Development of a Persian Academic Word List Based on an Academic Corpus

Morteza Rezaei Sharifabadi
PhD Candidate in Linguistics; Shiraz University; Shiraz, Iran;
Email: mrezaeis@gmail.com

Amirsaeid Moloodi*
PhD in Linguistics; Assistant Professor; Shiraz University;
Shiraz, Iran Email: amirsaeid.moloodi@shirazu.ac.ir

Alireza Ahmadi
PhD in TEFL; Professor; Shiraz University; Shiraz, Iran;
Email: arahmadi@shirazu.ac.ir

Alireza Khormaei
PhD in Linguistics; Associate Professor; Shiraz University;
Shiraz, Iran Email: akhormaee@rose.shirazu.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 38 | No. 3 | pp. 901-926
Spring 2023
https://doi.org/jipm.38.3

Received: 10, Jun. 2022 **Accepted: 12, Jul. 2022**

Abstract: Academic words occur with high frequency in texts from a wide range of scientific fields, and their frequency in academic texts is much higher than in general texts. Academic wordlists can facilitate the learning and teaching of scientific language. In this research, we have developed a frequency list of Persian academic words. The word list includes 307 word lemmas with a high frequency in academic texts. Creating a balanced corpus of Persian academic texts was the prerequisite for developing such a list. For this purpose, we collected scientific texts published in Persian scientific journals and built a balanced corpus containing more than 51 million words. The corpus includes texts of academic papers in four general categories, i.e., basic sciences and engineering; humanities, arts, and architecture; medicine and veterinary medicine; and agriculture and natural resources. We used four different criteria for lemmas to be included in our wordlist. 1- frequency: The lemmas should have a relative frequency of at least 30 per million words. 2- ratio: The relative frequency of the lemmas in the academic corpus should be two times greater than their frequency in a 10 million word general corpus. 3- dispersion: Juilland's D value of the lemmas in the four sections should be at least 0.5. 4- range: the observed frequency of the lemma should not be less than a third of its expected frequency in any of the four sections of the corpus. We evaluated the wordlist by measuring its coverage in our corpus's train

* Corresponding Author

and test sections. The wordlist covers 16.69 percent of the train subset and 16.13 percent of the test subset.

Keywords: Frequency List, Academic Wordlist, Academic Corpus, Persian Language, Corpus Linguistics

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

تهیهٔ فهرست بسامدی واژگان علمی فارسی با بهره‌گیری از پیکرهٔ علمی

مرتضی رضائی شریف‌آبادی | دانشجوی دکتری زبان‌شناسی؛ دانشگاه شیراز؛ شیراز، ایران mrezaeis@gmail.com

امیرسعید مولودی | دکتری زبان‌شناسی؛ استادیار؛ بخش زبان‌های خارجی و زبان‌شناسی؛ دانشگاه شیراز؛ شیراز، ایران؛ پدیدآور رابط amirsaeid.moloodi@shirazu.ac.ir

علیرضا احمدی | دکتری آموزش زبان انگلیسی؛ استاد؛ بخش زبان‌های خارجی و زبان‌شناسی؛ دانشگاه شیراز؛ شیراز، ایران؛ arahmadi@shirazu.ac.ir

علیرضا خرمایی | دکتری زبان‌شناسی؛ دانشیار؛ بخش زبان‌های خارجی و زبان‌شناسی؛ دانشگاه شیراز؛ شیراز، ایران؛ akhormaee@rose.shirazu.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۱/۰۳/۲۰ | **پذیرش:** ۱۴۰۱/۰۴/۲۱ | مقاله برای اصلاح به مدت ۶ روز نزد پدیدآوران بوده است.

چکیده: واژه‌های علمی واژه‌هایی هستند که در طیف وسیعی از رشته‌های علمی بسامد بالایی دارند و بسامدشان در متون علمی بسیار بیشتر از بسامدشان در سایر متون است. دسترسی به فهرستی بسامدی از واژه‌های علمی می‌تواند به یادگیری سریع‌تر زبان علمی کمک کند. پیش‌نیاز تهیهٔ چنین فهرستی ایجاد پیکره‌ای متوازن از متون علمی فارسی است. برای این منظور، متون علمی منتشرشده در نشریات علمی فارسی با برنامه‌نویسی رایانه‌ای و توسعهٔ خزندهٔ وب جمع‌آوری شد. سرانجام، پیکره‌ای شامل بیش از ۵۱ میلیون واژه با حجم متوازنی از داده در چهار حوزهٔ موضوعی شامل «علوم پایه و فنی و مهندسی»، «علوم انسانی و هنر و معماری»، «پزشکی و دامپزشکی» و «کشاورزی و منابع طبیعی» ساخته شد. در این پژوهش پس از بررسی ملاحظات مربوط به توسعهٔ فهرست‌های بسامدی علمی، فهرستی از واژه‌های علمی فارسی تهیه شد که شامل ۳۰۷ بن‌واژه است که در متون علمی بسامد بالایی دارند، بسامدشان در متون علمی به مراتب بیشتر از متون عمومی است و پراکندگی مناسبی در موضوعات مختلف علمی دارند. نتایج ارزیابی فهرست بسامدی تهیه‌شده نشان‌دهندهٔ پوشش بیش از ۱۶ درصدی فهرست روی متون علمی است که این درصد با پوشش فهرست‌های جمع‌آوری‌شده برای زبان‌های دیگر مطابقت دارد. همچنین، توزیع مناسب

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۸ | شمارهٔ ۳ | ص‌ص ۹۰۱-۹۲۶
بهار ۱۴۰۲
https://doi.org/jipm.38.3

واژه‌های فهرست جمع‌آوری‌شده در چهار حوزهٔ موضوعی پیکره باعث شده است که میزان پوشش در هر یک از این حوزه‌های موضوعی نیز عددی قابل قبول بین ۱۲ تا ۱۸ درصد باشد. فهرست واژه‌های علمی استخراج‌شده می‌تواند برای اهداف مختلف آموزشی و پژوهشی مورد استفاده قرار گیرد. همچنین، پیکرهٔ علمی تهیه‌شده نیز منبع ارزشمندی برای مطالعات حوزهٔ زبان‌شناسی پیکره‌ای و پژوهش‌های مرتبط با پردازش زبان است.

کلیدواژه‌ها: فهرست بسامدی، واژگان علمی، پیکرهٔ علمی، زبان فارسی، زبان‌شناسی پیکره‌ای

۱. مقدمه

امروزه، یکی از مسائل مهم پیرامون زبان فارسی، بهبود وضعیت این زبان در جایگاه «زبان علم» است؛ چنان‌که در اسناد بالادستی، به‌ویژه در نقشهٔ جامع علمی کشور بر این امر تأکید فراوان شده است (دبیرخانهٔ شورای‌عالی انقلاب فرهنگی ۱۳۸۹). از جمله مسائلی که در این زمینه در نقشهٔ جامع علمی به آن پرداخته شده، «توسعه و ابداع روش‌های سهل و سریع فارسی‌آموزی» است. یکی از جنبه‌های مهم یادگیری و آموزش هر زبانی، یادگیری و آموزش واژه‌های آن زبان است. این مسئله به‌ویژه دربارهٔ آموزش زبان برای اهداف علمی و دانشگاهی صادق است؛ چرا که توسعهٔ سریع علم و فناوری و گسترش روزافزون حوزه‌های دانشی باعث شده است که واژه‌ها و اصطلاحات جدید علمی با سرعت زیادی تولید و وارد ادبیات علمی شوند. از سوی دیگر، در حال حاضر بخش قابل توجهی از فارسی‌آموزان دانشجویانی هستند که برای تحصیل در رشته‌های مختلف علمی در ایران نیاز به آشنایی با زبان فارسی دارند. طبیعی است که برای این دسته از زبان‌آموزان، یادگیری هرچه سریع‌تر واژه‌های مهم زبان علمی از اهمیت ویژه‌ای برخوردار است؛ چرا که آن‌ها باید بلافاصله با استفاده از آموخته‌هایشان مشغول به تحصیل به زبان فارسی شوند. با توجه به این توضیحات، پژوهش در زمینهٔ روش‌های روزآمد برای شناسایی واژه‌های علمیِ مهم، آموزش آن‌ها و تدوین مواد آموزش و ارزیابی برای این منظور، اهمیت ویژه‌ای دارد.

اطلاعات بسامدی[1] مستخرج از پیکره‌های زبانی[2] منبع بسیار ارزشمندی برای تدوین مواد یادگیری زبان[3] و ارزیابی مهارت‌های زبانی است (Leech 2011). رایج‌ترین نوع اطلاعات بسامدی مورد استفاده برای اهداف ذکرشده، اطلاعات مربوط به فراوانی

1. frequency information 2. language corpora 3. language learning materials

واژه‌هاست. به گفتهٔ «نیشِن»، «گروه کوچکی از واژه‌های پربسامد هستند که بسیار اهمیت دارند؛ چرا که همین واژه‌ها بخش بسیار بزرگی از واژه‌های مورد استفاده در نوشتار و گفتار را پوشش می‌دهند ...» (Nation 2013, 22) و «واژه‌های پربسامد زبان آن‌قدر مهم هستند که معلمان و زبان‌آموزان باید وقت قابل توجهی را صرف آن‌ها کنند» (ibid 24). یکی از انواع واژگان بسامدی پرکاربرد، فهرست بسامدی واژه‌های علمی است. نتایج پژوهش‌های فراوان نشان‌دهندهٔ اهمیت مهارت زبانی علمی برای درک متون علمی، توانایی اندیشیدن مانند محققان، و در کل، موفقیت علمی است و می‌توان گفت واژگان علمی مشهودترین جنبهٔ زبان علمی است و عدم تسلط بر واژگان علمی به‌عنوان مانعی برای موفقیت دانشجویان شناخته می‌شود (Nagy and Townsend 2012).

واژه‌های علمی را باید از واژگان پربسامد عمومی و واژگان فنی[1] یا واژگان تخصصی[2] مجزا دانست. بر اساس تقسیم‌بندی «کاکسهد و نیشن»، واژه‌های پربسامد[3] (عمومی) در طیف وسیعی از متون، بسامد بالایی دارند و برای تمام کاربردهای زبان ضروری هستند (برای مثال، واژه‌های «کشور» و «سال»). از سوی دیگر، واژه‌های فنی یا واژگان تخصصی پربسامد از رشته‌ای به رشتهٔ دیگر متفاوت هستند (برای مثال، واژه‌های «لاشبرگ» و «صفاقی»، به‌ترتیب در «کشاورزی و منابع طبیعی» و «پزشکی و دام‌پزشکی») (Coxhead and Nation 2001). در این میان، واژگان علمی[4] در دامنهٔ گسترده‌ای از متون علمی بسامد بالایی دارند، اما رخدادشان در انواع دیگر متون آن‌چنان زیاد نیست و از آنجا که حدود ۸/۵ تا ۱۰ درصد از واژه‌های متون علمی را شامل می‌شوند، برای زبان‌آموزان با اهداف علمی و دانشگاهی بسیار مهم هستند (برای مثال، واژه‌های «پراکندگی» و «همگن»). هدف از تهیهٔ فهرست‌های بسامدی واژه‌های علمی تسهیل یادگیری واژه‌های علمی برای زبان‌آموزانی است که در طیف وسیعی از رشته‌های علمی مشغول به تحصیل هستند (برای مثال، در کلاس‌های زبان عمومی که به‌طور معمول، در ترم‌های نخست برگزار می‌شود و یا در دروس زبان که به‌صورت پیش‌نیاز برای دانشجویان رشته‌های مختلف برگزار می‌شود). بنابراین، سعی می‌شود واژه‌های پربسامد عمومی (که زبان‌آموزان قاعدتاً باید در آموزش‌های ابتدایی زبان با آن آشنا شده باشند) یا واژه‌های فنی و تخصصی پربسامد (که دانشجویان در ادامهٔ

1. technical vocabulary 2. specialized vocabulary 3. high frequency words
4. academic vocabulary

تحصیل و در دروس تخصصی خود با آن آشنا می‌شوند) در فهرست‌های بسامدی علمی قرار نگیرد.

ضرورت پژوهش حاضر آن است که در زبان فارسی، به‌رغم اهمیت پرداختن به زبان فارسی در جایگاه زبان علم و به‌رغم اهمیت فهرست‌های بسامدی در آموزش و یادگیری زبان علمی، شوربختانه پژوهش‌های انجام‌شده در زمینهٔ تهیهٔ فهرست‌های بسامدی واژگان علمی بسیار محدود و انگشت‌شمار است. در این پژوهش تلاش می‌کنیم، گامی در جهت رفع خلأ موجود در این زمینه در زبان فارسی برداریم.

در پژوهش حاضر، در صدد پاسخ به این پرسش هستیم که کدام واژه‌های فارسی در طیف وسیعی از متون علمی دارای بسامد بالایی هستند و بسامد آن‌ها در متون علمی به مراتب بالاتر از بسامدشان در سایر انواع متون است. همچنین، بررسی خواهیم کرد که بر اساس معیارهای ارزیابی به‌کاررفته در پیشینهٔ پژوهش، فهرست بسامدی واژه‌های علمی ارائه‌شده تا چه حد قابل اتکاست؟ افزون بر این، به این موضوع خواهیم پرداخت که برای استخراج فهرست‌های بسامدی علمی چه ملاحظاتی را باید مد نظر داشت. نظر به آنکه پیکرهٔ استاندارد متون علمی پیش‌نیاز تهیهٔ فهرست‌های بسامدی در این حوزه است و شوربختانه پیکرهٔ علمی متناسب با اهداف پژوهش حاضر در دسترس نبود، بخش مهمی از این پژوهش به تولید پیکرهٔ علمی متوازن با حجم قابل قبول اختصاص داده شده است.

۲. پیشینهٔ پژوهش

تهیهٔ فهرست بسامدی از واژه‌های علمی دارای ملاحظات مختلفی است که ابتدا در این بخش با نگاه به پیشینهٔ پژوهش‌های انجام‌گرفته در این حوزه به این ملاحظات خواهیم پرداخت. پس از آن، پژوهش‌های انجام‌شده در زبان فارسی در زمینهٔ تهیهٔ فهرست‌های بسامدی و پیکره‌های علمی را بررسی خواهیم کرد.

مسئلهٔ اول در ارتباط با تهیهٔ فهرست‌های بسامدی علمی، نحوهٔ دسته‌بندی واژه‌ها برای شمارش است. ساده‌ترین شیوه برای دسته‌بندی واژه‌ها برای تهیهٔ فهرست‌های بسامدی، دسته‌بندی بر اساس نوع واژه[1] است. در این روش، برای مثال، وقوع‌های «کتاب»، «کتاب‌ها»، «کتابی»، «کتاب‌هایشان» و ... هر کدام به‌صورت جداگانه شمارش می‌شوند.

1. word type

برخی پژوهشگران از جمله Ward (2009) برای تهیهٔ فهرست‌های بسامدی خود از این نوع دسته‌بندی استفاده می‌کنند. روشی دیگر برای دسته‌بندی واژه‌ها، دسته‌بندی آن‌ها بر اساس مفهوم «خانوادهٔ واژگانی»[1] است. Coxhead (2000) که یکی از پژوهش‌های شناخته‌شده در زمینهٔ تهیهٔ فهرست‌های بسامدی علمی است، برای تهیهٔ فهرست خود از همین شیوهٔ دسته‌بندی استفاده می‌کند. مزیت این نوع دسته‌بندی از نظر «کاکسهد»، ارتباط قوی میان واژه‌های هم‌خانواده است و اینکه خانواده‌های واژگانی، واحد مهمی در واژگان ذهنی هستند و درک مفهوم اعضای مختلف یک خانوادهٔ واژگانی کار دشواری نیست. تعریف وی از واژه‌های هم‌خانواده نیز تصریف‌ها و اشتقاق‌های متداول از یک ریشهٔ آزاد است و از دسته‌بندی سطح ۶ Bauer and Nation (1993) برای خانواده‌های واژگانی استفاده می‌کند. طبق این تعریف واژه‌های concepts، conception و conceptually در یک خانواده قرار دارند، اما specify و special در یک خانواده نیستند؛ چرا که spec ریشهٔ آزاد نیست. در آثار قدیمی‌تر مانند فهرست واژگان پایهٔ انگلیسی West (1953) و فهرست واژگان دانشگاهی Xue and Nation (1984) و بسیاری از پژوهش‌های بعدی از جمله Konstantakis (2007)، Hsu (2013)، Dang, Coxhead and Webb (2017) نیز واژه‌های هم‌خانواده در یک دسته قرار گرفته و شمارش می‌شوند. شیوهٔ رایج سوم برای دسته‌بندی واژه‌ها، دسته‌بندی آن‌ها بر اساس مفهوم «بن‌واژه» است. برخی پژوهشگران مانند Paquot (2007)، Gardner and Davies (2013) و Lei, Lei and Liu (2016) منتقد استفادهٔ Coxhead (2000) و سایرین از خانواده‌های واژگانی برای دسته‌بندی واژه‌های فهرست بسامدی‌اند. «گاردنر و دیویس» معتقدند که در مجموعهٔ واژه‌های هم‌خانواده، واژه‌هایی با معانی بسیار متنوع حضور دارند که قاعدتاً نباید در یک دسته شمارش شوند. برای مثال، مجموعهٔ هم‌خانواده‌های واژهٔ react (واکنش نشان دادن) در انگلیسی شامل واژه‌هایی چون reactionary (مرتجع)، reactivation (فعال‌سازی مجدد) و reactor (رآکتور) می‌شود که همان‌طور که مشخص است، ارتباط معنایی شفافی با هم ندارند و این‌گونه نیست که زبان‌آموز بتواند تمام اعضای یک خانوادهٔ واژگانی را با این تنوع از معانی به‌راحتی به‌خاطر بسپارد. به عقیدهٔ «گاردنر و دیویس» بسیاری از مشکلات معنایی این نوع دسته‌بندی ناشی از آن است که در آن به مقوله یا برچسب اجزای کلام[2] واژه‌ها توجهی نمی‌شود. بنابراین، برای تهیهٔ فهرست خود، واژه‌های با «بن‌واژهٔ» یکسان،

1. word family
2. part of speech (POS)

یعنی واژه‌هایی را که برچسب اجزای کلام یکسانی دارند، ریشهٔ مشترک دارند و حداکثر در وندهای تصریفی با هم تفاوت دارند، در یک دسته قرار می‌دهند و شمارش می‌کنند (Gardner and Davies 2013). در تهیهٔ فرهنگ‌های لغت از جمله فرهنگ‌های زبان‌آموز نیز به‌طور معمول، از همین مبنا برای تعیین مدخل‌ها استفاده می‌شود. بر اساس این دسته‌بندی، واژه‌های «کتاب»، «کتاب‌ها» و «کتاب‌هایشان» در یک دسته قرار می‌گیرند، اما واژه‌های «کتاب» (اسم) و «کتابی» (صفت) در دسته‌های جدا قرار می‌گیرند.

مسئلهٔ بعد، حداقل بسامد قابل قبول برای قراردادن واژه‌ها در فهرست بسامدی است. در بسیاری از فهرست‌های بسامدی از جمله Paquot (2007)، Wang, Liang and Ge (2008)، و Lei, Lei and Liu (2016) سعی شده است که حداقل بسامد نسبی واژه‌ها نزدیک به حداقلِ در نظر گرفته‌شده برای تهیهٔ فهرست Coxhead (2000) باشد (بسامد نسبی حدود ۲۹ واژه در یک میلیون واژه). البته، بعضی از پژوهشگران این حد از بسامد را پایین می‌دانند. برای مثال، «هایلند و تسه» معتقدند که حداقل بسامدی که Coxhead (2000) استفاده کرده، یعنی حداقل ۱۰۰ وقوع در پیکرهٔ ۳/۵ میلیون واژه‌ای، مقدار پایینی است (Hyland and Tse 2007). از سوی دیگر، Gardner and Davies (2013) هنگام انتخاب بن‌واژه‌ها برای فهرست بسامدی خود به بسامد آن‌ها اصلاً توجهی ندارند و صرفاً فهرست نهایی خود را که با استفاده از معیارهای دیگر به دست آمده بر اساس بسامد مرتب‌سازی می‌کنند.

مسئلهٔ دیگر در تهیهٔ فهرست‌های بسامدی واژه‌های علمی، نحوهٔ ایجاد تمایز میان واژه‌های پربسامد علمی و واژه‌های پربسامد عمومی است. یکی از رویکردهای رایج در این زمینه، استفاده از فهرست آماده از واژه‌های پربسامد عمومی و حذف واژه‌های آن از فهرست مورد نظر است. برای مثال، در پژوهش‌هایی مانند Praninskas (1972)، Coxhead (2000)، و Wang, Liang and Ge (2008)، ۲۰۰۰ خانوادهٔ واژگانی اول فهرست واژه‌های پربسامد عمومی West (1953) از فهرست بسامدی علمی حذف شده. در این رویکرد فرض بر آن است که زبان‌آموزان برای درک مطالب علمی باید پیش از یادگیری واژه‌های علمی، با واژه‌های پربسامد عمومی که در فهرست‌های عمومی ارائه شده، آشنا شده باشند. برخی پژوهشگران، منتقد استفاده از فهرست West (1953) برای حذف واژه‌های عمومی هستند. به اعتقاد «پکو»، برخی واژه‌ها ممکن است در فهرست West (1953) به‌کار رفته باشند، اما در متون علمی کاربرد متفاوتی داشته باشند (Paquot 2007). وی واژه‌های example، problem، reason، argument و result را به‌عنوان نمونه‌هایی نام می‌برد که با آنکه بسامد

و توزیع آن‌ها به شکل معناداری در متون علمی زیاد است، به‌خاطر وقوع در فهرست West (1953) از فهرست Coxhead (2000) حذف شده‌اند. از سوی دیگر، «گاردنر و دیویس» مطرح می‌کنند که فهرست West (1953) فهرستی قدیمی مبتنی بر پیکره‌ای مربوط به اوایل قرن بیستم است و بنابراین، فهرست مناسبی از واژه‌های عمومی پرکاربرد برای زبان انگلیسی امروز به حساب نمی‌آید (Gardner and Davies 2013). از نظر «گاردنر و دیویس» همین مسئله باعث شده است که از طرفی فهرست Coxhead (2000) شامل بسیاری از واژه‌های پربسامد پیکره‌های عمومی جدید باشد و از طرف دیگر، بسیاری از واژه‌های علمی به‌خاطر حضور در فهرست West (1953)، در فهرست Coxhead (2000) جایی نداشته باشند (Gardner and Davies 2013). برخی پژوهشگران برای حذف واژه‌های پرکاربرد عمومی از فهرست خود از روش فوق استفاده می‌کنند، اما به‌جای فهرست West (1953)، از فهرست‌های بسامدی جدیدتر استفاده می‌کنند. برای مثال، Hsu (2013)، ۳۰۰۰ خانوادهٔ واژگانی پربسامد پیکرهٔ ملی بریتانیا[1] را از فهرست خود حذف می‌کند. همچنین، پژوهشگرانی مانند Browne, Culligan and Phillips (2013) و Brezina and Gablasova (2015) فهرست‌های بسامدی عمومی جدیدی در زبان انگلیسی توسعه داده‌اند که می‌توانند برای این منظور استفاده شود. از رویکردهای رایج دیگر در مواجهه با این مسئله، مقایسهٔ بسامد واژه‌ها در پیکره‌هاست. به‌عبارت دیگر، واژه‌هایی که بسامد آن‌ها در پیکرهٔ علمی بسیار بیشتر از بسامدشان در پیکرهٔ عمومی (یا پیکرهٔ غیرعلمی دیگر مانند پیکرهٔ متون داستانی) باشد، در فهرست واژه‌های علمی قرار می‌گیرد، اما واژه‌هایی که بسامدشان در پیکره‌های علمی و عمومی به یک نسبت زیاد باشد، واژه‌های عمومی پربسامد تلقی می‌شوند. در پژوهش‌هایی مانند Paquot (2007)، Gardner and Davies (2013) و (Lei, Lei and Liu 2016) از این روش استفاده شده است.

مسئلهٔ دیگری که در تهیهٔ فهرست واژه‌های علمی مورد توجه قرار می‌گیرد، انتخاب واژه‌هایی است که در طیف وسیعی از رشته‌های علمی، و نه صرفاً یک یا دو رشتهٔ تخصصی، بسامد بالایی دارند. در همین راستا، از معیارهای مختلفی برای اطمینان از توزیع مناسب واژه‌ها در بخش‌های مختلف پیکره و در نتیجه، درج نشدن واژه‌های تخصصی در فهرست‌های بسامدی واژگان علمی استفاده می‌شود. یکی از معیارهای مهم مورد استفاده

1. British National Corpus (BNC)

برای اطمینان از توزیع مناسب واژه‌های علمی در پیکره، معیار پراکندگی[1] است. برای این منظور، در بسیاری از پژوهش‌های انجام‌شده در زمینهٔ تهیهٔ فهرست بسامدی واژه‌های علمی، از ضریب پراکندگی Juilland and Chang-Rodriguez (1964) استفاده می‌شود. این ضریب که نشان می‌دهد یک واژه در متون مختلف یک پیکره تا چه حد متوازن به‌کار رفته است، مقداری بین ۰ (واژه تنها در بخش بسیار کوچکی از پیکره به‌کار رفته است) تا ۱ (واژه به‌طور کامل در بخش‌های مختلف پیکره توزیع شده است) می‌تواند داشته باشد. ضریب پراکندگی «جویلاند» به‌صورت زیر محاسبه می‌شود.

$$\text{ضریب پراکندگی جویلاند} = 1 - \frac{\text{ضریب تغییرات}^{1}}{\sqrt{\text{تعداد بخش‌های پیکره} - 1}}$$

در پژوهش‌های انجام‌شده، کمینه مقدارهای مختلفی برای ضریب «جویلاند» واژه‌ها جهت اضافه‌شدن به فهرست بسامدی علمی تعیین شده است؛ برای مثال، در Paquot (2007) و (Lei, Lei and Liu 2016) ۰/۵، در Dang, Coxhead and Webb (2017) ۰/۶، و در Gardner and Davies (2013) ۰/۸. «گاردنر و دیویس» توضیح می‌دهند که برای تعیین مقادیر حداقلی برای معیارهای خود (از جمله پراکندگی و دامنه)، مقداری استاندارد مبتنی بر پژوهش‌های قبلی وجود ندارد و برای مثال، تعیین مقدار حداقل ۰/۸ برای ضریب پراکندگی «جویلاند» صرفاً بر اساس آزمایش‌های «گاردنر و دیویس» بوده. ایشان در آزمایش‌های خود مشاهده کرده‌اند که واژه‌های دارای ضریب پراکندگی «جویلاند» کمتر از ۰/۸ واژه‌هایی بودند که از نظر آن‌ها بیشتر جنبهٔ تخصصی داشتند.

معیار دامنه[2] یکی دیگر از معیارهای مورد استفاده برای اطمینان از توزیع مناسب واژه‌هاست. بر اساس این معیار، واژه‌ها برای قرارگرفتن در فهرست بسامدی، باید در تعداد مشخصی از زیرموضوعات پیکره به میزان مشخصی به‌کار رفته باشند. علت استفاده از این معیار، همان‌گونه که در Ward (2009) اشاره شده، آن است که اگر بر اساس ضریب «جویلاند» (یا معیارهای پراکندگی دیگر) واژه‌ها را برای قرارگرفتن در فهرست بسامدی در نظر بگیریم، ممکن است واژه‌ها در بیشتر بخش‌های پیکره بسامد بالایی داشته باشند؛ اما در یکی از بخش‌ها بسامدشان بسیار کمتر از حد انتظار یا حتی صفر باشد. معیار دامنه کمک می‌کند که مطمئن شویم که واژه‌های مورد نظر در تمام بخش‌های پیکره به‌اندازه‌ای

1. dispersion
2. range

قابل قبول به‌کار رفته‌اند. برای مثال، در Coxhead (2000)، اعضای خانواده‌های واژگانی باید حداقل ۱۰ بار در هر یک از ۴ رشتهٔ پیکره و در حداقل ۱۵ موضوع از ۲۸ موضوع پیکره تکرار می‌شدند؛ در Paquot (2007) واژه‌ها باید در تمام زیرموضوع‌های پیکره استفاده شده باشند؛ در Gardner and Davies (2013) رخداد هر واژه در دست کم ۷ دستهٔ موضوعی از مجموع ۹ موضوع پیکره باید به اندازهٔ حداقل ۲۰ درصد از بسامد مورد انتظار باشد؛ در Hsu (2014)، اعضای خانوادهٔ واژگانی باید در تمام ۲۰ زیربخش پیکره و در حداقل ۹۵ کتاب از ۱۰۰ کتاب مورد استفاده برای تهیهٔ پیکره به‌کار رفته باشند؛ در (Lei, Lei and Liu 2016) بن‌واژه‌های مورد نظر باید در حداقل ۱۲ زیرموضوع از مجموع ۲۱ زیرموضوع پیکره به اندازهٔ ۲۰ درصد بسامد مورد انتظار وقوع داشته باشند؛ و در Coxhead and Hirsch (2007)، Wang, Liang, and Ge (2008)، Hsu (2013) و Dang, Coxhead and Webb (2017) واژه‌ها باید در حداقل نیمی از بخش‌های موضوعی پیکره به‌کار رفته باشند.

برای واژگان علمی در زبان فارسی به نظر می‌رسد که تنها فهرست بسامدی موجود، فهرست Rezvani, Gholtash and Zamani (2016) است. پیکره‌ای که در این پژوهش استفاده شده، شامل متون علمی از ۷ دستهٔ موضوعی بوده است. موضوعات مورد استفاده عبارت‌اند از هنر و معماری، مهندسی و فناوری، کشاورزی و منابع طبیعی، علوم پزشکی، علوم دامپزشکی، علوم انسانی و علوم پایه. مقالات چهار نشریه در هر دستهٔ موضوعی (در مجموع، ۱۱۲ مقاله) و همچنین یک کتاب در هر موضوع در پیکره گنجانده شده‌اند. تاریخ متون مربوط به سال‌های ۲۰۱۲ تا ۲۰۱۵ است. معیار انتخاب واژه‌ها برای قرارگرفتن در فهرست بسامدی این بوده است که حداقل ۲۰ بار در پیکره تکرار شده باشند و در حداقل ۳ موضوع از ۷ دستهٔ موضوعی پیکره وقوع داشته باشند. با این دو معیار در مجموع، ۵۳۹ مدخل برای فهرست بسامدی انتخاب شده است. پژوهش Rezvani, Gholtash and Zamani (2016) از جنبه‌های مختلف دارای نقایص و ابهاماتی است که در ادامه، به آن پرداخته می‌شود. مسئلهٔ اول، موضوع اندازهٔ کوچک پیکرهٔ مورد استفاده (۹۲۷ هزار واژه) نسبت به پیکره‌هایی است که در پژوهش‌های دیگر به آن‌ها اشاره شد. مشکل دومی که می‌توان در اینجا در خصوص روش تهیهٔ این فهرست بسامدی به آن اشاره کرد، این است که در این پژوهش هیچ اقدامی برای جلوگیری از قرارگرفتن واژه‌های عمومی پربسامد در پیکره انجام نشده است؛ نه استفاده از فهرست واژگان عمومی پربسامد و نه استفاده از پیکرهٔ متون عمومی. بحث سوم ابهاماتی است که در این پژوهش وجود دارد؛ از جمله اینکه گفته شده

است که در این فهرست واژه‌های هم‌خانواده در یک دسته قرار گرفته و شمارش شده‌اند، اما هیچ توضیحی در خصوص روش این کار ارائه نشده است. یا اینکه گفته شده است که تمام منابع مورد استفاده پی‌دی‌اف بوده‌اند. مشخص نیست که برای یک پژوهش بسامدی که قاعدتاً باید مبتنی بر منابع ماشین‌خوان باشد، چگونه از فایل‌های پی‌دی‌اف استفاده شده است. تنها اشاره‌ای شده است به اینکه برای شمارش واژه‌ها از نرم‌افزاری به نام «اکسپرت پی‌دی‌اف»[1] استفاده شده است. طبق بررسی‌ها، این نرم‌افزار یک نرم‌افزار عمومی برای کار با فایل‌های پی‌دی‌اف است و امکانات ویژه‌ای برای نویسه‌خوانی متون فارسی یا تهیهٔ فهرست‌های بسامدی ارائه نمی‌کند. البته، پژوهشگران خود در پایان مقاله اشاره به این موضوع می‌کنند که نرم‌افزار مناسب برای شمارش واژه‌های فارسی در دسترس نبوده است. همچنین، یکی از مهم‌ترین نقایص این کار آن است که هیچ‌گونه ارزیابی بر روی فهرست تهیه‌شده انجام نشده است. یعنی مشخص نیست که برای مثال، فهرست تهیه‌شده تا چه میزان متون علمی و متون غیرعلمی را پوشش می‌دهد.

البته، پژوهش‌های بسامدی دیگری نیز در زبان فارسی انجام شده است که به‌طور مستقیم به تهیهٔ واژگان علمی مربوط نیستند. برای مثال، «واژه‌های پرکاربرد فارسی امروز بر مبنای پیکرهٔ یک میلیون لغتی» «حسنی» (۱۳۸۴)، «فرهنگ بسامدی بر اساس پیکرهٔ متنی زبان فارسی امروز» «بی‌جن‌خان و محسنی» (۱۳۹۷) و «فرهنگ بسامدی زبان فارسی» Miller and Aghajanian-Stewart (2017) از جمله فهرست‌های بسامدی عمومی برای واژگان فارسی هستند. «فرهنگ زبان‌آموز پیشرفتهٔ فارسی» «عاصی» (۱۳۹۸) را هرچند نمی‌توان فرهنگی بسامدی دانست (چرا که مدخل‌ها در آن بر اساس بسامد مرتب‌سازی نشده‌اند)، اما این فرهنگ نیز با اتکا به داده‌های پیکرهٔ زبانی تهیه شده و در بخش‌هایی نیز اطلاعات بسامدی برای واژه‌ها ارائه شده است. «بی‌جن‌خان، نصری و جلایی» (۱۳۹۳) بسامد واژه‌ها در پیکرهٔ رسمی را با بسامدشان در پیکره‌ای محاوره‌ای مقایسه می‌کنند. «جهانگرد» و همکاران (۱۳۹۵) و «نویدی، عامری و ابوالحسن چیمه» (۱۴۰۰) نیز به بررسی بسامدی واژه‌ها در کتاب‌های آموزش زبان فارسی به غیرفارسی‌زبانان می‌پردازند. همچنین «صحرایی، مجیری فروشانی و طالبی» (۱۳۹۸) فهرستی بسامدی از واژه‌های پایهٔ زبان فارسی ارائه می‌کنند که البته، پیکره‌ای از متون مطبوعاتی مبنای تهیهٔ این فهرست است و «ذوالفقار»

1. expert pdf

و همکاران (۱۳۹۹) نیز پژوهشی در زمینهٔ واژگان پایهٔ علوم پزشکی انجام می‌دهند. تهیهٔ واژه‌های پایهٔ کودکان فارسی‌زبان بر اساس اطلاعات بسامدی از دیگر پژوهش‌های رایج در این حوزه است که در پژوهش «عامری و ذوالفقاری» (۱۳۹۱) به مقایسهٔ چند نمونه از تحقیقات مهم در این زمینه پرداخته می‌شود. پژوهش «صحرایی، طالبی و مجیری فروشانی» (۱۳۹۶) پژوهش دیگری است که در آن به مقایسهٔ واژه‌های پایهٔ زبان فارسی در شش پژوهش پرداخته می‌شود. یک دسته از فهرست‌های بسامدی که در زبان فارسی به آن‌ها بسیار پرداخته شده، فرهنگ‌های بسامدی ادبی است که برای یک اثر ادبی و یا تمام آثار شاعر یا نویسنده‌ای مشخص تهیه می‌شوند. این فرهنگ‌ها به خوانندگان آثار ادبی کمک می‌کنند که هر چه سریع‌تر با واژه‌های مورد نیاز برای مطالعهٔ متن مورد نظرشان آشنا شوند. «مجلی‌زاده» (۱۳۹۴) فهرستی کتاب‌شناختی از حدود ۱۰۰ «پژوهش بسامدی» فارسی ارائه می‌کند که اغلب از جنس واژه‌نامه‌های بسامدی برای آثار ادبی هستند. برای نمونه، می‌توان به «فرهنگ بسامدی و تصویری دیوان حافظ» «انوری، معین‌الدینی و معین‌الدینی» (۱۳۸۵) و «دیوان پروین اعتصامی به انضمام فرهنگ بسامدی» «دانشگر» (۱۳۹۰) اشاره کرد.

پیش از آنکه به روش انجام پژوهش حاضر بپردازیم، ضروری است که اشاره‌ای به وضعیت پیکره‌ها در زبان فارسی داشته باشیم. پیکره‌های حاوی متون علمی در زبان فارسی به‌طور عام، با اهدافی غیر از تهیهٔ فهرست‌های بسامدی واژه‌های علمی تهیه شده و به‌دلایل مختلف قابل استفاده در پژوهش حاضر نیستند؛ از جمله اینکه محدود به یک یا چند رشتهٔ تخصصی هستند یا برای اهداف خاص پردازش روی متون علمی (مانند ارزیابی سامانه‌های مشابهت‌یاب و خلاصه‌ساز) توسعه داده شده‌اند و ساختار آن‌ها برای تهیهٔ فهرست‌های بسامدی مناسب نیست یا «متوازن» نیستند و بنابراین، نمی‌توان از آن‌ها برای توسعهٔ فهرست واژگان علمی که در موضوعات مختلف توزیع مناسبی دارند بهره گرفت؛ و یا حجم کوچکی دارند و در دسترس نیستند. پیکره‌های «نور» Elahimanesh et al. (2012)، «محک سمیم» Rezaei Sharifabadi and Eftekhari (2016)، «کامیابی گل» و همکاران (۱۳۹۷)، «شریفی و مهدوی» (۱۳۹۷)، «محرابی، محبی و احمدی» (۱۴۰۰)، «علایی ابوذر» و همکاران (۱۴۰۰)، «صدیقی‌فر، رحیمیان و خرمایی» (۱۴۰۰)، «قیومی و موسویان» (۱۴۰۱)، و «رضایی دینانی» (۱۴۰۱) از جمله پیکره‌هایی هستند که برای تهیهٔ آن‌ها از متون علمی استفاده شده است.

۲. روش پژوهش

۲-۱. تولید پیکرهٔ علمی زبان فارسی

به‌منظور استخراج واژگان پربسامد علمی به پیکره‌ای بزرگ از متون علمی نیاز داریم. در پژوهش حاضر منبع اصلی برای جمع‌آوری متون علمی، نشریات علمی دانشگاهی است. مقالات منتشرشده در چنین نشریاتی مزیت‌های مختلفی دارند، از جمله اینکه به‌طور معمول به‌صورت آزاد منتشر می‌شوند و بنابراین، هم از نظر دسترس‌پذیر بودن، گزینهٔ مناسبی برای جمع‌آوری و استفاده در پیکره هستند و هم از نظر حقوق مالکیت معنوی و امکان بازنشر. افزون بر این، مقالات علمی داوری می‌شوند. بنابراین، از نظر سطح علمی و همچنین رعایت استانداردهای نگارش علمی مورد ارزیابی قرار می‌گیرند و از این جهت تا حدی قابل اطمینان هستند. در پژوهش‌های پیشین در زمینهٔ تهیهٔ فهرست‌های بسامدی علمی نیز پیکره‌های مورد نیاز در موارد متعددی با استفاده از مقالات نشریات علمی تهیه شده‌اند؛ برای نمونه Wang, Liang, and Ge (2008)، Lei, Lei and Liu (2016) و Gilmore and Millar (2018).

برای دسترسی به فهرست نشریات علمی، از سامانه‌های نشریات علمی وزارت علوم، تحقیقات و فناوری و وزارت بهداشت، درمان و آموزش پزشکی استفاده شد[۱]. برای دریافت متن مقالات نشریات علمی، با استفاده از برنامه‌نویسی رایانه‌ای یک خزندهٔ وب[۲] توسعه دادیم. ابتدا، فهرست کامل نشریات و مشخصات آن‌ها (شامل عنوان، موضوع و زیرموضوع، زبان، url و ...) را از سامانه‌های نشریات وزارتخانه‌های مزبور استخراج کردیم. سپس، با بررسی وبگاه تک‌تک نشریات با استفاده از خزنده، مشخص کردیم که سامانهٔ مدیریت نشریات مورد استفادهٔ هر کدام چیست. بعد از آن، خزنده‌هایی را برای استخراج مقاله از نشریات فارسی که از دو مورد از رایج‌ترین سامانه‌های مدیریت نشریات استفاده می‌کردند، توسعه دادیم. بیش از ۹۲ درصد از نشریات وزارت علوم، تحقیقات و فناوری و بیش از ۸۸ درصد از نشریات وزارت بهداشت، درمان و آموزش پزشکی از این دو

---

۱. پرتال نشریات علمی وزارت علوم، تحقیقات و فناوری به آدرس https://journals.msrt.ir و بانک اطلاعات نشریات علوم پزشکی کشور (وابسته به وزارت بهداشت، درمان و آموزش پزشکی) به آدرس https://journals.research.ac.ir.

2. web crawler

سامانه[1] استفاده کرده‌اند. هر جا که به متن کامل ماشین‌خوان مقاله دسترسی داشتیم، آن را دریافت و در بقیهٔ موارد به دریافت چکیدهٔ ماشین‌خوان مقاله بسنده کردیم. لازم به ذکر است که برای تهیهٔ پیکره از متونی استفاده شد که به‌طور مستقیم در وبگاه نشریات منتشر شده بود و با توجه به چالش‌های استفاده از فایل‌های پی‌دی‌اف فارسی، سراغ آن‌ها نرفتیم. با این حال، استفاده از متونی که به‌طور مستقیم در وبگاه نشریات منتشر شده بود نیز بدون چالش نبود. با توجه به تعدد و تنوع پایگاه‌های مورد استفاده برای استخراج متون مقالات، شوربختانه نایکدستی‌های زیادی در متون مقالات و در نتیجه، داده‌های جمع‌آوری‌شده برای تولید پیکره دیده می‌شد. یکی از مهم‌ترین این نایکدستی‌ها در نویسه‌های[2] به‌کاررفته در متن مقالات بود. برای مثال، ۲۰ نوع نویسه «ی» و ۱۹ نوع نویسهٔ «ک» در داده‌های جمع‌آوری‌شده دیده شد که همگی با استفاده از کدنویسی به نویسه‌های استاندارد تبدیل شد. افزون بر دو حرف فوق (که نگارش آن‌ها در متون فارسی با نویسهٔ غیراستاندارد به‌نوعی رایج است)، برای تمام حروف دیگر الفبای فارسی نیز با همین مسئلهٔ عدم یکدستی در نویسه‌ها مواجه بودیم. برای مثال، در مقالات جمع‌آوری‌شده برای نگارش حرف «چ» از نویسه‌های «چ» «چ» و «چ» نیز استفاده شده است. لازم به توضیح است که اگرچه ممکن است واژه‌هایی که با این نویسه‌های غیراستاندارد نگاشته می‌شوند، در ظاهر تفاوتی با واژه‌های نوشته‌شده با نویسه‌های استاندارد نداشته باشند، اما این‌ها از نظر رایانه واژه‌هایی کاملاً متفاوت از هم هستند و بنابراین، هنگام تهیهٔ فهرست بسامدی نیز به‌عنوان واژه‌های مستقل شناسایی و شمارش می‌شوند. به‌منظور رفع این مشکل، یک نگاشت کامل از نویسه‌های غیراستاندارد به نویسه‌های استاندارد در کد پیش‌پردازش تهیه شد و نویسه‌های تمام متون جمع‌آوری‌شده به این وسیله به نویسه‌های استاندارد تبدیل شد. یکی از نویسه‌هایی مهمی که در این نگاشت اصلاح شد، نویسهٔ نیم‌فاصله بود که با بیش از ۱۰ نوع نویسهٔ غیراستاندارد در متون جمع‌آوری‌شده به‌کار رفته بود (از جمله نویسه‌های دارای یونی‌کد u2005، u200a، u200b، u200d، u200e، u200f، u202b، u202c، u202d، u202f، ue825، xad). افزون بر این، نویسهٔ نیم‌فاصله در بسیاری از موارد در جایگاه نامناسب، یعنی بلافاصله بعد از حروف منفصل به‌کار رفته بود و همچنین گاهی

۱. سیناوب و یکتاوب

2. charachters

به‌جای یک نیم‌فاصله از چند نیم‌فاصله استفاده شده بود که این موارد نیز همگی در کد پیش‌پردازش اولیهٔ متون پیکره اصلاح شدند. گذشته از موارد فوق، با توجه به اینکه هدف اولیهٔ تهیهٔ پیکرهٔ حاضر استخراج فهرست بسامدی واژه‌های علمی فارسی بود، اعداد و کاراکترهای غیرفارسی نیز از متون پیکره حذف شد.

نکتهٔ دیگر اینکه از نظر سال انتشار، به‌منظور اطمینان از به‌روز بودن متون پیکره و واژه‌های مورد استفاده، مقالات منتشرشده از سال ۱۳۷۰ تا پایان سال ۱۳۹۹ دریافت شدند و مقالات دوره‌های قبلی نشریات در آن‌ها به پیکره اضافه نشدند. سرانجام، پس از جمع‌آوری متون و متوازن‌سازی آن‌ها بر اساس موضوعات و زیرموضوع‌ها، پیکره‌ای با مشخصات زیر به‌دست آمد:

جدول ۱. آمار نهایی پیکرهٔ علمی فارسی

موضوع	↓ تعداد واژه
علوم پایه و فنی و مهندسی	۱۲۷۹۳۲۵۳
علوم انسانی و هنر و معماری	۱۲۷۸۴۶۰۰
پزشکی و دامپزشکی	۱۲۷۵۵۱۳۱
کشاورزی و منابع طبیعی	۱۲۷۵۲۷۰۳
جمع	۵۱۰۸۵۶۸۷

به‌منظور تهیهٔ فهرست بسامدی، به ریشه و برچسب اجزای کلام واژه‌های پیکره نیز نیاز داشتیم. برای این منظور، از ابزار متن باز «هضم»[1] که یک کتابخانهٔ پایتون[2] برای پردازش‌های پایه روی زبان فارسی است، استفاده شد و تمام واژه‌های پیکره ریشه‌یابی شدند و برچسب اجزای کلام دریافت کردند (با دقت حدود ۹۷ درصد)[3]. پیش از این مرحله نیز از همین ابزار «هضم» برای نرمال‌سازی و تقطیع واژه‌های پیکره استفاده شد.

در پایان، پیکره به دو بخش تقسیم شد. یک بخش بزرگ‌تر با حجم حدود ۴۰ میلیون و ۷۰۰ هزار واژه که برای استخراج فهرست بسامدی واژه‌های علمی استفاده شد، و یک بخش کوچک‌تر با حجم حدود ۱۰ میلیون و ۳۰۰ هزار واژه که جهت ارزیابی مستقل فهرست بسامدی استخراج‌شده مورد استفاده قرار گرفت. لازم به ذکر است که

1. https://github.com/sobhe/hazm 2. Python library 3. https://www.roshan-ai.ir/hazm

هنگام تقسیم پیکره به این دو بخش، توازن پیکره در هر یک از این دو بخش حفظ شد. به‌عبارتی، کدنویسی رایانه‌ای به‌نحوی انجام شد که به‌میزان متوازن از متون هر حوزهٔ موضوعی برای تهیهٔ بخش‌های حدود ۴۰ و ۱۰ میلیون واژه‌های استفاده شود.

۲-۲. استخراج فهرست واژه‌های پربسامد علمی فارسی

پس از تهیهٔ پیکره، نوبت به استخراج فهرست واژه‌های علمی پربسامد فارسی می‌رسد. در پژوهش حاضر، مانند پژوهش‌های Paquot (2007)، Gardner and Davies (2013)، Brezina and Gablasova (2015) و Lei, Lei and Liu (2016) دسته‌بندی واژه‌ها بر اساس بن‌واژه و با توجه به برچسب اجزای کلام و ریشهٔ واژه‌ها انجام شد. برای مثال، واژه‌های «شاخص، شاخص‌ها، شاخص‌های و ...» در دستهٔ «شاخص|اسم» قرار می‌گیرند و واژه‌های «شاخص، شاخص‌تر، شاخص‌ترین و ...» در دستهٔ «شاخص|صفت». بنابراین، معیار «هم‌خانواده بودن» برای دسته‌بندی واژه‌ها استفاده نمی‌شود و برای مثال، واژهٔ «غیرشاخص» در دسته‌های نمونهٔ گفته شده قرار نمی‌گیرد. علت این مسئله، افزون بر معایبی که برای دسته‌بندی بر اساس خانوادهٔ واژگانی ذکر و در پیشینهٔ پژوهش به آن‌ها اشاره شد، آن است که تا جایی که نگارندگان این مقاله اطلاع دارند، در زبان فارسی فهرست واژه‌های هم‌خانوادهٔ فارسی با ساختار مناسب برای پژوهش حاضر در دسترس نیست؛ البته داده‌هایی مانند «فارس‌نت»[1] که در آن‌ها انواع روابط میان واژه‌ها ثبت شده است، می‌توانند مبنایی برای استخراج چنین فهرست‌هایی قرار گیرند.

مسئلهٔ بعد تعیین معیارهایی است برای افزودن واژه‌ها به فهرست بسامدی واژه‌های علمی. معیارهایی که در این پژوهش مد نظر قرار گرفت، به‌شرح زیر است:

معیار اول، بسامد. بر اساس این معیار تعداد رخداد هر بن‌واژه در کل پیکره شمارش شده و بن‌واژه‌های دارای بسامد بیشتر امتیاز بالاتری برای قرارگرفتن در فهرست بسامدی خواهند داشت. بن‌واژه‌ها برای قرار گرفتن در فهرست بسامدی این پیکره باید به‌طور متوسط در هر یک میلیون واژه ۳۰ بار به‌کار رفته باشند (حدود حداقل ۱۲۰۰ بار در بخش ۴۰ میلیون واژه‌ای پیکره). لازم به ذکر است که همان‌گونه که در پیشینه اشاره شد، معیار بسامد مورد استفاده در پژوهش (Coxhead 2000) که مبنای تعیین حداقل بسامد در بسیاری

1. http://farsnet.nlp.sbu.ac.ir

از پژوهش‌های بعدی بود، حدود ۲۸/۵۷ وقوع در یک میلیون واژه (۱۰۰ وقوع در پیکرهٔ ۳/۵ میلیون واژه‌ای) بوده است.

معیار دوم، نسبت. از آنجا که هدف از این پژوهش تهیهٔ فهرست واژه‌های «علمی» است، یعنی واژه‌هایی که فراوانی آن‌ها در متون علمی به شکل معناداری بالاتر از فراوانی‌شان در سایر متون است، از معیار نسبت استفاده شد. برای لحاظ کردن این معیار، نسبت بسامد نسبی هر بن‌واژه در پیکرهٔ عمومی به بسامد نسبی آن در پیکرهٔ علمی محاسبه شد، و برای قرار گرفتن بن‌واژه‌ها در فهرست بسامدی علمی باید بسامد نسبی بن‌واژه در پیکرهٔ علمی حداقل ۲ برابر بسامد نسبی آن در پیکرهٔ عمومی باشد. Gardner and Davies (2013) ضریب ۱/۵ برابر را برای این معیار در نظر می‌گیرند، اما با توجه به بررسی‌های انجام‌شده روی داده‌های فارسی جمع‌آوری‌شده، ضریب سخت‌گیرانه‌تر دوبرابر برای تعیین واژه‌های علمی فارسی مناسب تشخیص داده شد (با در نظر گرفتن ضریب ۱/۵، واژه‌هایی که مانند «روستا»، «حس» و «تابستان» عمومی به نظر می‌رسند، در فهرست قرار می‌گرفت). برای اطمینان از قرار نگرفتن واژه‌های عمومی در فهرست علمی راه‌حل دیگر آن بود که مانند «کاکسهد» از واژه‌های پربسامد فهرست‌های عمومی موجود، مانند فرهنگ بسامدی پیکرهٔ متنی «بی‌جن‌خان و محسنی» استفاده و آن‌ها را از فهرست خود حذف کنیم. اما با توجه به عدم دسترسی به نسخهٔ ماشین‌خوان فرهنگ‌های بسامدی عمومی موجود برای زبان فارسی و همچنین، اختلاف در برخی برچسب‌ها، سراغ این روش نرفتیم، بلکه برای این منظور پیکرهٔ متنی ۱۰ میلیون واژه‌ای فارسی Bijankhan et al. (2011) را با استفاده از همان ابزار «هضم» برچسب‌گذاری کرده و بسامد بن‌واژه‌ها در پیکرهٔ خودمان را با بسامد آن‌ها در این پیکره مقایسه نمودیم. لازم به ذکر است که پیکرهٔ متنی مورد استفاده، خود پیکره‌ای برچسب‌خورده است، اما با هدف هماهنگی بیشتر در برچسب‌های پیکرهٔ علمی و پیکرهٔ متنی برای برچسب‌گذاری هر دو داده از ابزاری مشترک استفاده شد. برای این منظور پس از استخراج بسامد تمام بن‌واژه‌ها در پیکرهٔ علمی و پیکرهٔ عمومی با استفاده از برنامه‌نویسی رایانه‌ای، این اطلاعات بسامدی را به نرم‌افزار «مایکروسافت اکسل»[1] منتقل کردیم و در آنجا با استفاده از فرمول‌نویسی «اکسل»، بن‌واژه‌هایی را که بسامد نسبی آن‌ها (به ازای یک میلیون واژه) در پیکرهٔ علمی حداقل دوبرابر بسامد نسبی آن‌ها (به ازای یک

1. Microsoft Excel

میلیون واژه) در پیکرهٔ عمومی بود، مشخص کردیم. لازم به ذکر است که علت استفاده از برنامه‌نویسی رایانه‌ای جهت بسامدشماری به‌جای نرم‌افزارهای تحلیل پیکره‌ای آماده مانند «انت‌کانک»[1] و ... حجم بسیار بالای پیکره‌های مورد استفاده و چالش‌های سرعت است که این حجم از داده برای چنین نرم‌افزارهایی ایجاد می‌کند.

معیار سوم، پراکندگی. این معیار نشان می‌دهد که یک واژه تا چه حد به‌صورت متوازن در متون مختلف یک پیکره به‌کار رفته است. این معیار برای اطمینان از تخصصی نبودن واژهٔ مورد بررسی استفاده می‌شود. توضیح اینکه ممکن است واژه‌ای بسامد بالایی داشته باشد و عمومی هم نباشد (دو معیار قبل)، اما تنها در یکی از دسته‌های موضوعی پیکرهٔ علمی بسامد بالایی داشته باشد. در این صورت، دیگر نمی‌توان آن را واژهٔ «علمی» با تعریف ارائه‌شده در نظر گرفت که در رشته‌های مختلف علمی به‌کار می‌رود، بلکه واژه‌ای تخصصی است. در پژوهش حاضر برای اندازه‌گیری پراکندگی از معیار «ضریب پراکندگی جویلاند» با حداقل مقدار ۰/۵ استفاده شد. لحاظ کردن کمینهٔ قابل قبول بالاتر برای این معیار باعث حذف واژه‌های علمی مانند «عملکرد»، «نشانگر» و «بیشینه» از فهرست می‌شود.

معیار چهارم، دامنه. بر اساس این معیار، وقوع بن‌واژهٔ مورد نظر در هر یک از ۴ دستهٔ موضوعی پیکره نباید کمتر از یک‌سوم بسامد مورد انتظار[2] آن در این دسته‌ها باشد. همان‌طور که در پیشینهٔ پژوهش اشاره شد، این معیار در پژوهش Coxhead (2000) به‌صورت حداقل ۱۰ وقوع خانوادهٔ واژگانی در هر یک از ۴ رشتهٔ پیکره و در پژوهش Gardner and Davies (2013) به‌صورت وقوع بن‌واژه به اندازهٔ حداقل ۲۰ درصد از بسامد مورد انتظار در دست کم ۷ دستهٔ موضوعی از مجموع ۹ موضوع پیکره تعریف شده است.

بر اساس معیارهای فوق، داده‌های بخش آموزش پیکرهٔ علمی امتیازبندی شدند و بن‌واژه‌های دارای حداقل‌های مورد نظر در فهرست واژه‌های پربسامد علمی قرار گرفتند.

۳. یافته‌ها

در این بخش ابتدا، فهرست بسامدی استخراج‌شده از پیکرهٔ علمی ارائه می‌شود و در ادامه نیز به روش و نتایج ارزیابی فهرست استخراج‌شده خواهیم پرداخت.

1. AntConc
2. expected frequency

۳-۱. فهرست بسامدی واژه‌های علمی زبان فارسی

بر اساس معیارهای تشریح‌شده در بخش ۲-۲، فهرست واژه‌های علمی فارسی استخراج شد (شامل ۳۰۷ بن‌واژه) که ۱۰۰ بن‌واژهٔ نخست آن[1] عبارت است از:

۱. استفاده ۲. روش ۳. نتیجه ۴. بررسی ۵. مورد ۶. مطالعه ۷. نشان ۸. افزایش ۹. میزان ۱۰. انجام ۱۱. پژوهش ۱۲. سطح ۱۳. گشت#گرد ۱۴. هدف ۱۵. تحقیق ۱۶. اثر ۱۷. نمونه ۱۸. مختلف ۱۹. تحلیل ۲۰. نسبت ۲۱. عامل ۲۲. عملکرد ۲۳. تأثیر ۲۴. محیط ۲۵. مدل ۲۶. شاخص ۲۷. جهت ۲۸. تعداد ۲۹. داده ۳۰. مقایسه ۳۱. تعیین ۳۲. دارا(ی) ۳۳. آزمون ۳۴. تولید ۳۵. میانگین ۳۶. حاضر ۳۷. بالا ۳۸. شامل ۳۹. ارزیابی ۴۰. شرایط ۴۱. نوع ۴۲. آمار ۴۳. منظور ۴۴. کنترل ۴۵. منبع ۴۶. رشد ۴۷. انتخاب ۴۸. متغیر ۴۹. حاصل ۵۰. ترتیب ۵۱. شکل ۵۲ـ مرحله ۵۳. مناسب ۵۴. مشاهده ۵۵. تحت ۵۶. طول ۵۷. کیفیت ۵۸. مدیریت ۵۹. فرایند ۶۰. ضریب ۶۱. تجزیه ۶۲. تفاوت ۶۳. ویژگی ۶۴. سپس ۶۵. رفتار ۶۶. بهبود ۶۷. محاسبه ۶۸. شناسایی ۶۹. باعث ۷۰. درجه ۷۱. ساختار ۷۲. اهمیت ۷۳. معنی ۷۴. الگو ۷۵. ترکیب ۷۶. عدم ۷۷. سیستم ۷۸. جدول ۷۹. بنابراین ۸۰. اندازه ۸۱. معیار ۸۲. مثبت ۸۳. وضعیت ۸۴. طراحی ۸۵. مشخص ۸۶. همبستگی ۸۷. متوسط ۸۸. اختلاف ۸۹. کاربرد ۹۰. پوشش ۹۱. پایه ۹۲. ناشی ۹۳. کلی ۹۴. متفاوت ۹۵. ابزار ۹۶. بافت ۹۷. شدت ۹۸. استاندارد ۹۹. سبب ۱۰۰. استخراج

با نگاه به فهرست فوق ممکن است این پرسش پیش بیاید که واژه‌ای مانند «استفاده» که واژهٔ نخست فهرست بسامدی واژه‌های علمی است، به نظر می‌رسد که واژه‌ای عمومی با بسامد بالا باشد، برای چه در فهرست واژه‌های علمی آمده است؟ باید توجه داشت که همان‌طور که در روش پژوهش توضیح دادیم، مبنای حذف واژه‌های عمومی پربسامد از فهرست حاضر، مانند دسته‌ای از پژوهش‌های پیشین، مقایسهٔ بسامد نسبی واژه‌ها در پیکرهٔ

۱. علاقه‌مندان می‌توانند جهت دسترسی به فهرست کامل واژه‌های علمی به همراه بسامد و همچنین پیکرهٔ علمی، درخواست خود را به نویسندهٔ اول مقاله ارسال کنند.

علمی و پیکرهٔ عمومی بوده است. برای مثال، بسامد واژهٔ «استفاده» در پیکرهٔ علمی ۵۵۷۱ وقوع به ازای هر یک میلیون واژه بود، در حالی که در پیکرهٔ عمومی ۱۰۰۰ وقوع به ازای هر یک میلیون واژه بوده است؛ یعنی بسامد نسبی «استفاده» در پیکرهٔ علمی حدود ۵/۵۷ برابر بسامد نسبی آن در پیکرهٔ عمومی است. به‌عبارت دیگر، اگرچه واژهٔ «استفاده» در متون عمومی هم بسامد بالایی دارد، اما همان‌طور که مشخص است در متون علمی بسیار برجسته‌تر است و بنابراین، هنگام آموزش زبان فارسی برای اهداف علمی حتماً باید در اولویت قرار گیرد. لازم به ذکر است که در فهرست‌های شناخته‌شده از واژه‌های علمی انگلیسی مانند Coxhead (2000) واژه‌هایی چون aware، brief، classic، edit، gender و در Gardner and Davies (2013) نیز واژه‌هایی مانند provide، however، table، both، need که شاید بیشتر واژه‌های عمومی به نظر برسند، درج شده‌اند.

۳-۲. ارزیابی فهرست بسامدی

روش معمول برای ارزیابی فهرست‌های بسامدی این است که بررسی شود که فهرست استخراج‌شده تا چه حد متون علمی و عمومی را پوشش می‌دهد. منظور از پوشش، آن است که محاسبه شود که واژه‌هایی که در فهرست بسامدی قرار گرفته‌اند، چند درصد از کل واژه‌های پیکره را تشکیل می‌دهند. برای این منظور سه آزمایش انجام شد:

۱. بررسی میزان پوشش روی بخش حدود ۴۰ میلیون واژه‌ای پیکرهٔ علمی. این ارزیابی نشان داد که ۳۰۷ بن‌واژهٔ انتخاب‌شده برای فهرست بسامدی، ۱۶/۶۹ درصد از این بخش از پیکرهٔ علمی را که برای استخراج فهرست بسامدی استفاده شده بود، تشکیل می‌دهد. این میزان از پوشش با توجه به درصدهای گزارش‌شده در پژوهش‌های پیشین، عدد قابل قبولی است. برای مثال، فهرست بسامدی Coxhead (2000)، شامل ۵۷۰ خانوادهٔ واژگانی، ۱۰ درصد از واژه‌های پیکرهٔ ۳/۵ میلیون واژه‌ای را که فهرست بسامدی بر مبنای آن تهیه شده بود، پوشش می‌دهد و ۵۷۰ خانوادهٔ واژگانی نخست فهرست بسامدی Gardner and Davies (2013)، ۱۳۸ درصد از پیکره‌ای را که با آن ساخته شده، پوشش می‌دهد.

۲. بررسی میزان پوشش روی بخش مستقل حدود ۱۰ میلیون واژه‌ای پیکرهٔ علمی. ارزیابی فهرست واژه‌های علمی پربسامد روی این بخش از پیکرهٔ علمی نشان داد که این

فهرست شامل ۱۶/۱۳ درصد از کل واژه‌های این بخش می‌شود. این نتیجه با نتایج حاصل از ارزیابی فهرست‌های پیشین روی پیکره‌های مستقل انطباق دارد و نشان می‌دهد که درصد پوشش در داده‌های غیر از داده‌ای که فهرست از آن استخراج شده نیز مناسب است. در آزمایش مشابهی که «کاکسهد» روی پیکره‌ای مستقل انجام داده بود، پوشش فهرست واژگان علمی او ۸/۵ درصد بوده است.

۳. بررسی میزان پوشش روی پیکرهٔ عمومی (۱۰ میلیون واژه). در این بخش همان‌گونه که گفته شد، از پیکرهٔ متنی زبان فارسی به‌عنوان پیکرهٔ عمومی استفاده شد. میزان پوشش فهرست بسامدی استخراج‌شده در این پیکره ۴/۱۳ درصد بود. این پوششِ پایین نشان‌دهندهٔ آن است که فهرست انتخاب‌شده به‌درستی شامل واژه‌های علمی پربسامد است و فهرستی عمومی از واژه‌های پربسامد نیست.

یکی از مسائلی که در ارزیابی فهرست‌های بسامدی می‌توان در نظر گرفت، میزان پوشش فهرست در بخش‌های مختلف پیکره یا به‌عبارت دیگر، موضوعات مختلف علمی است. نتایج این بررسی روی فهرست واژه‌های علمی استخراج‌شده در این پژوهش به‌شرح زیر است:

جدول ۲. میزان پوشش فهرست واژه‌های علمی در بخش‌های مختلف پیکره

موضوع	↓ میزان پوشش فهرست واژه‌های علمی (درصد)
کشاورزی و منابع طبیعی	۱۸/۶۸
علوم پایه و فنی و مهندسی	۱۷/۷۸
پزشکی و دامپزشکی	۱۷/۴۱
علوم انسانی و هنر و معماری	۱۲/۸۲

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

همان‌طور که در جدول ۲، مشخص است، در سه موضوع کشاورزی و منابع طبیعی، علوم پایه و فنی و مهندسی، و پزشکی و دامپزشکی میزان پوشش فهرست بسامدی واژه‌های علمی با اختلاف کمتر از ۱/۵ درصد بسیار نزدیک به هم و به‌ترتیب ۱۸/۶۸، ۱۷/۷۸ و ۱۷/۴۱ درصد است. از سوی دیگر، میزان پوشش در علوم انسانی، هنر و معماری قدری پایین‌تر از موضوعات دیگر و حدود ۱۲/۸۲ درصد است. این اختلاف ممکن است به‌دلیل تفاوت روش‌های علمی مورد استفاده در دستهٔ علوم انسانی، هنر و معماری نسبت به سایر

علوم که بیشتر جنبهٔ تجربی دارند، باشد. در سایر پژوهش‌ها هم شاهد اختلاف در میزان پوشش فهرست‌ها در بخش‌های مختلف پیکره هستیم. برای مثال، در پژوهش Coxhead (2000)، میزان پوشش فهرست برای بخش تجارت پیکره ۱۲ درصد، حقوق ۹/۴ درصد، هنر ۹/۳ درصد و علوم ۹/۱ درصد بوده است.

٤. نتیجه‌گیری

پژوهش حاضر حداقل دو دستاورد مهم داشته است: یکی تهیهٔ پیکره‌ای به‌روز و متوازن از متون علمی فارسی شامل بالغ بر ۵۱ میلیون واژه از متون مقالات در ۴ دستهٔ موضوعی. دوم استخراج فهرستی بسامدی از واژه‌های پرکاربرد در رشته‌های مختلف علمی که بر اساس ارزیابی‌های انجام‌شده بیش از ۱۶ درصد متون علمی را پوشش می‌دهد. ابتدا با بررسی پیشینه نشان دادیم که چه ملاحظاتی در تهیهٔ فهرست‌های بسامدی وجود دارد. این ملاحظات عبارت‌اند از نحوهٔ دسته‌بندی واژه‌ها برای شمارش، حداقل بسامد قابل قبول برای قرارگرفتن واژه‌ها در فهرست بسامدی، نحوهٔ ایجاد تمایز میان واژه‌های پربسامد علمی و واژه‌های پربسامد عمومی، و سرانجام، نحوهٔ انتخاب واژه‌هایی که در طیف وسیعی از موضوعات علمی بسامد بالایی دارند. در اینجا لازم به ذکر است که متون فارسی، حتی متون علمی که موضوع بررسی این پژوهش هستند، دارای نایکدستی‌های فراوانی هستند که این نایکدستی‌ها برای دسته‌بندی و شمارش واژه‌ها چالش‌های جدی ایجاد می‌کنند. از جملهٔ مهم‌ترین نایکدستی‌های مشاهده‌شده در داده‌ها می‌توان به مشکلات فاصله‌گذاری (پیوسته‌نویسی، نزدیک‌نویسی، جدانویسی) و مسائل دیگر مربوط به رسم‌الخط، از جمله نحوهٔ به‌کارگیری همزه و مد و همچنین استفاده از انواع نویسه‌های غیراستاندارد اشاره کرد. طبیعی است هرچه ابزارهای پیش‌پردازشی دقت بالاتری در یکدست‌سازی متون فارسی داشته باشند، فهرست‌های بسامدی استخراج‌شده نیز کیفیت بهتری خواهند داشت.

از فهرست بسامدی علمی استخراج‌شده می‌توان برای اهداف مختلف آموزشی و پژوهشی از جمله تدوین مواد آموزشی برای آموزش و یادگیری زبان فارسی برای اهداف دانشگاهی استفاده کرد. Coxhead (2011) در مقالهٔ خود مجموعه‌ای از کاربردهای آموزشی و پژوهشی را که فهرست بسامدی او (Coxhead 2000) طی بیش از ۱۰ سال داشته است، تشریح می‌کند که می‌تواند ایده‌های خوبی برای چگونگی استفاده از فهرست بسامدی علمی فارسی به دست دهد. از سوی دیگر، پیکرهٔ علمی تهیه‌شده نیز منبع ارزشمندی برای

مطالعات حوزهٔ زبان‌شناسی پیکره‌ای و پژوهش‌های مرتبط با پردازش زبان است و می‌توان از آن به‌عنوان داده‌ای زیرساختی در حوزه‌هایی چون بهبود بازیابی اطلاعات، خلاصه‌سازی متون، استخراج کلیدواژه، نمایه‌سازی خودکار، مشابهت‌یابی خودکار و ... استفاده کرد.

یکی از کارهایی که در پژوهش‌های آینده در این حوزه می‌توان انجام داد، عبارت است از استفاده از منابع علمی غیر از مقالات علمی (شامل پایان‌نامه‌ها، کتاب‌های علمی، مجلات علمی و ...) و همچنین منابع علمی گفتاری (مکالمات در کلاس‌های دانشگاهی، سخنرانی‌های علمی و ...) برای تکمیل پیکرهٔ علمی. همچنین، بهبود ابزارهای پایهٔ پردازشی فارسی شامل نرمالایزر[1]، تقطیع‌گر، ریشه‌یاب و برچسب‌زن اجزای کلام می‌تواند به استخراج فهرست‌های بسامدی دقیق‌تر برای زبان فارسی کمک شایانی کند.

قدردانی

بدین‌وسیله از جناب آقای مهندس سپهر رضائی و جناب آقای مهندس حسن صادقی که در بخش‌هایی از برنامه‌نویسی رایانه‌ای این طرح کمک کردند، صمیمانه تشکر می‌کنیم.

فهرست منابع

انوری، حسن، احمد معین‌الدینی، و فاطمه معین‌الدینی. ۱۳۸۵. *کلک خیال‌انگیز: فرهنگ بسامدی و تصویری دیوان حافظ*. تهران: سخن.

بی‌جن‌خان، محمود، و مهدی محسنی. ۱۳۹۷. *فرهنگ بسامدی بر اساس پیکرهٔ متنی زبان فارسی امروز*. تهران: دانشگاه تهران.

بی‌جن‌خان، محمود، عباس نصری، و شهره جلایی. ۱۳۹۳. نقش واژگان بسامدی در ارزیابی مهارت واژگانی فارسی‌آموزان. *پژوهشنامهٔ آموزش زبان فارسی به غیرفارسی‌زبانان* ۳ (۲): ۲۵-۴۵.

جهانگرد، کیومرث، مصطفی عاصی، آزیتا افراشی، و امیررضا وکیلی‌فرد. ۱۳۹۵. واژه در کتاب‌های آموزش زبان فارسی به غیرفارسی‌زبانان: پژوهش‌ی پیکره‌بنیاد. *پژوهشنامهٔ آموزش زبان فارسی به غیرفارسی‌زبانان* ۵ (۲): ۳-۲۶.

حسنی، حمید. ۱۳۸۴. *واژه‌های پرکاربرد فارسی امروز بر مبنای پیکره یک میلیون لغتی*. تهران: کانون زبان ایران.

1. normalizer

دانشگر، احمد. ۱۳۹۰. *دیوان پروین اعتصامی به انضمام فرهنگ بسامدی اشعار*. تهران: جهان‌تاب.

دبیرخانهٔ شورای‌عالی انقلاب فرهنگی. ۱۳۸۹. *نقشهٔ جامع علمی کشور*. تهران: دبیرخانهٔ شورای‌عالی انقلاب فرهنگی.

ذوالفقار، زهره، طیبه موسوی میانگاه، بلقیس روشن، و امیررضا وکیلی‌فرد. ۱۳۹۹. بررسی تکنیک‌های بهبود عملکرد روش‌های بسامدشماری پیکره‌بنیاد در استخراج خودکار واژگان. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۵ (۴): ۱۰۳۹-۱۰۶۳.

رضایی دینانی، مینا. ۱۴۰۱. واکاوی تأثیر برچسب‌گذاری معنایی در ابهام‌زدایی هم‌نویسه‌های تخصصی از نظر کیفیت بازیابی (معیار F) در بازیابی متون علمی. پژوهشنامهٔ پردازش و مدیریت اطلاعات. دریافت از https://jipm.irandoc.ac.ir (دسترسی در ۱۴۰۱/۳/۱۰).

شریفی، عطیه، و محمدامین مهدوی. ۱۳۹۷. رویکردی با ناظر در استخراج واژگان کلیدی اسناد فارسی با استفاده از زنجیره‌های لغوی. *پردازش علائم و داده‌ها* ۱۵ (۴): ۹۵-۱۱۹.

صحرایی، رضامراد، امیرحسین مجیری فروشانی، و مروارید طالبی. ۱۳۹۸. واژه‌های پایهٔ زبان فارسی مبتنی بر متون مطبوعاتی. *زبان‌پژوهی* ۱۱ (۳۳). ۳۵۳-۳۷۸.

صحرایی، رضامراد، مروارید طالبی، و امیرحسین مجیری فروشانی. ۱۳۹۶ مقایسه واژه‌های پایهٔ زبان فارسی در شش پژوهش. پژوهشنامهٔ آموزش زبان فارسی به غیرفارسی‌زبانان ۶ (۱): ۱۱۵-۱۳۴.

صدیقی‌فر، زهره، جلال رحیمیان، و علیرضا خرمایی. ۱۴۰۰. شناسایی پیکره‌بنیاد الگوهای انسجام‌بخشی در گفتمان علمی زبان فارسی: رویکردی کاربردی در آموزش زبان فارسی برای اهداف دانشگاهی. پژوهشنامهٔ آموزش زبان فارسی به غیرفارسی‌زبانان ۱۰ (۲): ۱۸۳-۲۱۲ .

عاصی، مصطفی. ۱۳۹۸. *فرهنگ زبان‌آموز پیشرفته فارسی*. تهران: سمت.

عامری، حیات و حسن ذوالفقاری .۱۳۹۱. واژگان پایه و واژگان‌نگاری کودک در زبان فارسی. *مطالعات برنامه درسی ایران* ۲۷: ۱۵۹-۱۷۴.

علایی ابوذر، الهام، نصرالله پاک‌نیت، علی‌اصغر حجت‌پناه، مجتبی زالی، و محمدهای آقالویی آغمیونی. ۱۴۰۰. معرفی یک پیکره متنی تخصصی: پیکره پژوهشنامه. *نشریه پژوهش‌های زبان‌شناسی تطبیقی* ۱۱ (۲۲): ۲۷۱-۲۸۹.

قیومی، مسعود، و مریم موسویان. ۱۴۰۱. کاربرد یادگیری ماشینی مبتنی بر شبکه عصبی برای دسته‌بندی مستندات علمی. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۷ (۴): ۱۲۱۷-۱۲۴۴.

کامیابی گل، عطیه، الهام اخلاقی باقوجری، احسان عسگریان، و هانیه حبیبی. ۱۳۹۷. استخراج اطلاعات از پیکرهٔ زبانی: معرفی پیکرهٔ مقاله‌های علمی‌پژوهشی دانشگاه فردوسی. *نشریهٔ کتابداری و اطلاع‌رسانی* ۲۱ (۲): ۳-۲۵.

مجلی‌زاده، امین. ۱۳۹۴. کتاب‌شناسی پژوهش‌های بسامدی در زبان و ادب فارسی. *فرهنگ‌نویسی* ۹: ۹۵-۱۰۴.

محرابی، الهه، آزاده محبی، و عباس احمدی. ۱۴۰۰. بهبود الگوریتم RAKE برای استخراج کلیدواژه از متون علمی فارسی؛ مطالعهٔ موردی: پایان‌نامه‌ها و رساله‌های فارسی. پژوهشنامه پردازش و مدیریت اطلاعات ۳۷ (۱): ۱۹۷-۲۲۸.

نویدی، امین، حیات عامری، و زهرا ابوالحسنی چیمه. ۱۴۰۰. مقایسهٔ واژه‌های آموزشی کتاب‌های آموزش زبان فارسی با واژه‌های پربسامد. پژوهشنامهٔ آموزش زبان فارسی به غیرفارسی‌زبانان ۱۰ (۲): ۲۱۳-۲۳۵.

References

Bauer, Laurie, and Paul Nation. 1993. Word families. *International journal of Lexicography* 6 (4): 253-279.

Bijankhan, M., J. Sheykhzadegan, M. Bahrani, and M. Ghayoomi. 2011. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation* 45 (2): 143-164.

Brezina, Vaclav, and Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics* 36 (1): 1-22.

Browne, C., B. Culligan, and J. Phillips. 2013. *New general sevice list*. Retrieved from http://www.newgeneralservicelist.org (accessed Apr. 9, 2022)

Coxhead, Averil, and David Hirsch. 2007. A pilot science-specific word list. Revue française de linguistique appliquée 12 (2): 65-78.

Coxhead, Averil and Paul Nation. 2001. The Specialised Vocabulary of English for Academic Purposes. In Flowerdewand Peacock. *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

Coxhead, Averil. 2000. A new academic word list. *TESOL quarterly* 34 (2): 213-238.

Coxhead, Averil. 2011. The academic word list 10 years on: Research and teaching implications. *Tesol Quarterly* 45 (2): 355-362.

Dang, T. N. Y., A. Coxhead, and S. Webb. 2017. The academic spoken word list. *Language Learning* 67 (4): 959-997.

Elahimanesh, M. H., B. Minaei-Bidgoli, M. J. Gholami, and H. Juzi. 2012. An Introduction to Noor Corpus and its Language Model. *In Proceedings of the 1st international conference on Persian language processing*. Semnan, Iran.

Gardner, Dee, and Davies, Mark. 2013. A new academic vocabulary list. *Applied Linguistics* 35 (3): 305-327.

Gilmore, Alexander, and Neil Millar. 2018. The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes* 51: 1-17.

Hsu, Wenhua. 2013. Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research* 17 (4): 454-484.

Hsu, Wenhua. 2014. Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes* 33: 54-65.

Hyland, Ken, and Polly Tse. 2007. Is there an "academic vocabulary"? *TESOL quarterly* 41 (2): 235-253.

Juilland, Alphonse, & Eugenio Chang-Rodríguez. 1964. *Frequency dictionary of Spanish words*. The Hague: Mouton.

Konstantakis, Nikolaos. 2007. Creating a business word list for teaching business English. *ELIA: Estudios de Lingüística Inglesa Aplicada* 7: 79-102.

Leech, Geoffrey. 2011. Frequency, corpora and language learning. *A Taste for Corpora: In Honour of Sylviane Granger* 7: 32.

Lei, Lei, and Dilin Liu. 2016. A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes* 22: 42-53.

Miller, Corey, and Karineh Aghajanian-Stewart. 2017. *A Frequency dictionary of Persian: Core vocabulary for learners*. New York: Routledge.

Nagy, William, and Dianna Townsend. 2012. Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly* 47 (1): 91-108.

Nation, Paul. 2013. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Paquot, Magali. 2007. Towards a productively-oriented academic word list. In J. Walinski, K. Kredens, and S. Gozdz-Roszkowski (Eds.), *Corpora and ICT in language studies. PALC 2005. Lodz studies in LANGUAGE 13* (pp. 127–140). Frankfurt am main: Peter Lang.

Praninskas, Jean. 1972. *American university word list*. Harlow: Longman.

Rezaei Sharifabadi, Morteza, and Ahmad Eftekhari. 2016. Mahak Samim: A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems. *In FIRE (Working Notes)*190-192 ..

Rezvani, Reza, Abbas Gholtash, and Gerannaz Zamani. 2016. The first corpus-based Persian academic word list: Development and pedagogical implications. *Journal of Teaching Persian to Speakers of Other Languages* 5 (1): 43-64.

Wang, J., S. L. Liang, and G. C. Ge. 2008. Establishment of a medical academic word list. *English for Specific Purposes* 27 (4): 442-458.

Ward, Jeremy. 2009. A basic engineering English word list for less proficient foundation engineering undergraduates. *English for specific purposes* 28 (3: 170-182.

West, Michael. 1953. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans, Green.

Xue, Guoyi., and Paul Nation. 1984. A university word list. *Language Learning and Communication* 3 (2): 215-229.

مرتضی رضائی شریف‌آبادی

دارای مدرک کارشناسی ارشد زبان‌شناسی رایانشی از دانشگاه صنعتی شریف است. ایشان هم‌اکنون دانشجوی دکتری در رشتهٔ زبان‌شناسی همگانی در دانشگاه شیراز است.

از جمله علایق پژوهشی وی می‌توان به زبان‌شناسی رایانه‌ای و زبان‌شناسی پیکره‌ای اشاره نمود.

امیرسعید مولودی

دارای مدرک دکتری در رشتهٔ زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون استادیار بخش زبان‌های خارجی و زبان‌شناسی دانشگاه شیراز است.

از جمله علایق پژوهشی وی می‌توان به زبان‌شناسی رایانه‌ای، زبان‌شناسی پیکره‌ای و مطالعات استعاری در معنی‌شناسی شناختی اشاره نمود.

علیرضا احمدی

دارای مدرک دکتری در رشتهٔ آموزش زبان انگلیسی از دانشگاه اصفهان است. ایشان هم‌اکنون استاد بخش زبان‌های خارجی و زبان‌شناسی دانشگاه شیراز است.

از جمله علایق پژوهشی وی می‌توان به ارزشیابی و آزمون‌سازی اشاره نمود.

علیرضا خرمایی

دارای مدرک دکتری در رشتهٔ زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون دانشیار بخش زبان‌های خارجی و زبان‌شناسی دانشگاه شیراز است.

تحلیل گفتمان و زبان‌شناسی شناختی از جمله علایق پژوهشی وی است.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی