**نام مقاله: شناسايي واژه ها ي غير مفهومي (رايج) در نمايه سازي خودكار مدارك فارسي**

**نام نشريه: فصلنامه كتابداري و اطلاع رساني (اين نشريه در www.isc.gov.ir نمايه مي شود)**

**شماره نشريه: 48 \_ شماره چهارم،جلد12**

**پديدآور: مجيده سنجي،محمدرضا داورپناه**

**چكيده:**

**پژوهش حاضر با هدف شناسايي واژه­هاي غيرمفهومي در زبان فارسي و تهـيه سياهه­اي از اين واژه­ها براي نمايه­سازي خودكار متنهاي فارسي در رشته­هاي روانشناسي، علوم­تربيتي و كتابداري و اطلاع­رساني انجام شده است. اين پژوهش با روش تحليل محتوا صورت گرفته­است. جامعه آماري اين پژوهش را مقاله‌هاي مندرج در آخرين شماره منتشر­شده در مجله‌هاي علمي و پژوهشي ِ رشته­هاي علـوم­تربيتي، روانشناسي و كتابداري و اطلاع­رساني در سال 1385 تشكيل مي‌دهد. نمونه شامل 63 مقاله است. گردآوري داده­ها با استفاده از تفكيك واژگان به صورت ماشيني و دستي صورت گرفت. نتايج پژوهش نشان داد: 1- افعال ( معين و همراه­شونده)، قيدها، ضماير، حروف، اصوات، اعداد و علائم سجاوندي به­عنوان واژه نمايه­ها ظاهر نمي­شوند، بنابراين، آنها را مي­توان واژه­هاي غيرمفهومي يا به اصطلاح واژه­هاي بازدارنده تلقي كرد. 2- بدون احتساب علائم سجاوندي، در رشتة علوم تربيتي 96/39%، در رشته روانشناسي 57/38% و در رشـــته كتابداري 12/38% از حجم متون را واژه­هاي غيرمفهومي تشكيل مي­دهد. 3- واژه­هاي بازدارندة پربسامد در هرسه حوزه تقريبا مشابه است. 4- از تعداد 248552 واژه (بدون احتساب علائم سجاوندي) كه پيكرة زباني مورد مطالعه را تشكيل مي­داد، 97280 واژه كه 94/38% كل واژه‌ها را شامل مي­شود، جزء واژه­هاي بازدارنده هستند.5- نتيجة مقايسه فهرست فارسي حاصل از پژوهش با فهرست واژه­هاي بازدارنده انگليسي نشان داد بين اين دو فهرست به ميزان 5/28% همپوشاني وجود دارد. 6. همچنين 38/20% از واژه­ها فاقد توزيع بسامدي يكسان در سه رشتة مورد مطالعه مي­باشند.**

**كليدواژه­ها: واژه­هاي غيرمفهومي، واژه­هاي بازدارنده، نمايه­سازي خودكار، ساخت واژگاني زبان فارسي.**

**مقدمه**

نمايه­سازي فرايند تحليل محتواي اطلاعاتي، پيشينه­اي از دانش و بيان­كردن محتواي اطلاعاتي در زبان نمايه­سازي از طريق اصطلاحات نمايه­اي است. به اين ترتيب، نمايه­سازي حداقل سه مرحله دارد:

1.                   انتخاب مفاهيم قابل نمايه در يك مدرك

2.                   بيان­كردن اين مفاهيم در زبان نمايه­سازي ( به صورت واژه نمايه­ها[[1]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn1" \o ") )

3.                   تهيه يك فهرست مرتب­شده از اين شناسه­ها (لاتينن[[2]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn2" \o ")، 2000:ص. 29)

نمايه­سازي يا به­صورت دستي و يا به­صورت ماشيني[[3]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn3" \o ") است. هنگامي­كه اصطلاحات نمايه­اي توسط نمايه­ساز انتخاب شود، نمايه­سازي دستي است و چنانچه همه امور انتخاب كليدواژه، ترتيب­بندي و ... توسط رايانه انجام شود، نمايه­سازي ماشيني است. اكثر روشهاي نمايه­سازي خودكار موجود، براي انتخاب اصطلاحات نمايه­اي از زبان طبيعي[[4]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn4" \o ") استفاده مي‌كنند. در اين شيوه تكواژه­ها و عبارتهاي چندواژه­اي براي انعكاس محتواي متن، مستقيماً از عنوان، چكيده و متن كـامل يك مدرك انتخاب مي­شوند(موئنز[[5]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn5" \o ")،2003: ص. 24).

در مجموع، در نظامهاي نمايه­سازي رايانه­اي سعي نشده­است تا عملكرد ذهني شخص نمايه­ساز تقليد شود. برنامه­ريزي رايانه­اي به منظور گزينش اصطلاحات حايز اهميت از متني با زبان طبيعي، مستلزم اين است كه برنامه درك خوبي از زبانشناسي و دانش لازم در مورد موضوعي كه تحت بررسي است، داشته­باشد. البته، اين امر در حال حاضر براي تمام و يا اكثر نظامهاي بازيابي كار بسيار بزرگي است. در عوض، نمايه­سازي خودكار به روشهايي كه بر فراواني نسبي كلمات در متن مبتني است، تكيه دارد (ويكري و يكري[[6]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn6" \o ")،1380:ص180).

مي­توان مراحل زير را در فرايند نمايه­سازي خودكار درنظر گرفت:

1.    شناسايي واژه­هاي انفرادي از متن كه تحليل واژگان[[7]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn7" \o ") ناميده­مي­شود (موئنز، 2003: ص. 77)

2.    برداشتن واژه­هاي كاربردي و واژه­هاي با بسامد تكرار بالا كه در ارائه محتواي متن بــــي­تأثيرند، با استفاده از فهرست واژه­هاي غيرمجاز.

3.    تبديل واژه­هاي باقي‌مانده به شكل ريشه[[8]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn8" \o ") آنها؛ يعني حذف پسوند­ها يا پيشوندها تا هر كلمه تا حد ريشه­اش كوتاه شود.

4.    محاسبة رايانه­اي بسامد رخداد­هاي ريشه­هايي كه در متن تحليل شده­اند، به منظور تعيين تابع ارزشگذاري هر ريشه.

5.    ريشه­هايي كه نسبت به بعضي ارزشهاي قراردادي آستانه­اي[[9]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn9" \o ")، ارزشگذاري بزرگتري دارند، براي متني كه در آن ظاهر شده، به ­عنوان كليدواژه تعيين مي­شود. در بعضي نظامها، كليدواژه ممكن است ارزشي متناسب با ارزش تابع ارزشگذاري داشته باشد (ويكري و ويكري[[10]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn10" \o ")، 1380: ص. 183).

به هر حال، تعيين واحدهاي متني و مشخص كردن حد و حدود واژه براي ماشين از مسائل اساسي در گزينش اصطلاحات نمـايه­اي در نمـايه­سـازي خودكار است (گيلوري، 1379). به­ علاوه، امكان تشخيص واژه­هاي مفهومي از واژه­‌هاي غيرمفهومي، در فرايند انتخاب اصطلاحات نمايه تأثير بسزايي دارد. آنچه مسلم است، ماشين امكان تشخيص را تنها از طريق تطـبيق واژه­هاي استخراج شده از متن يا منتسب شده به متن با فهرستي كه واژه­هاي غيرمجاز ناميده مي­شـود، به دســـت مي­آورد. در اختيار داشتن فهرستي از اين واژه­ها و ارائه آنها به برنامه رايانه­اي براي ممانعت از ورود آنها به فهرست واژه­هاي مفهومي مطلوب براي نمايه­شدن، يكي از اقدامهاي سودمند در نمايه­سازي خودكار مبتني بر كليدواژه­هاست.

تحليل كلمات يك متن نشان مي­دهد گروهي از كلمات بي­اهميت وجـــــوددارد كه به­فراواني در متن ظاهر مي­شود (مانند يك، به، نه، براي، با، چه كسي، چه موقع، است، آن). گروهي نيز وجود دارد كه بندرت در متن مي­آيند و ممكن است نشان­دهنده محتواي اطلاعاتي متن نباشند (ويكري و ويكري، 1380:ص180). اين دسته از واژه­ها به تنهايي بارمعنايي ندارند (حاوي معنا و مفهوم نيست) بلكه در ارتباط با واژه­هاي ديگر معنا مي­يابند. به اين ترتيب، مفهوم و محتواي متن را نشان نمي‌دهند. از طرف ديگر، بود يا نبود آنها نه تنها در پرسش كاربر تأثيري ندارد، بلكه در ميزان ربط يا عدم ربط مدارك بازيابي شده نيز تأثيري نخواهدداشت. اين واژه­ها با­ عنوان واژه­هاي غيرمجاز براي ورود به نمايه معرفي مي‌شوند. درصورتي­كه واژه­هاي غيرمجاز قبل از فرايند نمايه­سازي مدارك مشخص و فهرست آنها براي كنترل به رايانه داده­شود، علاوه­بر صرفه­جويي در زمان و حجم بايگانيهاي نمايه، به ميزان زيادي از بازيابي مدارك نامرتبط و ريزش كاذب در جستجو جلوگيري خواهدشد (زو و ديگران[[11]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn11" \o ")، 2006).

با گسترش مدارك الكترونيكي فارسي و به تبع آن كاربران فارسي زبان و همچنين ويژگيهاي خاص زبان و خط فارسي، نياز به توجه و بهبود روشهاي نمايه­سازي بيش ­از پيش احساس مي‌شود. زبان فارسي مانند هر زبان ديگري واژه­هايي دارد كه هيچ گونه سهمي در ارائه بار معنايي مدرك ندارند، ولي فهرستي از پيش آماده از اين‌گونه واژه­ها در دسترس نبوده و مشخص نيست چگونه بايد انتخاب شوند. بنابراين، مسئله اساسي اين پژوهش آن است كه: معيارهاي شناسايي واژه­هاي غيرمفهومي در زبان فارسي كدام است؟ و آيا مي­توان سياهه­اي از اين واژه­ها را براي نمايه­سازي خودكار متنهاي فارسي در حوزه روانشــناسي، علوم­تربيتي و كتابداري و اطلاع­رساني تهيه كرد؟

**هدفهاي پژوهش**

اين پژوهش با بهره­گيري از اصول ساختاري زبان فارسي، درپي دستيابي به هدفهاي زير است:

·شناسايي معيارهاي نحوي و معنايي زبان فارسي در تشخيص واژه­هاي غيرمفهومي

·تهية سياهه‌اي از واژه­هاي غيرمفهومي در زبان فارسي

·مقايسه واژه­هاي غيرمفهومي فارسي و انگليسي

·بررسي چگونگي توزيع بسامد واژه­هاي غيرمفهومي فارسي در سه رشتة علوم تربيتي، روانشناســي و كتابداري و اطلاع­رساني.

**سؤالهاي پژوهش**

1.    معيار­هاي نحوي و معنايي شناسايي واژه­هاي غيرمفهومي در زبان فارسي كدام است؟

2.    واژه­هاي غيرمفهومي در زبان فارسي كه بتواند در نمايه­سازي خودكار مدارك فارسي هر يك از حوزه­هاي مورد مطالعه ناديده گرفته­شود، كدام است؟

3.    واژه­هاي غيرمفهومي چه حجمي از متون زبان فارسي هر يك از حوزه‌هاي مورد مطالعه را تشكيل مي­دهد؟

4.    بين سياهة واژه­هاي غيرمفهومي انگليسي و واژه­هاي غيرمفهومي به دست آمده در اين پژوهش، چه وجوه تشابه و تفاوتي وجود دارد؟

5.    توزيع بسامد واژه­­هاي غيرمفهومي در رشــته­هاي علوم­تربيتي، روانشناسي و كتابداري و اطلاع­رساني چگونه است؟

**پيشينة پژوهش**

تاريخچة نمايه­سازي خودكار بر مبناي بسامد تكرار واژه، به دهة 1950 و تحقيقات «لوان»[[12]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn12" \o ")(1975) و «باكسن دال»[[13]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn13" \o ") (1958) بر مي­گردد. پيش فرض كلي در بازيابي اطلاعات اين است كه اصطلاحات با بسامد متوسط، مناسب‌ترين اصطلاحات براي نمايه­سازي هستند. اين فرض اساس كار «زيف»[[14]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn14" \o ") (1949) قرار گرفت.

«فاكس»[[15]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn15" \o ") (1990) از يك شيوة رايج، يعني واژه­هاي پر­بسامد گروهي بزرگ از آثار نوشته شده در همان زبان، استفاده كرد. بسامد واژه­ها در يك مجموعه مواد انگليسي با عنوان پيكره زباني براون در حوزه­هاي موضوعي متنوع تعيين شده­است. مراحل كار فاكس را مي توان در مراحل زير خلاصه كرد:

·   تعيين بسامد واژه­ها در يك مجموعه از مدارك نوشته شده

·محدودكردن فهرست به اندازة مناسب براي استفاده (تعيين نقطه برش. فاكس    300 بار تكرار واژه را انتخاب نمود)

·   حذف واژه هاي بامعني اما با بسامد بالا در سياهة واژه­هاي بازدارندة اوليه.

·   اضافه كردن واژه­هاي پربسامد و كم­اهميتي كه نقطه برش را ازدست­داده­اند.

«فرانسيس و كوسرا»[[16]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn16" \o ") (نقل شده در : گيلوري 1379) در مرحله اول ده كلمه پر بسامد را تعيين و در مرحله دوم فهرستي از 425 كلمه ممنوعه را تهيه كردند . «رايجزبرگن»[[17]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn17" \o ") ( نقل شده در: گيلوري 1379) نيز فهرستي ارائه داد كه از 150 كلمه بازدارنده تشكيل شده بود. فهرست برگمن در سال 1975 منتشر شد.

در سال 1987، در كتابخانة بازمن[[18]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn18" \o ") دانشگاه ايالت مونتانا فهرستي از واژه­هاي بازدارنده نرم و سخت براي فهرست پيوسته آن با هدف جلوگيري از نمايه­سازي آنها و درنتيجه بازيابي آنها تهيه شد. در سال 1992 به دليل تغيير در نرم افزار پايگاه اطلاعاتي و نياز به نمايه­سازي مجدد مدخلهاي آن، بهترين فرصت جهت ارزيابي و بهبــود فهرست واژه­هاي بازدارنده به دست آمد.

«سيروتكين و ويلبور»[[19]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn19" \o ")(1992) با استفاده از اين پيش فرض كه واژه­هاي بازدارنده به عنوان واژه­هايي كه احتمال رخداد يكساني در مدارك مرتبط و نامرتبط با درخواست اطلاعاتي دارند، شناخته مي­شوند، پژوهشي را انجام دادند. آنها بيان مي­كنند كه اين احتمال وجود دارد واژه­هاي بازدارنده در يك مجموعه با استفاده از يك روش بازيابي برداري مبتني بر ضريب تأثير كسينوس مدارك و تشابه مدارك انجام ­گيرد. نتيجه بررسي سيروتكين و ويلبور در مجموعه­اي از مدارك مدلاين (71311 مدرك) در حوزة بيوتكنولوژي كاهـش 203040 تك­واژه در اين مدارك به 50508 ( كاهش 78 درصدي) تك­واژه است.

«يمـين و ويلبور»[[20]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn20" \o ") (1996) با استـفاده از روش ويلبور و سيروتكين (1992) و دو روش طبــقه­بندي آمــاري (Expert Network و LLSF ) براي بازيابــي و دسته­بندي مدارك و يك روش تطابق واژه­محور براي جستجو در پايگاه‌ها ([[21]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn21" \o ") WBM) استفاده نمودند. در اين بررسي، به­ عنوان مثال روش  Expert Network روي مدارك كتابخانه ملي پزشكي آمريكا، با حذف 87% واژه­هاي بازدارنـده، 8002 واژه به 1042 واژه كاهـش يافـت و در نتيـجه باعـث صرفه‌جويي 63 درصدي زمان، 74 درصدي در حجم فايل مقلوب و در نهايت بهبود دقت بازيابي تا 10% گرديد.

«برگ»[[22]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn22" \o ") (1997) با استفاده از روش «ادمونسون و وايلز»[[23]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn23" \o ") (1959) درمورد تعيين اهميت واژه، پژوهشي را در زمينه تعيين واژه‌هاي بازدارنده در يك حوزه موضوعي مشخص بر اساس مجموعه­اي از مدارك نيروي هوايي آمريكا انجام داد.

«ساووي»[[24]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn24" \o ")(1999) براي شناسايي واژه­هاي بازدارنده در زبان فرانسوي از پيكره­اي شامل دو مجموعه آزمايشي استفاده كرد. وي با پيروي از خط مشي فاكس ابتدا200 واژه پربسامد را استخراج كرد؛ سپس با بازبيني فهرست به دست­ آمده، تمام اسمها و صفتهايي كه بسامد بالايي داشتند ولي با موضوعات اصلي پيكره مرتبط بودند، از فهرست حذف شدند. ضمن اينكه بعضي واژه‌هاي غير اطلاعي كه در200 واژه پربسامد ظاهر نشده بودند مانند ضماير ملكي و شخصي، حروف اضافه، ربط و تعريف و بعضي از اشكال فعل Be به اين فهرست ابتدائي اضافه شد. فهرست نهايي شامل215 واژه است و هنگامي كه از چنين فهرستي استفاده مي شود، اندازه فايل مقلوب براي مجموعه آزمايشي اول تا حدود 21% و براي مجموعه آزمايشي دوم حدود 35% كاهش مي يابد.

«هو»[[25]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn25" \o ") (1999) پژوهشي را با استفاده از اين حقيقت زبانشناسي كه بيش از نيمي از واژه­هاي يك نمونه صفحه انگليسي از ميان 150 واژه­بازدارنده رايج هستند، يك شيوة سريع براي تعيين محل واژه­هاي بازدارنده ارائـه مي­دهد. اين روش از سنجش دامنه واژه‌هاي انفرادي و واژه­هاي همجوار آنها استفاده مي­كند. در يك آزمايش با استفاده از 400 تصوير از صفحات، اين روش 63% از واژه­هاي زايد متن را حذف كرد. وي بيان مي‌كند تمايز بين واژة بازدارنده و غيربازدارنده اغلب به تشــخيص كليدواژه كمك مي‌كند.

«ساووي» 2001، 2002 و 2007) براساس كار قبلي خود فهرستي از واژه­هاي بازدارنده را براي زبانهاي فنلاندي، ايتاليايي، آلماني، اسپانيايي و بلغاري شناسايي نمود. با استفاده از همين شيوه، «ساووي و راسولوفو»[[26]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn26" \o ") (2003) فهرست واژه­هاي بازدارنده در زبان عربي را نيز ايجاد كردند. فهرست ايجاد شده مبتني بر پيكره زبان عربي است كه توسط ديويد كرافت و كوين والكرد در كنسرسيوم اطلاعات زبانشناسي در فيلادلفيا شكل­گرفته و دربرگيرندة 383/872 مدرك (76 ميليون علامت) حاوي بيش از 094/666 تك واژه است.

«ابوالخير» در پايان­نامه دكتري خود با عنوان «اثربخشي روشهاي پردازش متن براي بازيابي زبان عربي» به فهرســتي از واژه­هاي بازدارنده نيازداشت. به اين منظور، با استفاده از سه شيوة متفاوت: 1- مبتني بر ساختار و ويژگيهاي زبان عربي 2- مبتني بر آمارهاي پيكره زباني و 3- تركيبي از اين دو روش، فهرستي از واژه­هاي بازدارنده زبان عربي را تهيه نمود.

در پژوهشي ديگر، «زو و ديگران»[[27]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn27" \o ") (2006) براي استخراج واژه­هاي بازدارنده در زبان چيني از يك روش خودكار يكپارچه مبتني بر الگوهاي اطلاعاتي و آماري استفاده كردند. الگوي آماري بر اساس احتمال و توزيع واژه و الگوي اطلاعاتي بر اساس اهميت واژه با استفاده از نظريه اطلاعـات عمـل مي­كند. در انـتها نتـايج به دست­آمـده از اين دو الگـو يكـپارچه مي­شوند.

«لازارينيس»[[28]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn28" \o ") (2007) مقاله اي را با هدف پردازش ساختمان فهرستي از واژه­هاي بازدارنده براي زبانهاي غير لاتين و ارزيابي تأثير حذف اين واژه ها از پرسش كاربران ارائه داد. براي انجام اين پژوهش، مجموعه اي از 32 پرسش موثق و صحيح كه توسط كاربران پيشنهاد­شده­بود، يك نوبت با درنـــظرگرفتن واژه­هاي بازدارنده و نوبت بعد بدون اين واژه­ها به موتور كاوش گوگل داده­ شد. به اين ترتيب، اهميت حذف واژه­هاي بازدارنده از عبارت پرسش بر حسب ربط در 10 نتيجه برتر گوگل ارزيابي شده­است.

اما بررسي متون و مرور نوشتار نشان مي­دهد در مورد واژه­هاي بازدارنــده در زبان فارسي كارهاي معدود و پراكنده­اي صورت پذيرفته است كه در ادامه به آن پرداخته خواهد شد.

«تقوا، بكلي و سده»[[29]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn29" \o ") (2003) مجموعه­اي متشكل از 1850 مدرك در يك دورة    6 ماهه از ميان وب­سايتهاي فارسي با حوزه­هاي موضوعي متنوع كه عمدتاً نسخــه الكترونيكي روزنامه­ها و مجـــله هاي مشهور ايراني بــودند، و يا وب­سايتهاي فارسي طراحي­شده در آمريكا را جمع­آوري­كردند. آنها فهرست خود را بر اساس پراكندگي واژه­ها تهيه نمودند. در فهرست نهايي 12 فعل وجود داشت كه «تقوا و همكارانش» به جاي فهرست­كردن همه وجوه و زمانهاي اين افعال، تنها زمان گذشته و حال بن فعل را در فهرست خود وارد كردند.

«پور اسماعيل و رستمي» (1384 ) ابتدا فهرست تقوا، بكلي و سده(2003) را براي صيغه­ها و زمانهاي مختلف كامل كردند. اين فهرست شامل 204 فعل فارسي است. سپس با كمك مجموعه آزمون استاندارد محك كه بالغ بر 3000 مستند را شامل مي­شود، بسامد كلمات مختلف را محاسبه نمودند و از ميان آنها در مرحله مقدماتي كلماتي را كه بسامدي بيشتر از 1000 بار داشته اند انتخاب و سپس فهرست نهايي را كه شامل 34 واژه مي­شود، استخراج كردند.

«داورپناه و بلنديان» (1386) پژوهشي را با موضوع نمايه­سازي ماشيني متون فارسي براساس قانون زيف انجام دادند. نتايج نشان داد توزيع فراواني واژگان در متون فارسي داراي الگوي پيش­بيني­پذير است. كاربرد واژه­هاي با بسامد بالا و بسامد پايين درمقاله هاي فارسي، از قانون زيف پيروي مي­كند. همچنين، مشخص شد بسامد واژگاني مي­تواند به عنوان معياري براي نمايه­سازي ماشيني متون فارسي در نظرگرفته ­شود. وضعيت همخواني كامل بين بسامد واژگاني و كليدواژه­هاي موضوعي در شيوة تفكيك صرفا ماشيني بدون دخالت عامل انساني به طور متوسط در كل مقاله هاي مورد بررسي به ميزان 50/21% است. در شيوة تفكيك ماشيني با دخالت عامل انساني، ميزان همخواني به 52% مي­رسد. وضعيت همخواني كامل بسامد واژگاني با كليدواژه­هاي عنواني در شيوة صرفاً ماشيني بدون دخالت عامل انساني، به طور متوسط در كل، مقاله هاي مورد بررسي 20/9% است كه در شيوة ماشيني با دخالت عامل انساني اين ميزان بيشتر از 5 برابر شده و به 14/54% مي­رسد.

**روش پژوهش، جامعه آماري و حجم نمونه**

اين پژوهش با استفاده از روش تحليل محتوا انجام پذيرفت**.** جامعه آماري اين پژوهش، مقاله هاي مندرج در آخرين شماره منتشر شده در مجله هاي علمي و پژوهشي در رشته­هاي علــــوم­تربيتي، روانشناسي و كتابداري و اطلاع­رساني در سال 1385 است. اين مجله ها از فهرست مجله هاي مورد تأييد وزارت علوم، تحقيقات و فناوري در همين سال شناسايي شد كه به شرح ذيل است:

·پژوهش در مسائل تعليم و تربيت/ انجمن ايراني تعليم و تربيت

·نوآوريهاي آموزشي/ وزارت آموزش و پرورش

·آموزش عالي ايران/ انجمن آموزش عالي ايران

·پژوهش و برنامه­ريزي در آموزش عالي/ مؤسسه پژوهش و برنامه­ريزي آموزش عالي

·روانشناسي و علوم­تربيتي/ دانشگاه تهران

·روانشناسي و علوم­تربيتي/ دانشگاه تبريز

·علوم تربيتي و روانشناسي/ دانشگاه شهيد چمران

·مطالعات تربيتي و روانشناسي/ دانشگاه فردوسي مشهد

·پژوهشهاي روانشناختي/ رضا زماني (بخش خصوصي)

·تازه­ها و پژوهشهاي مشاوره/ انجمن مشاوره ايران

·روانشناسي/ انجمن ايراني روانشناسي

·كتابداري و اطلاع­رساني/ كتابخانه مركزي و مركز اسناد آستان قدس رضوي

با استفاده از جدول تعيين حجم نمونه مورگان، براي 73 عنوان مقاله (كتابداري    20 عنوان، روانشناسي 22 عنوان و علوم­تربيتي 31 عنوان مقاله) حجم نمونه پژوهش         63 عنوان مقاله است؛ و از آنجا كه تعداد مقاله هاي سه رشته با هم برابر نبود، تعداد مقاله‌هاي هر رشته براساس فرمول زير محاسبه و تعيين شد:

|  |  |
| --- | --- |
| تعداد مقاله هاي هر رشته در نمونه=  حجم نمونه× | حجم جامعه آن طبقه |
| حجم كل جامعه |

با روش نمونه­گيري تصادفي طبقه­اي، در رشته كتابداري 17 عنوان مقاله، رشته روانشناسي 19عنوان و در رشته علوم­تربيتي 27 عنوان مقاله انتخاب شد.

**گردآوري داده­ها**

براي تهيه سياهة واژه­هاي غيرمجاز از متون مورد مطالعه، اولين گام، تفكيك واژگان اين متون بود. براي رسيدن به اين هدف، مراحل زير انجام پذيرفت:

**1. تهيه متن الكترونيكي**

در اولين گام نسخه الكترونيكي مقاله هاي منتشرشده در آخرين شماره مجله‌هاي علمي و پژوهشي در سه رشته مورد مطالعه در سال 1385 در محيط نرم­افزاري Word كه امكان تفكيك واژگان متن در آن وجود دارد، تهيه شد.

**2. تفكيك واژگان**

براي استفاده از اصول به دست آمده و استخراج واژه­هاي غيرمفهومي از ميان ديگر واژه­ها، واژگان هر يك از مقاله هاي حجم نمونه (63 مقاله) تفكيك شد. تفكيك واژگان متن هر يك از مقاله ها به صورت ماشيني و با استفاد از فرامين موجود در نرم افزار Word صورت پذيرفت. سپس واژگان تفكيك­ شده هر متن براساس معيارهاي زبانشناسي، قواعد دستوري و آيين نگارش فارسي ازلحاظ نوع و بار معنايي به ­صورت دستي بررسي و ويرايش گرديد. به اين ترتيب، در تايپ مجدد متن مقاله ها و در تفكيك واژگان، معيارهاي زير مورد استفاده قرارگرفت:

·همه صيغه­ها، وجوه و انواع فعل، به­صورت يك واژه ← فراهم شده­است، رفته بودم

·افعال مركب به­صورت يك واژه ← بناكرد

·افعال پيشوندي به­صــورت يـك واژه ← ازدســـت­داده­است

·مصدرهاي مركب به­صورت يك واژه ← پيش رفتن

·اسامي مركب به­صورت يك واژه ← آيين نامه

·اسامي پيشوندي به­صورت يك واژه ← بي­گناه ، به سرعت

·اسامي ميانوندي به­صورت يك واژه ← خودبه­خود

·پاره­هاي غيرمستقل و واژه­هاي تركيبي به­صورت يك واژه ← همكار

·عبارتهايي كه به­عنوان گروه اسمي، گروه قيدي، گروه حرف اضافه شناخته مي‌شوند، چون براساس مفهومي كه حامل آن هستند غيرقابل تفـكـيـك مي­باشند، به‌ صورت يك واژه ← محمدحسين دياني، شناي صدمتر

·نام­آواها به­صورت يك واژه ← جيك­جيك ، وزوز

·فاصله درمورد افعال حذف مي­شوند ← مي­گيرد

·فاصله درمورد علامت جمع (ها، هاي، هايي) حذف شده و شمارش نمي­شوند ← ماشين­ها

·فاصله درمورد تكواژهاي صرفي ( تر، ترين) حذف شده و شمارش نمي­شوند ← خوشبخت ترين

·آيه­هاي قرآن و واژه­هاي انگليسي (خارجي) درصورت وجود به دليل غيرفارسي بودنشان حذف مي­شوند.

·اختصارات حذف­شده و شمارش نمي­شوند ← ج. Cm   ( داورپناه و بلنديان، 1386)

**3.شمارش بسامد واژگان**

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

**يافته­هاي پژوهش**

با استفاده از داده­هاي گردآوري شده به سؤالهاي پژوهش پاسخ داده شد كه توضيح آن به شرح زير است:

**1. معيار­هاي نحوي و معنايي شناسايي واژه­هاي غيرمفهومي در زبان فارسي كدام است؟**

با مطالعه متون و كتابهاي مربوط به حوزه زبانشناسي[[30]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn30" \o ")، دستور زبان فارسي[[31]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn31" \o ") و متون مربوط به تهيه و تدوين اصطلاحنــامه­ها ساخت واژگاني زبان فارسي مورد مطالعه قرارگرفت؛ سپس اصول و قواعدي مشخص و مستدل استـــخراج شد كه با استناد به آنها، شناسايي و استخراج واژه­هاي كم­معنا يا بدون معنا در زبان فارسي امكان پذير خواهد بود. اين اصول و قواعد عبارتند از:

§به كوچكترين واحد معنادار كه در ساخت واژه مشخص مي گردد، تكواژ گفته مي­شود.

§تكواژها از ديد كم ­و ­بيشي در تعداد بسامد (كاربرد) به دو گروه محدود يا بسته و نامحدود يا باز تقسيم مي­شوند.

§تكواژهاي زبان از نظر ايفاي نقش به دو گروه تكواژهاي قاموسي و تكواژهاي دستوري تقسيم مي­شوند.

§تكواژهاي قاموسي معناي مستقلي داشته و بر اشيا، اعمال و كيفيات خاص كه قابل حس و لمس و درك هستند دلالت دارند. تعداد اجزا و آحاد اين گروه از واژه­ها، ثابت، معيّن و محدود نيست و فهرست آنها در زبان باز است.

§تكواژهاي دستوري اغلب به تنهايي به­كار نمي­روند (كاركرد دستوري دارند) و معناي آنها با پيوستن به تكواژهاي ديگر آشكار مي­شود. اين گروه داراي شمار معيّن و ثابتي از اعضا و اجزا هستند. فهرست اين تكواژها بسته و محدود است.

§ هرچه تعداد آحاد و تكواژها بيشتر باشد، بسامد آنها كمتر مي­شود.

§هرچه تعداد آحادِ تكواژها كمتر باشد، بسامـد آنها (يعني ميزان كاربرد آنها در جمله هاي مختلف) بيشتر مي­شود.

§ تكواژي كه متعلق به گروه محدود و بسامد آن بالا باشد، تكواژ قاموسي است.

§تكواژهاي دستوري شامل ضماير، قيود، حروف، اصوات، اعداد و افعال معين مي­باشند.

در پايان، بر اساس معيارهاي ذكر شده در بالا، مي­توان چنين استنباط نمود كه، افعال (معين و همراه شونده)، قيدها، ضماير، حروف، اصوات، اعداد و علايم سجاوندي به عنوان واژه نمايه ها ظاهر نمي شوند. اين قواعد مبنايي را براي شناسايي و تهيه فهرست واژه‌هاي بازدارنده در زبان فارسي فراهم­مي­كند.

**2.    واژه­هاي غيرمفهومي در زبان فارسي كه بتواند در نمايه­سازي خودكار مدارك فارسي هر يك از حوزه هاي مورد مطالعه ناديده گرفته­شود، كدام است؟**

به منظور شناسايي واژه هاي غيرمفهومي با توجه به معيارهاي استخراج شده ذيل سؤال اول پژوهش، ابتدا نوع دستوري واژگان بررسي شد. براي تعيين نوع دستوري واژه ها از فرهنگهاي لغت فارسي به فارسي ـ لغتنامه دهخدا، فرهنگ معين و فرهنگ سخن ـ استفاده شد. در تعيين نو ع دستوري واژه هايي كه در اين سه فرهنگ وجود نداشت، از كتابهاي دستور زبان فارسي و مشورت با صاحب نظران استفاده گرديد. چون برخي از واژه­ها داراي چندين نقش دستوري هستند، ملاك ما در انتخاب واژه بازدارنده آن نوع دستوري از واژه بود كه براساس اصول استخراج شده در سؤال اول پژوهش، جزء واژه­هاي غيرمفهومي زبان فارسي قرار مي­گيرند. به اين ترتيب، نوع دستوري تك­تك 248552 واژه تشكيل دهندة متون مورد مطالعه مشخص شد. فهرست درهم­كرد اين واژه­ها به جهت كوتاه تر شدن، بدون ذكر نوع دستوري، براساس بسامد واژه­ها به ترتيب از بيشترين به كمترين ميزان تكرار در جدول شماره 1 ارائه شده است. [جدول شماره 1. فهرست درهم­كرد واژه­هاي بازدارنده سه رشته مورد مطالعه](http://www.aqlibrary.org/UserFiles/File/48.pdf)

1.    و

2.    در

3.    به

4.    كه

5.    از

6.    اين

7.    را

8.    است

9.    با

10.              براي

11.              آن

12.              خود

13.              نيز

14.              آنها

15.              بر

16.              يا

17.              بين

18.              يك

19.              مي‌شود

20.              دو

21.              بود

22.              تا

23.              دارد

24.              ديگر

25.              شد

26.              شده­است

27.              هر

28.              هستند

29.              دارند

30.              مي‌باشد

31.              بنابراين

32.              بايد

33.              براساس

34.              آنان

35.              همچنين

36.              بيشتر

37.              يكي/ يكي­از

38.              ميان

39.              نسبت­­به

40.              يعني

41.              ما

42.              مي‌تواند

43.              مي‌توان

44.              سه

45.              نيست

46.              به­صورت

47.              يك

48.              ازنظر

49.              برخي/ برخي­از

50.              چنين

51.              به­عنوان

52.              اول

53.              درباره

54.              بسيار

55.              درمورد

56.              باشد

57.              چه

58.              شود

59.              اگر

60.              كلى

61.              مي‌شوند

62.              همين

63.              چون

64.              جهت

65.              زير

66.              زياد

67.              ديگري

68.              گرديد

69.              اما

70.              بسيارى/ بسيارى از

71.              دوم

72.              كمتر

73.              تنها

74.              وي

75.              هريك/ هريك­از

76.              لذا

77.              آنچه

78.              مي‌گردد

79.              بوده­است

80.              بلكه

81.              روي

82.              بالا

83.              حتي

84.              شده

85.              زيرا

86.              پس­از

87.              اينكه

88.              ولى

89.              بدون

90.              مستقيم

91.              بودند

92.              همان

93.              همه

94.              تمام

95.              نه

96.              يكديگر

97.              بهتر

98.              به­ترتيب

99.              شده‌اند

100.         درنتيجه

101.         كم

102.         مي‌توانند

103.         مشخص

104.         هم

105.         بدين

106.         به­ويژه

107.         پايين

108.         چگونه

109.         فقط

110.         البته

111.         بالاتر

112.         چهار

113.         سوم

114.         چند

115.         شدند

116.         آشكار

117.         زماني

118.         علاوه­بر

119.         بعضي/ بعضي­از

120.         كاملا

121.         همانطوركه

122.         فوق

123.         آيا

124.         بطوريكه

125.         مي‌باشند

126.         درخصوص

127.         ازلحاظ

128.         به­وسيله

129.         بيش­از

130.         كل

131.         هيچ

132.         برروي

133.         خارج/ خارج­از

134.         بعداز

135.         ازآنجاكه

136.         بوده

137.         مثلا

138.         پس

139.         درواقع

140.         درست

141.         نبود

142.         بدين­ترتيب/ به­اين­ترتيب

143.         عالي

144.         كامل

145.         عاشقانه

146.         ...1

[[32]](http://www.aqlibrary.org/modules/FCKEditor/pnincludes/editor/fckeditor.html?InstanceName=desc&Toolbar=Default" \l "_ftn32" \o ")نتايج بررسي اين سوال پژوهشي نشان­ داد از مجموع 248552 واژه به كار رفته در مقاله هاي مورد بررسي در هر سه رشته 97280 واژه (1291 واژه بدون احتساب بسامد)، به عنوان واژه­هاي غيرمفهومي در سه رشته مورد مطالعه شناخته شدند. از لحاظ نوع دستوري مي­توان بيان­داشت كه قيدها (9/45%)، افعال (04/14%)، حروف ربط (6/09%)، حروف اضافه (7/7%)، اعداد (25/4%)، ضماير (01/4%) و ادات (07/%) به ­ترتيب بيشترين حجم از واژه­هاي غيرمفهومي در سه رشته را به خود اختصاص داده اند.

**3. واژه­هاي غير مفهومي چه حجمي از متون زبان فارسي هر يك از حوزه‌هاي مورد مطالعه را تشكيل مي­دهند؟**

پس از شناسايي و استخراج فهرست واژه­هاي بازدارنده هر يك از مقاله هاي سه حوزة مورد مطالعه، نسبت واژه­هاي بازدارنده هر مقاله به تعداد كل واژه­هاي آن مقاله محاسبه گرديد كه نتايج به دست ­آمده در جدول شماره 2 نشان داده شده­است.

**جدول شماره 2 .درصد واژه­هاي بازدارنده هرمقاله درسه رشتة مورد مطالعه**

|  |  |  |
| --- | --- | --- |
| **رشته­هاي مورد مطالعه** | **بدون احتساب علايم سجاوندي** | **با احتساب علايم سجاوندي** |
| **واژه‌هاي متن** | **واژه‌هاي****بازدارنده** | **درصد واژه‌هاي بازدازنده** | **واژه‌هاي متن** | **واژه‌هاي بازدارنده** | **درصد واژه‌هاي بازدارنده** |
| رشته علوم تربيتي | 121963 | 48738 | 96/39 | 135956 | 63463 | 67/46 |
| رشته روانشناسي | 63112 | 24344 | 57/38 | 72037 | 33359 | 30/46 |
| رشته كتابداري | 63477 | 24198 | 12/38 | 72224 | 33239 | 02/46 |
| هر سه رشته | 248552 | 97280 | 94/38 | 380217 | 130061 | 41/46 |

يافته­هاي جدول بالا نشان مي­دهد در رشتة علوم تربيتي 96/39% (بدون احتساب علايم سجاوندي)، در رشته روانشناسي 57/38% (بدون احتساب علايم سجاوندي) و در رشته كتابداري 12/38% از تعداد 248552 واژه (بدون احتساب علايم سجاوندي) از متون اين رشته­ها را واژه­هاي بازدارنده تشكيل مي­دهد.

به­طوركلي، از تعداد 248552 واژه ( بدون احتساب علايم سجاوندي) كه پيكرة زباني مورد مطالعه را تشكيل مي­دهد، 97280 واژه كه 94/38% كل واژه‌ها را شامل مي‌شود، جزء واژه­هاي بازدارنده هستند. درصورتي كه با احتساب علايم سجاوندي، از 380217 واژه مورد بررسي، تعداد واژه بازدارنده به 130061 واژه خواهد رسيد كه     41/46% كل واژه­ها را تشكيل مي­دهد. به اين ترتيب، مشخص مي­شود كه علايم سجاوندي حدود 7% از كل يك متن را تشكيل مي­دهند.

**4. بين سياهه واژه­هاي غيرمفهومي انگليسي و واژه­هاي غيرمفهومي به دست آمده در اين پژوهش چه وجوه تشابه و تفاوتي وجود دارد؟**

همان گونه كه از پيشينة پژوهش برمي­آيد، دربارة واژ­ه­هاي بازدارنده زبان انگليسي مطالــعات گوناگوني صـورت گرفته­است. فهرست حاصل از مطالعة «فاكس» (1992) و فهــرست استاندارد SMART كه هر دو نمـونه­هايي از فهرست واژه­هاي بازدارنده در حوزه عمومي مي­باشند، بيشتر از فهرستهاي ديگر در تحقيقات بعدي مورد استناد قرارگرفته­است. ازآنجاكه فهرست SMART قابل دستيابي نبود، فهرست عمومي فاكس براي مقايسه بين فهرست واژه­هاي بازدارنده انگليسي و فارسي مورد استفاده قرارگرفت.

مقايسه صورت­گرفته بين فهرست پژوهش حاضر و فهرست فاكس نشان داد براي 231 واژه از 421 واژه بازدارنده فهرست فاكس، 363 معادل فارسي در فهرست به دست آمده وجود دارد. همان طوركه قبلاً ذكر شد، فهرست واژه­هاي بازدارنده حاصل از پژوهش حاضر از ميان واژه­هاي پيكره زباني سه رشته علوم تربيتي، روانشناسي و كتابداري و اطلاع­رساني استخراج شده­است؛ درحالي­كه فهرست حاصل از پژوهش فاكس يك پيكره عمومي است. از سوي ديگر، به دليل اينكه فاكس فهرست خود را براساس بسامد واژه تهيه كرده است، بسياري از صورتهاي مختلف فعلهاي انگليسي و صفتهاي ساده، تفضيلي و عالي نيز در فهرست او ديده مي­شود. ليكن پژوهش حاضر چون براساس قواعد دستور زبان تهيه شده است، تنها افعال كمكي و معين را در بين واژه­هاي بي­معنا و كم­معنا قرار داده ­است؛ به همين دليل اغــلب واژه­هايي كه در فهرست فاكس فاقد معادل فارسي هستند، جزء گروه افعال مي باشند. ضمن اينكه صفتها از اين حيث مستثنا بوده و همگي جزء كليدواژه‌ها محسوب مي‌شوند. بيشترين برابري بين واژه‌هاي فهرست فاكس و فهرست فارسي حاصل از پژوهش حاضر، به حروف ربط، اضافه (حروف اضافه ساده) و قيدهاي مختص اختصاص دارد. ضماير متصل «م، ت، ش، مان، تان، شان» معادل واژه هاي منفصل ... Me، His، Her مي‌باشند كه به دليل اينكه واژه مجزا نيستند، در فهرست واژه‌هاي بازدارنده فارسي قرار نمي‌گيرند.

**5. توزيع بسامد واژه­­هاي غيرمفهومي در رشــته­هاي علوم­تربيتي، روانشناسي و كتابداري و اطلاع­رساني چگونه است؟**

نتايج اجراي آزمون كاي اسكور بر روي فهرست حاصل از سه حوزة مورد مطالعه، نشان داد از ميان 1291 واژه بازدارنده، 379 واژه يعني 35/29% از كل واژه‌ها داراي توزيع يكساني بين سه رشته مورد مطالعه مي­باشند. به عبارت ديگر، در عين ­اينكه اين واژه­ها در هر سه رشته مورد مطالعه كاربرد دارند، ميزان تكرار آنها (بسامد واژه ها) در هر سه رشته تقريباً مشابه است. اين واژه­ها جزء واژه­هاي پربسامد در هر سه رشته مورد مطالعه مي‌باشند؛ به طوري كه واژه­هاي «و، در، كه، به، از، است، را، اين، با و براي» 10 واژة پربسامد در هر سه رشته است.

**ا**ز سوي ديگرP-Value  ي 269 واژه يعني 83/20% واژه­ها، كمتر از 0/05% مي‌باشد كه نشان­دهنده آن است كه اين تعداد واژه فاقد توزيع بسامدي يكسان در سه رشته مورد مطالعه مي­باشند. به عبارتي، احتمال رخداد هريك­از واژه­هاي اين گروه كه از ديد دستور زبان فارسي بيشتر از ميان حروف اضافه، ربط قيود خاص و افعال پربسامد زبان فارسي مي­باشند، در يكي از سه رشتة مورد مطالعه بيشتر از دو رشته ديگر است؛ به اين معنا كه احتمال استفاده و كاربرد اين واژه در يك رشته خاص بيشتر از رشته­هاي ديگر بوده و در آن رشته متداول­تر است. براي 644 واژه باقي مانده، به دليل اينكه تنها در يكي از سه رشته مورد مطالعه رخ داده بودند، آزمون كاي اسكور قابل اجرا نبود. اين دسته از واژه­ها منحصراً مربوط به همان رشته خاص مي­باشند.

**نتيجه­گيري**

از آنجا كه تهية فهرست واژه­هاي بازدارنده ­اي كه مبتني بر ساخت زبان مورد مطالعه باشد، مستلزم استخراج معيارهاي نحوي و معنايي زبان مورد مطالعه است، ابتدا اين معيارها شناسايي و مشخص گرديد افعال (معين و همراه شونده)، قيدها، ضماير، حروف، اصوات، اعداد و علايم سجاوندي به­عنوان واژه نمايه ها ظاهر نمي شوند. در ساير پژوهشهاي صورت گرفته، بيشتر از شيوة بسامد واژگاني استفاده شده و تنها ابوالخير (2003) در پايان­نامه خود فهرستي عمومي از واژه­هاي بازدارنده زبان عربي را بر اساس دستور زبان عربي، تهيه نموده است. اما در پژوهشهايي كه بر مبناي بسامد واژگاني بوده است نيز برخي از نقشهاي دستوري به عنوان واژه هاي بازدارنده معرفي شده اند كه با پژوهش حاضر تناسب دارد. به طوري­كه «لوان» (نقل شده در :نياكان، 1383)، در پژوهش خود حروف ربط و حروف تعريف را جزء واژه­هاي بي­معنايي مي داند كه بسامد بالايي دارند. «ساووي» (1999 و 2007) نيز در پژوهشهاي خود پس­از تعيين پربسامدترين واژه­ها و حذف تمامي اسامي و صفاتي كه با موضوعات اصلي پيكره هاي مورد مطالعه مرتبط بودند، حروف اضافه، ربط، تعريف، ضماير ملكي، شخصي و اشكال فعل Be را به عنوان واژه­هاي پربــسامد و بي­معنا معرفي مي­­­­­­­­­­­­­كند.

بررسي واژه­هاي غيرمفهـومي زبان فارسي در سه حوزة علوم تربيـتي، روانشــناسي و كتابـداري و اطلاع­رساني مشخص ساخت از بين مجموع 248552 واژه تشكيل دهنده متن مقاله هاي مورد بررسي در هر سه رشته، 97280 واژه (1291 واژه بدون احتساب بسامد)، به عنوان واژه­هاي غيرمفهومي در سه رشتة مورد مطالعه شناخته شدند. با مقايسه نتايج پژوهش حاضر و موارد ذكر شده مي­توان به اين مطلب پي­برد كه ميزان واژه­هاي غيرمفهومي معيّن شده با اين روش بسيار بيشتر از فهرستهايي است كه در ساير پژوهشها استخراج گرديده­است و اين نتيجه احتمالاً به دليل تفاوت در شيوة استخراج واژه­هاي غيرمفهومي است؛ به نحوي كه بيشترين ميزان واژه­هاي غيرمفهومي در زبان انگليسي توسط «فرانسيس و كوسرا» مشتمل بر 425 واژه و تقريباً 33% واژه­هاي غيرمفهومي شناسايي شده در مطالعه «تقوا» تنها ناظر بر گروه فعلي است و تعداد آنها نيز بسيار اندك مي­باشد. در پژوهش «پوراسماعيل و رستمي» نيز تنها 204 واژه فارسي به­عنوان واژه بازدارنده استخراج شده است. مي­توان چنين نتيجه گرفت كه با استفاده­ از اين شيوه، تعداد واژه­هايي كه به عنوان واژه­هاي غيرمفهومي شناسايي مي شوند، افزايش مي­يابد.

نتايج نشان داد از بين 100 واژه پربسامد در هر رشته، 67 واژه در بين هر سه رشته تكرار شده است و تنها ميزان تكرار آنها اندكي متفاوت است. بررسي حجم واژه­هاي غيرمفهومي متون زبان فارسي در هر يك­ از رشته­هاي مورد مطالعه نيز نشان داد واژه­هاي بازدارنده 94/38% كل واژه ها را شامل مي شود. «فراكز و بيزا ـ ياتس» (1992) عنوان كردند واژه­هاي بازدارنده احتمالا بين 20 تا 30% واژه­هاي درون يك متن انگليسي را شامل مي­شود. «سيروتكين و ويلبور»(1992) با اجراي آزمون آماري خود در مجموعه مدارك مورد بررسي تعداد 203040 واژه موجود در اين مدارك را به 50508 واژه كاهش دادند.«يمـين و ويلبور» (1996) با استـفاده ­از روش ويلبور و سيروتكين نشان دادند با حذف    87% واژه­هاي بازدارنده در يكي از چهار مجموعة مورد مطالعة خود، صرفه­جويي            63 درصدي زمان، 74 درصدي حجم فايل مقلوب و در نهايت بهبود دقت بازبابي تا      10% حاصل مي‌گردد. «ساووي»(1999) نشان داد با تعيين واژگان غيرمفهومي حجم فايل مقلوب بين 21 تا 35% كاهش مي­يابد.ازآنجاكه پژوهش حاضر بر اساس ساختار زباني و واژگاني زبان فارسي صورت­گرفته، تعداد واژه­هايي كه مي­توانند به­ عنوان واژة بازدارنده در نظرگرفته شود، افزايش مي­يابد. از سوي ديگر، مي­توان نتيجه­گرفت كه ميزان واژه­هاي بازدارنده در متون زبان فارسي بيشتر از متون زبان انگليسي است. اين نتايج نشان مي­دهد ميزان حشو و زوايد در متون زبان فارسي زياد است.

پرداختن به وجوه تشابه و تفاوت بين سياهة واژه­هاي غيرمفهومي انگليسي و واژه‌هاي غيرمفهومي فارسي نيز يكي ديگر از اجزاي پژوهش حاضر بود. نتايج به دست آمده نشان داد براي 231 واژه از 421 واژه بازدارنده فهرست فاكس، تعداد 363 معادل فارسي در فهرست به دست آمده وجود دارد. بيشترين برابري بين واژه­هاي فهرست فاكس و فهرست فارسي حاصل از پژوهش حاضر به حروف ربط، اضافه (حروف اضافه ساده) و قيدهاي مختص اختصاص دارد. ضماير متصل «م، ت، ش، مان، تان و شان» معادل واژه‌هاي منفصل ...Her, His, Me  است كه به دليل اينكه واژه مجزا نيستند، در فهرست واژه­هاي بازدارنده فارسي قرار نمي­گيرند.

بررسي توزيع بسامد واژه­هاي غيرمفهومي در رشته­هاي مورد مطالعه نشان داد P-Value ي 269 واژه يعني 83/20% واژه­ها، كمتر از 0/05 % است كه به اين معناست كه اين تعداد واژه فاقد توزيع بسامدي يكسان در سه رشته مورد مطالعه مي­باشند. به عبارتي، احتمال رخداد هر يك­ از واژه­هاي اين گروه كه از ديد دستور زبان فارسي بيشتر از ميان حروف اضافه، ربط قيود خاص و افعال پربسامد زبان فارسي مي­باشند، در يكي از سه رشته مورد مطالعه بيشتر از دو رشته ديگر است؛ به اين معنا كه احتمال استفاده و كاربرد اين واژه­ها در يك رشته خاص بيشتر از رشته­هاي ديگر بوده و در آن رشته متداول­تر است. بررسي توزيع بسامدي واژه هاي غيرمفهومي شناسايي شده نشان داد تعداد قابل توجهي واژه در اين فهرست وجود دارد كه بسامد پاييني دارند، بنابراين مي توان نتيجه گرفت كه استفاده از روش بسامد واژگاني در شناسايي واژه هاي بازدارنده احتمالا نتواند در زبان فارسي كارايي لازم را داشته باشد.

به طور كلي، مي­توان بيان داشت كه نتايج به دست­آمده از پژوهشهاي انجام شده در حوزة ذخيره و بازيابي اطلاعات نشان داد واژه­هاي بازدارنده به­عنوان يكي از ضروري ترين بخشها در نمايه­سازي و چكيده­نويسي پايگاه هاي اطلاعاتي، نقش مهمي در كاهش حجم پايگاه هاي اطلاعاتي و نرم­افزارهاي اطلاع­رساني ايفا مي­كنند و سبب تسهيل در امر بازيابي، افزايش ميزان مانعيت مدارك بازيابي شده، كاهش زمان جستجو و ... مي­شوند.

**منابع**

- پور اسماعيل، كيومرث و نسرين رستمي (1384). ليست كلمات ايست فارسي. بازيابي شده2 در آذر 1385 ازCcc.sharif.edu/~shesmail/resources/stopwords:pdf

- داورپناه، محمدرضا و صديقه بلنديان (1386). تحليل متن مقالات فارسي و امكان نمايه سازي ماشيني آنها براساس قانون زيف. فصلنامه پژوهش در مسائل تعليم و تربيت: ويژه نامه كتابداري و اطلاع رساني، دور دوم.

- گيلوري، عباس (1379). نمايه سازي خودكار: گذشته، حال، آينده. پيام كتابخانه، 10(4)، 25- 17.

- نياكان، شهرزاد (1383). نمايه سازي ماشيني. تهران: مركز اطلاعات و مدارك علمي ايران.

- ويكري، برايان و الينا ويكري (1380). علم اطلاع رساني در نظر و عمل. ترجمه عبدالحسين فرج پهلو، مشهد: انتشارات دانشگاه فردوسي.

- هادسن، گرور(1383). مباحث ضروري و بنيادين زبانشناسي مقدماتي (ضرورت زبانشناسي مقدماتي) (علي بهرامي، مترجم). تهران: رهنما.

- وحيديان كاميار، تقي و غلامرضا عمران (1385). دستور زبان فارسي (1). تهران: سازمان مطالعه و تدوين كتب علوم انساني(سمت).

- ناتل خانلري، پرويز (1359). دستور زبان فارسي (با تجديدنظر). تهران: توس.

- نجفي، ابوالحسن(1380). مباني زبانشناسي و كاربرد آن در زبان فارسي. تهران: نيلوفر.

- مشكوة الديني، مهدي (1382). دستور زبان فارسي برپايه نظريه گشتاري (ويرايش2). مشهد: فاطمي.

- ـــــــــــــــــــــــــــــ (1384). دستور زبان فارسي. واژگان و پيوندهاي ساختي. تهران: سازمان مطالعه و تدوين كتب علوم انساني (سمت).

- معين، محمد (1378). فرهنگ فارسي (متوسط): شامل يك مقدمه و سه بخش لغات، تركيبات خارجي، اعلام .... تهران: اميركبير.

- مرزبان راد، علي (1378). دستور سودمند. تهران: دانشگاه صنعتي اميركبير.

- محتشمي، بهمن (1370). دستور كامل زبان فارسي. تهران: اشراقي.

- صفوي، كورش (1360). درآمدي بر زبانشناسي. تهران: بنگاه ترجمه و نشر.

- صهبا، عبدالرشيد (1371). حرفهاي ربط، اضافه، نشانه در دستور زبان فارسي براي استفاده دانش آموزان، دانشجويان و پژوهندگان. تهران: غزل.

- غلامعلي زاده، خسرو (1374). ساخت زبان فارسي. تهران: احياء الكتاب.

- فرشيدورد، خسرو (1382). دستور مفصل امروز. تهران: سخن.

- فرشيدورد، خسرو (1386). دستور براي لغت سازي: فرهنگ پيشوندها و پسوندهاي فارسي به همراه گفتارهايي درباره دستور زبان فارسي. تهران: زوار.

- كلباسي، ايران (1380). ساخت اشتقاقي در فارسي امروز. تهران: پژوهشكده علوم انساني و مطالعات فرهنگي.

- شفاعي، احمد (1363). مباني علمي دستور زبان فارسي. تهران: نوين.

- دهخدا، علي اكبر (1383). لغتنامه. (با همكاري محمد معين، جعفر شهيدي). تهران: موسسه لغتنامه دهخدا.

- خطيب رهبر، خليل (1379). دستور زبان فارسي: كتاب حرف اضافه و ربط مشتمل بر تعريف و تقسيم و شرح اصطلاحات و معاني و كاربرد حروف. تهران: مهتاب.

- ـــــــــــــــــــــ (1381). دستور زبان فارسي: براي پژوهش دانشجويان و ادب دوستان در آثار شاعران و نويسندگان بزرگ ايران. تهران: مهتاب.

- بابك، علي (1383). دستور زبان فارسي پژوهشي معاصر. مشهد: سخن گستر.

- باطني، محمدرضا (1382). توصيف ساختاري دستوري زبان فارسي بر بنياد يك نظريه عمومي زبان. تهران: اميركبير

- انوري، حسن (1381). فرهنگ بزرگ سخن. تهران: سخن.

- انوري، حسن و حسن احمدي گيوي (1377). دستور زبان فارسي 2 (ويرايش 2). تهران: فاطمي

- احمدي گيوي، حسن (1380). دستور تاريخي فعل. تهران: قطره.

- Savoy, Jacues(1999). A stemming procedure and stop word list for general French corpora. 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](http://www-ca2.csa.com/ids70/p_search_form.php?field=au&query=sirotkin+karl&log=literal&SID=8b524f5bd7b596896b9037ea91dc6ed5); [Wilbur, W John](http://www-ca2.csa.com/ids70/p_search_form.php?field=au&query=wilbur+w+john&log=literal&SID=8b524f5bd7b596896b9037ea91dc6ed5)(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](http://www.isri.unlv.edu/publications/isripub/Taghva2003-01.ps)

- [Yang, Y](http://www-ca2.csa.com/ids70/p_search_form.php?field=au&query=yang+y&log=literal&SID=8b524f5bd7b596896b9037ea91dc6ed5)iming; [Wilbur, J](http://www-ca2.csa.com/ids70/p_search_form.php?field=au&query=wilbur+j&log=literal&SID=8b524f5bd7b596896b9037ea91dc6ed5)ohn(1996). Using corpus statistics to remove redundant words in text categorization. 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). DEVELOPINGA CORPUS SPECIFICSTOP-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](http://www.utdallas.edu/) /~fxz063000  /- 14k

- Fox, Cristopher(1990). [Astoplist forgeneral text](http://portal.acm.org/citation.cfm?id=378888&coll=portal&dl=ACM). 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

1. Index Term.

2. Timo Lahtinen.

3. Automatic Indexing.

4. Natural language.

5. Marie- Francine Moens.

1. Brian C. Vickery and Elian Vickery.

2. Lexical Analysis.

3. Stemming.

4. Threshold Value.

5. Brian C. Vickery and Elian Vickery.

1. Feng Zou and Others.

1. H. P. Luhn.

2. Boxendal.

3. George Zipf.

4. Christopher Fox.

5. Francis and Kucera.

6. Van Rijsbergen.

7. Bazemen.

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.

1. Jacques Savoy.

2. Tin Kam Ho.

3. Jacques Savoy and Rasolofo.

1. Feng Zou And Others.

2. Fotis Lazarinis.

3. Kazem Taghva, Russel Bechley and Mohammad Sadeh.

1. باقري، مهري(1367). «مقدمات زبانشناسي». تبريز: دانشگاه تبريز.

صفوي، كورش(1360). «درآمدي بر زبانشناسي». تهران: بنگاه ترجمه و نشر.

نجفي، ابوالحسن(1380). «مباني زبانشناسي و كاربرد آن در زبان فارسي». تهران: نيلوفر.

هادسن، گرور(1383). «مباحث ضروري و بنيادين زبانشناسي مقدماتي (ضرورت زبانشناسي مقدماتي)». ترجمه علي بهرامي. تهران: رهنما.

پالمر، فرانك(1366). «نگاهي تازه به معني شناسي». ترجمه كورش صفوي. تهران: مركز.

مشكوة الديني، مهدي(1382). دستور زبان فارسي بر پايه نظريه گشتاري (ويرايش 2). مشهد: فاطمي.

2. شفائي، احمد(1363). «مباني علمي دستور زبان فارسي». تهران: نوين.

بابك، علي(1383). «دستور زبان فارسي پژوهشي معاصر». تهران: دانشگاه آزاد اسلامي مشهد: سخن.

مشكوة­الديني، مهدي(1384). «دستور زبان فارسي (واژگان و پيوندهاي ساختي)». تهران: سمت.

وحيديان كامكار، تقي؛ عمران، غلامرضا، (1385). دستور زبان فارسي (1). تهران: سازمان مطالعه و تدوين كتب علوم انساني(سمت).

ناتل خانلري، پرويز (1359). دستور زبان فارسي (با تجديدنظر). تهران: توس.

مشكوة الديني، مهدي (1384). دستور زبان فارسي. واژگان و پيوندهاي ساختي. تهران: سازمان مطالعه و تدوين كتب علوم انساني(سمت).

معين، محمد(1378). فرهنگ فارسي (متوسط): شامل يك مقدمه و سه بخش لغات، تركيبات خارجي، اعلام .... تهران: اميركبير.

مرزبان راد، علي(1378). دستور سودمند. تهران: دانشگاه صنعتي اميركبير.

محتشمي، بهمن(1370). دستور كامل زبان فارسي. تهران: اشراقي.

صهبا، عبدالرشيد (1371). حرفهاي ربط، اضافه، نشانه در دستور زبان فارسي براي استفاده دانش آموزان، دانشجويان و پژوهندگان. تهران: غزل.

غلامعلي زاده، خسرو(1374). ساخت زبان فارسي. تهران: احياء الكتاب.

فرشيدورد، خسرو(1382). دستور مفصل امروز. تهران: سخن.

فرشيدورد، خسرو(1386). دستور براي لغت سازي: فرهنگ پيشوندها و پسوندهاي فارسي به همراه گفتارهايي درباره دستور زبان فارسي. تهران: زوار.

كلباسي، ايران(1380). ساخت اشتقاقي در فارسي امروز. تهران: پژوهشكده علوم انساني و مطالعات فرهنگي.

دهخدا، علي اكبر(1383). لغتنامه. (با همكاري محمد معين، جعفر شهيدي). تهران: موسسه لغتنامه دهخدا.

خطيب رهبر، خليل(1379). دستور زبان فارسي: كتاب حرف اضافه و ربط مشتمل بر تعريف و تقسيم و شرح اصطلاحات و معاني و كاربرد حروف. تهران: مهتاب.

خطيب رهبر، خليل(1381). دستور زبان فارسي: براي پژوهش دانشجويان و ادب دوستان در آثار شاعران و نويسندگان بزرگ ايران. تهران: مهتاب.

باطني، محمدرضا(1382). توصيف ساختاري دستوري زبان فارسي بر بنياد يك نظريه عمومي زبان. تهران: اميركبير

انوري، حسن(1381). فرهنگ بزرگ سخن. تهران: سخن.

انوري، حسن؛ احمدي گيوي، حسن (1377). دستور زبان فارسي 2 (ويرايش 2). تهران: فاطمي

احمدي گيوي، حسن(1380). دستور تاريخي فعل. تهران: قطره.

1. براي مشاهده ادامه اين فهرست به نسخه الكترونيكي قرار داده شده در سايت كتابخانه آستان قدس رضوي بخش نشريات مراجعه فرمائيد.