



فهرست پیکره - بنیاد واژه‌های دانشگاهی زبان فارسی: استخراج و اعتبارسنجی به روش ارزیابی عملکرد خارج از نمونه

معصومه استاجی^۱

مجتبی الماسی^{۲*}

رضا مرادصحرايي^۳

مقاله پژوهشی

چکیده

گسترش روز افزون تحصیلات دانشگاهی، ضرورت تدوین و استخراج فهرست‌های واژه‌های دانشگاهی را بیشتر نمایان کرد. مفهوم واژه‌های دانشگاهی، در زبان انگلیسی و توسط فهرست کلمات دانشگاهی (AWL) کاکسهد براساس معیارهای سه گانه بسامد، محدوده و وقوع تخصصی و سپس فهرست واژگان دانشگاهی (AVL) گاردنر و دیویس مطرح شد. این دو فهرست جایگاه واژه‌های دانشگاهی را در آموزش زبان دوم برجسته‌تر از قبل کردند. علی‌رغم اهمیت دانش واژگان دانشگاهی، خلاء پژوهش‌های تجربی دقیق برای استخراج و شناسایی واژگان دانشگاهی در زبان فارسی وجود دارد. رضوانی، گلناتش و زمانی، فهرستی ۵۳۹ سرواژه‌ای از یک پیکره ۹۲۷،۰۰۸ واژه‌ای پیشنهاد کرده‌اند. پیکره آن‌ها شامل هفت کتاب و ۱۱۲ مقاله از ۲۸ مجله مختلف تشکیل شده است. در پژوهش حاضر پس از تجزیه و تحلیل پیکره دانشگاه فردوسی مشتمل بر بیش از ۸ میلیون واژه از ۱۱۰۰ مقاله پژوهشی در پنج زیررشته (دامپزشکی، کشاورزی، علوم پایه، علوم انسانی و مهندسی) براساس معیارهای سه گانه بسامد، محدوده و وقوع تخصصی کاکسهد، واژه‌هایی که در هر پنج دسته، حداقل بسامد ۱۰۰ دارند، استخراج گردید و در نهایت یک فهرست ۵۸۷ خانواده واژه‌ای پیشنهاد شد. یک پیکره‌ی شاهد توسط محققین از ۸۷ مقاله در زمینه آموزش زبان و ادبیات فارسی در چارچوب این مقاله ایجاد گردید. سپس میزان پوشش فهرست پیشنهادی این پژوهش با فهرست رضوانی در پیکره شاهد محاسبه و بررسی گردید که نتایج نشان داد فهرست پیشنهادی این پژوهش میزان پوشش بیشتری را بر روی این پیکره شاهد دارد.

کلیدواژه‌ها: واژه، واژه‌های دانشگاهی، زبان با اهداف ویژه، فهرست واژه‌های دانشگاهی، پیکره.

✉ mestaji74@gmail.com

✉ mojtaba.almasi@yahoo.com

✉ sahraei@atu.ac.ir

۱- دانشیار آموزش زبان انگلیسی، گروه زبان انگلیسی، دانشکده

ادبیات فارسی و زبان‌های خارجی، دانشگاه علامه طباطبائی(ره)

۲- دانشجوی دکتری آموزش زبان فارسی به غیر فارسی زبانان،

دانشکده ادبیات و زبان‌های خارجی، دانشگاه علامه طباطبائی(ره)

۳- استاد زبان شناسی همگانی، گروه زبان شناسی، دانشکده

ادبیات فارسی و زبان‌های خارجی، دانشگاه علامه طباطبائی(ره)

تاریخ پذیرش: ۱۴۰۳/۰۸/۰۹

تاریخ دریافت: ۱۴۰۳/۰۲/۳۱

شناسه دیجیتال (DOI): 10.22084/RJHLL.2024.29387.2325

۱- مقدمه

یکی از لازمه‌های موفقیت در تحصیلات دانشگاهی برای دانشجویان غیرفارسی زبان شاغل به تحصیل در دانشگاه‌های ایران، توانایی خواندن و درک انواع متون دانشگاهی به زبان فارسی است، لذا ضرورت دارد دانشجویان خارجی که مشغول یادگیری زبان فارسی هستند، علاوه بر مهارت‌های چهارگانه خواندن، نوشتن، صحبت کردن و شنیدن و نیز مؤلفه‌های سه‌گانه (دستور، واژگان و تلفظ)، برای تحصیل و پژوهش به زبان فارسی، نیز آموزش ببینند. یکی از مهم‌ترین آموزش‌های تکمیلی، مفهوم فارسی دانشگاهی یا فارسی با اهداف دانشگاهی است.

آنچه زبان با اهداف دانشگاهی را از مطالعه عمومی زبان جدا می‌کند، تمرکز آن بر استفاده‌های خاص و هدفمند از زبان است. دانشجویان، زبان دانشگاهی را برای رفع یک سری نیازهای عملی و کاربردی خاص یاد می‌گیرند و این بدین معناست که باید طراحان برنامه درسی، ویژگی‌های زبان هدف را در بافت‌های دانشگاهی خاص مطالعه کنند و مدرسان در کلاس‌های خود به این ویژگی‌ها توجه کنند. بنابراین، این خاص و متمایز بودن زبان با اهداف دانشگاهی، در انواع داده‌های جمع‌آوری شده، شیوه‌های جمع‌آوری آن و نظریه‌هایی که برای درک آن استفاده می‌کنند، تأثیر می‌گذارد. به همان اندازه نیز، ویژگی خاص بودن، نوعی ضرورت کاربردی و پژوهشی را شکل داده است و منجر به بازبینی مفاهیم کلیدی مانند ژانر، اصالت^۱، گروه گفتمانی^۲، اهداف ارتباطی^۳ و مخاطبان^۴ شده است. زبان با اهداف ویژه به‌طور عام و زبان با اهداف دانشگاهی به‌طور خاص، در زبان‌های دیگر از جمله انگلیسی، مدت‌هاست که مدنظر پژوهشگران مختلف بوده است. یکی از مهم‌ترین دستاوردهای این پژوهشگران، طرح مفهوم «واژه‌های دانشگاهی» در مقابل واژه‌های عمومی زبان است.

پژوهش‌ها نشان داده‌اند که در زبان انگلیسی، فهرست واژه‌های عمومی دانشگاهی مجزا از واژه‌های پایه‌ی عمومی هستند (کاکسهد، ۲۰۰۰؛ گاردنر و دیویس، ۲۰۱۴) در زبان ترکی نیز نیاز به فهرست مجزای واژه‌های دانشگاهی وجود دارد (دولماسی^۵ و ارتاف^۶، ۲۰۱۶) ولی در زبانی مانند فرانسه واژه‌های پربسامد دانشگاهی با فهرست واژه‌های عمومی پایه همپوشانی زیادی دارند و نیازی به فهرست مجزای واژه‌های دانشگاهی نیست (کاب^۷ و هورست^۸، ۲۰۰۴).

-
1. genre
 2. authenticity
 3. discourse community
 4. communicative purpose
 5. audience
 6. Dolmacı
 7. Ertaf
 8. Cobb
 9. Horst

در زبان فارسی، تاکنون فهرستی که ارائه شده است، مبتنی بر پیکره‌ای شامل ۹۲۷،۰۰۸ واژه از هفت کتاب و ۱۱۲ مقاله بر مبنای دو اصل بسامد و محدوده استخراج و پیشنهاد کرده‌اند. لذا در این پژوهش برآنیم که با استفاده از یک پیکره با حجم بالا در مقایسه با پژوهش قبلی (بیش از ۸ میلیون واژه) و اعتبارسنجی خارج از نمونه، یک فهرست از واژه‌های دانشگاهی زبان فارسی را براساس سه معیار بسامد، محدوده و وقوع تخصصی ارائه دهیم. کاکسهد (۲۰۰۰) در مطالعه خود سه معیار را تعیین کرده است: ۱. بسامد: وقوع ۱۰۰ بار در کل مجموعه ۲. محدوده: وقوع حداقل ۱۰ بار در هر یک از رشته‌های فرعی ۳. وقوع تخصصی^۱: خارج از فهرست واژگان عمومی وست^۲ با بسامد بالا.

بنابراین، با فرض عدم وجود سازه معتبر واژگان دانشگاهی در قالب یک فهرست بسامدی در تحقیقات زبان فارسی، سؤالات ذیل در این پژوهش مطرح شد:

سؤال اول: مفهوم فارسی دانشگاهی در قیاس با فارسی عمومی چه جایگاهی دارد؟
سؤال دوم: میزان اعتبار فهرست پیشنهادی واژه‌های دانشگاهی در مقایسه با فهرست رضوانی به‌عنوان اولین فهرست واژه‌های دانشگاهی زبان فارسی چقدر است؟

هدف این پژوهش استخراج فهرست واژه‌های دانشگاهی در زبان فارسی براساس سه اصل بسامد، محدوده و وقوع تخصصی کاکسهد (۲۰۰۰) از یک پیکره مقالات دانشگاهی برای طراحی یک آزمون رایانه‌ای انطباقی واژه‌های عمومی دانشگاهی می‌باشد. اعتبارسنجی فهرست واژه‌های دانشگاهی زبان فارسی پیشنهاد شده توسط محققان این پژوهش به‌روش ارزیابی عملکرد خارج از نمونه (یک پیکره شاهد توسط محققان از ۸۷ مقاله در زمینه آموزش زبان و ادبیات فارسی در چهارچوب این پژوهش ایجاد شده است)، بررسی شده است.

۲- پیشینه پژوهش

در این بخش فهرست‌های واژه‌های دانشگاهی در زبان انگلیسی، فهرست گاردنر و دیویس (۲۰۱۴)، فهرست کاکسهد (۲۰۰۰)، تفاوت آن‌ها با یکدیگر و در بخش پایانی، فهرست واژه‌های دانشگاهی زبان فارسی، بررسی می‌شوند.

۲-۱- فهرست واژه‌های دانشگاهی

مفهوم و جایگاه واژه‌های دانشگاهی، به‌دلیل شرایط خاص زبان انگلیسی مانند: جایگاه زبان انگلیسی به‌عنوان زبان علم، روند افزایشی پدیده مهاجرت، مسائل اقتصادی و سیاسی و گسترده بودن آن، ابتدا برای زبان انگلیسی مطرح شده و مورد پژوهش قرار گرفته است. لذا عمده پژوهش‌ها در مورد واژه‌های دانشگاهی

1. specialised occurrence

2. West

بر روی زبان انگلیسی انجام شده است و خودبخود پژوهش‌ها درباره سایر زبان‌ها نیز با نگاه به دستاورد پژوهش‌های زبان انگلیسی انجام شده‌اند. لذا ضروری است که ابتدا فهرست‌های مطرح و مهم زبان انگلیسی را معرفی کنیم و در بخش دوم، فهرست‌های پیشنهاد شده برای زبان فارسی را مورد بررسی قرار دهیم.

۲-۱-۱- فهرست واژه‌های دانشگاهی در زبان انگلیسی

در پژوهش‌های انگلیسی به‌عنوان زبان دوم (ESL)، واژگان دانشگاهی عمدتاً توسط دو فهرست برجسته، مطرح می‌شوند: فهرست AVL کاکسهد (۲۰۰۰) و فهرست AVL توسط گاردنر و دیویس (۲۰۱۴). این دو فهرست به‌دنبال فهرست‌های دیگر کلمات دانشگاهی در دهه ۱۹۷۰ که توسط زو^۱ و نیشن^۲ (۱۹۸۴) به‌طور مفصل شرح داده شده است، ایجاد شده‌اند. با این وجود، تنها ظهور فناوری‌های رایانه‌ای، امکان جمع آوری متون بزرگ را برای ایجاد فهرست‌های واژه مبتنی بر پیکره و اعتبارسنجی تحقیقات قبلی در این زمینه فراهم کرده است.

فهرست‌های واژه‌های دانشگاهی در زبان انگلیسی که مورد استناد سایر محققان و مبنای تهیه و تدوین کتب آموزشی انگلیسی دانشگاهی قرار گرفته‌اند، فهرست کاکسهد (۲۰۰۰) و فهرست گاردنر و دیویس (۲۰۱۴) می‌باشند که در این بخش به‌ترتیب بررسی و تشریح می‌شوند. سپس مطالعاتی که به مقایسه این دو فهرست پرداخته‌اند، مرور می‌شوند. در ادامه مروری اجمالی بر فهرست حیدری تبریزی و جمال‌زاده (۲۰۲۰) که یک فهرست واژه‌های دانشگاهی گردشگری ارائه کرده‌اند، خواهیم داشت.

۲-۱-۱-۱- فهرست واژه‌های دانشگاهی (AWL) کاکسهد

AWL از مجموعه ۳,۵ میلیون کلمه از متون علمی نوشتاری استخراج شده است و شامل ۵۷۰ خانواده واژگانی است که تقریباً ۱۰٪ کل کلمات در متون دانشگاهی را تشکیل می‌دهند. کاکسهد در مطالعه خود سه معیار را تعیین کرد: ۱. بسامد: وقوع ۱۰۰ بار در کل مجموعه ۲. محدوده: وقوع حداقل ۱۰ بار در هر یک از رشته‌های فرعی ۳. وقوع تخصصی^۳: خارج از فهرست واژگان عمومی وست^۴ با بسامد بالا. در نهایت، او فهرستی متشکل از خانواده‌های ۵۷۰ کلمه‌ای ارائه کرد که به ۱۰ فهرست فرعی تقسیم شده است. هر زیرفهرست شامل خانواده‌های ۶۰ کلمه‌ای است به‌جز زیرفهرست ۱۰م که حاوی ۳۰ کلمه است. کاکسهد (۲۰۰۰) ادعا کرد که فهرست او ۱۰ درصد از کل متون دانشگاهی را تشکیل می‌دهد. پوشش فهرست کاکسهد در چهار زیرشاخه هنر، تجارت، حقوق و علوم به‌ترتیب ۹,۳٪، ۱۲٪، ۹,۴٪ و ۹,۱٪ بود. او استدلال کرد که فهرست کلمات دانشگاهی (AWL) برای دانشجویان رشته‌های مختلف دانشگاهی از اهمیت حیاتی

1. Xue
2. Nation
3. specialised occurrence
4. West

برخوردار است، زیرا می‌تواند پس از ترکیب با فهرست واژگان عمومی وست، تقریباً ۹۰ درصد از کلمات در حال استفاده در متون دانشگاهی را پوشش دهد. نتایج مثبت AWL توسط برخی از پژوهشگران به‌طور کامل اثبات و تأیید شده است (هانتلی^۱، ۲۰۰۵؛ لی^۲ و کیان^۳، ۲۰۱۰).

هوانگ، چانگ، وون‌پومیتویچ^۴ (۲۰۰۹) نشان دادند که میزان پوشش فهرست AWL در زبان‌شناسی کاربردی ۱۱٪ و بالاتر از پوششی معادل ۹٪ در رشته هنر است که به‌طور کلی توسط کاکسهد (۲۰۰۰) بررسی شده است؛ بنابراین، این مطالعات ثابت کرد که AWL برای دانشجویان این رشته‌ها کاربردی است. به‌همین ترتیب، سان و لی (۲۰۱۳) اظهار داشتند که AWL می‌تواند برای دانشجویان اقتصاد با در نظر گرفتن استانداردهای رشته‌ای مفید باشد. مودرایا (۲۰۰۶) نشان داد واژگان دانشگاهی در کلاس‌های زبان با اهداف ویژه برای دانشجویان مهندسی نقش مهمی دارد. علاوه بر این، تاد (۲۰۱۷) در رشته مهندسی، بر لزوم توجه بیشتر به واژه‌های چندمعنایی که بسامد بالایی دارند، تأکید دارد و بیان می‌کند که معنای موردنظر این نوع واژه‌ها، معنی معمول آن کلمه نیست. در مقابل، مطالعات در چندین رشته با تأکید بر محدودیت‌های AWL، لزوم ارزیابی فهرست خاص هر رشته را گوشزد می‌کنند. اسلامی‌زاده و معینی (۲۰۱۶) در مورد ضرورت فهرست جداگانه‌ای برای رشته زبان‌شناسی بحث کردند.

لی^۵ و لی‌یو^۶ (۲۰۱۶) شواهدی را درباره ضرورت وجود فهرستی برای زبان‌آموزان انگلیسی در رشته پزشکی ارائه دادند. لیو و هان (۲۰۱۵) به نفع فهرست فنی برای دانشجویان علوم محیطی استدلال کردند. مارتینز، بک، و پانزا (۲۰۰۹) به‌همراه مونوز (۲۰۱۵) فهرستی را برای رشته کشاورزی پیشنهاد کردند. سایر مطالعات فهرست‌های جدیدتر ویژه رشته‌های دانشگاهی را توصیف می‌کنند: فهرست واژه‌های دانشگاهی پزشکی (MAWL) توسط وانگ و لیانگ (۲۰۰۸). فهرست کلمات علمی پرستاری توسط یانگ (۲۰۱۵)؛ و فهرست واژه‌های علمی شیمی (CAWL) توسط ولی‌پوری و ناساجی (۲۰۱۳)؛ بنابراین، مطالعات متنوعی شواهدی را بر این ادعا ارائه کرده‌اند که واژگان دانشگاهی درجه مختلفی از ارتباط با رشته‌های مختلف را دارند؛ درحالی‌که در بعضی از آن‌ها بسیار کاربرد دارد، در برخی دیگر کافی نیست و باید با فهرست فنی، تقویت یا جایگزین شود.

-
1. Huntley
 2. Li
 3. Qian
 4. Vongpumivitch
 5. Lei
 6. Liu

۱-۲-۱-۲- فهرست واژه‌های دانشگاهی گاردنر و دیویس (AVL)

AVL از یک مجموعه بزرگ‌تر شامل ۱۲۰ میلیون کلمه دانشگاهی تهیه می‌شود و حدود ۱۴٪ از مطالب دانشگاهی را در مجموعه انگلیسی معاصر آمریکایی^۱ (COCA) و در پیکره ملی انگلیسی^۲ (BNC) پوشش می‌دهد. فهرست لغات دانشگاهی AVL، فهرست واژگان دانشگاهی AVL کاکسهد (۲۰۰۰) را از چند جهت بهبود بخشیده است. فهرست واژگان دانشگاهی AVL بر اساس ۳٫۵ میلیون واژه از ۴۱۴ متن دانشگاهی، در سال ۱۹۹۸ استخراج شده است، ولی AVL بر اساس ۱۲۰ میلیون واژه از ۱۳،۰۰۰ متن دانشگاهی (که خود برگرفته شده از پیکره‌ی ۴۲۵ میلیون واژه COCA است)، در سال ۲۰۱۱ استخراج شده است. فهرست لغات دانشگاهی AVL شامل اطلاعات کامل‌تری از قبیل بسامد و مقوله نحوی است و در سه گروه: دانشگاهی پایه، فنی تخصصی و عمومی دانشگاهی در نه حوزه «آموزش، علوم انسانی، الهیات، علوم اجتماعی، تاریخ، حقوق، علوم، پزشکی و تجارت» طبقه‌بندی شده است. گاردنر و دیویس (۲۰۱۴) مدعی هستند که فهرست واژگان دانشگاهی ایجاد شده توسط آن‌ها (AVL)، به چند دلیل نسبت به فهرست کلمات دانشگاهی (AVL) کاکسهد به‌طور قابل‌توجهی بهبود یافته است. دلیل اول اینکه، AVL تنها بر اساس ۳٫۵ میلیون کلمه از دهه ۱۹۹۰ ایجاد و استخراج شده است، ولی فهرست AVL بر اساس ۱۲۰ میلیون کلمه و در ۱۳۰۰۰ متن دانشگاهی و در مجموعه ۴۲۵ میلیون کلمه‌ای انگلیسی معاصر آمریکایی، با متن‌هایی جدیدتر که از سال ۲۰۱۱ به بعد منتشر شده‌اند، ایجاد شده است. دوم، در AVL «خانواده واژه‌ها» حاوی اطلاعات زیادی است که در AVL موجود نیست:

۱. در AVL ریشه کلمات به ترتیب فراوانی فهرست شده‌اند. در فهرست خانواده‌های واژه‌ای AVL، هیچ نشانه‌ای وجود ندارد که کدام کلمات از آن ریشه، پربسامدتر هستند و کدام‌ها کم بسامدتر.
۲. در AVL ریشه‌ها بر حسب مقوله نحوی^۳ جدا شده‌اند و فراوانی هر ریشه و مقوله نحوی فهرست شده‌اند. به‌عنوان مثال، در خانواده واژگانی [effect]، effect به‌عنوان یک اسم (۶۰۰۷۸ نشانه) و effect به‌عنوان فعل (۱۵۸۱؛ یعنی بسیار کمتر رایج) فهرست شده است.
۳. کلمات بر اساس ریشه گروه‌بندی می‌شوند و نه صورت‌های تصریفی مختلف.
۴. کلمات طوری دسته‌بندی می‌شوند که مشخص شود آیا آن‌ها بخشی از فهرست واژگان عمومی دانشگاهی هستند یا فنی‌تر هستند و عمدتاً در یک حوزه (مانند حقوق یا پزشکی) رخ می‌دهند، یا اینکه در واقع یک کلمه دانشگاهی نیستند، بلکه فقط عضوی از خانواده کلمات هستند. برای کلمات فنی، نشان داده می‌شود که در کدام زیر - ژانر(های) بیشترین فراوانی را دارند.

1. Corpus of Contemporary American English

2. British National Corpus

3. POS

چندین مطالعه دو فهرست کاکسهد و گاردنر را با هم مقایسه کرده‌اند (هارستون^۱ و هارت^۲؛ ۲۰۱۶ و نیومن^۳، ۲۰۱۶ و کی^۴، ۲۰۱۶) و متفق‌القول هستند که هر دو فهرست نمایانگر ساختار واژگان دانشگاهی هستند. دوران^۵ (۲۰۱۶) در یک پژوهش اعتبارسنجی درباره محدودیت‌های AVL نشان می‌دهد تقریباً نیمی از کلمات فهرست به‌ندرت در مجموعه مقالات و متون نوشته شده توسط دانشجویان به کار برده می‌شوند و مواردی که به‌طور مکرر استفاده می‌شوند در رشته‌های مختلف متفاوت است. در همان زمان، این مطالعه ثابت کرد که به‌نظر می‌رسد ۴۲۷ مورد در ۹۰٪ رشته‌ها پرتکرار است و این یافته منجر به این نتیجه گیری شد که "یک واژگان عمومی تولیدی وجود دارد، اما دامنه آن از فهرست کامل واژه‌های دانشگاهی کوچک‌تر است" (دوران، ۲۰۱۶: ۴۹). در مقابل، طراحان AVL رویکرد دوران را در استفاده از مطالب نوشته شده توسط فراگیران به‌عنوان یک ابزار تأیید روایی واحد را زیر سؤال بردند و بر اهمیت یادگیری کلمات اصلی دانشگاهی (AVL) برای دانشجویان رشته‌های تحصیلی دانشگاهی تأکید کردند و بر اهمیت موضوع لغات دریافتی^۶ و مولد^۷ و ضرورت فهرست‌ها در افزایش سطح بالاتر مهارت زبانی را مورد تأکید قرار دادند (گاردنر و دیویس، ۲۰۱۶).

حیدری تبریزی و جمال‌زاده (۲۰۲۰) به بررسی مجموعه نوشتاری مقالات پژوهشی دانشگاهی به زبان انگلیسی، با هدف ایجاد فهرست واژه‌های دانشگاهی گردشگری^۸ (TAWL) در گرایش‌های مختلف این رشته پرداخته‌اند. این نویسندگان درصدد پاسخ به این سؤال بودند که تا چه حد کلمات شناسایی شده در مجموعه گردشگری، در فهرست خدمات عمومی وست (GSL) و فهرست کلمات دانشگاهی کاکسهد (AWL) شناسایی شده‌اند. آنها با تجزیه و تحلیل یک مجموعه ۳,۷ میلیون کلمه‌ای، نشان دادند که کلمات AWL، به اندازه ۱۲,۳۴٪ از مجموعه مقالات پژوهشی گردشگری^۹ (TRAC) را تشکیل می‌دهند. بیشتر فرم‌های کلمه AWL با خانواده‌های واژه‌ای که در فهرست‌های فرعی اول و دوم کاکسهد گنجانده شده‌اند، قرار می‌گیرند. فراوانی کلمات بالا و پوشش متنی گسترده TAWL در سراسر مقالات پژوهشی گردشگری بیانگر این است TAWL نقش مهمی در مقالات پژوهشی گردشگری دارد. موارد AVL پرتکرار، ترتیب بسامد متفاوتی با موارد موجود در AVL کاکسهد نشان می‌دهند. علاوه بر این، بسیاری از خانواده‌های واژه‌های محتوای غیر AVL وجود داشتند که با فراوانی بالا در مجموعه رخ می‌دادند.

-
1. Hartshorn
 2. Hart
 3. Newman
 4. Qi
 5. Durrant
 6. Receptive
 7. Productive
 8. Tourism Academic Word List
 9. Tourism Research Articles Corpus (TRAC)

۲-۲. فهرست‌های واژه‌های فارسی دانشگاهی

رضوانی، گلشن و زمانی (۲۰۱۶) یک فهرست واژه‌های دانشگاهی از یک پیکره ۹۲۷،۰۰۸ واژه‌ای بر مبنای دو اصل بسامد و محدوده استخراج و پیشنهاد کرده‌اند. پیکره آن‌ها شامل هفت کتاب و ۱۱۲ مقاله که از ۲۸ مجله مختلف به صورت تصادفی انتخاب شده بودند، تشکیل شده است. زمینه‌های علمی مقالات این پیکره به هفت رشته دانشگاهی: هنر و معماری، مهندسی و فناوری، کشاورزی و علوم طبیعی، علوم پزشکی، دامپزشکی، علوم انسانی و علوم پایه تعلق دارند. فهرست پیشنهادی آن‌ها ۵۳۹ سرواژه^۱ است. هر واژه با بسامد بیش از ۲۰ و تکرار در سه زیررشته از هفت رشته اصلی اشاره شده، مبنای انتخاب آن‌ها بوده است. فهرست رضوانی، علی‌رغم اینکه اولین فهرست واژه‌های دانشگاهی در زبان فارسی است، اما دارای چند ایراد است. یکی از این ایرادها که خود محققان به آن اشاره کرده‌اند، اندازه پیکره جمع‌آوری شده است یعنی یک پیکره مشتمل بر ۹۲۷،۰۰۸ واژه. ایراد دیگر این فهرست، عدم ارائه بسامد هر یک از کلمات خانواده واژه است مثلاً درباره واژه سر خانواده «روش» که متشکل از دو واژه روش و روش‌ها است، مشخص نشده است که هر یک از این دو، در پیکره چه بسامدی داشته‌اند. از طرف دیگر فهرست واژه‌های دانشگاهی رضوانی به جای بسامد به صورت الفبایی ارائه شده است.

زرزآ (۱۳۹۳) صرفاً فهرست واژه‌های دانشگاهی کاکسپد را ترجمه و معادل‌سازی کرده است و هیچ نوع اعتبارسنجی و یا بسط و گسترشی برای فهرست معادل خود انجام نداده است. زرزآ (۱۳۹۳) صرفاً فهرست معادل فارسی کاکسپد (۲۰۰۰) را در کتاب‌های خواندن سطوح میانی و پیشرفته و کتاب‌های فارسی ویژه علوم پزشکی و ویژه فنی مهندسی که در مرکز آموزش زبان فارسی به غیرفارسی‌زبانان دانشگاه بین‌المللی قزوین تألیف و تدریس می‌شود، را بررسی و تحلیل کرده است.

۳- چهارچوب نظری

در این بخش مفهوم واژه‌های دانشگاهی و تفاوت آن با واژه‌های فنی و تخصصی و نیز جایگاه مفهوم «واژه دانشگاهی» در زبان انگلیسی، فرانسه و ترکی و پژوهش‌های انجام شده در مورد آن‌ها بررسی می‌شوند.

۳-۱- سازه واژه‌های دانشگاهی

اولین قدم در ایجاد فهرست واژه‌ها، تعریف صریح ساختار واژه‌ها است (بریزنا^۲، گابلاسوا^۳، ۲۰۱۷). اصطلاح «واژه دانشگاهی» یک مفهوم پیچیده و مرتبط با زبان دانشگاهی است. زبان دانشگاهی به طور معمول در بافت دانشگاهی و در ارتباطات بین افراد در اشکال مختلفی همانند سخنرانی، تدریس، مقالات و کتب

1. Headword
2. Brezina
3. Gablasova

دانشگاهی با محوریت رشته‌ای خاص و با هدف انتقال ایده‌های انتزاعی، فنی و ظریف باشد، شکل می‌گیرد (باومن^۱، گریوز^۲؛ ۲۰۱۰ و نگی^۳ و تاون سند^۴، ۲۰۱۲). کاکسهد و نیشن^۵ ترکیب «واژه‌ی دانشگاهی» را ناظر به کلماتی می‌دانند که در طیف گسترده‌ای از ژانرهای دانشگاهی به‌طور منطقی پربسامد و پرتکرار هستند اما در انواع دیگر متون نسبتاً غیرمعمول هستند (کاکسهد و نیشن، ۲۰۰۱). اصطلاحات جایگزین «واژه دانشگاهی» اصطلاحاتی مانند «واژگان فرعی» (کوان^۶، ۱۹۷۴) یا «تیمه فنی» (فارل^۷، ۱۹۹۰) است. طبق گفته اشمیت (۲۰۱۰)، واژه‌های دانشگاهی در حدود ۹ الی ۱۰٪ کلمات در حال استفاده در یک متن دانشگاهی به زبان انگلیسی را تشکیل می‌دهند و از این‌رو برای افرادی که در زمینه‌های دانشگاهی فعالیت می‌کنند، یادگیری آن‌ها ضروری است. یکی از مشخصه‌های کلیدی واژگان دانشگاهی، بسامد و فراوانی آن است و به‌عنوان مبنایی منطقی جهت کسب اطمینان از یادگیری بهتر دانشجویان در نظر گرفته می‌شود (نیشن، وارینگ^۸، ۱۹۹۷: ۱۵) اطلاعات مربوط به بسامد، زمینه‌ساز توسعه فهرست‌های واژگان دانشگاهی است و به‌عنوان راهنما در طراحی مطالب آموزشی زبان دوم نیز عمل می‌کند. واژه‌های دانشگاهی به‌دلیل تأثیر بر کسب نتیجه مطلوب توسط ذی‌نفعان از جمله دانشجویان، مدرسان و طراحان برنامه درسی، در آموزش زبان دوم مورد توجه است.

۳-۲- تفاوت بین واژگان دانشگاهی و واژه‌های فنی

واژگان دانشگاهی با واژگان عمومی متفاوت هستند و در گروه واژگان با اهداف ویژه دسته‌بندی می‌شوند (نیشن، کیونگهو^۹، ۱۹۹۵). در حالی که واژگان عمومی کاملاً ساده هستند و هسته اصلی برنامه‌های درسی زبان دوم را تشکیل می‌دهند (نیشن، وارینگ، ۱۹۹۷)، تعریف و شناسایی دقیق واژگان با اهداف ویژه دانشگاهی با چالش‌هایی همراه می‌باشد. به‌عنوان مثال، فلاوردرو^{۱۰} (۱۹۹۳) بین کلمات فنی و فرعی فنی تفاوت قائل می‌شود و اظهار می‌کند کلمات فنی مختص یک‌رشته خاص هستند، درحالی‌که دسته دوم مختص رشته‌ای خاص و یا هسته مشترک نیستند، چرا که کلماتی با کاربردهای عمومی هستند و دارای معنایی خاص در آن حوزه فنی می‌باشند، و لذا، باید در کلاس‌های زبان با اهداف ویژه ظاهر شوند. این دسته از کلمات، در طبقه‌بندی چهار سطحی واژگان چانگ^{۱۱} و نیشن (۲۰۰۳) بخشی از واژگان دانشگاهی

1. Baumann
2. Graves
3. Nagy
4. Townsend
5. Nation
6. Cowan
7. Farrell
8. Waring
9. Kyongho
10. Flowerdew
11. Chung

در سه‌گانه واژگان عمومی (بسامد زیاد و بسامد کم)، واژگان فنی و واژگان دانشگاهی می‌باشند. اصطلاحات جایگزین مورد استفاده گاردنر و دیویس (۲۰۱۴)، اصطلاحات «اصلی دانشگاهی» و «فنی دانشگاهی» است و این ایده را برجسته می‌کند که هر دو سازه به یک مجموعه دانشگاهی اشاره دارند، اما درجه کاربرد آن‌ها برای رشته‌های مختلف، متفاوت است.

چندین مطالعه به تمایز واژگان دانشگاهی و فنی پرداخته‌اند، زیرا هر دو گروه به فضای دانشگاهی تعلق دارند و به واژگان با اهداف خاص اشاره می‌کنند. عمدتاً، این مطالعات به دو دسته تقسیم می‌شوند: مواردی که مؤید فهرست واژگان دانشگاهی برای عموم رشته‌های دانشگاهی هستند و مواردی که لزوم فهرست واژگان خاص هر رشته را ثابت می‌کنند، و بنابراین واژگان فنی را در اولویت قرار می‌دهند. واژگان فنی «نشان‌دهنده حدود ۵٪ کلمات در حال استفاده در متون تخصصی است و شامل کلماتی است که غالباً در متون تخصصی یا موضوعی وجود دارند اما در سایر زمینه‌ها بسیار کم هستند و یا اصلاً استفاده نمی‌شوند» (نیشن، ۲۰۰۱: ۱۸).

کاب و هورست (۲۰۰۴) این سؤال را مطرح کردند که آیا فهرست واژگان دانشگاهی مبتنی بر بسامد در زبان فرانسه قابل استفاده است. مطالعه مبتنی بر پیکره آن‌ها نشان داد که بر خلاف انگلیسی، ۲۰۰۰ کلمه اول کلمات عمومی زبان فرانسوی بخش قابل توجهی از واژگان دانشگاهی را پوشش می‌دهد، و بنابراین فهرست جداگانه کلمات دانشگاهی مورد نیاز نیست. آن‌ها توضیح می‌دهند که هر زبان مشخصات واژگانی خاص خود را دارد که در مورد زبان فرانسه در ژانرهای مختلف چندان متمایز نیست تا نیازی به فهرست واژگان جداگانه وجود داشته باشد؛ بنابراین، کاب و هورست (۲۰۰۴) فهرست واژگان دانشگاهی جداگانه‌ای را پیشنهاد نکرده است اما دولماسی و ارتاف (۲۰۱۶) فهرست واژگان زبان ترکی را از مجموعه کتاب‌های درسی دوره کارشناسی به زبان ترکی و بدون در نظر گرفتن ۳۰۰ کلمه اول از فهرست کلمات عمومی^۱ که قبلاً تولید شده بود، تهیه کرده‌اند و بیان کرده‌اند که در زبان ترکی، مفهوم واژه‌های عمومی دانشگاهی وجود دارد.

۴- روش‌شناسی

در این بخش ابتدا پیکره مقالات دانشگاهی زبان فارسی و فهرست‌های واژه که در این پژوهش مورد استفاده قرار گرفته است، تشریح می‌گردد و سپس روش جمع‌آوری و پردازش داده‌های مورد نیاز برای استخراج فهرست نهایی بیان می‌گردد.

1 General service list

۴-۱- بیکره‌ها و فهرست‌های واژه

فهرست‌های واژه، عمدتاً از بیکره‌ها استخراج می‌شوند و تا زمان انجام این پژوهش، تنها بیکره دانشگاهی منتشر شده‌ی در دسترس، بیکره دانشگاه فردوسی مشهد می‌باشد. دلیل اصلی انتخاب بیکره فردوسی در این پژوهش، این است که از ۱۱۰۰ مقاله علمی در حوزه‌های کشاورزی، دامپزشکی، علوم پایه و نیز علوم انسانی و مهندسی گردآوری شده است و تنوع لازم در رشته‌های دانشگاهی را دارد. بیکره فردوسی در حساب کاربری گیت‌هاب^۱ یکی از مجریان طرح به صورت ۹۱ فایل مجزا و در قالب CONLL منتشر شده است. همه این مقالات تألیف اعضای هیأت علمی دانشگاه فردوسی مشهد، در پنج زیررشته (دامپزشکی، کشاورزی، علوم پایه، علوم انسانی و مهندسی) بوده و شامل رشته‌های زیرمجموعه علوم پزشکی نیستند. توزیع واژه‌های موجود در این بیکره برحسب مقوله دستوری واژه در جدول ۱ نشان داده شده است. لازم به توضیح است که این مقوله‌ها براساس آنچه توسط گردآوردندگان بیکره برچسب‌گذاری و اظهار شده است، محاسبه گردیده است.

جدول ۲: توزیع مقوله‌های نحوی مختلف در بیکره فردوسی

مقوله دستوری واژه	تعداد واژه یکتا	تعداد واژه‌ها	مقوله دستوری واژه	تعداد واژه یکتا	تعداد واژه‌ها
POSTP	۵۹۷	۷۱۳۷۶	ADJ	۱۳۷۷۵	۷۰۵۹۳۵
PR	۱۵۲۶	۱۱۴۵۰۳	ADR	۵۴۳۳	۴۴۳۸۱
PREM	۱۱۷۹	۱۳۶۶۶۲	ADV	۱۰۴۸	۹۴۴۷۲
PRENUM	۲۸۵۱	۱۰۱۴۰۹	CONJ	۱۱۹۱	۳۸۹۵۵۶
PREP	۸۱۹۵	۹۱۲۰۱۶	IANM	۳	۹۵
PSUS	۷۸۹۰	۴۱۷۴۴	IDEN	۱۳۸۲	۱۲۴۲۵
PUNC	۲۴۰۵	۹۹۳۴۲۳	N	۱۴۲۱۰۸	۳۵۶۶۸۶۷
SUBR	۸۵۱	۱۲۸۱۳۲	PART	۸۸۶	۶۲۳۷
V	۳۶۵۷۹	۵۷۴۰۳۵	POSNUM	۱۲۷	۲۱۲۹۵
مجموع				۷۰۹۱۴۰۵۷۰	

به موازات، یک بیکره شاهد مشتمل بر ۸۷ مقاله در زمینه آموزش زبان و ادبیات فارسی با هدف اعتبارسنجی فهرست نهانی، در چهارچوب این مقاله توسط محققان گردآوری و تهیه شد. این بیکره شاهد، دارای ۵۳۸۰۰۶۱ واژه است که مطابق بیکره فردوسی از نظر نویسه‌ها و قالب، استانداردسازی و تجمیع

گردید. هدف از این پیکره شاهد، مقایسه فهرست رضوانی و فهرست واژه‌های دانشگاهی پیشنهادی این کار پژوهشی می‌باشد. آمار این پیکره در جدول ۲ آمده است.

جدول ۲- آمار پیکره شاهد

۷۶۲۰	۲۰۰۷	۵۹
۳۶۸۳	۸۹۵	۶۰
۳۰۴۵	۹۳۰	۶۱
۶۶۶۱	۹۵۰	۶۲
۳۸۸۱	۱۰۹۱	۶۳
۳۸۷۰	۹۵۱	۶۴
۱۰۸۱۵	۲۸۵۳	۶۵
۵۲۱۵	۱۳۸۰	۶۶
۳۴۰۶	۸۰۶	۶۷
۳۷۸۶	۶۰۶	۶۸
۶۲۹۷	۱۷۹۷	۶۹
۳۳۴۲	۹۵۸	۷۰
۸۶۲۹	۱۵۹۵	۷۱
۶۹۸۲	۱۴۶۰	۷۲
۸۰۴۷	۱۸۸۱	۷۳
۷۹۶۶	۲۰۸۱	۷۴
۳۴۳۸	۳۴۳۸	۷۵
۸۱۹۷	۱۱۵۵	۷۶
۴۹۸۲	۱۱۱۱	۷۷
۳۸۹۴	۹۵۶	۷۸
۳۱۶۰	۷۰۷	۷۹
۴۸۳۴	۱۲۷۰	۸۰
۶۴۶۷	۲۰۸۴	۸۱
۷۴۰۳	۱۵۹۵	۸۲
۶۸۶۰	۱۶۱۱	۸۳
۴۵۸۴	۸۱۶	۸۴
۷۵۵۲	۱۲۶۸	۸۵
۶۳۳۹	۱۷۰۶	۸۶
۶۶۵۹	۱۱۵۴	۸۷
۵۳۸۰۶۱	۱۲۶۱۹۴	مجموع

۴۰۴۷	۱۰۷۴	۲۹
۳۱۹۸	۷۰۵	۳۰
۴۴۳۳	۱۲۵۵	۳۱
۶۶۰۹	۱۴۲۸	۳۲
۸۲۶۳	۲۰۳۶	۳۳
۹۸۳۴	۱۲۴۹	۳۴
۵۵۶۵	۱۳۳۷	۳۵
۳۰۵۳	۹۲۸	۳۶
۶۳۸۱	۱۷۷۹	۳۷
۵۲۱۶	۹۶۲	۳۸
۳۳۰۱	۱۲۳۹	۳۹
۷۸۳۸	۲۲۲۵	۴۰
۴۱۵۳	۱۲۹۸	۴۱
۱۰۳۳۴	۱۷۰۵	۴۲
۵۴۱۷	۱۶۸۲	۴۳
۴۴۵۱	۱۱۸۳	۴۴
۹۵۱۴	۱۵۷۲	۴۵
۴۷۰۲	۱۰۶۵	۴۶
۳۴۵۰	۱۲۶۷	۴۷
۱۱۳۷۷	۲۵۴۴	۴۸
۹۴۷۰	۱۶۴۰	۴۹
۹۰۹۸	۱۹۲۹	۵۰
۳۸۷۰	۹۸۱	۵۱
۶۳۳۹	۱۷۵۰	۵۲
۳۷۶۳	۸۱۰	۵۳
۹۹۹۵	۲۱۹۷	۵۴
۷۴۵۴	۱۶۵۶	۵۵
۵۷۲۰	۱۵۷۹	۵۶
۳۱۷۷	۱۰۱۱	۵۷
۳۶۱۳	۱۱۵۷	۵۸

فایل	تعداد واژه یکتا	تعداد واژه
۱	۱۸۸۱	۹۱۶۸
۲	۱۳۰۰	۵۸۱۴
۳	۱۷۶۴	۸۶۰۵
۴	۱۳۷۱	۷۸۵۲
۵	۱۳۴۰	۵۱۷۲
۶	۲۰۴۷	۷۶۴۱
۷	۱۰۹۷	۴۴۱۶
۸	۱۴۲۱	۵۰۷۳
۹	۱۴۸۰	۴۵۲۲
۱۰	۱۰۸۶	۴۲۴۹
۱۱	۶۲۳	۲۶۴۶
۱۲	۱۷۵۲	۷۵۲۲
۱۳	۲۱۴۱	۷۴۱۹
۱۴	۱۵۹۲	۶۱۴۳
۱۵	۱۶۲۷	۵۰۳۶
۱۶	۱۸۱۱	۶۰۵۴
۱۷	۱۳۲۲	۴۰۳۵
۱۸	۱۶۱۸	۵۰۹۸
۱۹	۶۹۰	۳۰۵۱
۲۰	۱۷۳۸۴	۵۱۷۴
۲۱	۱۲۹۷	۴۶۶۸
۲۲	۲۰۸۹	۶۲۵۸
۲۳	۱۸۲۵	۱۴۶۳۳
۲۴	۱۹۵۷	۸۳۸۱
۲۵	۲۰۱۱	۹۹۴۲
۲۶	۲۲۲۷	۱۶۷۷۹
۲۷	۱۰۶۵	۴۴۱۳
۲۸	۱۴۷۶	۸۹۶۵

با هدف بررسی میزان همپوشانی واژه‌های دانشگاهی با واژه‌های عمومی زبان فارسی، دو پژوهش علمی واژگان پایه زبان فارسی (نعمت‌زاده و همکاران، ۱۳۸۰) و واژه‌های پایه زبان فارسی (صحرائی و همکاران، ۱۳۹۶) انتخاب گردیدند. پس از بررسی واژگان این دو مجموعه، میزان همپوشانی آن‌ها با یکدیگر بیش از ۷۰ درصد محاسبه گردید و ۳۷۲ واژه‌ی مشترک در هر دو فهرست، به‌عنوان واژه‌های عمومی زبان فارسی در استخراج نهایی فهرست واژه‌های دانشگاهی مبنای مقایسه قرار گرفتند. فهرست این واژه‌های مشترک در پیوست ۱ ارائه شده است.

۲-۴- روش جمع‌آوری و پردازش داده‌ها

در گام اول، کلیه فایل‌های پیکره فردوسی از فرمت CONLL به فایل اکسل تبدیل شده و در یک بانک اطلاعاتی تجمیع شدند. در گام دوم، کلیه‌ی نویسه‌های واژه‌های موجود در فایل تجمیع شده، توسط نرم‌افزار ویراستیار^۱ یکسان و استاندارد شدند تا در جستجوها و پردازش‌های بعدی، مشکلاتی مانند نویسه‌های متفاوت حرف «ی» وجود نداشته باشد. این اصلاحات شامل: «اصلاح انواع حرف 'ک' - اصلاح انواع حرف 'ی' - تبدیل ارقام عربی به معادل فارسی - اصلاح نویسه‌ی نیم‌فاصله - اصلاح اعراب - حذف نیم‌فاصله‌های تکراری» می‌باشد.

در گام سوم، جهت اطمینان از توازن پراکندگی مقالات در رشته‌های متنوع دانشگاهی، در چند مرحله، عنوان اصلی ۱۱۰۰ مقاله استخراج گردید و براساس عنوان مقاله و نیز مراجعه به سامانه مدیریت نشریات دانشگاه فردوسی مشهد به آدرس: <https://jm.um.ac.ir> زمینه اصلی مقاله از بین پنج حوزه کلی: «دامپزشکی، کشاورزی، علوم پایه، علوم انسانی و مهندسی» به‌صورت دستی برچسب‌گذاری شد. لازم به یادآوری است که این دسته‌بندی از ابتدا در پیکره و توسط گردآوردندگان انجام شده است و در این پژوهش نیز همین دسته‌بندی ملاک عمل قرار گرفت. میزان پراکندگی موضوعی کل مقالات در جدول ۳ نشان داده شده است. البته لازم به یادآوری است که گاهی موضوع برخی مقالات به گونه‌ای بوده است که امکان قرار گرفتن در دو حوزه مختلف وجود داشته است و در صورت تغییر حوزه انتسابی مقاله، آمار ارائه شده در جدول ۲ ممکن است دستخوش تغییر گردد.

جدول ۳: توزیع مقالات در رشته‌های دانشگاهی

تعداد واژه‌های یکتا	تعداد مقالات	حوزه دانشگاهی
۵۲۷۸۰	۱۲۰	دامپزشکی
۱۸۹۱۳۱۱	۴۸۰	علوم انسانی
۳۴۲۹۲۷	۱۳۱	علوم پایه

مهندسی	۱۱۸	۳۳۷۶۸۰
کشاورزی، منابع طبیعی و آبخیزداری	۲۵۱	۸۸۴۲۱۲
مجموع	۱۱۰۰	۳،۵۰۸،۸۱۰

در گام چهارم، کلیه اعداد، کلمات انگلیسی، علائم سجاوندی و حروف ربط فارسی موجود در پیکره، از فرآیند شمارش و دسته‌بندی حذف گردیدند.

استخراج واژه‌ها براساس بسامد، بدون در نظر گرفتن دو اصل محدوده و وقوع تخصصی به یک فهرست قابل اعتماد منجر نمی‌شود. اما به صورت نمونه تعدادی از واژه‌های این فهرست اولیه در جدول شماره ۴ ارائه شده است. به طور مثال سه نوع از صورت‌های صرفی و یا اشتقاقی واژه «سطح» و در مجموع ۱۷۲۹۳ بار و یا واژه «نتیجه» در چهار صورت صرفی یا اشتقاقی و در کل ۱۵۶۰۳ بار در پیکره فردوسی شمارش شده است.

در گام پنجم، کلیه مقالات برحسب عنوان و موضوع در یکی از دسته‌های (علوم انسانی، کشاورزی، علوم پایه، مهندسی و دامپزشکی) برچسب‌گذاری شدند. علت دسته‌بندی در ۵ زیررشته این است که گردآورندگان پیکره مقالات دانشگاهی فردوسی کلیه مقالات را مطابق دسته‌بندی تعریف شده در آن دانشگاه به این صورت دسته‌بندی کرده‌اند.

جدول ۴: واژه‌های پربسامد پیکره فردوسی برحسب بسامد و تعداد صورت‌های واژه (بن واژه)

(تعداد هم خانواده) (بسامد)

کرد# کن (۱۳) (۲۷۰۶۶)، سطح (۳) (۱۷۲۹۳)، استفاده (۱) (۱۷۲۲۶)، نتیجه (۴) (۱۵۶۰۳)، درصد (۲) (۱۵۵۲۲)، میزان (۱) (۱۴۳۳۰)، افزایش (۱) (۱۴۲۲۰)، اثر (۳) (۱۳۶۸۵)، بود#باش (۵) (۱۳۴۶۲)، داشت#دار (۱۱) (۱۳۳۵۴)، بیشتر (۴) (۱۳۲۶۰)، جدول (۲) (۱۳۲۴۳)، بررسی (۱) (۱۲۳۱۵)، توجه (۱) (۱۲۲۹۷)، مطالعه (۲) (۱۲۰۶۷)، روش (۳) (۱۱۸۲۰)، عملکرد (۱) (۱۱۳۸۳)، اساس (۳) (۱۱۰۸۸)، شکل (۲) (۱۰۹۵۴)، مختلف (۲) (۱۰۷۸۷)، گشت#گرد (۳) (۱۰۷۵۰)، پژوهش (۵) (۱۰۳۷۸)، مدل (۴) (۱۰۳۵۱)، عنوان (۱) (۱۰۲۹۷)، عامل (۴) (۱۰۲۸۸)، مقدار (۳) (۱۰۲۴۷)، آب (۱) (۹۹۶۲)، تأثیر (۲) (۹۸۰۴)، نظر (۳) (۹۶۲۱)، رابطه (۳) (۹۵۲۸)، کاهش (۱) (۹۴۷۵)، فرد (۴) (۹۳۲۳)، شاخص (۳) (۹۱۶۳)، صورت (۱) (۹۰۲۹)، رشد (۲) (۸۸۱۹)، شرط (۳) (۸۷۳۸)، می‌توان (۱) (۸۷۱۷)، سال (۱) (۸۷۱۱)، مورد (۵) (۸۷۰۳)، شهر (۲) (۸۴۵۵)، علم (۳) (۸۴۳۰)، کشور (۳) (۸۳۰۸)، نوع (۳) (۸۲۹۲)، تغییر (۲) (۸۲۷۷)، منطقه (۲) (۸۲۴۰)، زمان (۲) (۸۲۰۶)، تحقیق (۲) (۷۹۶۹)، تعداد (۱) (۷۹۳۸)، دلیل (۳) (۷۷۹۹)، منبع (۳) (۷۷۷۰)، تیمار (۳) (۷۶۷۱)، دارا (۳) (۷۵۳۴)، تولید (۲) (۷۳۳۶)، تحلیل (۲) (۷۲۳۰)، ماده (۳) (۷۲۲۶)، وجود (۱) (۷۰۷۲)، اجتماع (۱) (۷۰۴۴)، همکار (۲) (۷۰۰۲)، بن (۱) (۶۶۲۶)، داده (۶) (۶۵۹۴)، دوره (۴) (۶۵۱۳)، اقتصادی (۱) (۶۴۸۸)، مهم (۵) (۶۴۶۸)، نمونه (۳) (۶۳۶۶)، عدم (۱) (۶۳۳۲)، انجام (۱) (۶۲۴۹)، خاک (۱) (۶۱۶۶)، زبان (۲) (۶۰۵۱)
--

در گام ششم واژه‌های پربسامد به تفکیک هر یک از این ۵ زیررشته استخراج شد.

در گام پایانی، با توجه به عدم توازن تعداد مقالات در هر پنج زیرشاخه دانشگاهی (جدول ۱-۱) ابتدا بسامد وقوع واژه‌ها در زیرشاخه‌های دانشگاهی برحسب تعداد بسامد واژه در آن زیرشاخه و تعداد کل مقالات و نیز واژه‌های پیکره، با فرض به هنجار بودن توزیع گردید. هدف از این به هنجارسازی، فراهم کردن امکان مقایسه‌ی نسبت‌ها جهت استخراج نهایی واژه‌ها می‌باشد. با توجه به اینکه تعداد مقالات در پنج حوزه دانشگاهی اشاره شده، نامتوازن است (در حوزه‌های دامپزشکی، مهندسی، علوم پایه، کشاورزی و علوم انسانی به ترتیب ۱۲۰، ۴۸۰، ۱۳۱، ۱۱۸ و ۲۵۱ مقاله گردآوری شده است) و به همین علت تعداد واژه‌ها و تعداد واژه‌های یکتا در این پنج بخش تفاوت زیادی دارند (به‌طور مثال خانواده «اثر» در حوزه‌های دامپزشکی، مهندسی، علوم پایه، کشاورزی و علوم انسانی به ترتیب ۳۱۶، ۱۵۰۴، ۲۰۸۳، ۹۰۱۴ و ۷۳۵۸ بار رخ داده است) برای اطمینان بیشتر از صحت فهرست پیشنهادی، از بسامد نسبی استفاده شده است. بسامد کلمه تعداد دفعاتی است که یک کلمه در یک متن یا پیکره ظاهر می‌شود. می‌توان آن را به‌عنوان یک بسامد مطلق، که تعداد خام رخدادها است بیان کرد، یا به‌عنوان یک بسامد نسبی، که نسبت وقوع به تعداد کل کلمات است، بیان کرد. ۱۰ خانواده واژگانی برحسب این به هنجارسازی در جدول شماره ۵ آمده است. برای برآورده شدن شرط پوشش (یعنی وجود واژه در هر پنج زیررشته)، واژه‌هایی که حداقل بسامد ۱۰۰ در همه زیررشته‌ها را داشتند، مجدداً استخراج شدند. در نهایت، با توجه به اینکه هدف ما شناسایی خانواده‌های واژگانی است، همه واژه‌های هم‌خانواده از نظر صرفی و معنایی در یک مرحله دیگر تجمیع و دسته‌بندی و نهایی شدند.

جدول ۵: ۱۰ خانواده واژگانی اول به‌صورت هنجار شده

خانواده واژگانی	بسامد در زیر شاخه‌های دانشگاهی					بسامد هنجار شده					
	علوم انسانی	کشاورزی	علوم پایه	مهندسی	دامپزشکی	بسامد	علوم انسانی	کشاورزی	درصد علوم	مهندسی	دامپزشکی
اثر، اثرات، اثرگذار، اثرگذاری، تأثیر، تأثیرگذار، متأثر	۷۳	۹۰۱۴	۲۰	۱۵۰	۳۱۶	۲۰۲	۱۳۶۵	۳۵۷۷	۲۱۳۱	۱۵۶۲	۲۱۰۰
داشتن، دارا، دارایی، دارا بودن	۸۳	۵۰۶۵	۱۶	۱۶۸	۳۱۰	۱۷۰	۱۵۵۲	۲۰۰۹	۱۶۵۵	۱۷۴۸	۲۰۶۰
	۷۲		۱۸	۳		۴۸	۳،۷۸	۹،۳۹	۵،۲۹	۷،۹۴	۸،۷۷

داد، دادن، داده، داده شدن، داده‌ها	۶۱ ۱۸	۵۶۰۵	۱۶ ۶۱	۱۷۴ ۸	۲۶۴	۱۵۳ ۹۶	۱۱۳۵ ۰٫۱۷	۲۲۲۴ ۲٫۲۷	۱۶۹۹ ۵٫۲۶	۱۸۱۶ ۳٫۳۵	۱۷۵۵ ۰٫۷
نتیجه، نتایج، نتیجه‌گیری، نتیجه گرفتن	۶۱ ۴۸	۵۵۴۳	۱۵ ۱۷	۱۷۲ ۸	۳۱۵	۱۵۲ ۵۱	۱۱۴۰ ۶٫۵۳	۲۱۹۹ ۶٫۲۳	۱۵۵۲ ۱٫۸۶	۱۷۹۵ ۵٫۵۳	۲۰۹۴ ۱٫۱۷
استفاده، استفاده کردن، استفاده شدن	۵۶ ۱۳	۵۲۹۱	۱۶ ۱۷	۲۲۷ ۲	۲۳۱	۱۵۰ ۹۴	۱۰۴۱ ۳٫۹۴	۲۰۹۹ ۶٫۲۲	۱۷۲۶ ۱٫۲۹	۲۳۶۰ ۸٫۲	۱۵۳۵ ۶٫۸۶
افزایش، افزودن، افزایش یافتن، افزایش دادن	۳۰ ۰۵	۸۴۸۸	۱۶ ۵۳	۱۳۲ ۳	۳۳۴	۱۴۸ ۰۳	۵۵۷۵، ۲۵	۳۳۶۸ ۲٫۸۵	۱۶۹۱ ۳٫۴	۱۳۷۴ ۷٫۲	۲۲۲۰ ۴٫۲۹
سطح، سطحی، سطوح	۳۸ ۶۱	۷۵۲۵	۱۸ ۴۴	۱۲۹ ۳	۲۶۳	۱۴۷ ۸۶	۷۱۶۳، ۴۰٫۸	۲۹۸۶ ۱٫۳۹	۱۸۸۶ ۷٫۷۱	۱۳۴۳ ۵٫۴۸	۱۷۴۸ ۴٫۲۲
متغیر، تغییر، تغییرات	۷۵ ۳۷	۳۴۳۴	۱۹ ۱۲	۱۶۹ ۹	۱۸۳	۱۴۷ ۶۵	۱۳۹۸ ۳٫۵۸	۱۳۶۲ ۷٫۱۱	۱۹۵۶ ۳٫۴۸	۱۷۶۵ ۴٫۱۹	۱۲۱۶ ۵٫۸۲
کار، کارکرد، کارایی، کاربر، کاربرد، کارکنان، کار کردن، کار داشتن	۷۴ ۳۲	۴۵۵۹	۱۷ ۹	۱۱۲ ۴	۱۲۲	۱۴۱ ۱۶	۱۳۷۸ ۸٫۷۷	۱۸۰۹ ۱٫۴۴	۸۹۹۳، ۸۷۹	۱۱۶۷ ۹٫۴۱	۸۱۱۰، ۵۵
عمل، عملکرد، عملیات	۴۰ ۳۸	۸۵۹۲	۶۲ ۲	۵۹۳	۲۳۸	۱۴۰ ۸۳	۷۴۹۱، ۸	۳۴۰۹ ۵٫۵۵	۶۳۶۴، ۲۶۹	۶۱۶۱، ۸۲۳	۱۵۸۲ ۲٫۲۲

۵- نتایج

هدف این پژوهش استخراج فهرست واژه‌های دانشگاهی در زبان فارسی براساس سه اصل بسامد، محدوده و وقوع تخصصی کاکسهد (۲۰۰۰) از یک پیکره مقالات دانشگاهی برای طراحی یک آزمون رایانه‌ای

انطباقی واژه‌های عمومی دانشگاهی می‌باشد. بدین منظور، ابتدا، میزان پوشش فهرست رضوانی در پیکره فردوسی توسط ابزار انتورد پروفایلر^۱ محاسبه گردید. نرم‌افزار انتورد پروفایلر، یک نرم‌افزار رایگان برای تجزیه و تحلیل سطح واژگان و پیچیدگی متون می‌باشد.

همان‌گونه که در جدول ۶ نشان داده شده است، میزان پوشش درصد واژه‌های یکتا^۲ فهرست رضوانی در پیکره شاهد در حدود ۲,۰۸ درصد محاسبه گردیده است که با نسبت ارائه شده برای واژه‌های دانشگاهی در مقالات این حوزه که در بخش پیشینه بررسی شد، تا حدودی همخوانی دارد. میزان پوشش فهرست پیشنهادی بر روی پیکره شاهد، معادل ۳,۸۶ درصد و بیشتر از میزان ۲,۰۸ پوشش فهرست رضوانی می‌باشد. در فرآیند استخراج سه اصل بسامد، محدوده و وقوع تخصصی مبنای استخراج قرار گرفتند و مشخص شد که میزان پوشش فهرست پیشنهادی بر روی پیکره شاهد بیشتر از همپوشانی فهرست رضوانی می‌باشد.

جدول ۶: اعتبارسنجی فهرست پیشنهادی و فهرست رضوانی

درصد واژه یکتا	واژه یکتا	درصد تعداد واژه‌ها	تعداد واژه‌ها		
۳,۸۶	۹۲۸	۴۵,۶۵	۲۴۵۶۳۶	موجود در فهرست پیشنهادی	فهرست پیشنهادی
۹۶,۱۴	۲۳۱۰۱	۵۴,۳۵	۲۹۲۴۲۵	ناموجود در فهرست پیشنهادی	
۱۰۰,۰	۲۴۰۳۵	۱۰۰,۰	۵۳۸۰۶۸	مجموع	
۲,۰۸	۵۰۱	۹,۳۳	۵۰۱۹۴	موجود در فهرست رضوانی	فهرست رضوانی
۹۷,۹۲	۲۳۵۳۴	۹۰,۶۷	۴۸۷۸۷۴	ناموجود در فهرست رضوانی	
۱۰۰,۰	۲۴۰۳۵	۱۰۰,۰	۵۳۸۰۶۸	مجموع	

با بررسی فهرست واژه‌های رضوانی و فهرست پیشنهادی این پژوهش و نیز بررسی واژه‌های ۵٪ اول در پیکره فردوسی، می‌توان ادعا کرد که در زبان فارسی، سازه و مفهوم «واژه‌های دانشگاهی» برخلاف زبان فرانسه و مانند زبان‌های انگلیسی و ترکی مفهومی مجزا از واژه‌های عمومی است و یک دانشجوی غیرفارسی‌زبان که قصد ادامه تحصیل در دانشگاه‌های ایران را دارد باید علاوه بر واژه‌های عمومی زبان فارسی، واژه‌های دانشگاهی عمومی را نیز آموزش ببیند.

از طرف دیگر با هدف بررسی تعلق یا عدم تعلق واژه‌های مستخرج این پژوهش به فهرست واژه‌های عمومی، دو پژوهش علمی واژگان پایه زبان فارسی (نعمت‌زاده و همکاران، ۱۳۸۰) و واژه‌های پایه زبان فارسی (صحرايي و همکاران، ۱۳۹۶) انتخاب گردیدند. پس از بررسی واژگان این دو مجموعه، میزان همپوشانی آن‌ها بایکدیگر بیش از ۷۰ درصد محاسبه گردید و ۳۷۲ واژه‌ی مشترک در هر دو فهرست،

1. AntwordProfiler

2. type

به‌عنوان واژه‌های عمومی زبان فارسی در استخراج نهایی فهرست واژه‌های دانشگاهی مبنای مقایسه قرار گرفتند. فهرست این واژه‌های مشترک در پیوست (۱) ارائه شده است. نتایج بررسی نشان داد که فهرست پیشنهادی این پژوهش، اشتراکی با واژه‌های عمومی زبان فارسی ندارند.

به‌عبارت دیگر در پاسخ به سؤال اول پژوهش «مفهوم فارسی دانشگاهی در قیاس با فارسی عمومی چه جایگاهی دارد؟» می‌توان ادعا کرد که واژگان عمومی دانشگاهی در زبان فارسی همانند زبان انگلیسی از فهرست واژه‌های عمومی زبان فارسی مجزا هستند و باید فهرست‌های واژه‌های دانشگاهی به‌عنوان یک مؤلفه مجزا و مهم در آموزش زبان فارسی به غیرفارسی‌زبانان اهداف دانشگاهی و نیز در فرآیند تهیه و تدوین مواد آموزشی، مدنظر پژوهشگران و مدرسان قرار بگیرد.

در پاسخ به سؤال دوم پژوهش «میزان اعتبار فهرست پیشنهادی در مقایسه با فهرست رضوانی به‌عنوان اولین فهرست واژه‌های دانشگاهی زبان فارسی چقدر است؟» با استناد به جدول ۴ می‌توان بیان کرد که فهرست پیشنهادی این پژوهش با درصد پوشش ۳۸٫۶٪ بر روی پیکره شاهد در مقایسه با فهرست رضوانی با پوششی معادل ۲۸٪، از اعتبار بیشتری برخوردار می‌باشد.

۶- تحلیل داده‌ها

استخراج یک فهرست از خانواده واژگانی پربسامد در متون علمی دانشگاهی از یک پیکره با هدف طراحی یک آزمون انطباقی رایانه‌ای، دلیل اصلی این پژوهش بوده است. لذا پس از اتمام فرآیند مطالعه، فهرستی از واژه‌های زبان فارسی که در متون علمی مانند مقالات و کتب دانشگاهی پربسامد هستند پیشنهاد گردید. ویژگی این فهرست پوشش بیشتر آن در مقایسه با فهرست مشابه می‌باشد. این فهرست در جدول ۷ ارائه شده است.

پس از بررسی دو فهرست مشاهده شد که حدود ۰/۰۴ درصد واژه‌های فهرست رضوانی در پیکره فردوسی وجود ندارند که در برخی از این واژه‌ها می‌توان به عدم استاندارد بودن نویسه‌ها و یا چندبخشی بودن واژه‌ها اشاره کرد. این واژه‌ها عبارتند از: «ارائه، ازطرف، به‌دست‌آمدن، به‌دست‌آوردن، بسنده کردن، به‌صورت، به‌علاوه، به‌ندرت، جابه‌جا، درمعرض، حاصل از، حاصل شدن، حاکی از، در خلال، سلسله مراتب، قابل استفاده، قابل توجه، قائل، مربوط به، مسئول، مورد توجه، نسبت دادن، وقوح»^۱.

در عین حال با بررسی بیشتر معلوم شد که موارد متعددی از فهرست رضوانی در پیکره فردوسی بسیار کم بسامد هستند و برخی از این واژه‌های کم بسامد در جدول شماره ۷ ارائه شده‌اند. کم‌بسامد بودن این واژه‌ها

۱. در پیوست مقاله اصلی رضوانی و همکاران، این واژه درج شده است که احتمالاً واژه «وقوع» مدنظر بوده و به‌علت غلط تایپی «وقوح» درج شده است.

ضریب، ضرایب (۴۰۴۳)	مصرف، مصرفی (۴۸۴۷)	دوره، دوران، دوره (۶۶۰۷)	قرار دادن، قرار گرفتن، قرار، قرار داد (۱۰۰۳۹)
ترتیب (۴۰۴۳)	وزن (۴۸۲۸)	اقتصاد، اقتصادی (۶۵۰۴)	توجه (۹۸۷۳)
امور، امر، امری (۳۹۸۷)	نمود، نمودار، نمودن، نموده (۴۷۹۹)	رابطه، رابطه داشتن (۶۲۳۰)	تعداد، تعدادی، متعدد، متعددی، عدد، عددی (۹۷۳۴)
مسئول، مسیول، مسیولیت، مساله، مسائل (۳۹۷۷)	آزمون، آزمون‌های (۴۷۸۷)	نمونه، نمونه برداری، نمونه گیری (۶۲۲۳)	پژوهش، پژوهشگر (۹۷۳۰)
آمار، آماره، آماری (۳۹۵۷)	بخش، بخش‌های، بخشی (۴۷۵۶)	دلیل، دلایلی، دلیلی، دلیل آوردن، دلیل داشتن (۶۱۳۳)	مقدار، مقادیر (۹۷۳۰)
کل (۳۸۸۶)	منطقه (۴۶۵۴)	همکار، همکاری، همکاران (۶۰۸۱)	کشت، کشاورز، کشاورزی، کشاورزان (۹۵۲۹)
باعث (۳۸۷۱)	فرهنگ، فرهنگی (۴۶۰۰)	مربوط، مرتبط، مربوطه (۶۰۴۲)	شکل، شکل گرفتن، شکل گیری (۹۵۲۳)
تحت (۳۸۷۰)	بیان (۴۵۹۹)	مجموع، مجموعه، جمع، جمع آوری، تجمع، جمعیت (۶۰۳۳)	اجتماع، اجتماعی، جامعه، جوامع (۹۱۶۷)
کیفی، کیفیت (۳۸۶۷)	انسان، انسانی (۴۵۹۳)	نظر (۵۹۲۹)	مختلف، مختلفی (۹۱۴۶)
منظور (۳۸۲۵)	افراد (۴۵۴۴)	تحلیل، تحلیلی (۵۹۱۲)	صورت گرفتن (۸۸۵۵)
آموزش، آموزشی (۳۸۲۲)	مقایسه، مقایسه کردن (۴۵۳۱)	متفاوت، متفاوت بودن، تفاوت، تفاوت داشتن (۵۹۰۳)	عامل، عاملی (۸۵۴۷)
محصول، محصولات (۳۷۶۳)	فعل، فعالیت، فعل و انفعال، فعال (۴۴۳۳)	حالی، حالت، حالی، حالی که (۵۸۳۶)	شاخص، شاخص سازی، شاخص بودن (۸۳۹۱)
ارزیابی (۳۷۱۱)	طول، طولانی، طولی (۴۴۲۸)	گروه، گروهی (۵۷۸۷)	عنوان (۸۱۳۶)
درجه (۳۶۵۸)	آزمایش، آزمایشی، آزمایشگاه، آزمایشات، آزمایشگاهی (۴۴۱۶)	جهت (۵۷۵۰)	مدل، مدل‌های (۷۹۵۹)
شماره، شماره (۳۶۴۹)	ارتباط، ارتباطات، ارتباطی (۴۴۰۲)	نوع، نوعی (۵۵۸۳)	رشد، رشدی، رشد کردن (۷۷۰۶)
تنش (۳۶۳۶)	شناخت، شناسایی، شناسی، شناختی، شناختن (۴۳۵۶)	ویژه، ویژگی (۵۴۲۶)	اساس، اساسی (۷۶۰۴)
نقش، نقشی (۳۶۳۱)	شامل (۴۳۴۶)	منابع (۵۴۱۶)	حد، حداکثر، حداقل، حدود، حدودی (۷۴۳۲)
اندازه، اندازه گیری، اندازه گیری شد (۳۵۶۲)	تعیین، تعیین کننده (۴۳۴۱)	محیط زیست، محیط، محیطی (۵۳۰۴)	شرایط (۷۳۸۷)
مانند (۳۵۳۷)	مرحله، مراحل (۴۲۴۴)	مهم، مهمترین، مهمی (۵۲۱۹)	خشک، خشکی، خشکسالی (۷۲۳۷)
مبنی، مبنا، مبنی بر، مبتنی، مبانی (۳۵۳۴)	مشاهده، مشاهدات، مشاهده شده (۴۲۲۶)	اطلاعات، اطلاعاتی (۵۱۹۳)	وجود، وجودی، وجود داشتن (۷۰۰۰)
الگو، الگوهای، الگوی (۳۵۲۴)	توسعه (۴۲۱۴)	عدم (۵۱۰۸)	تولید، تولیدات، تولیدی (۶۸۸۷)
طریق (۳۵۱۹)	حضور، حاضر (۴۱۳۸)	توسط (۵۰۷۱)	علم، علمی، علوم (۶۸۵۶)
خصوص، خصوصی، خصوصیات، خواص (۳۴۹۴)	مناسب، مناسبی، نامناسب (۴۰۴۴)	بالا، بالاتر، بالاترین، بالای، بالایی (۴۸۸۶)	اصل، اصلی، اصول (۶۷۰۳)

محاسبه، محاسبات، محاسبه کردن (۳۴۱۷)	کلی (۲۵۳۹)	استراتژی، راه، راهبرد، راهبردها (۲۹۵۷)	محاسبه، محاسبات، محاسبه کردن (۳۴۱۷)
نیاز، نیازمند (۳۳۹۳)	همبستگی (۲۵۱۸)	مشخص، مشخصات (۲۹۶۱)	نیاز، نیازمند (۳۳۹۳)
فرد، فردی (۳۳۳۸)	بدست آوردن (۲۵۱۵)	تنها، تنهایی (۲۹۴۵)	فرد، فردی (۳۳۳۸)
رفتار، رفتارها، رفتار کردن (۳۳۳۸)	سیاست، سیاسی، سیاستگذاری (۲۵۰۵)	ارایه شده است، ارایه شده (۲۹۴۱)	رفتار، رفتارها، رفتار کردن (۳۳۳۸)
سیستم (۳۳۲۲)	زیست، زیستی، زیست محیطی (۲۴۴۲)	یافتن، یافته، یافته‌ها، یافته‌های (۲۹۳۹)	سیستم (۳۳۲۲)
موضوع، موضوعات، موضوعی (۳۳۰۶)	طی (۲۴۳۸)	ماده (۲۹۱۳)	موضوع، موضوعات، موضوعی (۳۳۰۶)
مؤثر (۲۲۹۴)	لذا (۲۴۲۱)	جهان، جهانی (۲۸۸۵)	مؤثر (۲۲۹۴)
ترجمه (۳۲۸۴)	هوا، هوایی، هوای (۲۴۱۷)	چند، چندان، چندانی، چندگانه، چندین (۲۸۵۹)	ترجمه (۳۲۸۴)
واقع، واقعی، واقعیت (۳۲۸۰)	موجب (۲۴۱۳)	سال، سالانه (۲۸۲۵)	واقع، واقعی، واقعیت (۳۲۸۰)
تاریخ، تاریخی (۳۲۴۷)	فضا، فضایی (۲۴۱۰)	پردازش، پرداختن، پرداخته (۲۸۱۱)	تاریخ، تاریخی (۳۲۴۷)
عبارت بودن از، عبارت (۳۲۰۰)	خاص، خاصی (۲۴۰۳)	موردبررسی (۲۸۱۱)	عبارت بودن از، عبارت (۳۲۰۰)
انتشار، انتشارات (۳۱۹۷)	بهبود (۲۳۹۷)	اهمیت (۲۷۷۴)	انتشار، انتشارات (۳۱۹۷)
گونه، گونه‌ها (۳۱۸۰)	شخص، شخصی، شخصیت (۲۳۹۵)	شاهد، شواهد (۲۷۶۳)	گونه، گونه‌ها (۳۱۸۰)
مفهوم، مفاهیم، مفهومی (۳۱۶۵)	پیش فرض، فرض، فرضیه (۲۳۸۷)	هدف (۲۷۵۲)	مفهوم، مفاهیم، مفهومی (۳۱۶۵)
ترکیب، ترکیبات، ترکیبی (۳۱۴۰)	داخل، داخلی (۲۳۵۹)	وضع، وضعیت (۲۷۵۱)	ترکیب، ترکیبات، ترکیبی (۳۱۴۰)
برنامه، برنامه‌های، برنامه ریزی (۳۱۳۷)	مناطق (۲۳۴۶)	معیار (۲۷۴۸)	برنامه، برنامه‌های، برنامه ریزی (۳۱۳۷)
طرح، طراحی (۳۱۰۳)	صفت، صفات (۲۳۳۴)	سرعت (۲۷۳۷)	طرح، طراحی (۳۱۰۳)
مثبت، مثبتی (۳۰۹۷)	گزارش، گزارش کردن (۲۳۱۲)	اختلاف، اختلافات (۲۷۰۲)	مثبت، مثبتی (۳۰۹۷)
حق، حقیقت، حقوق (۳۰۸۹)	موارد، مواردی (۲۳۱۰)	انرژی (۲۷۰۰)	حق، حقیقت، حقوق (۳۰۸۹)
ایجاد، ایجاد می شود (۳۰۸۳)	لازم (۲۲۹۵)	انتخاب (۲۶۹۹)	ایجاد، ایجاد می شود (۳۰۸۳)
دید، دیدگاه، دیدن (۳۰۷۷)	کنترل (۲۲۷۹)	محدود، محدود بودن، محدود، محدوده، محدودیت (۲۶۹۲)	دید، دیدگاه، دیدن (۳۰۷۷)
احتمال، احتمالاً، احتمالی (۳۰۶۵)	برآورد (۲۲۶۰)	ریشه (۲۶۶۴)	احتمال، احتمالاً، احتمالی (۳۰۶۵)
برابر، برابری (۳۰۱۱)	جریان (۲۲۳۷)	جز، جزء، جزو، جزیبی، جزییات (۲۶۱۷)	برابر، برابری (۳۰۱۱)
کامل، کاملاً (۲۹۸۸)	از جمله (۲۲۲۴)	اجرا، اجرا شد، اجرای، اجرایی (۲۶۱۲)	کامل، کاملاً (۲۹۸۸)
جنس، جنسیت (۵۹۹۹)	ابتدا، ابتدای، ابتدایی (۲۲۰۶)	مرکز، مرکزی (۲۵۹۱)	جنس، جنسیت (۵۹۹۹)
	جدید، جدیدی (۲۱۹۵)	متوسط (۲۵۸۳)	
		سبب (۲۵۵۱)	
برخوردار، برخورداری، برخورداری (۲۱۹۰)			
بعد (۲۱۸۶)			
پایدار، پایداری (۲۱۶۴)			
فراوان بودن، فراوانی (۲۱۶۲)			
پاسخ، پاسخگویی (۲۱۴۴)			
به علاوه، علاوه بر (۲۱۴۲)			
عناصر (۲۱۳۹)			
طبیعی (۲۱۰۲)			
منفی (۲۰۹۵)			
فاصله، فاصله داشتن، فاصله دار، فواصل (۲۰۸۷)			
پارامتر، پارامترهای (۲۰۸۰)			
رسیدن (۲۰۷۲)			
علت، علل (۲۰۶۲)			
همراه (۲۰۶۰)			
حوزه، حوزه‌های (۲۰۵۱)			
زمینه، زمینه داشتن (۲۰۴۷)			
منجر (۲۰۴۲)			
جای، جایگزین، جایگاه، جایگزینی، جای، جایی (۲۰۳۹)			
مشکل، مشکلات (۲۰۳۲)			
بزرگ، بزرگتر، بزرگ بودن (۲۰۳۲)			
ارزش، ارزش‌های (۲۰۲۰)			
خدمات (۱۹۹۵)			
سرمایه، سرمایه گذاری (۱۹۸۸)			

انواع (۱۴۱۲)	سن (۱۵۴۶)	فن، فناوری، فنی (۱۷۶۱)	دقت کردن، دقیق، دقیق
عمق، عمیق (۱۳۹۹)	چهار (۱۵۴۳)	خط، خطی (۱۷۴۷)	شدن، دقت (۱۹۷۲)
مقدمه (۱۳۹۶)	مکان، مکانی، مکان	آزاد، آزادی (۱۷۴۵)	تصویر، تصاویر (۱۹۷۱)
صنعت، صنعتی (۱۳۹۳)	یابی (۱۵۳۳)	پایه (۱۷۳۹)	جذب (۱۹۵۸)
برش، برشی (۱۳۹۳)	معادله (۱۵۲۶)	صحت، صحیح (۱۷۳۴)	چاپ (۱۹۴۶)
تصادفی (۱۳۹۲)	حفظ، حفظ کردن، حفاظت (۱۵۱۹)	برداشت (۱۷۲۸)	بهره، بهره برداری، بهره گیری، بهره وری (۱۹۴۵)
نیرو (۱۳۸۸)	توصیف، توصیفی (۱۵۱۵)	شرکت (۱۷۲۵)	امکان، امکانات (۱۹۴۴)
ابزار، ابزارهای، ابزاری (۱۳۸۴)	ضروری، ضرورت (۱۵۰۷)	ثابت (۱۷۲۰)	روابط (۱۹۳۶)
حس، حساس، حساسیت (۱۳۸۲)	متقابل (۱۵۰۷)	قدرت (۱۷۰۵)	توزیع (۱۹۳۶)
قاعده، قواعد، علی القاعده (۱۳۷۳)	قابل (۱۵۰۶)	مورد استفاده (۱۷۰۰)	بحث (۱۹۲۳)
درک (۱۳۶۹)	درآمد (۱۵۰۴)	نفر (۱۶۹۶)	فصل، فصلنامه (۱۹۱۸)
سمت (۱۳۶۵)	تشکیل (۱۵۰۴)	رویگرد، رویگردهای (۱۶۷۹)	تجزیه (۱۹۱۲)
دریافت، دریافتن (۱۳۵۰)	بهینه، بهینه سازی (۱۵۰۲)	فرآیند (۱۶۶۸)	ناشی (۱۹۰۲)
وسیله، وسایل (۱۳۴۸)	امنیت، امنیتی (۱۴۹۸)	شدت (۱۶۶۳)	درون، درونی (۱۹۰۱)
ناحیه (۱۳۳۸)	مطرح (۱۴۹۷)	براساس (۱۶۶۲)	آثار (۱۸۸۷)
پیش‌بینی (۱۳۳۰)	نهایی، نهایت (۱۴۹۱)	جمله، جملات (۱۶۳۹)	غذایی (۱۸۷۲)
ادبیات (۱۳۲۸)	دولت (۱۴۹۰)	حجم (۱۶۱۹)	هزینه، هزینه‌ها (۱۸۷۰)
یافت (۱۳۲۷)	انتقال (۱۴۸۷)	لحاظ (۱۶۱۲)	عمومی (۱۸۵۸)
مؤلفه، مؤلفه‌های (۱۳۲۱)	وابسته (۱۴۸۶)	مستقیم (۱۶۰۸)	نرخ (۱۸۵۰)
اهداف (۱۳۱۴)	تجربه، تجربی (۱۴۸۳)	هر دو (۱۶۰۰)	مردم (۱۸۴۹)
معمول، معمولاً، معمولی (۱۳۱۲)	شبکه (۱۴۸۳)	دسترس، دسترسی داشتن (۱۵۸۳)	ممکن (۱۸۴۶)
معتقد، معتقدند (۱۳۱۰)	واریانس (۱۴۷۸)	طبقه، طبقه بندی، طبقه بندی کردن، طبقات (۱۵۸۲)	مزرعه، مزارع (۱۸۴۴)
حرکت، حرکات، حرکتی (۱۳۰۵)	یکدیگر (۱۴۶۷)	مقاله (۱۵۸۱)	مالی (۱۸۳۲)
قابلیت (۱۳۰۵)	نقل (۱۴۵۱)	رقم (۱۵۷۱)	مدت، مدت زمان (۱۸۲۳)
ارقام (۱۲۹۷)	مذکور (۱۴۵۰)	خوب، خوبی (۱۵۶۵)	قانون، قوانین، قانون گذاری (۱۸۰۷)
مقیاس (۱۲۹۵)	تفسیر (۱۴۴۷)	مشابه، مشابهی (۱۵۵۷)	اشاره، اشاره کرد، اشاره دارد (۱۸۰۷)
خالص (۱۲۹۱)	تامین (۱۴۴۱)	اقليم، اقلیمی (۱۵۵۶)	گسترش، گسترده، گستره (۱۸۰۴)
	نرم‌افزار (۱۴۳۵)	مشارکت (۱۵۴۶)	توان، توانایی (۱۷۹۵)
	ابعاد (۱۴۳۳)		
	کمی (۱۴۱۴)		

سنجش (۹۲۴)	درواقع (۱۰۵۷)	پیوند، پیوسته،	کلمه، کلمات (۱۲۹۰)
وقوع (۹۲۱)	پوشش (۱۰۵۵)	پیوستگی (۱۱۷۵)	مطلوب (۱۲۸۹)
بکار، بکارگیری (۹۱۹)	حاکمی (۱۰۵۲)	اعمال (۱۱۷۳)	به نظر می‌رسد (۱۲۷۰)
موفق، موفقیت (۹۱۶)	رفتن (۱۰۴۹)	فیزیکی (۱۱۶۸)	فشار، فشاری (۱۲۶۹)
نقاط (۹۱۵)	ظرفیت (۱۰۴۷)	پیشنهاد،	معادل (۱۲۶۷)
نقطه (۹۱۰)	پتانسیل (۱۰۴۵)	پیشنهادی (۱۱۶۵)	فراهم (۱۲۶۴)
مورد نیاز (۹۰۸)	در نهایت (۱۰۴۰)	مورد نظر (۱۱۶۳)	محور، محوری (۱۲۶۴)
تشخیص (۹۰۲)	سبک (۱۰۲۷)	چگونه، چگونگی (۱۱۵۸)	آگاه (۱۲۵۸)
نحوه (۸۹۷)	بحران، بحرانی (۱۰۱۹)	تجارت، تجاری (۱۱۳۴)	ضعف، ضعیف (۱۲۵۵)
نمایان، نمایش،	اخیر (۱۰۱۳)	تایید (۱۱۳۳)	نشر، نشریه (۱۲۴۹)
نماید (۸۸۸)	قطع، قطعه،	طرف (۱۱۳۱)	موسسه (۱۲۴۷)
خطای (۸۸۰)	قطعات (۱۰۱۲)	استخراج، استخراج	موقعیت (۱۲۴۷)
دستگاه (۸۸۰)	اگرچه، اگر چه (۱۰۰۴)	شده (۱۱۳۰)	بیانگر (۱۲۴۲)
توده (۸۷۱)	دقیقه (۱۰۰۴)	فقط (۱۱۲۶)	فوق (۱۲۳۹)
رنگ، رنگی (۸۶۹)	نکته (۱۰۰۲)	چکیده (۱۱۲۲)	مثل، مثال، مثلاً (۱۲۳۴)
اثبات (۸۶۵)	محل (۱۰۰۰)	بین المللی، بین	دسته (۱۲۳۳)
پدیده (۸۶۵)	اقدام، اقدامات (۹۹۳)	المللی (۱۱۱۶)	حل، حلال (۱۲۳۱)
قوت، قوی (۸۵۱)	ماه (۹۹۲)	نظیر (۱۱۰۵)	اجزای، اجزاء (۱۲۳۱)
مواجهه، مواجهه (۸۴۳)	شیب (۹۹۲)	نویسنده،	منبع (۱۲۲۸)
بشر (۸۴۲)	تیبین (۹۸۸)	نویسندگان (۱۰۹۶)	مطابق، مطابق (۱۲۲۴)
قابل توجه، قابل	تعریف (۹۸۱)	تهیه، تهیه شده (۱۰۹۵)	قالب (۱۲۱۸)
توجهی (۸۴۰)	ظاهر، ظاهری (۹۸۰)	پیشین، پیشینه (۱۰۹۴)	به دست آمد، به دست
تأکید (۸۳۷)	خاوی (۹۷۳)	مقاومت (۱۰۹۲)	آمده، به دست
حمایت (۸۳۵)	استاندارد (۹۶۷)	جغرافیایی (۱۰۹۱)	می‌آید (۱۲۱۸)
جو (۸۳۵)	کافی (۹۶۴)	کلیدی (۱۰۸۹)	شرح (۱۲۱۱)
مجله (۸۳۳)	ذکر (۹۶۲)	تمایل (۱۰۸۷)	ذرات (۱۱۹۸)
اختیار (۸۳۲)	نگاه، نگاهی (۹۴۶)	رگرسیون (۱۰۸۷)	مستقل (۱۱۹۶)
تقریباً (۸۳۲)	سلامت (۹۴۳)	مشترک (۱۰۸۵)	تکرار (۱۱۹۶)
هدایت (۸۲۹)	کوتاه، کوتاه مدت (۹۴۳)	مسیر (۱۰۷۶)	عمده، عمدتاً (۱۱۸۸)
بانک (۸۲۷)	مثال (۹۳۵)	کمک (۱۰۷۵)	رقابت، رقابتی، رقابت
لایه (۸۲۶)	اطمینان (۹۲۵)	محتوی، محتوا (۱۰۶۹)	کردن (۱۱۸۷)
		خروج، خروجی (۱۰۶۰)	طبق (۱۱۸۷)

مقابل (۵۵۵)	موردی (۶۷۰)	بروز (۷۶۷)	پذیرش (۸۲۵)
غیر مستقیم (۵۵۲)	راستای (۶۶۴)	نواحی (۷۵۸)	وسیع، وسعت، وسیعی (۸۲۳)
رخ، رخ دادن، رخداد (۵۵۱)	درخصوص (۶۵۱)	مفید (۷۵۷)	صفحه، صفحات (۸۲۲)
تمرکز (۵۴۹)	گام (۶۵۱)	نگهداری (۷۵۶)	آلودگی (۸۱۷)
تفکیک (۵۴۹)	فاکتور (۶۴۶)	سخت، سخت بودن (۷۵۳)	جدا، جداگانه (۸۱۵)
حاشیه (۵۴۹)	آسیب، آسیب پذیری (۶۴۰)	بعضی (۷۵۲)	سازگار، سازگاری داشتن (۸۱۲)
نفوذ (۵۴۲)	ملاحظه (۶۱۴)	انحراف (۷۵۲)	فرعی (۸۰۹)
راستا (۵۳۶)	تلاش (۶۱۰)	درزمینه (۷۴۸)	ذیل (۸۰۳)
غالب (۵۳۵)	خسارت (۶۱۰)	همزمان (۷۴۷)	کسب (۷۹۷)
انتظار (۵۲۶)	ریسک (۶۱۰)	صرفاً (۷۴۵)	کمبود (۷۹۵)
بسته (۵۲۲)	همواره (۶۰۷)	حساب (۷۴۵)	ضمن (۷۹۴)
بازگشت (۵۱۸)	به‌عنوان (۶۰۷)	حذف (۷۳۷)	تعرق (۷۹۲)
متناسب (۵۱۱)	گوناگون (۶۰۵)	مطلب (۷۳۴)	برازش (۷۹۱)
تعلق (۴۹۳)	رشته (۵۹۶)	تحصیلات (۷۳۴)	بستر (۷۹۱)
اشکال (۴۹۰)	ماتریس (۵۹۲)	واکنش (۷۳۱)	ورود (۷۸۹)
زاویه (۴۹۰)	ستنی (۵۸۹)	مقطع (۷۳۰)	جنبه، جنبه‌های (۷۸۸)
ثبات (۴۸۶)	اصلاح (۵۸۸)	جانب، جانبی (۷۲۸)	نرمال (۷۸۷)
پیشرفت (۴۸۵)	برخلاف (۵۸۷)	ذات (۷۲۴)	روشن (۷۸۳)
نرم (۴۸۵)	خوشه (۵۸۶)	حاکم (۷۱۶)	درمقابل (۷۸۳)
تمایز (۴۸۴)	قرن (۵۸۵)	مانع، موانع، مانع شدن (۷۱۶)	دستیابی (۷۸۳)
فرمول (۴۸۳)	مساحت (۵۸۴)	رایج (۷۱۵)	تبدیل (۷۸۳)
مهندسی (۴۸۲)	اعتقاد (۵۸۳)	جامع (۷۱۴)	ستون (۷۸۲)
برخورد (۴۸۱)	پرورش (۵۸۱)	اختلال (۷۱۴)	منحنی (۷۸۱)
تقسیم (۴۷۸)	ارتقاء، ارتقای (۵۷۴)	ماهیت (۷۱۱)	دنبال (۷۷۵)
ازای، ازاء (۴۷۶)	فاقد (۵۷۴)	معرفی (۶۹۲)	جرم (۷۷۴)
مباحث (۴۷۲)	جلوگیری (۵۷۲)	توضیح (۶۹۱)	قسمت (۷۷۱)
رعایت (۴۷۱)	تعادل (۵۷۲)	اشکار (۶۷۸)	پیدایش، پیدا کردن (۷۶۹)
ثانی، ثانویه (۴۷۰)	خطر (۵۶۱)	تخمین (۶۷۷)	دامنه (۷۶۹)
منطقی (۴۶۱)	دچار (۵۵۷)	اندک (۶۷۴)	اعتبار (۷۶۸)
ادامه (۴۶۱)	بهداشت (۵۵۷)		

مجدد (۴۶۰)	اطراف (۴۴۶)	اتصال (۴۳۳)	در نظر گرفته شد (۴۱۶)
تنظیم (۴۵۵)	واسطه (۴۴۱)	بایستی (۴۳۲)	انتهای (۴۰۷)
تدریج (۴۵۱)	معین (۴۴۱)	مدنظر (۴۲۷)	در صورتی (۴۰۵)
خلاءصه (۴۴۹)	پیچیده (۴۴۰)	وابستگی (۴۲۰)	ذخیره (۴۰۰)
تکنیک (۴۴۹)	تقویت (۴۳۸)	تاکنون (۴۲۰)	

استخراج فهرست واژه‌های دانشگاهی در زبان فارسی با چندین محدودیت ناشی از ویژگی رسم‌الخط و ساختار زبان فارسی مواجه است. به‌طور مثال می‌توان به‌مواردی چون افعال مرکب، چندبخشی بودن واژه‌ها، مقوله شعاعی یا معنی سرنمون واژه، عدم نمایش نویسه‌ها، واژه‌های مرکب، ایست واژه‌ها، طبقه‌بندی کلان رشته‌های دانشگاهی اشاره کرد. برای تشریح دقیق این محدودیت‌ها در ادامه، مثال‌های واقعی ارائه می‌شوند.

در مورد افعال مرکب، می‌توان واژه «تحقیق» که هم به‌صورت اسم و هم به‌صورت همراه فعل مرکب «تحقیق کردن» در متون علمی به زبان فارسی به‌صورت مکرر استفاده می‌شود و ۸۱۰۸ بار در پیکره فردوسی به‌کار رفته است اما جز با رجوع به بافت یا برچسب‌گذاری نحوی با ضریب اطمینان بالا که بتواند آن را به‌عنوان اسم یا همراه فعل مرکب برچسب‌گذاری کند، نمی‌توان درباره آن تصمیم‌گیری کرد. لذا ضرورت دارد که در ارایه فهرست واژه‌های دانشگاهی در مورد اینکه واژه تحقیق با مقوله اسم چه رتبه‌ای دارد و به‌عنوان هم‌کرد فعل مرکب چه مقدار بسامد دارد، بررسی و بحث بیشتری شکل بگیرد. بنا به محدودیت‌های این پژوهش، مقوله دستوری واژه (POS) جز ملاک‌های دسته‌بندی نبوده است، و لذا «تحقیق» را با «محقق، محققان، محققین، تحقق، تحقق یافتن (۱۱۹۹۰)» در یک خانواده قرار داده‌ایم. سرواژه «فرض» در فهرست رضوانی برای سه صورت «فرضی»، «فرضیه»، «فرضیه‌ها» و «فرض‌ها» انتخاب شده است در حالی که از نظر شمول و بار معنایی بین «فرض» و «فرضیه» و کاربردهای آن‌ها تفاوت‌های مهمی وجود دارد. از دیگر سو در زیر خانواده‌ی «فرض»، کلمه مرکب «پیش‌فرض» که با رجوع به پیکره به‌نظر می‌رسد در متون علمی بسامد بالایی داشته باشد، وجود ندارد. ترکیب «پیش‌فرض» در پیکره فردوسی ۲۳۹ بار ظاهر شده است و از طرف دیگر واژه «پیش» بیش از ۲۲۶۱ به‌تنهایی و بدون حضور در یک واحد چند واژه‌ایی با مقوله‌هایی مانند حرف اضافه، صفت، اسم در این پیکره شمارش شده است. «پیشینه» که در بسیاری از مقالات علمی غالباً یک بخش مشخص و ویژه برای اشاره به مطالعات انجام شده قبلی دارد، در پیکره فردوسی ۶۴۰ بار در صورت‌های مختلف تصریفی تکرار شده است اما در فهرست رضوانی وجود ندارد. با توجه به کاربرد آتی این فهرست در بافت دانشگاهی و کاربردهای آن، واژه «پیش‌فرض» با «فرضیه، فرض کردن، فرض» در یک دسته قرار گرفتند. فعل مرکب «فرض کردن» ۱۴۲ بار در پیکره فردوسی تکرار شده است اما در زیرخانواده «فرض» در فهرست رضوانی قید نشده است. دلیل

این عدم شمول «پیش فرض» یا «فرض کردن» در این زیر خانواده، می‌تواند استاندارد نگارش خط فارسی و به‌ویژه درج یا عدم درج نویسه نیم فاصله باشد که در تقطیع باعث شده این دو بخش، دو واژه مجزا لحاظ شوند. همین بحث درباره سرواژه «ظاهر» و زیر خانواده «ظاهر شدن» در فهرست رضوانی می‌توان اقامه کرد.

از مشکلات دیگر مرتبط با رسم‌الخط زبان فارسی، عدم نمایش واکه‌هاست که باعث می‌شود واژه‌های هم‌نویسه را به‌درستی نتوان تحلیل و بررسی کرد و دسته‌بندی و احصای آن‌ها بدون در نظر گرفتن بافت، مشکل‌ساز خواهد بود. مانند واژه «بعد» که در پیکره فردوسی ۳۰۹۷ بار تکرار شده است، اما بدون رجوع به بافت نمی‌توان گفت منظور «بُعد» است یا «بَعْد». «گرم» نیز ۲۵۲۱ تکرار شده است. به‌طور مثال واژه «بعد» بدون اعراب را اگر «بُعد» در نظر بگیریم باید با «ابعاد» هم خانواده شوند و اگر «بَعْد» به معنای «پس از» باشد باید جز ایست واژه‌ها در نظر گرفته شود. واژه‌های «سمت» یا «سِمَت»، «مسکن» یا «مُسْکَن»، «نرم» یا «نُرم» نیز هم نویسه هستند و همین مشکل در مورد آن‌ها نیز صدق می‌کند. با توجه به عدم نمایش این واکه‌ها، این موارد در فهرست پیشنهادی با رجوع به بافت جمله در مورد آن‌ها تصمیم‌گیری شد.

بنا به محدودیت‌های این پژوهش، حوزه بررسی در این پژوهش، صرفاً واژه‌های بسیط بوده است. به طور مثال «بکارگیری» که در فهرست رضوانی ارائه شده است، علاوه بر صورت‌های نوشتاری «به کار گیری» و «بکارگیری» که در آن‌ها درج نیمه فاصله یا عدم درج نیمه فاصله در انتخاب آن تأثیرگذار است. واژه «آزمون» در پیکره فردوسی ۵۴۴۲ بار تکرار شده است که هم می‌تواند به عنوان واژه تخصصی در رشته‌های مرتبط با سنجش و اندازه‌گیری تلقی شود و هم به‌عنوان یک واژه دانشگاهی عمومی. با توجه به بسامد بالای این واژه و تکرار شدن آن در هر پنج حوزه دانشگاهی (دامپزشکی، کشاورزی، علوم پایه، علوم انسانی و مهندسی) واژه «آزمون» جزء واژه‌های عمومی دانشگاهی قرار گرفته است. همچنین واژه‌های «محیط»، «خاک» و «آب» «غلظت» ۱۹۸۶ بار تکرار شده‌اند، در برخی رشته‌های دانشگاهی مانند محیط‌زیست و عمران و ... واژه تخصصی محسوب می‌شوند. در این پژوهش به علت در دسترس نبودن فهرست مدون واژه‌های تخصصی رشته‌های مختلف دانشگاهی این واژه‌ها به‌صورت واژه عمومی دانشگاهی در نظر گرفته شد که لازم است در پژوهش‌های آتی به این موضوع با دقت بیشتری پرداخته شود.

به‌علاوه باید به مباحث سبکی نیز در کنار سایر مسائل ذکر شده در بالا نیز اشاره کرد. به‌طور مثال دو واژه اولین (۱۲۵۸) و نخستین (۵۷۶) از نظر معنایی یکسان هستند و جزء واژگان پایه محسوب می‌شوند ولی کاربرد واژه «نخستین» در متون علمی به‌دلایل سبکی بیشتر از «اولین» است. همین دلیل کار انتخاب یا رد آن را تاحدی دشوار می‌کند. این موضوع زمانی اهمیت خود را نشان می‌دهد که فهرست نهایی برای آموزش به غیرفارسی زبانان مورد استفاده قرار بگیرد.

نکته حایز اهمیت دیگر انتخاب واژه‌های دخیلی هستند که علیرغم داشتن معادل فارسی، از بسامد بیشتری برخوردار هستند، به‌طور مثال استراتژی (۲۳۱ بار) و راهبرد (۱۴۴۸) در پیکره وجود دارند که از نظر معنا تقریباً یکسان هستند و به جای یکدیگر نیز به کار می‌روند و اگر فقط بسامد معیار انتخاب باشد، واژه‌هایی مانند «استراتژی» فهرست نخواهند شد. در این پژوهش واژه «استراتژی» با «راهبرد» در یک دسته قرار گرفت.

در استخراج فهرست‌های واژه مطابق کاکسهد و گاردنر سه معیار بسامد، پوشش و وقوع تخصصی در انتخاب واژه دخیل هستند و لذا ضرورت دارد که در تدوین پیکره دانشگاهی، تنوع لازم از مقالات، کتب و پایان‌نامه‌های دانشگاهی در رشته‌های مختلف رعایت شود تا فهرست پیشنهادی از اعتبار کافی برخوردار باشد. از طرف دیگر باید یک طبقه‌بندی کلان ولی محدود برای رشته‌های دانشگاهی وجود داشته باشد. در برخی پژوهش‌ها، این طبقه‌بندی ارائه شده توسط وزارت علوم آن کشور ملاک عمل قرار گرفته است. برای زبان فارسی تا آنجا که نگارندگان اطلاع دارند طبقه‌بندی جامعی وجود ندارد. به‌علاوه طبقه‌بندی تولیدات مکتوب رشته‌های متنوع دانشگاهی در چند حوزه کلان و کلی گاهی با مشکل روبرو می‌شود. به‌طور مثال موضوع مقاله به گونه‌ای است که امکان طبقه‌بندی در دسته‌های مختلفی وجود دارد و مشاهده می‌شود که مقاله بنا به موضوع اصلی آن، در مجله‌ای متفاوت از رشته دانشگاهی نویسندگان منتشر شده است.

۶- نتیجه‌گیری

در این پژوهش یک فهرست واژه‌های دانشگاهی زبان فارسی براساس سه اصل کاکسهد یعنی بسامد، محدوده و وقوع تخصصی از یک پیکره مقالات دانشگاهی با هدف طراحی یک آزمون رایانه‌ای انطباقی پیشنهاد گردیده است. فهرست پیشنهادی شامل ۵۷۸ خانواده واژگانی است که نسبت به فهرست رضوانی بیش از ۱,۷۴ درصد پوشش بیشتری بر روی پیکره شاهد دارد. استخراج پیکره - بنیاد فهرست واژه‌ها و اعتبارسنجی به‌شيوه خارج از نمونه، با مدنظر قرار دادن واژه‌های پایه و عمومی زبان فارسی عمده دستاورد این پژوهش می‌باشد که با نگاهی دقیقتر به مسأله و ویژگی‌های زبان فارسی تلاش نموده است، ابعاد مسأله و دشواری‌های استخراج یک فهرست دانشگاهی در زبان فارسی را بشکافد و تاحدی روشن‌کننده مسیر پژوهش‌های آتی باشد. پژوهشگران علاقمند به فهرست‌های واژگانی می‌توانند با شناخت ابعاد این مسأله و دشواری‌های موجود که تاحدی بحث شده است، با طراحی ابزارهای دقیق نسبت به استخراج فهرست‌های معتبرتر اقدام نمایند. در پژوهش‌های آتی، پیشنهاد می‌شود که برای اعتبارسنجی فهرست‌های واژه‌ای از پیکره‌های متوازن استفاده شود. پژوهشگران و مدرسان آفا می‌توانند از این فهرست پیشنهادی، در تهیه و تدوین کتب آموزشی فارسی دانشگاهی، طراحی آزمون‌های تعیین سطح، آزمون‌های بسندگی و

مهارت استفاده کنند. به‌علاوه طراحان کتب درسی نیز می‌توانند از این فهرست در طراحی کتب درسی مقطع دبیرستان استفاده کنند.

یکی از محدودیت‌های اصلی این پژوهش، عدم دسترسی به پیکره فردوسی به‌صورت تفکیک شده برحسب رشته دانشگاهی بود که با استفاده از نرم‌افزارهای مربوط، فهرست واژه‌ها برحسب بسامدشان در زیررشته‌های دانشگاهی علوم انسانی، کشاورزی، علوم پایه، مهندسی پزشکی و دامپزشکی مجدداً استخراج و دسته‌بندی شدند و فهرست نهایی بر اساس پارامتر پوشش همانند فهرست کاکسهد استخراج شد. در صورت ایجاد پیکره‌های آتی، پیشنهاد می‌شود که تنوع و تعادل متون علمی در زیررشته‌های مختلف دانشگاهی مدنظر قرار بگیرد تا بتوان معیار پوشش در کنار بسامد را با دقت بالاتری به کار گرفت و از این طریق فهرست‌های معتبرتر و دقیقتری استخراج و پیشنهاد شوند.

چالش‌های دیگری که باید در استخراج فهرست واژه‌های دانشگاهی به‌طور دقیق، مورد بررسی قرار بگیرند به‌صورت اجمالی عبارتند از: در دسترس نبودن نرم‌افزارهای دقیق برچسب‌گذاری مقوله نحوی (POS)، رسم‌الخط زبان فارسی و استاندارد و یکسان نبودن نویسه‌ها در فایل‌های دیجیتالی، عدم نمایش واژه‌ها، مرزبندی بین واژه‌های دانشگاهی عمومی و واژه‌های فنی یا تخصصی، در دسترس نبودن یک فهرست یا مدخل روزآمد از مقوله‌های شعاعی یا معنی سرنمون واژه، عدم وجود دسته‌بندی وندهای پرتکرار، افعال مرکب، ترکیب‌های چند کلمه‌ای^۱ در واژه‌های دانشگاهی، فهرست ایست واژه‌ها و عدم وجود یک طبقه‌بندی کلان ولی محدود برای رشته‌های دانشگاهی در ایران.

منابع

- صحرایی، رضامراد و همکاران (۱۳۹۶). آموزش کاربردی واژه. ایران، دانشگاه علامه طباطبائی.
- نعمت‌زاده، شهین؛ دادرس، محمد؛ دستجردی کاظمی، مهدی؛ منصوریزاده، محرم (۱۳۹۱). واژگان پایه زبان فارسی. تهران: انتشارات برهان.
- Anthony, L. (2022). AntWordProfiler, Version 2.0. 0. [Computer software]. Waseda University.
- Baumann, J. F., & Graves, M. F. (2010). What is academic vocabulary? *Journal of Adolescent & Adult Literacy*, 54(1), 4-12.
- Brezina, V., & Gablasova, D. (2017). How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. A response to Gabriele Stein. *Applied Linguistics*, 38(5), 764-767.
- Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, 15(2), 103-116.
- Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? *Vocabulary in a Second Language*, 23, 15-38.
- Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly*, 8(4), 389-399.
- Coxhead A., Nation P. (2001). The specialized vocabulary of English for academic purposes. In Flowerdew J., Peacock M. (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). Cambridge University Press.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.
- Dolmacı, M., & Ertal, A. (2016). Developing a textbook-based academic Turkish wordlist. *Procedia-Social and Behavioral Sciences*, 232, 821-827.
- Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for specific purposes*, 43, 49-61.
- Farrell, P. (1990). *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary*. CLCS Occasional Paper No. 25.
- Flowerdew, J. (1993). Concordancing as a tool in course design. *System*, 21(2), 231-244.
- Flowerdew, J., & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 8-24). Cambridge University Press. <http://dx.doi.org/10.1017/CBO9781139524766.004>
- Gardner, D. (2013). *Exploring vocabulary: Language in action*. Routledge.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied linguistics*, 35(3), 305-327.
- Gardner, D., & Davies, M. (2016). A response to "To what extent is the Academic Vocabulary List relevant to university student writing?". *English for Specific Purposes*, 43, 62-68.
- Graves, M. F. (2006). *The vocabulary book: Learning and instruction*. Newark, DE: International Reading Association.
- Hartshorn, K. J., & Hart, J. M. (2015, 27-31 March). *Comparing the academic word list with the academic vocabulary list: Analyses of frequency and ESL learner performance*. Paper presented at the 2015 AAAL, Toronto, Canada.

- Heidari Tabrizi, H., & Jamalzadeh, M. (2020). Academic Vocabulary in Tourism Research Articles: A Corpus-Based Study. *Journal of Language and Discourse Practice*, 1(2), 25-44.
- Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for academic purposes*, 22, 42-53.
- Li, Y., & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System*, 38(3), 402-411.
- Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, 39, 1-11.
- Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for specific purposes*, 28(3), 183-198.
- Moini, R., & Islamizadeh, Z. (2016). Do we need discipline-specific academic word lists? Linguistics Academic Word List (LAWL). *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, 35(3), 65-90.
- Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25(2), 235-256.
- Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, 39, 26-44.
- Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1), 91-108.
- Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Heinle, Cengage Learning.
- Nation, P., & Kyongho, H. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, 23(1), 35-41.
- Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, Acquisition and Pedagogy*, 14(1), 6-19.
- Newman, J. A. (2016). *A corpus-based comparison of the Academic Word List and the Academic Vocabulary List*. Brigham Young University.
- Rezvani, R., Gholtash, A., & Zamani, G. (2016). The first corpus-based Persian academic word list: Development and pedagogical implications. *Journal of Teaching Persian to Speakers of Other Languages (JTPSOL)*, 5(1), 43-64.
- Sun, Y., & Lei, L. E. I. (2013). Academic lexical items in economics journal papers. *Chinese Journal of Applied Linguistics*, 36(3), 354-367.
- Talalakina, E., Stukal, D., & Kamrotov, M. (2020). Developing and validating an academic vocabulary list in Russian: A computational approach. *The Modern Language Journal*, 104(3), 618-646.
- Todd, R. W. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes*, 45, 31-39.
- Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes*, 12(4), 248-263.
- Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, 27(4), 442-458.
- Xue, G., & Nation, I. S. (1984). A university word list. *Language Learning and Communication*, 3(2), 215-229.
- Yang, M. N. (2015). A nursing academic word list. *English for specific purposes*, 37, 27-38.

-
- Zimmerman, C. B. (2009). Word knowledge: A vocabulary teacher's handbook. *The Electronic Journal for English as a Second Language*, 4(13), 1-2.

پیوست ۱: واژه‌های مشترک نعمت زاده و صحرائی

امتحان	الاغ	استکان چای	اسب	اتوبوس	اتو	اتاق خواب	اتاق پذیرایی	ابر	امروز
آرد	آخر	آتش سوزی	آبی	ایستادن	ایران	انگور	انگشت	انسان	آسان
باز	باران	باد	بابا	آینه	آواز	آفتاب	آشپز	آسمان	باز کردن
برف	بردن	برادر	بد	بچه	بالای ...	بال	باغچه	بازی کردن	برنج
بی ادب	بلند	بشقاب	بسته	بستی	بستن	بزرگ	بز	برنده	بیمارستان
پرده	پدر بزرگ	پدر	پایین ...	پاکت	پاک کن	پارک	پارچه	پا	پزندگان
پنیر	پنجره	پلیس	پله	پسر	پزشک	پریدن	پروانه	پرواز کردن	پوست
ترازو	تخم مرغ	تخته سیاه	تخت	تاکسی	پیراهن	پیاز	پیاده شدن	پوشیدن	تراش
تند(مزه)	تنبل	تمیز	تمیز	تمبر	تلویزیون	تلفن	تعطیل	تشنه	تنگ
جیب	جوراب شلواری	جوان	جمع آوری کردن	جلوی	جمعه	جارو کردن	جارو	توت فرنگی	چادر
چوب	چنگال	چشم	چسب	چرخ خیاطی	چراغ خواب	چای	چاقو	چاق	چوراب
خسته	خراب	خدا	خانه	خاموش	خاک	حیات	حوله	حمام	خط کش
خوشمزه	خوشحال	خورشید	خوردن	خودکار	خواهر	خوابیدن	خندیدن	خمیر دندان	خون امدن
در	در	دبستان	دانه	دانش آموز	دامن	خیس	خیاط	خیابان	درخت
دوچرخه	دهان	دندان	دکمه	دستمال	دستکش	دست	دریا	درس دادن	دور از...
راننده	رانندگی کردن	رادیو	دیوار	دیدن	دویدن	دوستانه	دوست داشتنی	دوست داشتن	راه پله
ریش	ریختن	روغن	روشن	روسری	روستا	روز	رودخانه	رنگ	زیان
سبزی	سبز	ساعت(زمان)/ وسیله	ساعت چنده؟	ساعت چنده؟	ساختمان	زن(جنسیت)/همسر	زمین خوردن	زرد	سبزی
سفید	سفره خانه	سطل زباله	سطل	سرما	سر	سبزیجات	سبزی فروشی	سبزی پلو با ماهی	سگ
سبب زمینی	سیب	سیاه	سوزن	سواری کردن	سوار کاری	سوار شدن	سنگ	سلام	سبب زمینی سرخ کرده
شکرپاش	شکر (خورندنی)	شبانه روز	شب به خیر	شب	شانه(بزار)	شامپو	شام	سیر	شکم
صابون	شیرینی پزی	شیرینی	شیرین	شیرآب	شهر	شناگر	شنا	شلوار	صبح

عینک زدن	صبح به خیر	صبحانه	صندلی	عسل	عکاسی	عکس	عید شما مبارک	عینک	عینک
قاشق	غذا پختن	غذا درست کردن	غذای اصلی	فارسی	فارسی گفتاری	فارسی نوشتاری	فرش	فوتبال	فیلم
قندان	قالی	قالیبافی	قد	قرآن خواندن	قرمز	قطار	قلم	قلم مو	قند
کنتری	قوری	قیچی	کارگر	کارنامه	کبوتر	کت شلوار	کتاب	کتابخانه	کتابخانه
کوتاه	کتیف	کفش	کلاغ	کلاه	کلید	کم رنگ	کم هوش	کمد	کمک کردن به
گردو	کوچک	کوچه	کوه	کوهستان	کیف(اسم ابزار)	گاو	گج	گربه	گردن
گوشت قرمز	گرسنه	گریه کردن	گل دار	گل فروش	گل(گیاه)	گلدان	گنجشک	گوش کردن	گوش کردن
لباس پسرانه	گوشته‌خوار	گوشفند	گیاه	گیاهان	گیاهخوار	لاغر	لانه	لب	لباس بچگانه
ماژیک	لباس پوشیدن	لباس دخترانه	لباس زنانه	لباس مردانه	لیوان	مادر	مادرن	مادر شوهر	مادربزرگ
مداد رنگی	ماست	ماشین ظرفشویی	ماشین لباسشویی	ماتو	ماه	ماهی	ماهی تابه	متر	مداد
مگس	مدرسه	مربا	مرد(جنسیت)	مردم	مرغ	مسجد	مسواک	مسواک زدن	معلم
ناخن گیر	مهمان	مهمانی رفتن	مو	موتور سیکلت	میز	میوه	میوه فروشی	ناخن	ناخن گرفتن
نزدیک ...	ناراحت	نامه	نان	نانوا	نانوایی	ناهار	نخ	نخ دادن	نخی
نوک	نشستن	نقاش ساختمان	نقاشی	نقشه	نماز خواندن	نمره	نمک	نور خورشید	نوشابه
یخچال	نیمکت	همسایه	هندوانه	ورزش	ورزش کردن	ورزشکار	ورزشگاه	یخ زدن	یخبندان