به فارسی ـ لغتنامه دهخدا، فرهنگ معین و فرهنگ سخن ـ استفاده شد. در تعیین نوع دستوری واژه هایی که در این سه فرهنگ وجود نداشت، از کتابهای دستور زبان فارسی و مشورت با صاحب نظران استفاده گردید. چون برخی از واژه‌ها دارای چندین نقش دستوری هستند، ملاک ما در انتخاب واژه بازدارنده آن نوع دستوری از واژه بود که براساس اصول استخراج شده در سؤال اول پژوهش، جزء واژه‌های غیرمفهومی زبان فارسی قرار می‌گیرند. به این ترتیب، نوع دستوری تک‌تک ۲۴۸۵۵۲ واژه تشکیل دهندهٔ متون مورد مطالعه مشخص شد. فهرست درهم‌کرد این واژه‌ها به جهت کوتاه تر شدن، بدون ذکر نوع دستوری، براساس بسامد واژه‌ها به ترتیب از بیشترین به کمترین میزان تکرار در جدول شماره ۱ ارائه شده است.

جدول شماره ۱. فهرست درهم‌کرد واژه‌های بازدارنده سه رشته مورد مطالعه

۱. و	۱۱. آن
۲. در	۱۲. خود
۳. به	۱۳. نیز
۴. که	۱۴. آنها
۵. از	۱۵. بر
۶. این	۱۶. یا
۷. را	۱۷. بین
۸. است	۱۸. یک
۹. با	۱۹. می‌شود
۱۰. برای	۲۰. دو

۲۱. بود
۲۲. تا
۲۳. دارد
۲۴. دیگر
۲۵. شد
۲۶. شده‌است
۲۷. هر
۲۸. هستند
۲۹. دارند
۳۰. می‌باشد
۳۱. بنابراین
۳۲. باید
۳۳. براساس
۳۴. آنان
۳۵. همچنین
۳۶. بیشتر
۳۷. یکی/ یکی‌از
۳۸. میان
۳۹. نسبت‌به
۴۰. یعنی
۴۱. ما
۴۲. می‌تواند
۴۳. می‌توان
۴۴. سه
۴۵. نیست
۴۶. به‌صورت
۴۷. یک
۴۸. ازنظر
۴۹. برخی/ برخی‌از
۵۰. چنین
۵۱. به‌عنوان
۵۲. اول
۵۳. درباره
۵۴. بسیار
۵۵. درمورد
۵۶. باشد
۵۷. چه
۵۸. شود
۵۹. اگر
۶۰. کلی
۶۱. می‌شوند
۶۲. همین
۶۳. چون
۶۴. جهت
۶۵. زیر
۶۶. زیاد
۶۷. دیگری
۶۸. گردید

۶۹. اما
۷۰. بسیاری/ بسیاری از
۷۱. دوم
۷۲. کمتر
۷۳. تنها
۷۴. وی
۷۵. هریک/ هریک‌از
۷۶. لذا
۷۷. آنچه
۷۸. می‌گردد
۷۹. بوده‌است
۸۰. بلکه
۸۱. روی
۸۲. بالا
۸۳. حتی
۸۴. شده
۸۵. زیرا
۸۶. پس‌از
۸۷. اینکه
۸۸. ولی
۸۹. بدون
۹۰. مستقیم
۹۱. بودند
۹۲. همان
۹۳. همه
۹۴. تمام
۹۵. نه
۹۶. یکدیگر
۹۷. بهتر
۹۸. به‌ترتیب
۹۹. شده‌اند
۱۰۰. درنتیجه
۱۰۱. کم
۱۰۲. می‌توانند
۱۰۳. مشخص
۱۰۴. هم
۱۰۵. بدین
۱۰۶. به‌ویژه
۱۰۷. پایین
۱۰۸. چگونه
۱۰۹. فقط
۱۱۰. البته
۱۱۱. بالاتر
۱۱۲. چهار
۱۱۳. سوم
۱۱۴. چند
۱۱۵. شدند
۱۱۶. آشکار

۱۱۷. زمانی
۱۱۸. علاوه‌بر
۱۱۹. بعضی/ بعضی‌از
۱۲۰. کاملا
۱۲۱. همانطورکه
۱۲۲. فوق
۱۲۳. آیا
۱۲۴. بطوریکه
۱۲۵. می‌باشند
۱۲۶. درخصوص
۱۲۷. ازلحاظ
۱۲۸. به‌وسیله
۱۲۹. بیش‌از
۱۳۰. کل
۱۳۱. هیچ
۱۳۲. برروی
۱۳۳. خارج/ خارج‌از
۱۳۴. بعداز
۱۳۵. ازآنجاکه
۱۳۶. بوده
۱۳۷. مثلا
۱۳۸. پس
۱۳۹. درواقع
۱۴۰. درست
۱۴۱. نبود
۱۴۲. بدین‌ترتیب/ به‌این‌ترتیب
۱۴۳. عالی
۱۴۴. کامل
۱۴۵. عاشقانه
۱۴۶. ...[1]

نتایج بررسی این سوال پژوهشی نشان داد از مجموع ۲۴۸۵۵۲ واژه به کار رفته در مقاله های مورد بررسی در هر سه رشته ۹۷۲۸۰ واژه (۱۲۹۱ واژه بدون احتساب بسامد)، به عنوان واژه‌های غیرمفهومی در سه رشته مورد مطالعه شناخته شدند. از لحاظ نوع دستوری می‌توان بیان‌داشت که قیدها (۴۵/۹٪)، افعال (۱۴/۰۴٪)، حروف ربط (۰۹/۶٪)، حروف اضافه (۷/۷٪)، اعداد (۴/۲۵٪)، ضمایر (۴/۰۱٪) و ادات (۰۷/٪) به ترتیب بیشترین حجم از واژه‌های غیرمفهومی در سه رشته را به خود اختصاص داده اند.

۱. برای مشاهده ادامه این فهرست به نسخه الکترونیکی قرار داده شده در سایت کتابخانه آستان قدس رضوی بخش نشریات مراجعه فرمائید.

۳. واژه‌های غیر مفهومی چه حجمی از متون زبان فارسی هر یک از حوزه‌های مورد مطالعه را تشکیل می‌دهند؟

پس از شناسایی و استخراج فهرست واژه‌های بازدارنده هر یک از مقاله های سه حوزهٔ مورد مطالعه، نسبت واژه‌های بازدارنده هر مقاله به تعداد کل واژه‌های آن مقاله محاسبه گردید که نتایج به دست آمده در جدول شماره ۲ نشان داده شده‌است.

جدول شماره ۲ .درصد واژه‌های بازدارنده هرمقاله درسه رشتهٔ مورد مطالعه

رشته‌های مورد مطالعه	بدون احتساب علایم سجاوندی			با احتساب علایم سجاوندی		
	واژه‌های متن	واژه‌های بازدارنده	درصد واژه‌های بازدازنده	واژه‌های متن	واژه‌های بازدارنده	درصد واژه‌های بازدارنده
رشته علوم تربیتی	۱۲۱۹۶۳	۴۸۷۳۸	۳۹/۹۶	۱۳۵۹۵۶	۶۳۴۶۳	۴۶/۶۷
رشته روانشناسی	۶۳۱۱۲	۲۴۳۴۴	۳۸/۵۷	۷۲۰۳۷	۳۳۳۵۹	۴۶/۳۰
رشته کتابداری	۶۳۴۷۷	۲۴۱۹۸	۳۸/۱۲	۷۲۲۲۴	۳۳۲۳۹	۴۶/۰۲
هر سه رشته	۲۴۸۵۵۲	۹۷۲۸۰	۳۸/۹۴	۳۸۰۲۱۷	۱۳۰۰۶۱	۴۶/۴۱

یافته‌های جدول بالا نشان می‌دهد در رشتهٔ علوم تربیتی ۳۹/۹۶٪ (بدون احتساب علایم سجاوندی)، در رشته روانشناسی ۳۸/۵۷٪ (بدون احتساب علایم سجاوندی) و در رشته کتابداری ۳۸/۱۲٪ از تعداد ۲۴۸۵۵۲ واژه (بدون احتساب علایم سجاوندی) از متون این رشته‌ها را واژه‌های بازدارنده تشکیل می‌دهد.

به‌طورکلی، از تعداد ۲۴۸۵۵۲ واژه (بدون احتساب علایم سجاوندی) که پیکرهٔ زبانی مورد مطالعه را تشکیل می‌دهد، ۹۷۲۸۰ واژه که ۳۸/۹۴٪ کل واژه‌ها را شامل می‌شود، جزء واژه‌های بازدارنده هستند. درصورتی که با احتساب علایم سجاوندی، از ۳۸۰۲۱۷ واژه مورد بررسی، تعداد واژه بازدارنده به ۱۳۰۰۶۱ واژه خواهد رسید که ۴۶/۴۱٪ کل واژه‌ها را تشکیل می‌دهد. به این ترتیب، مشخص می‌شود که علایم سجاوندی حدود ۷٪ از کل یک متن را تشکیل می‌دهند.

٤. بین سیاهه واژه‌های غیرمفهومی انگلیسی و واژه‌های غیرمفهومی به دست آمده در این پژوهش چه وجوه تشابه و تفاوتی وجود دارد؟

همان گونه که از پیشینهٔ پژوهش برمی‌آید، دربارهٔ واژه‌های بازدارنده زبان انگلیسی مطالعات گوناگونی صورت گرفته‌است. فهرست حاصل از مطالعهٔ «فاکس» (۱۹۹۲) و فهرست استاندارد SMART که هر دو نمونه‌هایی از فهرست واژه‌های بازدارنده در حوزه عمومی می‌باشند، بیشتر از فهرستهای دیگر در تحقیقات بعدی مورد استناد قرارگرفته‌است. ازآنجاکه فهرست SMART قابل دستیابی نبود، فهرست عمومی فاکس برای مقایسه بین فهرست واژه‌های بازدارنده انگلیسی و فارسی مورد استفاده قرارگرفت.

مقایسه صورت‌گرفته بین فهرست پژوهش حاضر و فهرست فاکس نشان داد برای ۲۳۱ واژه از ۴۲۱ واژه بازدارنده فهرست فاکس، ۳۶۳ معادل فارسی در فهرست به دست آمده وجود دارد. همان طورکه قبلاً ذکر شد، فهرست واژه‌های بازدارنده حاصل از پژوهش حاضر از میان واژه‌های پیکره زبانی سه رشته علوم تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی استخراج شده‌است؛ درحالی‌که فهرست حاصل از پژوهش فاکس یک پیکره عمومی است. از سوی دیگر، به دلیل اینکه فاکس فهرست خود را براساس بسامد واژه تهیه کرده است، بسیاری از صورتهای مختلف فعلهای انگلیسی و صفتهای ساده، تفضیلی و عالی نیز در فهرست او دیده می‌شود. لیکن پژوهش حاضر چون براساس قواعد دستور زبان تهیه شده است، تنها افعال کمکی و معین را در بین واژه‌های بی‌معنا و کم‌معنا قرار داده است؛ به همین دلیل اغلب واژه‌هایی که در فهرست فاکس فاقد معادل فارسی هستند، جزء گروه افعال می باشند. ضمن اینکه صفتها از این حیث مستثنا بوده و همگی جزء کلیدواژه‌ها محسوب می‌شوند. بیشترین برابری بین واژه‌های فهرست فاکس و فهرست فارسی حاصل از پژوهش حاضر، به حروف ربط، اضافه (حروف اضافه ساده) و قیدهای مختص اختصاص دارد. ضمایر متصل «م، ت، ش، مان، تان، شان» معادل واژه های منفصل Me، His، Her ... می‌باشند که به دلیل اینکه واژه مجزا نیستند، در فهرست واژه‌های بازدارنده فارسی قرار نمی‌گیرند.

۵. توزیع بسامد واژه‌های غیرمفهومی در رشته‌های علوم‌تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی چگونه است؟

نتایج اجرای آزمون کای اسکور بر روی فهرست حاصل از سه حوزهٔ مورد مطالعه، نشان داد از میان ۱۲۹۱ واژه بازدارنده، ۳۷۹ واژه یعنی ۲۹/۳۵٪ از کل واژه‌ها دارای توزیع یکسانی بین سه رشته مورد مطالعه می‌باشند. به عبارت دیگر، در عین اینکه این واژه‌ها در هر سه رشته مورد مطالعه کاربرد دارند، میزان تکرار آنها (بسامد واژه ها) در هر سه رشته تقریباً مشابه است. این واژه‌ها جزء واژه‌های پربسامد در هر سه رشته مورد مطالعه می‌باشند؛ به طوری که واژه‌های «و، در، که، به، از، است، را، این، با و برای» ۱۰ واژهٔ پربسامد در هر سه رشته است.

از سوی دیگر P-Value ی ۲۶۹ واژه یعنی ۲۰/۸۳٪ واژه‌ها، کمتر از 0/05٪ می‌باشد که نشان‌دهنده آن است که این تعداد واژه فاقد توزیع بسامدی یکسان در سه رشته مورد مطالعه می‌باشند. به عبارتی، احتمال رخداد هریک‌از واژه‌های این گروه که از دید دستور زبان فارسی بیشتر از میان حروف اضافه، ربط قیود خاص و افعال پربسامد زبان فارسی می‌باشند، در یکی از سه رشتهٔ مورد مطالعه بیشتر از دو رشته دیگر است؛ به این معنا که احتمال استفاده و کاربرد این واژه در یک رشته خاص بیشتر از رشته‌های دیگر بوده و در آن رشته متداول‌تر است. برای ۶۴۴ واژه باقی مانده، به دلیل اینکه تنها در یکی از سه رشته مورد مطالعه رخ داده بودند، آزمون کای اسکور قابل اجرا نبود. این دسته از واژه‌ها منحصراً مربوط به همان رشته خاص می‌باشند.

نتیجه‌گیری

از آنجا که تهیهٔ فهرست واژه‌های بازدارنده ای که مبتنی بر ساخت زبان مورد مطالعه باشد، مستلزم استخراج معیارهای نحوی و معنایی زبان مورد مطالعه است، ابتدا این معیارها شناسایی و مشخص گردید افعال (معین و همراه شونده)، قیدها، ضمایر، حروف، اصوات، اعداد و علایم سجاوندی به‌عنوان واژه نمایه ها ظاهر نمی شوند. در سایر

پژوهشهای صورت گرفته، بیشتر از شیوهٔ بسامد واژگانی استفاده شده و تنها ابوالخیر (۲۰۰۳) در پایان‌نامه خود فهرستی عمومی از واژه‌های بازدارنده زبان عربی را بر اساس دستور زبان عربی، تهیه نموده است. اما در پژوهشهایی که بر مبنای بسامد واژگانی بوده است نیز برخی از نقشهای دستوری به عنوان واژه های بازدارنده معرفی شده اند که با پژوهش حاضر تناسب دارد. به طوری‌که «لوان» (نقل شده در :نیاکان، ۱۳۸۳)، در پژوهش خود حروف ربط و حروف تعریف را جزء واژه‌های بی‌معنایی می داند که بسامد بالایی دارند. «ساووی» (۱۹۹۹ و ۲۰۰۷) نیز در پژوهشهای خود پس‌از تعیین پربسامدترین واژه‌ها و حذف تمامی اسامی و صفاتی که با موضوعات اصلی پیکره های مورد مطالعه مرتبط بودند، حروف اضافه، ربط، تعریف، ضمایر ملکی، شخصی و اشکال فعل Be را به عنوان واژه‌های پربسامد و بی‌معنا معرفی می کند.

بررسی واژه‌های غیرمفهومی زبان فارسی در سه حوزهٔ علوم تربیتی، روانشناسی و کتابداری و اطلاع‌رسانی مشخص ساخت از بین مجموع ۲۴۸۵۵۲ واژه تشکیل دهنده متن مقاله های مورد بررسی در هر سه رشته، ۹۷۲۸۰ واژه (۱۲۹۱ واژه بدون احتساب بسامد)، به عنوان واژه‌های غیرمفهومی در سه رشتهٔ مورد مطالعه شناخته شدند. با مقایسه نتایج پژوهش حاضر و موارد ذکر شده می‌توان به این مطلب پی‌برد که میزان واژه‌های غیرمفهومی معیّن شده با این روش بسیار بیشتر از فهرستهایی است که در سایر پژوهشها استخراج گردیده‌است و این نتیجه احتمالاً به دلیل تفاوت در شیوهٔ استخراج واژه‌های غیرمفهومی است؛ به نحوی که بیشترین میزان واژه‌های غیرمفهومی در زبان انگلیسی توسط «فرانسیس و کوسرا» مشتمل بر ۴۲۵ واژه و تقریباً ۳۳٪ واژه‌های غیرمفهومی شناسایی شده در مطالعه «تقوا» تنها ناظر بر گروه فعلی است و تعداد آنها نیز بسیار اندک می‌باشد. در پژوهش «پوراسماعیل و رستمی» نیز تنها ۲۰۴ واژه فارسی به‌عنوان واژه بازدارنده استخراج شده است. می‌توان چنین نتیجه گرفت که با استفاده از این شیوه، تعداد واژه‌هایی که به عنوان واژه‌های غیرمفهومی شناسایی می شوند، افزایش می‌یابد.

نتایج نشان داد از بین ۱۰۰ واژه پربسامد در هر رشته، ۶۷ واژه در بین هر سه رشته تکرار شده است و تنها میزان تکرار آنها اندکی متفاوت است. بررسی حجم واژه‌های غیرمفهومی متون زبان فارسی در هر یک از رشته‌های مورد مطالعه نیز نشان داد واژه‌های بازدارنده ۳۸/۹۴٪ کل واژه ها را شامل می شود. «فراکز و بیزا ـ یاتس» (۱۹۹۲) عنوان کردند واژه‌های بازدارنده احتمالا بین ۲۰ تا ۳۰٪ واژه‌های درون یک متن انگلیسی را شامل می‌شود. «سیروتکین و ویلبور» (۱۹۹۲) با اجرای آزمون آماری خود در مجموعه مدارک مورد بررسی تعداد ۲۰۳۰۴۰ واژه موجود در این مدارک را به ۵۰۵۰۸ واژه کاهش دادند. «یمین و ویلبور» (۱۹۹۶) با استفاده از روش ویلبور و سیروتکین نشان دادند با حذف ۸۷٪ واژه‌های بازدارنده در یکی از چهار مجموعهٔ مورد مطالعهٔ خود، صرفه‌جویی ۶۳ درصدی زمان، ۷۴ درصدی حجم فایل مقلوب و در نهایت بهبود دقت بازیابی تا ۱۰٪ حاصل می گردد. «ساووی» (۱۹۹۹) نشان داد با تعیین واژگان غیرمفهومی حجم فایل مقلوب بین ۲۱ تا ۳۵٪ کاهش می‌یابد. ازآنجاکه پژوهش حاضر بر اساس ساختار زبانی و واژگانی زبان فارسی صورت‌گرفته، تعداد واژه‌هایی که می‌توانند به عنوان واژهٔ بازدارنده در نظرگرفته شود، افزایش می‌یابد. از سوی دیگر، می‌توان نتیجه‌گرفت که میزان واژه‌های بازدارنده در متون زبان فارسی بیشتر از متون زبان انگلیسی است. این نتایج نشان می‌دهد میزان حشو و زواید در متون زبان فارسی زیاد است.

پرداختن به وجوه تشابه و تفاوت بین سیاههٔ واژه‌های غیرمفهومی انگلیسی و واژه‌های غیرمفهومی فارسی نیز یکی دیگر از اجزای پژوهش حاضر بود. نتایج به دست آمده نشان داد برای ۲۳۱ واژه از ۴۲۱ واژه بازدارنده فهرست فاکس، تعداد ۳۶۳ معادل فارسی در فهرست به دست آمده وجود دارد. بیشترین برابری بین واژه‌های فهرست فاکس و فهرست فارسی حاصل از پژوهش حاضر به حروف ربط، اضافه (حروف اضافه ساده) و قیدهای مختص اختصاص دارد. ضمایر متصل «م، ت، ش، مان، تان و شان» معادل واژه‌های منفصل ... Her, His, Me است که به دلیل اینکه واژه مجزا نیستند، در فهرست واژه‌های بازدارنده فارسی قرار نمی‌گیرند.

بررسی توزیع بسامد واژه‌های غیرمفهومی در رشته‌های مورد مطالعه نشان داد P-Value ی ۲۶۹ واژه یعنی ۲۰/۸۳٪ واژه‌ها، کمتر از 0/05 ٪ است که به این معناست که این تعداد واژه فاقد توزیع بسامدی یکسان در سه رشته مورد مطالعه می‌باشند. به عبارتی، احتمال رخداد هر یک از واژه‌های این گروه که از دید دستور زبان فارسی بیشتر از میان حروف اضافه، ربط قیود خاص و افعال پربسامد زبان فارسی می‌باشند، در یکی از سه رشته مورد مطالعه بیشتر از دو رشته دیگر است؛ به این معنا که احتمال استفاده و کاربرد این واژه‌ها در یک رشته خاص بیشتر از رشته‌های دیگر بوده و در آن رشته متداول‌تر است. بررسی توزیع بسامدی واژه های غیرمفهومی شناسایی شده نشان داد تعداد قابل توجهی واژه در این فهرست وجود دارد که بسامد پایینی دارند، بنابراین می توان نتیجه گرفت که استفاده از روش بسامد واژگانی در شناسایی واژه های بازدارنده احتمالا نتواند در زبان فارسی کارایی لازم را داشته باشد.

به طور کلی، می‌توان بیان داشت که نتایج به دست‌آمده از پژوهشهای انجام شده در حوزهٔ ذخیره و بازیابی اطلاعات نشان داد واژه‌های بازدارنده به‌عنوان یکی از ضروری ترین بخشها در نمایه‌سازی و چکیده‌نویسی پایگاه های اطلاعاتی، نقش مهمی در کاهش حجم پایگاه های اطلاعاتی و نرم‌افزارهای اطلاع‌رسانی ایفا می‌کنند و سبب تسهیل در امر بازیابی، افزایش میزان مانعیت مدارک بازیابی شده، کاهش زمان جستجو و ... می‌شوند.

منابع

- پور اسماعیل، کیومرث و نسرین رستمی (۱۳۸۴). *لیست کلمات ایست فارسی*. بازیابی شده۲ در آذر ۱۳۸۵ از Ccc.sharif.edu/~shesmail/resources/stopwords:pdf

- داورپناه، محمدرضا و صدیقه بلندیان (۱۳۸۶). تحلیل متن مقالات فارسی و امکان نمایه سازی ماشینی آنها براساس قانون زیف. *فصلنامه پژوهش در مسائل تعلیم و تربیت: ویژه نامه کتابداری و اطلاع رسانی*، دور دوم.

- گیلوری، عباس (۱۳۷۹). نمایه سازی خودکار: گذشته، حال، آینده. *پیام کتابخانه*، ۱۰(۴)، ۲۵- ۱۷.

- نیاکان، شهرزاد (۱۳۸۳). *نمایه سازی ماشینی*. تهران: مرکز اطلاعات و مدارک علمی ایران.

- ویکری، برایان و الینا ویکری (۱۳۸۰). *علم اطلاع رسانی در نظر و عمل*. ترجمه عبدالحسین فرج پهلو، مشهد: انتشارات دانشگاه فردوسی.

- هادسن، گرور(۱۳۸۳). *مباحث ضروری و بنیادین زبانشناسی مقدماتی (ضرورت زبانشناسی مقدماتی)* (علی بهرامی، مترجم). تهران: رهنما.

- وحیدیان کامیار، تقی و غلامرضا عمران (۱۳۸۵). *دستور زبان فارسی (۱)*. تهران: سازمان مطالعه و تدوین کتب علوم انسانی(سمت).

- ناتل خانلری، پرویز (۱۳۵۹). *دستور زبان فارسی* (با تجدیدنظر). تهران: توس.

- نجفی، ابوالحسن(۱۳۸۰). *مبانی زبانشناسی و کاربرد آن در زبان فارسی*. تهران: نیلوفر.

- مشکوة الدینی، مهدی (۱۳۸۲). *دستور زبان فارسی برپایه نظریه گشتاری* (ویرایش۲). مشهد: فاطمی.

- ـــــــــــــــ (۱۳۸۴). *دستور زبان فارسی. واژگان و پیوندهای ساختی*. تهران: سازمان مطالعه و تدوین کتب علوم انسانی (سمت).

- معین، محمد (۱۳۷۸). *فرهنگ فارسی (متوسط): شامل یک مقدمه و سه بخش لغات، ترکیبات خارجی، اعلام ....* تهران: امیرکبیر.

- مرزبان راد، علی (۱۳۷۸). *دستور سودمند*. تهران: دانشگاه صنعتی امیرکبیر.

- محتشمی، بهمن (۱۳۷۰). *دستور کامل زبان فارسی*. تهران: اشراقی.

- صفوی، کورش (۱۳۶۰). *درآمدی بر زبانشناسی*. تهران: بنگاه ترجمه و نشر.

- صهبا، عبدالرشید (۱۳۷۱). *حرفهای ربط، اضافه، نشانه در دستور زبان فارسی برای استفاده دانش آموزان، دانشجویان و پژوهندگان*. تهران: غزل.

- غلامعلی زاده، خسرو (۱۳۷۴). *ساخت زبان فارسی*. تهران: احیاء الکتاب.

- فرشیدورد، خسرو (۱۳۸۲). *دستور مفصل امروز*. تهران: سخن.

- فرشیدورد، خسرو (۱۳۸۶). *دستور برای لغت سازی: فرهنگ پیشوندها و پسوندهای فارسی به همراه گفتارهایی درباره دستور زبان فارسی*. تهران: زوار.

- کلباسی، ایران (۱۳۸۰). *ساخت اشتقاقی در فارسی امروز*. تهران: پژوهشکده علوم انسانی و مطالعات فرهنگی.

- شفاعی، احمد (۱۳۶۳). *مبانی علمی دستور زبان فارسی*. تهران: نوین.

- دهخدا، علی اکبر (۱۳۸۳). *لغتنامه*. (با همکاری محمد معین، جعفر شهیدی). تهران: موسسه لغتنامه دهخدا.

- خطیب رهبر، خلیل (۱۳۷۹). *دستور زبان فارسی: کتاب حرف اضافه و ربط مشتمل بر تعریف و تقسیم و شرح اصطلاحات و معانی و کاربرد حروف*. تهران: مهتاب.

- ــــــــــــ (۱۳۸۱). *دستور زبان فارسی: برای پژوهش دانشجویان و ادب دوستان در آثار شاعران و نویسندگان بزرگ ایران*. تهران: مهتاب.

- بابک، علی (۱۳۸۳). *دستور زبان فارسی پژوهشی معاصر*. مشهد: سخن گستر.

- باطنی، محمدرضا (۱۳۸۲). *توصیف ساختاری دستوری زبان فارسی بر بنیاد یک نظریه عمومی زبان*. تهران: امیر کبیر

- انوری، حسن (۱۳۸۱). *فرهنگ بزرگ سخن*. تهران: سخن.

- انوری، حسن و حسن احمدی گیوی (۱۳۷۷). *دستور زبان فارسی ۲* (ویرایش ۲). تهران: فاطمی

- احمدی گیوی، حسن (۱۳۸۰). *دستور تاریخی فعل*. تهران: قطره.

- Savoy, Jacues(1999). *A stemming procedure and stop word list for general French corpor*a. Journal of the American society for information science; 50(1),p. 944-952.

- Savoy, Jacues(2006). *Searching strateies for the Bulgarian language*. Information retrieval; 10(6), p. 509-529.

- Sirotkin, Karl; Wilbur, W John(1992). The *automatic identification of stop words*. Journal of Information Science; 18 (1) , p.45-55.

- Taghva, Kazem; Bechley, Russel; Sadegh, Mohammad(2003). Alist of farsi stop words. Retrieved November 29, 2006, from: www.isri.unlv.edu/publications/isripub/Taghva2003-01.ps

- Yang, Yiming; Wilbur, John(1996). *Using corpus statistics to remove redundant words in text categorizatio*n. Journal of the American Society for Information Science; 47 (5), p.357-69.

- Lahtinen, T. (2000). *Automatic Indexing: an approach using an index term corpus and combining linguistic and statistical methods*. PhD thesis, University of Helsinki. Retrieved November 29, 2006, from,

- Lazarinis, Fotis(2007). *Engineering and utilizing a stop word list in Greek web*. Journal of the American society for information science and technology;58(11), p. 1645-1652

- Moens, Marie - Francine(2003). *Automation indexing and abstracting of document texts*. Second edition. Massachusetts: Kluwer academic publisher.

- Berg, Criage N.(1997). *DEVELOPING A CORPUS SPECIFIC STOP-LIST USING QUANTITATIVE COMPARISON*. PhD thesis, Graduate school of Logistics and acquisition management, Retrieved November 20, 2006, from, research.airuniv.edu/papers/ay1997/afit/berg_cn.pdf/ ethesis.helsinki.fi/julkaisut/hum/yleis/vk/lahtinen/ - 3k

- zou, Feng; wang, Fu lee; Deng,Xiaotie; Han, Song; Wang, Lusheng (2006). *Stop word list construction in Chinese Langugege Processing*. Retrieved November 20, 2006, from, WWW.utdallas.edu /~fxz ۰۶۳۰۰۰ /- ۱۴k

- Fox, Cristopher(1990). A stop list for general text. Retrieved November 20, 2006, from, www.informatik.uni-trier.de/ley/indice/a-tree.pdf

- Abu-El Khair, Ibrahim Hassan(2003). PhD thesis, University of Pittsburg, Retrieved June 18 , 2007, from, www.mons.edu.eg. pcvs/13702/13102.asp

- Ho. Tin Kam(1999). *Fast identification of stop words for font learning and keyword spotting*. Retrieved November 22, 2006, from http://netlib.bell-labs.com/who/tkh/papers/wordlength.pdf