پژوهش‌نامه‌ی آموزش زبان فارسی به غیرفارسی‌زبانان
(علمی- پژوهشی)
سال ششم، شماره‌ی اول (پیاپی ۱۳)، بهار و تابستان ۱۳۹۶

مقایسه‌ی واژه‌های پایه‌ی زبان فارسی در شش پژوهش

رضامراد صحرایی
دانشیار زبان‌شناسی- دانشگاه علامه طباطبائی
مروارید طالبی
دانش آموخته‌ی کارشناسی ارشد آموزش زبان فارسی به غیرفارسی‌زبانان
امیرحسین مجیری فروشانی
دانش آموخته‌ی کارشناسی ارشد آموزش زبان فارسی به غیرفارسی‌زبانان

چکیده

یکی از مهم‌ترین مؤلفه‌ها در آموزش زبان خارجی/ دوم، واژه‌های زبان است. بسیاری از کارشناسان آموزش زبان معتقدند که اهمیت این مؤلفه‌ی زبانی به-حدی است که یادگیری زبان با یادگیری واژه مترادف شده است؛ به همین جهت چگونگی گزینش و سطح‌بندی محتوای واژگانی در منابع درسی از مسائل مهم در حوزه‌ی زبان‌دوم‌آموزی به شمار می‌رود. استفاده از صورت نشان‌دار واژه‌های زبان و گنجاندن سلیقه‌ای واژه‌ها در منابع آموزش زبان فارسی، یکی از بزرگ‌ترین چالش‌ها در سر راه یادگیری زبان فارسی است. در پژوهش پیش‌رو سعی ما بر این بوده است که با استفاده از متون مطبوعاتی و استخراج و بسامدگیری واژه‌های موجود در این متون، به فهرستی از واژه‌های پرکاربرد زبان فارسی دست پیدا کنیم. بدین منظور پیکره‌ای بیش از یک‌میلیون و دویست هزار واژه در ۷ زمینه‌ی موضوعی فرهنگی، اجتماعی، سیاسی، ورزشی، علمی، اقتصادی و ادبیات داستانی در طول ۱۰۰ روز کاری از روزنامه‌های کثیرالانتشار استخراج شد و در پایگاه داده‌هایی که جهت انجام این پژوهش تولید شده‌است ثبت شد. در مرحله‌ی بعد واژه‌ها برچسب‌گذاری و بسامدگیری شدند که پیرو آن به جدول «جمع بسامدها» نیز دست یافتیم. پس از دستیابی به واژه‌های پربسامد استخراج شده از پژوهش حاضر، به مقایسه‌ی این واژه‌ها با نتایج سایر پژوهش‌ها، از جمله حمید حسنی (۱۳۸۴)، بی‌جن‌خان (۱۳۹۰)، محک‌وب (۱۳۸۷)، همشهری (۲۰۰۹) و نعمت‌زاده و همکاران (۱۳۹۰) و اعتبارسنجی یافته‌های پژوهش پرداخته شد. نتیجه‌ی این مقایسه حاکی از اختلاف حدود ۳۰ درصدی با پژوهش‌های مشابه است. این اختلاف طبیعی به نظر می‌رسد؛ زیرا منبع داده‌های این پژوهش‌ها یکسان نیست.

کلیدواژه‌ها: زبان فارسی، واژه‌های پربسامد، واژه‌های پایه، متون مطبوعاتی، پیکره زبانی

تاریخ دریافت مقاله:۱۳۹۵/۱۰/۰۲ تاریخ پذیرش نهایی مقاله: ۱۳۹۶/۴/۲۰
رایانشانی نویسنده مسئول (رضامراد صحرایی): sahraeereza@gmail.com

۱. مقدمه

فهم و درک واژه‌های زبان از مهم‌ترین گام‌های یادگیری زبان است. یادگیری واژه‌های زبان دارای چنان اهمیتی است که ویلکینز (۱۹۷۶: ۱۱۱) در این باره چنین نظر می‌دهد: «بدون دانستن دستور زبان می‌توان اطلاعات اندکی را انتقال داد، اما بدون آشنایی با واژه‌ها، هیچ‌گونه اطلاعاتی را نمی‌توان منتقل کرد».

به اعتقاد مک‌کارتی (۱۹۹۰) هرقدر زبان‌آموز در زمینه‌ی دستور و آواهای زبان دوم/خارجی تبحر داشته باشد، بدون به‌کارگیری واژه، هرگز نمی‌تواند ارتباط معناداری با گویشوران زبان مورد نظر برقرار سازد. ورمیر (۱۹۹۲) نیز معتقد است برای آن‌که گویشور سخن دیگران را درک کند و سخنش نیز برای آنان قابل فهم باشد باید واژه‌ها را بداند. درواقع، مهم‌ترین رکن یادگیری یک زبان، یادگیری واژه‌های زبان بوده و آشنایی زبان‌آموز با دستور زبان به تنهایی نمی‌تواند او را در استفاده از زبان یاری نماید.

گاس و سلینکر نیز می‌گویند: «واژگان مهم‌ترین بخش برای زبان‌آموز است» (لوفر، ۱۹۹۷: ۱۴۰). همچنین تحقیقات میرا (۱۹۸۰) و نیشن (۱۹۹۰) نشان می‌دهد بسیاری از مشکلات زبان‌آموزان در تولید و دریافت زبان، ناشی از دانش واژه‌های اندک آنان است.

میلتون (۲۰۰۹) تعداد واژه‌هایی را که یک زبان‌آموز قادر به ادراک آن است، اندوخته‌ی واژگانی[1] می‌نامد. مطابق نظر نیشن (۲۰۰۶) درصورتی‌که اندوخته‌ی واژگانی حاوی ۸۰۰۰ تا ۹۰۰۰ واژه باشد؛ فرد قادر به درک متون ساده‌نشده‌ی زبان مقصد است. یادگیری این تعداد واژه برای یک بومی‌زبان بسیار ساده است و به‌راحتی و تنها در اثر مجاورت و برخورد زبانی، ذخیره‌ی واژگانی فرد در دوران کودکی مملو از این تعداد واژه خواهد شد، اما ازآنجاکه یک زبان‌آموز از این شرایط برخوردار نیست، ضرورت می‌یابد در یادگیری واژه‌های زبان مقصد مطابق روش و در نظر گرفتن اولویت‌های خاصی اقدام نماید؛ به این معنا که یادگیری واژه‌های کاربردی و مفیدتر باید در اولویت آموزشی زبان‌آموز قرار گیرد. با توجه به آن‌چه در بالا ذکر شد، اهمیت یادگیری واژه‌های زبان به‌خوبی مشخص شده است، اما پرسشی که مطرح می‌شود این است که «چگونه می‌توان دانش واژگانی زبان‌آموزان را ارزیابی نمود؟» و یا بهتر است بگوییم «واژه‌های یک زبان بر چه اساس و به چه ترتیب و اولویتی می‌بایست آموزش داده شوند؟».

جواب این پرسش عمدتاً با دستیابی به فهرست واژه‌های پایه‌ی زبان به‌دست می‌آید؛ به‌عنوان مثال نیشن (۲۰۰۱) آزمونی طراحی کرده است که نتایج آن می‌تواند نمایان‌گر سطح واژگانی زبان‌آموزان باشد، به این صورت که موارد آزمون براساس بسامد، به ۵ دسته‌ی زیر تقسیم می‌شوند:

سطح ۱: واژگانی که جزء ۱۰۰۰ تا ۲۰۰۰ واژه‌ی پربسامد زبان هستند.

سطح ۲: واژگانی که جزء ۲۰۰۰ تا ۳۰۰۰ واژه‌ی پربسامد زبان هستند.

سطح ۳: واژگانی که جزء ۳۰۰۰ تا ۵۰۰۰ واژه‌ی پربسامد زبان هستند.

---

[1]. vocabulary size

سطح ۴: واژگانی که جزء فهرست واژه‌های دانشگاهی هستند.

سطح ۵: واژگانی که جزء ۱۰۰۰۰ واژه‌ی پربسامد زبان هستند.

بنابراین اگر زبان‌آموزی در این آزمون موفق به رسیدن به سطح زبانی سه شود، به این معنا است که بین ۳۰۰۰ تا ۵۰۰۰ واژه از زبان مقصد را می‌داند.

در زمینه‌ی سطح‌بندی واژه‌های زبان، پژوهش‌های بسیاری در زبان‌های مختلف صورت گرفته است. نتیجه‌ی بسیاری از این پژوهش‌ها، حاکی از اهمیت استخراج واژه‌های پایه‌ی زبان است (نیشن، ۲۰۰۷). درواقع با استفاده از فهرست واژه‌های زبان است که می‌توان به تولید محتوا جهت آموزش و یا ارزیابی زبان‌آموزان اقدام کرد؛ اما با وجود پژوهش‌هایی که پژوهشگران ایرانی در جهت استخراج واژه‌های پایه و پربسامد زبان فارسی انجام داده‌اند، تاکنون تلاش‌های اندکی به‌منظور سطح‌بندی محتوای آموزشی براساس بسامد واژه‌های زبان صورت گرفته است و این مسئله از خلأهای اساسی در حوزه‌ی آموزش زبان فارسی به حساب می‌آید؛ به همین سبب در پژوهش پیش‌رو سعی شده است که واژه‌های پایه‌ی زبان فارسی از متون روزنامه‌ای با موضوعات اجتماعی، سیاسی، ورزشی، علمی، اقتصادی، ادبیات داستانی و فرهنگی، استخراج شود و سپس با مقایسه‌ی نتایج پژوهش حاضر با دیگر پژوهش‌های انجام شده در همین زمینه، اعتبار نتایج به‌دست‌آمده سنجیده شود.

فهرست واژه‌های پربسامد یا فرهنگ بسامدی مجموعه‌ای از واژه‌ها است که در مجموعه‌ای از داده‌های زبانی (پیکره‌ی زبانی) فراوانی بیشتری داشته‌اند؛ فرض بر این است که «هر چه فراوانی کلمه‌ای در زبان بیشتر باشد، آن کلمه از اهمیت بیشتری برخوردار است و درنتیجه، کلمات پربسامد در فعالیت‌هایی چون آموزش زبان به خارجیان باید در اولویت اول یادگیری قرارگیرند» (بی‌جن‌خان، ۱۳۹۰). روش‌های متفاوتی برای تعریف واژه‌های پایه وجود دارد. از میان آن‌ها می‌توان به سه مورد زیر اشاره کرد:

۱) واژه‌های پایه شامل واژه‌های پربسامد یک زبان است.

۲) واژه‌های پایه متشکل از واژه‌هایی است که میان همه‌ی گویشوران زبان مشترک است.

۳) واژه‌های پایه‌ی معنایی شامل واژه‌هایی است که برای توصیف سایر واژه‌های یک زبان کافی است.

دو روش اول را می‌توان به عنوان روش‌های آماری و روش سوم را به عنوان رویکردی معنایی تبیین کرد (بارنت و همکاران، ۱۹۸۶). واژه‌های پایه عموماً بسیار ساده و اساسی به نظر می‌رسند. معلمان زبان در زمره‌ی اولین کسانی هستند که برای تأمین اهداف یادگیری به تعریف واژه‌های پایه می‌پردازند. این واژه‌های پایه فاقد درون‌مایه‌ی عاطفی، ورزشی، ادبی و عاری از مظاهر فرهنگی هستند (کارتر، ۲۰۰۲: ۳۴). طبق نظریه‌ی به‌نمون[1]راش (۱۹۷۳)، کودک در ابتدا واژه‌های اصلی و پایه‌ی زبان را فرامی‌گیرد؛ زیرا این واژه‌ها نمونه‌های کامل و منعکس‌کننده‌ی دنیای اطراف ما هستند. پس از این مرحله نوبت به یادگیری واژه‌های کلی‌تر و

[1]. prototype theory

خاص‌تر می‌رسد. کوک (۱۹۸۲) ثابت می‌کند که روند یادگیری واژه‌های زبان دوم/ خارجی نیز به همین صورت است. درواقع، ذهن بشر به‌طورکلی یادگیری را از سطح ملموس و عینی شروع می‌کند نه از سطح انتزاعی یا ویژه (کوک ۱۹۹۱: ۳۹-۴۰).

جا دارد در این‌جا به دو روش رایج در شمارش واژه اشاره کنیم: شمارش تک‌به‌تک[1] و نوع[2] واژه. در شمارش تک‌به‌تک هر بار حضور واحدهای زبانی منفرد در متن، که معمولاً واژه هستند (با توجه به نظام رمزگذاری مورد مطالعه) شمرده می‌شود؛ اما در شمارش نوع هر واژه‌ی منحصربه‌فرد صرف نظر از تعداد وقوعش در متن تنها یک بار شمرده می‌شود. در فرایند جمع‌آوری پیکره‌های زبانی و بسامدگیری واژه‌های متون مختلف، مفهوم دیگری با عنوان لِما[3] اهمیت پیدا می‌کند. منظور از لما مجموعه صورت‌های تصریفی‌ای است که دارای ستاک[4] مشترک و وندهای تصریفی متفاوت هستند و به یک خانواده‌ی واژگانی[5] تعلق دارند؛ به‌عنوان مثال واژه‌های می‌آموزد، بیاموز، نیاموزی و غیره تنها دارای یک لما هستند. در اکثر پژوهش‌های پیکره‌بنیاد مانند پژوهش پیش‌رو، از شمارش لما به‌جای شمارش نوع استفاده شده است (بیکر و همکاران، ۲۰۰۶).

برخی پژوهشگران بر این باورند که بسامدگیری و شمارش واژه‌ها به‌تنهایی نمی‌تواند معیار مناسبی جهت تعیین واژه‌های پایه‌ی زبان باشند و تا به حال مؤلفه‌های مختلفی برای تشخیص و انتخاب واژه‌های پایه معرفی شده‌اند، از قبیل قابلیت جایگزینی نحوی، وجود واژه‌های متضاد، موارد همنشینی متعدد، هسته‌ای بودن، شامل بودن، عاری بودن از مظاهر فرهنگی، عدم تعلق به سبکی خاص، سرعت در تداعی و خنثایی زمینه‌ی گفتمانی. همچنین معیارهای افزوده‌ای جهت استفاده‌های آموزشی تعیین شده‌اند که دشواری واژه، تعداد واژه و هدف زبان‌آموز از این دست‌اند. با این حال طبق نظر ویلکینز (۱۹۷۲:۱۱۸) یکی از مهم‌ترین معیارها بسامد است و معمولاً مفیدترین واژه‌ها، پرکاربردترین آن‌ها هستند. دیکسون (۱۹۷۱: ۴۴۱) نیز بسامد وقوع را به‌عنوان معیاری از پایه بودن واژه ذکر می‌کند و تورنبری (۲۰۰۴: ۱۳۸) معتقد است که واژه‌های پربسامد نیز در هر زبانی حاوی معنای پربسامد زبانی است.

البته همیشه لزوماً واژه‌های پربسامد مفیدتر نیستند، اما به جرأت می‌توان گفت که متعارف‌ترین روش برای تخمین واژه‌های پایه، روش بسامدبنیاد است و هم‌اکنون برخی از معروف‌ترین فرهنگ‌های لغت در جهان مانند آکسفورد و لانگمن در انتخاب و چینش واژه‌های خود بر مبنای پیکره‌های زبانی و بسامد واژه‌ها عمل می‌کنند. در ادامه به بررسی پژوهش‌های انجام شده در زمینه‌ی واژه‌های پربسامد زبان، در ایران و خارج از ایران خواهیم پرداخت.

---

1. token
2. type
3. lemma
4. stem
5. word class

مطالعه در زمینه‌ی فهرست واژه‌های پربسامد، از سال ۱۸۹۷ توسط کدینگ و برای زبان آلمانی آغاز گردید. در این مطالعه با استفاده از پیکره‌ای شامل ۱۱ میلیون واژه، کدینگ فرهنگ بسامدی زبان آلمانی را تدوین نمود. وی جزء اولین کسانی است که به گردآوری پیکره‌های زبانی و شمارش واژگانی پرداخته است. پس از آن مطالعات بسیاری درمورد پیکره‌های زبانی و فهرست‌های واژگانی در زبان‌های گوناگون صورت گرفت.

از جمله این موارد می‌توان به "کتاب واژه‌ی معلم" اثر ثورندایک اشاره کرد. این کتاب درسال ۱۹۴۴ تألیف شد و ۳۰ هزار واژه‌ی پایه‌ی زبان انگلیسی، برگرفته از پیکره‌ای حاوی ۱۸میلیون واژه را در خود جای داد.

در سال ۱۹۲۳ آگدن و ریچاردز "فهرست پایه‌ی انگلیسی" را تدوین نمودند که حاوی ۸۵۰ واژه‌ی پربسامد در زبان انگلیسی است. دلچ (۱۹۳۶) فهرستی ۲۲۰ واژه‌ای از واژه‌های پایه‌ی زبان انگلیسی در کتاب خود ارائه داد. پیکره‌ی مورد استفاده‌ی دلچ حاوی حدوداً ۴۵۰۰ واژه و دربرگیرنده‌ی متون گروه سنی کودک نیز بود. وست در سال ۱۹۵۳، فهرستی از ۲۰۰۰ واژه‌ی پایه در زبان انگلیسی ارائه داد. این فهرست از پیکره‌ای نوشتاری شامل ۵ میلیون واژه استخراج شده بود. لازم به ذکر است که عوامل استخراج و شمارش واژه‌ها در کلیه‌ی پژوهش‌های فوق، انسانی بوده‌اند.

از دیگر موارد در این زمینه می‌توان به اثر کارول و همکاران (۱۹۷۱) با عنوان "کتاب بسامد واژه[1]" اشاره کرد. در این اثر از پیکره‌ای که شامل متون آموزشی مدارس آمریکا و حاوی ۵ میلیون واژه بود استفاده شد. همچنین در این کتاب متناسب با هر موضوع و سطح زبانی، یک فهرست واژه‌ای ارائه شد. کاکس‌هد در سال ۲۰۰۰ در پژوهشی با عنوان «فهرست واژه‌ای علمی جدید»، اقدام به استخراج واژه‌های پایه در چهار زمینه‌ی هنر، بازرگانی، حقوق و علوم نمود. در این پژوهش، کاکس‌هد ۵۷۰ خانواده‌ی واژگانی را در بین هر چهار حوزه مشترک یافت.

در ۲۰۰۱ ورلیند و سلوا فهرستی از واژه‌های پربسامد در زبان فرانسه ارائه دادند. پیکره‌ی مورد استفاده در این پژوهش دو روزنامه‌ی فرانسوی و بلژیکی و تعداد واژه‌های موجود در پیکره در کل ۵۰ میلیون واژه بوده است. ۱۰۰ و ۱۰۰۰ واژه‌ی پایه‌ی زبان انگلیسی به کوشش فرای (۲۰۰۰) از پیکره‌ای حاوی پنج میلیون واژه استخراج شده است.

«فرهنگ بسامدی آلمانی» اثر جونز و شیرنر (۲۰۰۶) از دیگر فرهنگ‌های بسامدی است که به واژه‌های پربسامد در زبان آلمانی اختصاص دارد. این فرهنگ دارای ۴۰۳۷ مدخل، و پیکره‌ی مورد استفاده در این پژوهش حاوی ۴میلیون واژه بوده است. دیویس و گاردنر (۲۰۱۰) از پیکره‌ای حاوی بیش از ۴۰۰ میلیون واژه به استخراج ۱۰۰۰-۵۰۰۰ واژه‌ی پایه در زبان انگلیسی پرداختند. علاوه بر این می‌توان به فهرست بسامدی ۱۰۰ واژه‌ی پایه به کوشش لغت‌نامه‌ی انگلیسی آکسفورد و ۳۰۰۰ واژه‌ی پایه توسط لغت‌نامه‌ی لانگ‌من و بسیاری موارد دیگر نیز اشاره نمود.

---

[1]. Word Frequency Book

اولین پژوهش انجام شده در زمینه‌ی واژه‌های پربسامد در زبان فارسی به کار فریدون بدره‌ای (۱۳۵۰) بازمی‌گردد. پس از او دیگران کار در این زمینه را ادامه دادند، همانند براهنی (۱۳۵۴)، ایمن (۱۳۵۷)، صفارپور (۱۳۷۰)، عاصی (۱۳۷۳)، تحریریان (۱۳۷۳) و غروی قوچانی (۱۳۸۵). اما ازآنجاکه برخی از پژوهش‌های فوق صرفاً در قالب جمع‌آوری پیکره‌ی زبانی و بعضاً نامرتبط با آموزش زبان فارسی بوده است و مهم‌تر از آن دسترسی آزاد به داده‌های این پیکره‌ها وجود نداشته است، به‌اجبار از نتایج این پیکره‌ها در پژوهش پیش‌رو استفاده نشده است. در ادامه با اهم پژوهش‌های انجام‌شده در این زمینه آشنا می‌شویم:

پیکره‌ی حمید حسنی (۱۳۸۴) که شامل ۱۰۰۲۳۹۴ واژه است از ۸۰ کلان‌متن (متن مادر) با بیش از ۵۰۰ خردمتن فارسی امروز استخراج شده‌اند. کلان‌متن‌ها شامل کتاب، مجله و روزنامه بوده‌اند و خردمتن‌ها شامل فصل‌ها و مقاله‌های بزرگ و کوچک این کلان‌متن‌ها هستند. در فهرست بسامدی این پیکره، ۸۴۳۸ واژه که بسامدشان ده بار یا بیشتر بوده آورده شده است.

پیکره‌ی محمود بی‌جن‌خان (۱۳۹۰) شامل ۱۰۶۱۸۱۸۷ قطعه‌ی نوشتاریِ پراکنده در ۲۹۹۰ پرونده‌ی متنی است که پس از تقطیع و برچسب‌گذاری به ۹۸۸۰۴۰۰ کلمه‌ی نوشتاری رسیده است. متون این پیکره از منابع مختلف (کتاب، روزنامه، مجله، یادداشت‌های روزانه و غیره) انتخاب شده‌اند و موضوعات متنوعی را پوشش می‌دهند، اما حجم واژه‌های هر پرونده از یک حد آستانه بیشتر نیست تا جلوی ورود واژه‌های کم‌بسامد خاص یک متن به پیکره گرفته شود.

احسان درودی و همکاران (۱۳۸۸) در پروژه‌ی محک وب دات آی آر، از پیکره‌های اینترنتی برای تهیه‌ی پیکره استفاده کرده‌اند. این کار با تصفیه‌ی صفحات و وبگاه‌ها و دستیابی به ۱۹۹۶۱ صفحه‌ی وب انجام شده است.

مجموعه‌ی اسناد همشهری با خزش[1] وب‌سایت همشهری و چندین مرحله پیش‌پردازش و برچسب‌گذاری به‌دست آمده است. حجم نسخه‌ی یک این پیکره[2] ۷۰۰ مگابایت و نسخه‌ی دو ۱۴۰۰ مگابایت است. اسناد از ۱۶۰ هزار عدد در نسخه‌ی یک به ۳۱۸ هزار در نسخه‌ی دو رسیده‌اند و بازه‌ی زمانی جمع‌آوری اسناد از «۴ اردیبهشت ۱۳۷۵ تا ۲۲ بهمن ۱۳۸۱» (۷ سال) در نسخه‌ی یک به «۴ اردیبهشت ۱۳۷۵ تا ۲۳ اردیبهشت ۱۳۸۶» (۱۲ سال) افزایش یافته است.

طرح نعمت‌زاده و همکاران (۱۳۹۰) براساس «طرح شناسایی واژه‌های پایه‌ی دانش‌آموزان ایرانی در دوره‌ی ابتدایی» است که در سال ۱۳۸۰ در سازمان پژوهش و برنامه‌ریزی آموزشی آغاز گردید. حاصل کار بایگانی واژگانی و دادگان واژگانی، نرم‌افزار و گزارشی در ۵ جلد و ۱۰۰۰ صفحه بود و در مجموع ۴۹۷ واژه‌ی پایه به‌دست آمده است.

---

[1]. crawl

[2]. یونیکد در قالب CLEF

در این پژوهش شیوه‌ی به‌کارگرفته شده جهت دستیابی به واژه‌های پایه‌ی زبان فارسی، شیوه‌ی پیکره‌بنیاد است؛ به این معنی که برای انجام پژوهش از پیکره‌های معیار زبانی (همچون متون نوشتاری و یا مکالمات صورت گرفته بین فارسی‌زبانان) استفاده شده است. در تحقیق پیش‌رو داده‌های پیکره از متون روزنامه‌ای انتخاب شده‌اند. در این پژوهش روزانه ۲۱ متن با میانگین ۵۰۰ واژه از متون موجود در ۷ زمینه‌ی موضوعی اجتماعی، سیاسی، ورزشی، علمی، اقتصادی، فرهنگی و ادبیات داستانی از میان روزنامه‌های ایران، همشهری، جام‌جم، اطلاعات، آفرینش و تهران امروز، استخراج شده و در پایگاه داده‌ها ثبت شده است که نتیجه‌ی آن پیکره‌ای بالغ بر ۲۴۰۱ متن با ۱۲۰۳۵۹۸ واژه است.

پایگاه داده‌ها درواقع نرم‌افزاری جهت ثبت و شمارش واژه‌های استخراج شده از متون روزنامه‌ای براساس قوانین مشخصی است و هر پژوهشگر می‌بایست واژه‌های استخراج شده از متن‌ها را به ترتیبی که نرم‌افزار مشخص کرده است، برچسب‌دهی و درصورت نیاز اصلاح کند تا به‌صورت نهایی در پایگاه داده‌ها و فهرست واژه‌ها ثبت شود. نتایج به‌دست آمده در این مرحله نمایانگر فهرست پربسامدترین و پرکاربردترین واژه‌های موجود در بخش‌های مختلف متون روزنامه‌ای زبان فارسی است. نرم‌افزار استفاده شده در این پژوهش به‌صورت برخط طراحی شده و زبان برنامه‌نویسی نرم‌افزارPHP بوده است، گرچه از HTML، زبان کدنویسی CSS و کتابخانه‌ی BootStrap، Javascript و jQuery نیز استفاده شده است.

هر یک از واردکننده‌ها ابتدا ملزم به تفکیک واژه‌ها بوده است. به این صورت که در میان حروف یک واژه‌ی واحد، نباید فاصله‌ای قرار گیرد؛ به‌عنوان مثال «می‌رفته است» به شکل «میرفتهاست» ویرایش می‌شود. سپس واردکننده باید فعل‌های مرکب را بین دو علامت ## قرار دهد؛ مانند #انجامگرفتهاست#. این کار برای تشخیص فعل‌های مرکب توسط نرم‌افزار انجام می‌شود. پس از آن واردکننده باید اسم‌های خاص را بین دو علامت ** قرار دهد؛ مانند *حسنروحانی*. این‌کار برای تشخیص اسم‌های خاص توسط نرم‌افزار انجام می‌شود.

پس از واردکردن یک متن واردکننده می‌بایست واژه‌ها را برچسب‌گذاری کرده و در بعضی موارد حالت ریشه‌ای واژه را وارد کند تا نرم‌افزار فرایند بسامدگیری را انجام دهد. ۷ برچسب نامشخص، اسم، فعل، حرف، اسم خاص، فعل مرکب و صفت در نرم‌افزار تعریف شده که از این میان برچسب‌های اسم، صفت، فعل و حرف توسط واردکننده وارد شده و دو برچسب اسم خاص و فعل مرکب توسط نرم‌افزار وارد شده است. با این حال واردکننده‌ها می‌بایست حالت مصدری افعال را مشخص می‌کردند. ازآنجاکه جمع‌های مکسر و وندها پیش از جمع‌آوری داده‌ها تعریف شده‌اند، نرم‌افزار در این بخش به‌صورت خودکار عمل کرده و با اخذ تأیید نهایی از واردکننده‌ها، به‌دسته‌بندی واژه‌ها می‌پردازد. حذف وندهای تصریفی، ثبت افعال به‌صورت مصدری، حذف افعال کمکی، عدم جداسازی جزء غیرفعلی افعال مرکب از ویژگی‌های این پژوهش است. پس از اصلاح داده‌ها، فهرست بسامدی ۲۰۰۰ واژه‌ی پایه‌ی زبان فارسی استخراج شد و به مقایسه‌ی این نتایج با دیگر پژوهش‌ها پرداخته شد.

۲. تجزیه و تحلیل داده‌ها

هدف در این بخش از پژوهش، مقایسه‌ی بین واژه‌های پربسامد پروژه‌ی حاضر و دیگر پروژه‌های مشابه است و نتایج حاصل از این مقایسه می‌تواند به بهبود نتایج همه‌ی پروژه‌ها (پروژه‌ی حاضر و دیگر پروژه‌های مشابه) بیانجامد و همچنین معیاری برای اعتبارسنجی نتایج باشد.

برای مقایسه ابتدا لازم بود فهرست‌های چاپی، تایپ شده و برای فهرست‌های رایانه‌هایی که تنها با رمز عبور قابل دسترسی بودند، رمز عبور دریافت شود. همچنین می‌بایست داده‌ها به نحوی سامان بیابند که قابل مقایسه باشند؛ زیرا در بیشتر این پروژه‌ها ریشه و مصدر واژه ثبت نشده است؛ یعنی «می‌کنم»، «کرده‌بودم» و «کرده‌اند» همگی به‌صورت مجزا ثبت شده‌اند؛ درحالی‌که در پروژه‌ی حاضر، همه‌ی این واژه‌ها ذیل مصدر فعل «کردن» ثبت شده‌اند. حذف «» در پروژه‌ی حمید حسنی (۱۳۸۴)، تبدیل جمع‌های مکسر به حالت مفرد، حذف «ات» از آخر جمع‌ها و تصحیح واژه در صورت لزوم و حذف فاصله‌ها و نیم‌فاصله‌ها از جمله‌ی دیگر اقداماتی است که پیش از مقایسه‌ی نتایج انجام شد.

مقایسه‌ی کیفی ویژگی‌های هر پروژه در جدول ۱ خلاصه شده است (علامت × نشانه‌ی نبود ویژگی در پروژه‌ی مذکور است و علامت √ نشانه‌ی حضور آن ویژگی در پروژه است). همچنین، تعداد واژه‌های موجود در پیکره‌ی هر یک از پژوهش‌ها در جدول ۲ آمده است.

در مرحله‌ی بعد ۱۰۰۰ واژه‌ی پربسامد هر یک از پژوهش‌ها (از جمله پژوهش حاضر) با ۱۰۰۰ واژه‌ی پربسامد پژوهش‌های دیگر مقایسه شد و میانگین درصد مشابهت نتایج پژوهش‌ها به‌دست آمد. نتیجه‌ی این مقایسه را می‌توان به‌صورت خلاصه در جداول ۳ و ۴ مشاهده کرد.

جدول ۱. مقایسه‌ی ویژگی‌های پژوهش حاضر و دیگر پژوهش‌ها

	پروژه‌ی حاضر	حمید حسنی	بی‌جن خان (اینترنتی)	بی‌جن خان (چاپی)	محک وب	همشهری	نعمت‌زاده
مفرد کردن جمع‌های مکسر	√	×	×	√	×	×	√
جداکردن «ات» از واژه‌ها	√	×	×	√	×	×	√
جدا کردن پسوندهای تصریفی	√	×	×	√	×	×	√
تفکیک واژه‌ها با مشابهت نوشتاری	√	×	×	×	×	×	√
برچسب‌گذاری صرفی	√	×	×	√	×	×	√
مصدری کردن فعل‌ها	√	×	×	√	×	×	√
برچسب‌گذاری موضوعی	√	×	√	√	×	×	√
برچسب‌گذاری زمانی	√	×	√	×	√	√	×
مدخل کردن فعل‌های مرکب	√	×	×	×	×	×	√
تصحیح شکل نوشتاری واژه‌ها	√	√	×	×	×	×	√
رعایت نیم‌فاصله	×	√	√	√	×	×	√

جدول۲. مقایسه‌ی تعداد کل واژه‌ها در پژوهش حاضر و دیگر پژوهش‌ها

	پروژه‌ی حاضر	حمید حسنی	بی‌جن‌خان (اینترنتی)	بی‌جن‌خان (چاپی)	محک‌وب	همشهری	نعمت‌زاده
تعداد کل واژه‌ها	۱۲۱۱۴۰۲	۱۰۰۲۳۹۴	۲۲۹۹۰۴۲	۹۸۸۰۴۰۰	۶۷۷۰۹۷۴۲۲	۱۲۴۰۹۰۸۲۷	-

جدول ۳. مقایسه‌ی ۱۰۰۰ واژه در پژوهش حاضر با دیگر پژوهش‌ها

	پروژه‌ی حاضر	حمید حسنی	بی‌جن‌خان (اینترنتی)	بی‌جن‌خان (چاپی)	محک وب	همشهری	نعمت‌زاده
پروژه‌ی حاضر	۱۰۰۰	۵۲۷	۶۶۹	۶۷۹	۶۰۸	۷۰۷	۵۷ *۱۰۳
حمید حسنی	۵۲۷	۱۰۰۰	۵۶۱	۵۴۵	۵۲۷	۵۴۰	۱۳۶ ۱۹۵
بی‌جن‌خان (اینترنتی)	۶۶۹	۵۶۱	۱۰۰۰	۸۱۴	۶۷۴	۸۳۱	۵۷ ۱۰۰
بی‌جن‌خان (چاپی)	۶۷۹	۵۴۵	۸۱۴	۱۰۰۰	۶۵۳	۷۹۰	۵۶ ۱۰۶
محک وب	۶۰۸	۵۲۷	۶۷۴	۶۵۳	۱۰۰۰	۷۱۵	۵۵ ۱۰۸
همشهری	۷۰۷	۵۴۰	۸۳۱	۷۹۰	۷۱۵	۱۰۰۰	۷۶ ۹۶
نعمت‌زاده	۵۷ *۱۰۳	۱۳۶ **۱۹۵	۵۷ ۱۰۰	۵۶ ۱۰۶	۵۵ ۱۰۸	۷۶ ۹۶	۴۹۷

جدول۴. میانگین درصد مشابهت ۱۰۰۰ واژه در پروژه‌ی حاضر و دیگر پروژه‌ها* **

نام پروژه	میانگین درصد مشابهت واژه‌ها (بر حسب ۱۰۰۰ واژه) با دیگر پروژه‌ها
پروژه‌ی حاضر	۶۳,۸٪ [۵۲,۷+ ۶۶,۹+ ۶۷,۹+ ۶۰,۹+ ۷۰,۷]
حمید حسنی	۵۴٪ [۵۲,۷+ ۵۶,۱+ ۵۴,۵+ ۵۲,۷+ ۵۴,۰]
بی‌جن خان (اینترنتی)	۷۰,۹٪ [۶۶,۹+ ۵۶,۱+ ۸۱,۴+ ۶۷,۴+ ۸۳,۱]
بی‌جن خان (چاپی)	۶۹,۶٪ [۶۷,۹+ ۵۴,۵+ ۸۱,۴+ ۶۵,۳+ ۷۹,۰]
محک وب	۶۳,۵٪ [۶۰,۹+ ۵۲,۷+ ۶۷,۴+ ۶۵,۳+ ۷۱,۵]
همشهری	۷۱,۶٪ [۷۰,۷+ ۵۴,۰+ ۸۳,۱+ ۷۹,۰+ ۷۱,۵]

* چون پیکره‌ی موجود در پژوهش نعمت‌زاده در کل دارای ۴۹۷ واژه بوده است، ۵۷ واژه از مقایسه‌ی ۴۹۷ واژه‌ی پروژه‌ی حاضر با کار نعمت‌زاده حاصل شده است و ۱۰۳ واژه از مقایسه ۱۰۰۰ واژه‌ی اول پروژه‌ی حاضر با ۴۹۷ واژه کل کار نعمت‌زاده. در مقایسه‌ی بقیه کارها با کار نعمت‌زاده هم به همین ترتیب عمل شده است.

** چون پروژه‌های دیگر برچسب‌گذاری نشده بودند، مقایسه به‌طور کلی بین آن‌ها و پروژه‌ی نعمت‌زاده انجام‌شد. اما باید توجه داشت که نعمت‌زاده، حرف‌ها و اسامی خاص را در کار خود وارد نکرده‌است؛ بنابراین مقایسه بین این کار و پروژه‌های دیگر چندان دقیق نیست.

جدول۵. مقایسه‌ی درصد مشابهت ۱۰۰۰ واژه در این پروژه و دیگر پروژه‌ها***

	پروژه‌ی حاضر	حمید حسنی	بی‌جن خان (اینترنتی)	بی‌جن خان (چاپی)	محک وب	همشهری	نعمت‌زاده
پروژه‌ی حاضر	۱۰۰٪	۵۲,۷٪	۶۶,۹٪	۶۷,۹٪	۶۰,۹٪	۷۰,۷٪	۱۱,۴٪ ۲۰,۶٪
حمید حسنی	۵۲,۷٪	۱۰۰٪	۵۶,۱٪	۵۴,۵٪	۵۲,۷٪	۵۴,۰٪	۲۷,۳٪ ۳۹,۲٪
بی‌جن خان (اینترنتی)	۶۶,۹٪	۵۶,۱٪	۱۰۰٪	۸۱,۴٪	۶۷,۴٪	۸۳,۱٪	۱۱,۴٪ ۲۰,۱٪
بی‌جن خان (چاپی)	۶۷,۹٪	۵۴,۵٪	۸۱,۴٪	۱۰۰٪	۶۵,۳٪	۷۹,۰٪	۱۱,۲٪ ۲۱,۳٪
محک وب	۶۰,۹٪	۵۲,۷٪	۶۷,۴٪	۶۵,۳٪	۱۰۰٪	۷۱,۵٪	۱۱,۰٪ ۲۱,۷٪
همشهری	۷۰,۷٪	۵۴,۰٪	۸۳,۱٪	۷۹,۰٪	۷۱,۵٪	۱۰۰٪	۱۵,۲٪ ۱۹,۳٪
نعمت‌زاده	۱۱,۴٪ ۲۰,۶٪	۲۷,۳٪ ۳۹,۲٪	۱۱,۴٪ ۲۰,۱٪	۱۱,۲٪ ۲۱,۳٪	۱۱,۰٪ ۲۱,۷٪	۱۵,۲٪ ۱۹,۳٪	۱۰۰٪

با توجه به داده‌های جدول‌های فوق، درصد اختلاف بین پژوهش حاضر و دیگر پژوهش‌ها بین ۷,۳۳٪ تا ۲۹,۳٪ تخمین زده شده است که با توجه به وجود همین میزان اختلاف بین نتایج سایر پژوهش‌ها با یکدیگر، میزان اختلاف قابل قبول ارزیابی شده است.

در ادامه ۱۰۰ واژه‌ی پربسامد از هر پروژه با یکدیگر مقایسه شدند. ازآنجاکه در پژوهش نعمت‌زاده (۱۳۹۰)، ترتیب واژه‌ها مشخص نشده‌است، امکان مقایسه‌ی ۱۰۰ واژه‌ی نخست مشترک بین دو پروژه وجود نداشته است. نتایج این مقایسه در جداول زیر قابل مشاهده است.

جدول ۶. میانگین درصد مشابهت ۱۰۰ واژه در پروژه‌ی حاضر و دیگر پروژه‌ها

نام پروژه	میانگین درصد مشابهت واژه‌ها (بر حسب ۱۰۰ واژه) با دیگر پروژه‌ها
پروژه‌ی حاضر	۵۹,۶٪
حمید حسنی	۴۷,۲٪
بی‌جن خان (اینترنتی)	۶۶,۶٪
بی‌جن خان (چاپی)	۶۱٪
محک‌وب	۵۱,۴٪
همشهری	۶۷,۴٪

*** در این محاسبه نتایج پژوهش نعمت‌زاده لحاظ نشده‌است.

جدول ۷. مقایسه‌ی ۱۰۰ واژه‌ی اول در پژوهش حاضر با دیگر پژوهش‌ها

	پروژه‌ی حاضر	حمید حسنی	بی‌جن خان (اینترنتی)	بی‌جن خان (چاپی)	محک‌وب	همشهری
پروژه‌ی حاضر	۱۰۰	۵۰	۶۹	۵۸	۵۰	۷۱
حمید حسنی	۵۰	۱۰۰	۴۸	۴۷	۴۲	۴۹
بی‌جن خان (اینترنتی)	۶۹	۴۸	۱۰۰	۷۶	۵۶	۸۴
بی‌جن خان (چاپی)	۵۸	۴۷	۷۶	۱۰۰	۵۰	۷۴
محک‌وب	۵۰	۴۲	۵۶	۵۰	۱۰۰	۵۹
همشهری	۷۱	۴۹	۸۴	۷۴	۵۹	۱۰۰

در آخر ده واژه‌ی پربسامد پژوهش حاضر و دیگر پژوهش‌ها مقایسه شده‌اند که در جداول زیر به‌صورت جداگانه به آن‌ها پرداخته می‌شود.

واژه‌های پربسامد به‌دست آمده از نتیجه‌ی این پژوهش و پژوهش حمید حسنی (۱۳۸۴) در جدول زیر مقایسه شده‌اند.

جدول ۸. مقایسه‌ی توزیع فراوانی واژه‌ها در پژوهش حاضر و پژوهش حمید حسنی

ردیف	واژه های این پروژه	بسامد واژه های این پروژه	واژه های حمید حسنی	بسامد واژه های حمید حسنی
۱	و	۶۱۰۸۱	و	۴۹۵۷۸
۲	در	۴۸۷۹۵	به	۳۲۵۹۳
۳	به	۳۸۳۱۲	را	۲۷۵۳۰
۴	از	۳۳۴۲۳	بودن	۲۵۵۵۵
۵	که	۲۹۴۳۰	از	۲۳۷۱۷
۶	این	۲۹۳۸۰	کردن	۲۲۹۵۹
۷	بودن	۲۶۰۴۶	که	۲۲۵۹۳
۸	را	۲۱۴۳۸	در	۲۱۶۷۱
۹	با	۱۹۹۰۹	شدن	۱۴۸۳۶
۱۰	آن	۱۰۴۵۴	این	۱۳۱۶۱

در جدول زیر ۱۰ واژه‌ی پربسامد به‌دست آمده از نتیجه‌ی این پژوهش و پژوهش بی‌جن‌خان (نسخه‌ی اینترنتی) مقایسه شده‌اند.

جدول ۹. مقایسه‌ی توزیع فراوانی واژه‌ها در پژوهش حاضر و پژوهش بی‌جن‌خان (نسخه‌ی اینترنتی)

ردیف	واژه های این پروژه	بسامد واژه های این پروژه	واژه های بیجن خان (اینترنتی)	بسامد واژه های بیجن خان (اینترنتی)
۱	و	۶۱۰۸۱	و	۱۲۰۵۵۸
۲	در	۴۸۷۹۵	در	۸۵۷۲۴
۳	به	۳۸۳۱۲	به	۶۸۹۷۰
۴	از	۳۳۴۲۳	بودن	۶۴۷۱۳
۵	که	۲۹۴۳۰	از	۶۰۶۹۴
۶	این	۲۹۳۸۰	که	۴۹۶۸۳
۷	بودن	۲۶۰۴۶	این	۴۲۷۵۶
۸	را	۲۱۴۳۸	را	۳۵۸۳۱
۹	با	۱۹۹۰۹	شدن	۳۳۴۱۷
۱۰	آن	۱۰۴۵۴	کردن	۳۲۴۴۴

در جدول زیر ۱۰ واژه‌ی پربسامد به‌دست آمده از نتیجه‌ی این پژوهش و پژوهش بی‌جن‌خان (نسخه‌ی چاپی) مقایسه شده‌اند.

جدول ۱۰. مقایسه‌ی توزیع فراوانی واژه‌ها در پژوهش حاضر و پژوهش بی‌جن‌خان (نسخه‌ی چاپی)

ردیف	واژه های این پروژه	بسامد واژه های این پروژه	واژه های بیجن خان (کتاب)	بسامد واژه های بیجن خان (کتاب)
۱	و	۶۱۰۸۱	و	۴۷۲۹۵۹
۲	در	۴۸۷۹۵	در	۳۱۶۹۷۹
۳	به	۳۸۳۱۲	به	۲۶۶۱۶۱
۴	از	۳۳۴۲۳	از	۲۲۱۸۰۹
۵	که	۲۹۴۳۰	که	۱۸۶۵۳۸
۶	این	۲۹۳۸۰	بودن	۱۸۰۶۷۵
۷	بودن	۲۶۰۴۶	این	۱۵۸۶۷۳
۸	را	۲۱۴۳۸	را	۱۴۳۵۸۹
۹	با	۱۹۹۰۹	کردن	۱۴۰۱۳۶
۱۰	آن	۱۰۴۵۴	شدن	۱۱۷۷۷۹

در جدول زیر ۱۰ واژه‌ی پربسامد به‌دست آمده از نتیجه‌ی این پژوهش و پژوهش محک‌وب مقایسه شده‌اند.

جدول ۱۱. مقایسه‌ی توزیع فراوانی واژه‌ها در پژوهش حاضر و پژوهش محک‌وب

ردیف	واژه های این پروژه	بسامد واژه های این پروژه	واژه های محک وب	بسامد واژه های محک وب
۱	و	۶۱۰۸۱	و	۳۰۵۷۲۴۸۷
۲	در	۴۸۷۹۵	در	۱۶۵۴۶۶۶۱
۳	به	۳۸۳۱۲	به	۱۴۱۵۹۱۳۹
۴	از	۳۳۴۲۳	بودن	۱۲۰۶۹۲۷۷
۵	که	۲۹۴۳۰	از	۱۱۷۷۰۲۲۵
۶	این	۲۹۳۸۰	شدن	۷۲۴۱۹۵۶
۷	بودن	۲۶۰۴۶	با	۷۲۰۷۸۱۰
۸	را	۲۱۴۳۸	این	۶۹۵۴۴۸۷
۹	با	۱۹۹۰۹	را	۶۷۵۳۴۱۵
۱۰	آن	۱۰۴۵۴	که	۵۴۸۲۱۷۱

در جدول زیر ۱۰ واژه‌ی پربسامد به‌دست آمده از نتیجه‌ی این پژوهش و پژوهش همشهری مقایسه شده‌اند.

جدول ۱۲. مقایسه‌ی توزیع فراوانی واژه‌ها در پژوهش حاضر و پژوهش همشهری

ردیف	واژه های این پروژه	بسامد واژه های این پروژه	واژه های همشهری	بسامد واژه های همشهری
۱	و	۶۱۰۸۱	و	۵۶۱۶۲۶۴
۲	در	۴۸۷۹۵	در	۴۲۴۵۴۶۹
۳	به	۳۸۳۱۲	به	۳۷۶۴۲۰۸
۴	از	۳۳۴۲۳	بودن	۳۳۶۴۷۴۰
۵	که	۲۹۴۳۰	از	۲۹۱۴۳۶۳
۶	این	۲۹۳۸۰	که	۲۴۱۶۵۹۷
۷	بودن	۲۶۰۴۶	این	۲۲۱۶۱۱۴
۸	را	۲۱۴۳۸	شدن	۱۸۹۴۱۰۶
۹	با	۱۹۹۰۹	کردن	۱۸۹۱۰۹۳
۱۰	آن	۱۰۴۵۴	را	۱۸۰۵۵۱۸

ازآنجاکه در پژوهش نعمت‌زاده (۱۳۹۰)، ترتیب واژه‌ها مشخص نیست، امکان مقایسه‌ی ۱۰ واژه‌ی ابتدایی مشترک بین دو پروژه وجود نداشت.

مقایسه‌ی نتایج فوق تنها براساس وجود یک واژه از یک فهرست در فهرست دیگر انجام شده است؛ اما جایگاه واژه در هرکدام از فهرست‌ها بررسی نشده است. برای انجام این بررسی به مقایسه‌ی فاصله‌ای داده‌ها پرداخته‌ایم. مقایسه‌ی فاصله‌ای به این دلیل انجام می‌شود که علاوه بر شناسایی تعداد واژه‌های مشترک بین دو فهرست، جایگاه هر واژه در دو فهرست و نسبت آن‌ها با یکدیگر را نیز به‌دست آوریم. به این ترتیب می‌توان اطلاعات دو فهرست (فهرست پژوهش حاضر و هر یک از دیگر پژوهش‌ها) را دقیق‌تر بررسی کرد.

از فرمول ۱ که توسط نویسندگان پژوهش طرح شده است، جهت دستیابی به درصد اختلاف میان نتایج پژوهش حاضر و دیگر پژوهش‌ها استفاده شده است:

$$M = \left(\frac{\sum_{i=1}^{n} |x_{1_i} - x_{2_i}| + \sum_{i=1}^{m_1} |1001 - y_{1_i}| + \sum_{i=1}^{m_2} |1001 - y_{2_i}|}{n + m_1 + m_2} \Big/ 1001 \right) \times 100$$

فرمول ۱. درصد اختلاف

در این فرمول M درصد میانگین فاصله‌ی کلی واژه‌های دو فهرست (هر چه مقدار آن بیشتر باشد، واژه‌های دو فهرست بیش‌تر با هم فاصله دارند و درنتیجه دو فهرست بیشتر با هم اختلاف دارند)، n تعداد واژه‌های مشترک بین دو فهرست، x_{1_i} جایگاه واژه‌ی مشترک در فهرست اول، x_{2_i} جایگاه واژه‌ی مشترک در فهرست دوم، m_1 تعداد واژه‌هایی که تنها در فهرست اول وجود دارند، y_{1_i} جایگاه واژه‌ی غیرمشترک در فهرست اول،

m_2 تعداد واژه‌هایی که تنها در فهرست دوم وجود دارند و y_{2_i} جایگاه واژه‌ی غیرمشترک در فهرست دوم است.

در این فرمول، رتبه‌ی واژه‌های مشترک در هر فهرست از یکدیگر تفریق می‌شوند. به‌عنوان مثال، اگر واژه‌ی «بودن» در یک فهرست جایگاه سوم را دارد و در فهرست دیگر، جایگاه هفتم، از حاصل تفریق هفت و سه، چهار به دست می‌آید. سپس حاصل این تفریق‌ها، که برای هرکدام از واژه‌های مشترک به‌دست آمده است، با یکدیگر جمع می‌شوند.

واژه‌هایی که تنها در فهرست اول وجود دارند در نظر گرفته می‌شوند؛ با فرض این‌که این واژه در جایگاه ۱۰۰۱ فهرست دوم (اولین جایگاه پس از هزار واژه‌ی مورد بررسی) قرار دارد. حال جایگاه هر واژه در فهرست اول از ۱۰۰۱ کم می‌شود؛ برای مثال اگر واژه‌ی «تیم» در جایگاه ۵۵۴ فهرست اول بوده و در فهرست دوم (۱۰۰۰ واژه‌ی نخست) نباشد، فرض می‌کنیم که این واژه در جایگاه ۱۰۰۱ فهرست دوم قرار دارد و حاصل تفریق ۵۵۴ و ۱۰۰۰ را به‌دست می‌آوریم (۴۴۶)؛ سپس حاصل این تفریق‌ها را که برای هر کدام از واژه‌های غیرمشترک فهرست اول به‌دست‌آمده با هم جمع می‌کنیم. در مورد واژه‌هایی نیز که تنها در فهرست دوم وجود دارند، همین فرایند را انجام می‌دهیم.

در مرحله‌ی چهارم حاصل سه تفریق را با هم جمع کرده و تقسیم بر تعداد کل واژه‌های مشترک و واژه‌های غیرمشترک دو فهرست می‌کنیم. گفتنی است تعداد واژه‌های غیرمشترک دو فهرست با هم برابر هستند؛ بنابراین حاصل نتایج مرحله‌ی چهار برابر تعداد کل واژه‌های مشترک به اضافه‌ی دو برابر تعداد واژه‌های غیرمشترک است و تعداد واژه‌های غیرمشترک، برابر با حاصل تفریق ۱۰۰۰ از تعداد واژه‌های مشترک است.

پس می‌توان با توجه به برابری‌های

$$m_1 = m_2$$

$$m_1 = 1000 - n$$

به برابری زیر رسید و فرمول را به این شکل تغییر داد:

فرمول ۲. درصد اختلاف جایگزین‌شده

$$M = \left(\frac{\sum_{i=1}^{n} |x_{1_i} - x_{2_i}| + \sum_{i=1}^{m_1} |1001 - y_{1_i}| + \sum_{i=1}^{m_2} |1001 - y_{2_i}|}{2000 - n} \Big/ 1001 \right) \times 100$$

در مرحله‌ی پنجم حاصل مرحله‌ی چهارم را تقسیم بر حداکثر فاصله‌ی ممکن میان یک واژه در دو فهرست (۱۰۰۱) می‌کنیم تا نسبت شباهت دو فهرست به‌دست آید. این تقسیم بر اساس برابری زیر قابل توجیه است:

$$\frac{x}{1001} = \frac{y}{100}$$

در این برابری نسبت‌ها، اگر میانگین فاصلهٔ واژه‌ها در دو فهرست؛ یعنی x برابر با ۱۰۰۱ باشد، دو فهرست صددرصد با هم اختلاف دارند. پس برای به‌دست آوردن درصد اختلاف در این‌جا، کافی است x را محاسبه کنیم که حاصل تقسیم بر ۱۰۰۱ و ضرب در صد است. در مرحله ششم، حاصل مرحله پنجم را در صد ضرب می‌کنیم تا درصد اختلاف دو فهرست به‌دست آید.

از فرمول زیر، که در آن Z درصد شباهت کلی واژه‌های دو فهرست است، به‌منظور دستیابی به درصد شباهت میان نتایج پژوهش حاضر و دیگر پژوهش‌ها استفاده شده است، در این فرمول، تنها عدد ۱۰۰ از درصد اختلاف کم می‌شود تا درصد شباهت دو فهرست به دست آید.

$$Z = 100 - M$$

نتایج بررسی و مقایسه پروژه‌ی حاضر با دیگر پروژه‌ها براساس دو فرمول فوق به شرح جدول زیر است. گفتنی است که با توجه به عدم وجود ترتیب در پژوهش نعمت‌زاده (۱۳۹۰) امکان اجرای این فرمول برای مقایسه‌ی پروژه‌ی حاضر با آن وجود نداشته است.

به‌منظور دستیابی به واژه‌های مشترک بین این پروژه و پروژه‌های دیگر، ۱۰۰۰ واژه‌ی پربسامد پژوهش حاضر و پژوهش‌های دیگر بررسی شد. سپس با مقایسه‌ی این واژه‌ها با یکدیگر (با روش Filter در Excel)، واژه‌های مشترک تمامی پروژه‌های نام برده استخراج گردید. درمجموع ۵۷ واژه از هزار واژه‌ی اول در همه‌ی پروژه‌ها (پروژه‌ی حاضر، حمید حسنی(۱۳۸۴)، بیجن‌خان (نسخه‌ی اینترنتی، ۱۳۹۰)، بی‌جن‌خان (نسخه‌ی چاپی، ۱۳۸۳)، محک‌وب، همشهری و نعمت‌زاده(۱۳۹۰)) مشترک بودند. این واژه‌ها در جدول زیر آمده‌اند:

جدول ۱۳. واژه‌های مشترک بین پروژه‌ی حاضر و سایر پروژه‌ها

در	دست	مرد	فارسی
روز	رفتن	درست	مادر
فیلم	دیدن	شکل	دفتر
کتاب	جوان	نقطه	ساختمان
مردم	آمدن	دوست	آوردن
بازی	خانه	گل	نامه
ماه	بالا	دکتر	خیابان
خواستن	آب	کوتاه	پدر
بزرگ	زبان	تلفن	خون
شهر	انسان	هوا	رنگ
خوب	تلویزیون	چشم	خدا
کم	ساعت	خط	صبح
صورت	دور	باز	
فوتبال	سر	کوچک	
زیاد	شب	مدرسه	

جدول ۱۴. مقایسه درصد شباهت و اختلاف میان پژوهش حاضر با سایر پژوهش‌ها

	مقایسه با پروژه‌ی حاضر	
	درصد اختلاف	درصد شباهت
حمید حسنی	۳۳،۸۸۹٪	۶۶،۱۱۰٪
بیجن‌خان (نسخه‌ی اینترنتی)	۲۲،۲۲۳٪	۷۵،۷۷۶٪
بی‌جن‌خان (نسخه‌ی چاپی)	۲۴،۸۵۳٪	۷۵،۱۴۶٪
محک وب	۲۹،۹۷۹٪	۷۰،۰۲۰٪
همشهری	۲۲،۵۸۱٪	۷۷،۴۱۹٪

اما با توجه به متفاوت بودن روش نعمت‌زاده (۱۳۹۰)، واژه‌های مشترک بین پروژه‌ها بدون احتساب پروژه‌ی نعمت‌زاده بررسی و استخراج شد. در مجموع ۳۶۱ واژه از ۱۰۰۰ واژه‌ی اول همه‌ی پروژه‌ها (به جز نعمت‌زاده (۱۳۹۰)) با هم مشترک بودند. این واژه‌ها در جدول زیر آمده‌اند.

جدول ۱۵. واژه‌های مشترک بین پروژه‌ی حاضر و سایر پروژه‌ها (بدون احتساب پروژه‌ی نعمت‌زاده (۱۳۹۰))

و	وقت	طرف	حد	ملی	طور	دستگاه	اختیار
در	صورت	چیز	هوا	رسیدن	خانه	ادبیات	شیوه
به	فوتبال	تغییر	سیستم	زمان	نفر	هنری	ساختمان
از	بار	تاریخ	جریان	عنوان	بیان	لازم	محیط
که	ایرانی	نویسنده	سمت	پس	اساس	مرد	حرف
این	جهان	خانواده	تصمیم	فرد	حق	پاسخ	آوردن
بودن	رئیس	عامل	تحت	مورد	عضو	درست	مرگ
را	جشنواره	معاون	ارائه	برنامه	چون	موفق	چاپ
با	زیاد	جمله	بدن	برخی	نتیجه	گذاشتن	پخش
آن	تنها	انسان	طول	بازی	شما	مربوط	نامه
برای	تمام	بررسی	سایر	ماه	خصوص	بحث	حفظ
گفتن	نظر	تلویزیون	خرید	همین	اصلی	استاد	ترتیب
سال	حتی	ساعت	چشم	سازمان	بالا	حاضر	خیابان
داشتن	جامعه	دادن	خط	حضور	هدف	شکل	دستور
خود	مسئله	سطح	علت	چند	آب	وضعیت	کنترل
هم	هفته	مناسب	ایشان	بعد	قبل	جمع	دل
کشور	زندگی	نیاز	یافتن	خواستن	چرا	همیشه	عالی
تا	شعر	وارد	سرعت	موضوع	چنین	اعلام	گاز
شدن	گرفتن	ارتباط	چگونه	گروه	فقط	آیا	بعضی
توانستن	دست	تلاش	خاطر	جدید	امر	کمک	موجب
اما	مسئول	شاید	روش	پیش	اسلامی	جنگ	دسته
نیز	منطقه	ماده	باز	اول	مانند	مسیر	مستقیم

دیگر	نوع	خارجی	مطالعه	بزرگ	استفاده	واحد	اندازه
ما	نسبت	سیاسی	منظور	همچنین	تعداد	انجام	پشت
تیم	خبر	نیرو	آقا	دلیل	واقع	موجود	انتقال
هر	رفتن	خاص	دوران	مشکل	اجتماعی	فکر	هنگام
او	نقش	نگاه	اصل	شهر	ولی	ورود	سپس
روز	مسابقه	طی	باعث	بین	هنر	نقطه	شهید
کار	نام	قابل	جز	چه	فصل	آغاز	پدر
وی	دیدار	انرژی	کامل	شرط	همان	دوست	تشکیل
یا	آینده	سفر	عمل	مهم	هنوز	تجربه	خون
دولت	امروز	رشد	غیر	بانک	وجود	گل	شروع
بیش	پایان	امکان	کوچک	مرکز	فضا	زیر	مشخص
فیلم	جهانی	دور	ساختن	تولید	موسیقی	ضمن	رنگ
کردن	زمینه	انقلاب	مدرسه	بازار	ادامه	دنبال	روزنامه
فرهنگ	گفتگو	سر	فارسی	البته	مراسم	مثل	خدا
کتاب	انتخاب	گونه	احساس	اجرا	برابر	رشته	حل
بخش	مرحله	محل	مادر	سو	زبان	سخن	صبح
اثر	دیدن	ایجاد	قالب	شرکت	مقابل	دکتر	آزادی
مردم	نه	آموزش	آماده	طرح	همراه	کوتاه	یاد
اگر	جوان	تازه	اهمیت	خوب	یعنی	حالا	تهیه
بسیار	اشاره	عمومی	انتظار	توجه	بدون	تلفن	پیام
حال	آمدن	کنار	دفتر	هیچ	مدت	خبرنگار	نمایش
من	دنیا	دریافت	طبیعی	کم	توسط	میدان	طریق
همه	داستان	شب	قسمت	مختلف	خیلی	قدرت	ویژه
گذشته							

داده‌های پروژه‌ی حاضر ممکن است در مراحل مختلف ورود اطلاعات، بسامدگیری و برچسب‌زنی، تصحیح اطلاعات، تهیه‌ی فهرست واژه‌ها، همچنین استخراج فهرست‌ها، تصحیح فهرست‌ها و تطبیق آن‌ها با یکدیگر، دچار اشتباهات انسانی شده باشد؛ با این حال با اعتبارسنجی دوباره داده‌ها می‌توان نتایج دقیق‌تری کسب کرد.

۳. نتیجه‌گیری

پژوهش پیش‌رو به سبب ماهیتی که از آن برخوردار است و مقایسه‌ای که بین نتایج مهم‌ترین پژوهش‌های بسامدی در آن صورت گرفته است، در نوع خود بی‌سابقه است. همچنین نتایج حاصل از مقایسه‌های گوناگون کیفی و کمّی حاکی از تمایز پروژه‌ی حاضر از لحاظ ویژگی‌های مختلف نسبت به اکثر پروژه‌های مشابه است. علاوه بر آن با در نظر گرفتن دو معیار «وجود واژه‌ها» و «اختلاف درجه‌ی واژه‌ها» و مقایسه‌ی پژوهش حاضر با دیگر پژوهش‌های مشابه، به اختلافی در حدود ۳۰ درصد در نتایج به‌دست آمده برخوردیم که در مقایسه‌ی

سایر فهرست‌ها با یکدیگر نیز حاصل شد؛ بنابراین می‌توان چنین نتیجه گرفت که این اختلاف، طبیعی بوده و یافته‌های فهرست واژه‌های پربسامد در این فهرست، اعتبار قابل قبولی دارد.

نتایج پژوهش حاضر می‌تواند دارای کاربردهای آموزشی زیر باشد:

۱. استفاده از واژه‌های پایه‌ی زبان فارسی جهت تولید منابع آموزشی: ازآنجاکه به‌منظور تولید هرگونه محتوای آموزشی در حوزه زبان، فهرست بسامدی و واژه‌های پایه‌ی آن زبان مورد نیاز است، تولید متون آموزش زبان فارسی بدون در دست داشتن فهرست واژه‌های پربسامد زبان فارسی ناممکن و یا نامعتبر است؛ بنابراین می‌توان از نتایج پژوهش پیش‌رو در تولید محتوای آموزش زبان فارسی (به فارسی‌زبانان و غیرفارسی‌زبانان) بهره برد.

۲. استفاده از واژه‌های پایه‌ی زبان فارسی جهت آزمون‌سازی؛ علاوه بر نقش آموزشی واژه‌های پایه‌ی زبان، می‌توان استفاده در آزمون‌سازی و ارزشیابی را به دیگر کارکردهای آن اضافه کرد. جهت سنجش دانش واژگانی زبان‌آموزان و سطح‌بندی آنان، می‌توان از آزمون‌هایی که حاوی سطوح مختلف دانش واژگانی است (۱۰۰۰ واژه‌ی پربسامد، ۲۰۰۰ واژه‌ی پربسامد و غیره) استفاده نمود.

۳. استفاده از واژه‌های پایه‌ی زبان فارسی توسط آموزگاران زبان فارسی: آشنایی آموزگاران زبان فارسی با فهرست بسامدی واژه‌های این زبان، آن‌ها را در امر انتخاب محتوا و سیر تدریس زبان فارسی یاری فراوان می‌رساند.

۴. استفاده از واژه‌های پایه‌ی زبان فارسی توسط فارسی‌آموزان: با توجه به پژوهش‌های انجام شده در زمینه‌ی واژه‌های پایه، آشنایی با ۲۰۰۰ واژه‌ی پایه در هر زبان، شخص را قادر به درک حدود ۹۰ درصد متون روزنامه‌ای در آن زبان می‌سازد (نیشن، ۲۰۰۱). ازاین‌رو، دسترسی فارسی‌آموزان به واژه‌های پایه‌ی زبان فارسی می‌تواند بسیار مفید باشد و آن‌ها را در امر یادگیری و فهم زبان فارسی یاری فراوان رساند.

منابع:

بی‌جن‌خان، م. (۱۳۹۰). *فرهنگ بسامدی براساس پیکره‌ی متنی زبان فارسی امروز*. تهران: مؤسسه‌ی انتشارات دانشگاه تهران.

حسنی، ح. (۱۳۸۴). *واژه‌های پرکاربرد فارسی امروز بر مبنای پیکره‌ی یک‌میلیون لغتی شامل بیش از ۸۰۰۰ لغت قاموسی و غیر قاموسی*. تهران: کانون زبان ایران.

درودی، ا.، برادران‌هاشمی، ه.، آل‌احمد، ا.، زارع بیدکی، ع.، حبیبیان، ا.، مهدیخانی، ف.، شاکری، آ.، و رهگذر، م. (۱۳۸۷). *مجموعه محک استاندارد برای تحقیقات بازیابی اطلاعات وب فارسی*. تهران: گزارش فنی گروه تحقیقاتی پایگاه داده‌ها دانشگاه تهران، شماره: DBRG-TR-138702.

نعمت‌زاده، ش.، دادرس، م.، دستجردی‌کاظمی، م.، و منصوری‌زاده، م. (۱۳۹۰). *واژه‌های پایه فارسی از زبان کودکان ایرانی. تهران:* مؤسسه‌ی فرهنگی مدرسه برهان (انتشارات مدرسه).

AleAhmad, A., Amiri H., Darrudi E., Rahgozar M., & Oroumchian F. (2009). Hamshahri A *Standard Persian Text Collection. Knowledge-Based Systems*, 22(5), pp. 382–387.

Baker, P., Hardie, A. & McEnery, T. (2006). *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Barnett, B., Lehmann, Hu. & Zoeppritz, M. (1986). A Word Database for Natural Language Processing. *Proceedings of the 11th International Conference on Computational Linguistics COLING86*.

Bullon, S. & Leech, G. (2007). *Longman Communication 3000*. Harlow: Pearson Longman.

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American Heritage Word Frequency Book*. Boston: Houghton Mifflin.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34, 2: 213-238.

Davies, M. & Gardner, D. (2010). *A Frequency Dictionary of Contemporary American English: Word Sketches, Collocates and Thematic Lists*. London: Routledge.

Dixon, R. M. W. (1971). *Method of Semantic description.* In L. Verhoeven and J.H.A.L de Jong (eds.)

Dolch, E.W. (1936). A Basic Sight Vocabulary. *Elementary School Journal*, 36,pp. 456-460.

Fry, E. B., Kress, J. E., & Fountoukidis, D. L. (2000). *The Reading Teachers Book of Lists*, 4th Edition. London: Pearson Ptr.

Jones, R. L., & Tschirner, E. (2006). *A Frequency Dictionary of German*. London: Routledge.

Käding, F.W. (1897). *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: no publ.

Laufer, B. (1997). The Lexical Plight in Second Language Reading: Words You Do Not Know, Words You Think You Know, and Words You Can not Guess. In J. Coady, & T. Huckin (Eds.). *Second language vocabulary acquisition* (pp 20-34). Cambridge: Cambridge University Press.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

Ogden, C. K., & Richards, I. A. (1923). *The Meaning of Meaning*. London: Kegan, Paul, Trench, Trubner.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching &. Linguistics Abstracts*, 13(4a),pp. 221-247.

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge, UK: Cambridge University Press.

Nation, P. (2006). How Large a Vocabulary Is Needed for Reading and Listening?. *The Canadian Modern Language Review*, 63(1), pp.59–82.

Nation, P. (2007).*Teaching Vocabulary: Strategies and Techniques*. New York: Thomson/Heinle.

Oxford (2008). *My Oxford Wordlist*. Oxford University Press.

Rosch, E. H. (1973). On the Internal Structure of Perceptual and Semantic Categories. In T. E. Moore (ed.). *Cognitive Development and the Acquisition of Language* (pp 111-144).

Thornbury, S. (2004). *How to Teach Vocabulary*. Essex: Pearson Education Limited.

Thorndike, E. L. (1921). *The Teacher's Word Book*. New York: Columbia University Press.

Verlinde S., Selva T. (2001). Nomenclature de Dictionnaire et Analyse de Corpus. *Cahiers de Lexicologie*, 79, 2 ,pp.113-139.

Vermeer, A. (1992). Exploring the Second Language Learner Lexicon. In L. Verhoeven and J.H.A.L de Jong (eds.). *The Construct of Language Proficiency: Applications of Psychological Models to Language Assessments* (pp 147-171). Amsterdam: John Benjamins.

Wilkins, D. A. (1972). *Linguistics in language teaching*. London: Edward Arnold.