فصلنامهٔ علمی - پژوهشی زبان‌پژوهی دانشگاه الزهرا(س)
سال یازدهم، شمارهٔ ۳۳، زمستان ۱۳۹۸

واژه‌های پایهٔ زبان فارسی مبتنی بر متون مطبوعاتی [1*]

رضامراد صحرایی [2]
امیرحسین مجیری فروشانی [3]
مروارید طالبی [4]

تاریخ دریافت: ۱۳۹۷/۰۳/۲۱
تاریخ پذیرش: ۱۳۹۷/۱۱/۰۳

چکیده

آموزش واژه‌های زبان، یکی از مهم‌ترین مؤلفه‌های آموزش زبان خارجی است که می‌تواند هر چهار مهارت اصلی زبان (شنیداری، گفتاری، خواندن و نوشتن) را تحت تأثیر قرار دهد. بر پایهٔ پژوهش‌هایی که در حوزهٔ آموزش واژه انجام شده‌است، واژه‌های پربسامد و پایهٔ زبان، به دلیل فراگیری آسان و کاربرد فراوان

---

[1] شناسه دیجیتال (DOI): 10.22051/jlr.2019.19564.1520

[*] به این وسیله از آقایان سعید کلاهی، امین نویدی، سهیل محمدیان و منوچهر نادری و خانم‌ها شیما حسینی، نسیبه مؤذنی و سمیه اسمعلی‌زاده به سبب شرکت در این پژوهش، گردآوری متن‌های مطبوعاتی و انجام برخی کارهای پژوهشی بسیار سپاسگزاریم. همچنین از آقای محمدمهدی مجیری به دلیل طراحی و پشتیبانی نرم‌افزار استخراج واژه‌های پایه، صمیمانه تشکر می‌کنیم.

[2] دکترای زبان‌شناسی، دانشیار گروه زبان‌شناسی دانشگاه علامه طباطبایی و مدیر هستهٔ گروه پژوهش های بنیادیِ توسعهٔ آموزش زبان فارسی به غیر فارسی‌زبان‌ها، دانشگاه علامه طباطبائی (نویسندهٔ مسئول)؛ sahraei@atu.ac.ir

[3] کارشناسی ارشد آموزش زبان فارسی به غیرفارسی‌زبانان، پژوهشگر بنیاد سعدی؛ amojiry@saadifoundation.ir

[4] کارشناسی ارشد آموزش زبان فارسی به غیرفارسی‌زبانان، عضو هسته پژوهش‌های بنیادیِ توسعه آموزش زبان فارسی به غیر فارسی‌زبان‌ها، دانشگاه علامه طباطبائی؛ morvarid_talebi@atu.ac.ir

در زبان روزمره، از اهمیت ویژه‌ای برخوردار هستند. فهرست واژه‌های پربسامد یا پایه، مجموعه‌ای از واژه‌هاست که در پیکره‌ای زبانی، فراوانی (تکرار) بیشتری داشته‌اند. برای دست‌یابی به پیکره‌ای مناسب، از متن‌های مطبوعاتی در هفت حوزهٔ گوناگون (مشتمل بر فرهنگی، اجتماعی، سیاسی، ورزشی، ادبیات داستانی، اقتصادی و علمی) استفاده شد. سپس به مدت ۱۰۰ روز کاری، پیکره‌ای ۲۴۰۰ متنی، مشتمل بر یک میلیون و دویست‌هزار واژه استخراج گردید. سپس، با استفاده از نرم‌افزاری که برای انجام این پژوهش طراحی شده بود، واژه‌ها براساس گونه برچسب‌گذاری شدند. در پایان، از میان واژه‌های برچسب‌گذاری‌شده، ۲۰۰۰ واژه‌ای که بیش از ۵۰ بار تکرار شده‌اند، به عنوان واژه‌های پایهٔ زبان فارسی مبتنی بر متون مطبوعاتی معرفی شدند.

واژه‌های کلیدی: پژوهش پیکره بنیاد، پیکرهٔ متون مطبوعاتی، آموزش واژه، واژه‌های پایه، واژه‌های پربسامد فارسی

۱. مقدمه

از مهم‌ترین مؤلفه‌های آموزش زبان خارجی، آموزش واژه‌های زبان است. بدون آشنایی با واژه‌های زبان، هیچ‌گونه اطلاعاتی را نمی‌توان انتقال داد (Wilkins, 1976) و یا ارتباط معناداری با گویشوران زبان مورد نظر برقرار کرد (McCarthy, 1990). آموزش واژه می‌تواند هر چهار مهارت اصلی زبان (گوش دادن، خواندن، صحبت کردن و نوشتن) را تحت تأثیر قرار دهد. بسیاری از مشکلات زبان‌آموزان در تولید و دریافت زبان، ناشی از دانش واژگانی اندک است (Meara, 1980) و مهم‌ترین بخش زبان برای زبان‌آموز، واژگان است (Laufer, 1997). در این مؤلفه، زبان‌آموز علاوه بر آموختن واژه، باید کارکرد واژه را نیز بیاموزد. نخستین گام در آموزش واژه، دسترسی به فهرست واژه‌های پایهٔ زبان است. فهرست واژه‌های پربسامد یا فرهنگ بسامدی، مجموعه‌ای از واژه‌هاست که در مجموعه‌ای از داده‌های زبانی (پیکرهٔ زبانی)، فراوانی (تکرار) بیشتری داشته‌اند.

پژوهش در زمینهٔ واژه‌های پایهٔ زبان و استخراج این واژه‌ها و به‌کاربستن آن‌ها در آموزش زبان، بسیار مورد توجه قرار گرفته‌است. زیرا در آموزش هر زبانی (خواه زبان اول فرد باشد و خواه زبان دوم)، علاوه بر دسته‌بندی و تشخیص محتوای مناسب دستوری بر پایهٔ سطح دانش زبان‌آموزان، باید از واژه‌های ضروری و مناسب همان سطح زبانی هم استفاده شود. از سال ۱۸۹۷ به طور جدی، فهرست‌های فراوانی از واژه‌های پایه در زبان‌های گوناگون ارائه شده‌است. در این میان، پژوهش‌های زبان انگلیسی بیش از سایر پژوهش‌ها است. به همین سبب، سال‌هاست که در

سطح‌بندی متن‌های آموزشی زبان انگلیسی از فهرست واژه‌های پربسامد این زبان استفاده می‌شود. از سال ۱۳۵۰، در زبان فارسی هم پژوهش‌هایی برای استخراج واژه‌های پربسامد انجام شده‌است. برای تهیهٔ متن‌های آموزش زبان فارسی به غیرفارسی‌زبانان، فهرست قابل قبولی از واژه‌های پایهٔ زبان فارسی مورد نیاز است. در واقع، چنین فهرستی از اولویت‌های آموزشی زبان فارسی است.

اصلی‌ترین هدف این پژوهش، استخراج فهرست واژه‌های پربسامد در زبان فارسی است. به این منظور، سعی شده‌است تا اندازهٔ ممکن، از متن‌های نوشتاری معیار که موضوع‌های متنوعی دارند، استفاده شود. این انتخاب سبب شکل‌گیری پیکره‌ای گسترده و فراگیر از واژه‌های گوناگون شده‌است. از یافته‌های پژوهش در زمینه‌های گوناگونی مانند تحلیل گفتمان انتقادی، استخراج اصطلاح‌های علمی و معادل‌سازی آن‌ها و بسیاری موارد دیگر، می‌توان بسیار بهره‌مند شد. این پژوهش، همچنین کاربردهای گوناگونی در زمینهٔ آموزشی دارد. از جمله این موارد می‌توان کمک به تولید منابع آموزشی، آزمون‌سازی، طراحی سیر آموزش و طرح درس‌نامه‌ها و کمک به زبان‌آموزان برای خواندن متون مطبوعاتی فارسی را نام برد. همچنین، واژه‌های پایه همگی از پژوهش‌های پیکره‌بنیاد استخراج می‌شوند و خروجی هر پیکرهٔ زبانی می‌تواند یک فهرست واژه‌های پایه (با توجه به زبان پیکره و نوع متون آن) باشد. با این وجود، نگارندگان در این بخش به مرور پژوهش‌هایی در ایران و خارج از ایران می‌پردازند که هدف اصلی آن‌ها تهیهٔ فهرست واژه‌های پایه بوده‌است.

نخستین فرهنگ بسامدی را کدینگ[1] در سال ۱۸۹۷ در زبان آلمانی با پیکره‌ای ۱۱ میلیون واژه‌ای تولید کرد. پس از آن افراد بسیاری دست به تولید فرهنگ‌های بسامدی برای زبان‌های گوناگون زدند که در ادامه به برخی از آن‌ها اشاره خواهیم کرد. ثورندایک (Thorndike, 1992)، در «کتاب واژهٔ معلم» (۱۱هزار واژهٔ پربسامد انگلیسی) ۳۰ هزار واژهٔ پایهٔ انگلیسی را ارائه داد که به شکل دستی از پیکره‌ای با ۱۸ میلیون واژه استخراج شده بود. در سال ۱۹۲۳، آگدن[2] و ریچاردز[3] «فهرست پایهٔ انگلیسی» را به شکل دستی مدون کردند که شامل ۸۵۰ واژهٔ پربسامد زبان انگلیسی است. دلچ (Dolch, 1936)، فهرستی ۲۲۰ واژه‌ای از واژه‌های پایهٔ زبان انگلیسی در کتاب خود ارائه داد. پیکرهٔ مورد استفادهٔ دلچ مشتمل بر حدود ۴۵۰۰ واژه و دربرگیرندهٔ متن‌های گروه سنی کودک هم بود و استخراج واژه‌ها به شکل دستی انجام شد. وست (West, 1953)، فهرستی از ۲۰۰۰ واژهٔ پایه در زبان انگلیسی ارائه داد. این فهرست به صورت دستی از پیکره‌ای نوشتاری شامل

---

[1] F. W. Käding
[2] C. K. Ogden
[3] I. A. Richards

۵ میلیون واژه استخراج شد. کارول و همکاران (Carroll et al., 1971) در «کتاب بسامد واژه[1]» از پیکره‌ای ۵ میلیون واژه‌ای، شامل متون آموزشی مدارس آمریکا، واژه‌های پایه را استخراج کردند. در این کتاب متناسب با هر موضوع و سطح زبانی، یک فهرست واژه ارائه شده‌است. کاکس‌هد (Coxhead, 2000)، در «فهرست واژگانی علمی جدید»، واژه‌های پایه را در چهار زمینهٔ هنر، بازرگانی، حقوق و علوم استخراج کرده‌است. در این پژوهش، کاکس‌هد ۵۷۰ خانوادهٔ واژگانی را در بین هر چهار حوزهٔ مشترک یافت. در ۲۰۰۱، ورلیند[2] و سلوا[3] فهرستی از واژه‌های پربسامد در زبان فرانسه ارائه دادند. پیکرهٔ مورد استفاده در این پژوهش، دو روزنامهٔ فرانسوی و بلژیکی و تعداد واژه‌های موجود در پیکره در کل ۵۰ میلیون واژه بود. ۱۰۰ و ۱۰۰۰ واژهٔ پایهٔ زبان انگلیسی به کوشش فرای و همکاران (Fry et al., 2000)، از پیکره‌ای حاوی پنج میلیون واژه استخراج شد. «فرهنگ بسامدی آلمانی» اثر جونز و شیرنر (Jones & Tschirner, 2006)، از دیگر فرهنگ‌های بسامدی است که به واژه‌های پربسامد در زبان آلمانی اختصاص دارد. این فرهنگ دارای ۴۰۳۷ مدخل، و پیکرهٔ مورد استفاده در این پژوهش حاوی ۴میلیون واژه بود. دیویس و گاردنر (Davies & Gardner, 2010)، از پیکره‌ای حاوی بیش از ۴۰۰ میلیون واژه به استخراج ۱۰۰۰–۵۰۰۰ واژهٔ پایه در زبان انگلیسی پرداختند. علاوه بر این، می‌توان به فهرست بسامدی ۱۰۰ واژهٔ پایه به کوشش لغت‌نامهٔ انگلیسی آکسفورد[4] و ۳۰۰۰ واژهٔ پایه توسط لغت‌نامهٔ لانگ‌من[5] و بسیاری موارد دیگر نیز اشاره کرد.

نخستین پژوهش در زمینهٔ واژه‌های پربسامد در زبان فارسی متعلق به فریدون بدره‌ای (Badreee, 1971) است. پس از او نیز پژوهش‌های دیگری در این زمینه انجام شد. براهنی (Baraheni,) اثر «بررسی میزان فراوانی واژه‌ها و تأثیر آن بر خواندن» و ایمن (Imen, 1978) مقالهٔ «فهرست پربسامدترین واژه‌های خردسالان ۹ تا ۱۲ سال» را ارائه کردند. صفارپور (Safarpour, 1991)، اثر «واژگان پایهٔ کلاس‌های اول تا پنجم دبستان» و تحریریان (Tahririyan, 1994) مقالهٔ «نگاهی به فراوانی واژه‌ها و واژگان پایه در فارسی» نگارش کرده‌اند. بی‌جن‌خان (Bijan Khan, 2011). «پیکرهٔ متنی زبان فارسی» را ارائه کرد که از پیکره‌ای با حدود ۱۰ میلیون قطعه نوشتار استخراج شده‌است. این فهرست شامل ۱۸۶۸ واژه با فراوانی ۵۰۰ بار یا بیش‌تر است. حسنی (Hasani, 2005). اثر «واژه‌های پرکاربرد فارسی امروز» را ارائه کرده‌است که از میان یک میلیون

---

[1] word frequency book
[2] S. Verlinde
[3] T. Selva
[4] Oxford English dictionary
[5] Longman communication 3000

واژه از پیکره‌ای شامل کتاب، مجله و روزنامه استخراج شده‌است. این فهرست شامل ۸۴۳۸ واژه با فراوانی ده بار یا بیش‌تر است. غروی قوچانی (Gourori Gouchani, 2006). مقالهٔ «تعیین واژگان پایهٔ فارسی معیار گفتاری در بزرگسالان و تجزیه و تحلیل تواتر آن‌ها بر اساس قانون زیف» را نگارش کرده‌اند. درودی و همکاران (Doroodi et al., 2009). اثر «محک استاندارد برای تحقیقات بازیابی اطلاعات وب فارسی» را ارائه کرده‌اند که از پیکره‌ای شامل حدود ۴ میلیون صفحهٔ اینترنتی استخراج شده‌است. آل‌احمد و همکاران (AleAhmad et al., 2009)، پیکرهٔ همشهری را ارائه کرده‌اند که از پیکره‌ای شامل ۳۱۸ هزار سند از ۱۲ سال چاپ روزنامهٔ همشهری برگرفته شده‌است. نعمت‌زاده و همکاران (Nematzadeh, 2011)، اثر «واژه‌های پایه فارسی از زبان کودکان ایرانی» را نگارش کرده‌اند که از آزمون‌های ادراکی، تولیدی (برای دانش‌آموزان) و آزمون‌های مخصوص معلمان از ۲۰ هزار دانش‌آموز و ۷۵۰ معلم ابتدایی در ۱۷۵ مدرسهٔ سراسر کشور استخراج شده‌است. این فهرست شامل ۴۹۷ واژه در ۲۴ موضوع گوناگون است.

پژوهش‌ها در زمینهٔ واژه‌های پایه اغلب به صورت دستی انجام می‌شده‌است. با ظهور سیستم‌های رایانه‌ای انجام این گونه پژوهش‌ها بسیار ساده‌تر و دقیق‌تر شده‌است. نخستین گام به منظور پژوهش در زمینهٔ واژه‌های پایه، به صورت دستی و رایانه‌ای، ارائهٔ تعریفی برای واژهٔ پایه است. سه رویکرد عمده برای تعریف واژه‌های پایه وجود دارد (Barentt, 1986): نخست، تعداد مشخصی از فهرست واژه‌های پربسامد (رویکرد آماری). دوم، واژه‌های مشترک بین همهٔ گویشوران زبان (رویکرد آماری) و سوم، واژه‌های کافی برای توصیف سایر واژه‌های زبان (رویکرد معنایی).

فهرست واژه‌های پربسامد مجموعه‌ای از واژه‌هاست که در مجموعه‌ای از داده‌های زبانی (پیکرهٔ زبانی)، فراوانی بیش‌تری داشته‌اند. پژوهشگران معتقدند واژه‌های پایه، فراوانی وقوع بیش‌تری نسبت به واژه‌های غیرپایه دارند (Dixon, 1971). برای نمونه، ویلکینز (Wilkins, 1972) یکی از مهم‌ترین معیارهای انتخاب واژه‌های پایه را فراوانی می‌داند. فرض بر این است که فراوانی بیش‌تر واژه، به معنی اهمیت بیش‌تر آن است و در نتیجه واژه‌های پربسامد باید در اولویت یادگیری قرار گیرند (Bijnkhan, 2011). زیرا کودک در ابتدا واژه‌های اصلی و پایهٔ زبان را فرا می‌گیرد (Rosch, 1973) و روند یادگیری واژه‌های زبان دوم/ خارجی نیز به همین ترتیب است (Cook, 1991).

دایرهٔ واژگانی انگلیسی‌زبانان در ۲۰ سالگی، حدود ۲۰ هزار خانوادهٔ واژگانی است، اما انگلیسی‌زبانان می‌توانند با دانستن ۸ تا ۹ هزار خانوادهٔ واژگانی، بیشتر متن‌های انگلیسی را بخوانند و درک کنند. همچنین آن‌ها با دانستن ۱۰۰۰ واژهٔ پایهٔ زبان مقصد، ۳.۸۲ متن‌های داستانی، ۶.۷۵

متن‌های روزنامه‌ای و ۵.۷۳ متن‌های علمی را درک می‌کنند. اگر فرایند یادگیری واژه در زبان خارجی/دوم هم‌چون فرایند یادگیری در زبان اول باشد، می‌توان مدت زمان یادگیری زبان خارجی/دوم را کاهش داد. این عمل از طریق آموزش واژه‌های ضروری و پرکاربرد صورت می‌گیرد (Nation, 2001; Nation, 2006).

برخی پژوهشگران اشاره کرده‌اند که بسامدگیری و شمارش واژه‌ها به تنهایی معیار مناسبی برای تعیین واژه‌های پایهٔ زبان نیست. وست (West, 1953)، علاوه بر مؤلفهٔ فراوانی واژه، موارد دیگری مانند سادگی، ضرورت، پوشش معنایی و ویژگی سبکی و سطح عاطفی واژه را در انتخاب واژه‌های پایهٔ زبان مؤثر می‌داند. همچنین، گروهی دیگر از پژوهشگران مانند شین و نیشن (Shin & Nation, 2008) استفاده از پیکره‌های گفتاری و در نظر گرفتن باهم‌آیی‌ها و جمله‌های قالبی را در استخراج واژه‌های پایهٔ زبان، ضروری دانسته و استفادهٔ صرف از پیکره‌های نوشتاری را ناکارآمد می‌دانند. علاوه بر این، برخی معیارهای رایج در تشخیص واژه‌های پایه از این قرارند (Carter, 2002):

الف) جایگزینی نحوی (ساختاری): واژه‌های پایه معمولاً می‌توانند به جای واژه‌های مشابه دیگر قرار گیرند (مانند «دادن» به جای «بخشیدن»، «اعطاکردن»، «هدیه دادن» و موارد مشابه)

ب) متضاد: معمولاً یافتن متضاد واژه‌های پایه راحت‌تر از واژه‌های غیرپایه است.

پ) باهم‌آییِ هم‌نشینی: واژه‌های پایه، معمولاً در محور هم‌نشینی، بیش‌تر کاربرد دارند (مانند «روشن» در «اتاق روشن»، «آینده روشن»، «آبی روشن» و موارد مشابه)

ت) گستردگی: واژه‌های پایه معمولاً بیش‌تر در مدخل‌های فرهنگ‌های لغت تکرار می‌شوند.

ث) شمول معنایی: واژه‌های پایه ویژگی‌های عام‌تری از واژه‌های غیرپایه دارند و معمولاً واژه‌های دیگر، زیرمجموعهٔ آن‌ها قرار می‌گیرند (مانند «گل» به جای «لاله»، «رز» و موارد مشابه)

ج) عاری بودن از مظاهر فرهنگی: واژه‌های پایه معمولاً کم‌تر به کاربردهای فرهنگی وابسته‌اند.

چ) خلاصه‌سازی: اگر از گویشوران زبان خواسته شود یک متن را خلاصه کنند، معمولاً از واژه‌های پایه استفاده می‌کنند.

ه) باهم‌آیی متداعی: واژه‌های پایه معمولاً راحت‌تر به ذهن سپرده می‌شوند.

خ) خنثایی زمینهٔ گفتمانی: زمینهٔ گفتمانی واژه‌های پایه معمولاً ویژه و مربوط به یک بافت خاص نیست (مانند «آشپزخانه»، «چپ» و «راست» در مقابل «بندر»، «لنگر» و مواردی از این قبیل).

د) خنثایی ارتباط گفتمان: واژه‌های پایه معمولاً از دید گویشوران زبان، واژه‌های خنثی به حساب می‌آیند (مانند «چاق» در مقابل «گوشتالو»، «تپلو»، «فربه»، «تنومند» و موارد مشابه).

واژه‌های پایه، بسته به بافت کاربردشان به چهار دستهٔ نوشتاری (فرد در نوشتن به کار می‌برد)، خوانداری (فرد در خواندن متن به سادگی می‌فهمد)، گفتاری (در کلام فرد یافت می‌شود) و شنیداری (فرد وقتی می‌شنود، می‌فهمد) دسته‌بندی می‌شوند. همچنین واژه‌های پایه، از جنبهٔ محتوا به دو دستهٔ عمومی (برای ارتباط متعارف) و تخصصی (زبان علم و حوزه‌ای تخصصی) گروه‌بندی می‌شوند (Nematzadeh et al., 2011).

۲. روش پژوهش

مراحل کار در این پژوهش به ترتیب زیر بوده‌است:

۲. ۱. استخراج متون و ثبت در پایگاه داده‌ها

۸ نفر در ۱۰۰ روز کاری (روزهایی که روزنامه‌های کشور منتشر می‌شوند)، هر روز ۳ متن با میانگین ۵۰۰ واژه از سه روزنامه (از میان روزنامه‌های ایران، همشهری، جام‌جم، اطلاعات، آفرینش و تهران امروز) در یکی از ۷ زمینهٔ موضوعی مشخص‌شده (اجتماعی، سیاسی، ورزشی، علمی، اقتصادی، ادبیات داستانی و فرهنگی) استخراج کرده و در پایگاه داده‌ها ثبت می‌کردند. برخی قانون‌های وارد کردن متن در این مرحله از این قرارند:

الف) متن‌های انتخابی باید به‌روز و گزارشی می‌بودند.

ب) گروه پژوهشی فعل‌های مرکب را بین دو علامت «#» و اسامی خاص را بین دو علامت «*» قرار می‌دادند تا نرم‌افزار این واژه‌ها را خود برچسب‌گذاری کند.

پ) معیار نرم‌افزار برای تفکیک واژه‌ها، «فاصله[1]» بود. بنابراین گروه پژوهشی باید توجه می‌کرد که بین هر واژه، فاصله باشد و بین بخش‌های مختلف یک واژه، فاصله‌ای نباشد.

برای ثبت اطلاعات در پایگاه داده و شمارش واژه‌های مستخرج از متن‌های روزنامه‌ای از نرم‌افزاری استفاده شد که مخصوص همین پژوهش و با استفاده از زبان برنامه‌نویسی پی.اچ.پی[2] و تحت وب نوشته شده‌است. مهم‌ترین ویژگی این نرم‌افزار که آن را از دیگر نرم‌افزارهای مدیریت پیکره مانند ورداسمیت[3] متمایز می‌کند، برخط بودن و امکان استفادهٔ همزمان چند کاربر از آن است. از دیگر قابلیت‌های این نرم‌افزار امکان تشخیص پیشوندها و پسوندهای واژه و جمع‌های بی‌قاعده است. همچنین، با اضافه شدن متن‌ها و ویرایش کاربرها، هوشمندی نرم‌افزار افزایش پیدا

---

[1] space
[2] PHP
[3] Wordsmith

کرده و قابلیت تشخیص وندها، جمع‌های مکسر و فعل‌های مرکب در آن بهبود می‌یابد و به صورت خودکار عمل می‌کند. خروجی نرم‌افزار فهرست‌های بسامدی بر اساس نوع متون و یا نوع واژه‌ها است.

انواع واژه‌ها در نرم‌افزار عبارت اند از «اسم»، «فعل» (و به طور ویژه «فعل مرکب»)، «حرف»، «اسم خاص»، «صفت» و «قید». همچنین برای تشخیص جمع‌های بی‌قاعده، فهرستی از این جمع‌ها تهیه شد. برای نمونه «آثار: اثر»، «اواخر: آخر»، «ادبا: ادیب»، «اساطیر: اسطوره» و «اسامی: اسم». همچنین فهرستی از وندهای تصریفی (پیشوندها و پسوندها) و قوانین حاکم بر آن‌ها تهیه شد. در این فهرست، مشخص است که هر نوع از واژه چه پسوند یا پیشوندهایی می‌تواند بگیرد. برای نمونه، فعل می‌تواند با پیشوند تصریفی «می» شروع شود. اسم می‌تواند به پسوند تصریفی «ها» ختم شود. پسوند تصریفی «ات»، می‌تواند به واژهٔ «آیه» اضافه شود اما پیش از آن باید حرف «ه» از آخر این واژه حذف شود. پسوند تصریفی «ان»، می‌تواند به واژهٔ آینده اضافه شود اما پیش از آن باید حرف «ه» از آخر این واژه حذف شده و حرف «گ» اضافه شود.

همچنین در این پیکره، هر متن دارای فراداده‌های «نوع متن» (فرهنگی، اجتماعی، سیاسی و موارد مشابه)، نام روزنامه، تاریخ چاپ، تاریخ تایپ متن، تاریخ بسامدگیری و نام پژوهشگر بررسی‌کنندهٔ متن است. هر واژه نیز در نرم‌افزار دارای ویژگی‌های پسوند، پیشوند، ریشهٔ واژه، مقولهٔ واژه، شمارهٔ متن (متن‌ها فقط به منظور تشخیص اینکه واژه مربوط به کدام متن است، شماره‌گذاری شده‌اند)، فراوانی واژه و اصل واژه (واژهٔ به‌کاررفته در متن) است. در پایان پژوهش، به طور کلی، ۲۴۰۱ متن با یک میلیون و ۲۰۳ هزار و ۵۹۸ واژهٔ منحصربه‌فرد در پایگاه داده‌ها ثبت شد.

۲. ۲. برچسب‌گذاری واژه‌ها

هر فرد واژه‌های متن‌های خود را به ترتیبی که نرم‌افزار مشخص کرده‌بود، برچسب‌گذاری و اصلاح کرده تا به صورت نهایی در پایگاه داده‌ها ثبت شود. برچسب‌گذاری شامل مشخص‌کردن مقولهٔ واژه (اسم، صفت، فعل یا حرف) و ریشهٔ واژه بود. برخی قانون‌های برچسب‌گذاری واژه‌ها عبارت بودند از:

الف) واژه‌ها به همراه وندهای اشتقاقی و با حذف وندهای تصریفی و حذف واژه‌بست‌ها ثبت می‌شدند (برای نمونه، «کارهایش» و «کاری» (در «کاری را که باید انجام می‌داد...») با نام «کار» ثبت می‌شدند اما «کاردان» و «کارمند» جداگانه ثبت می‌شدند).

ب) فعل‌ها به شکل مصدر ثبت می‌شدند.

پ) فعل‌های مرکب به شکل مصدر ثبت می‌شدند و بخش‌های فعلی و غیرفعلی آن‌ها جدا نمی‌شدند. (معیار تشخیص فعل مرکب بر اساس مقالهٔ دبیرمقدم (Dabirmoghaddam, 1997) انجام شده‌است).

۳. یافته‌ها

۳. ۱. به‌دست‌آمدن واژه‌ها و ویرایش آن‌ها

پس از پایان مرحلهٔ پیشین، واژه‌های متن‌ها (شامل ۱۲۰۳۵۹۸ واژه) به همراه ریشهٔ هر واژه به دست آمد. سپس اقدام‌هایی انجام شد؛ نخست، این واژه‌ها در جدولی به نام «بسامد ریشهٔ واژه‌ها به همراه اصل واژه» قرار داده شدند. این جدول شامل ۴۹۵۸۲ ردیف بود و در هر ردیف اصل واژه (مانند «خودشان»)، ریشهٔ واژه (مانند «خود») و فراوانی واژه (مانند «۵۸۳») وجود داشت. دوم، برچسب‌های هر واژه، بازنگری و در صورت لزوم اصلاح شد (برای نمونه، اگر «تا» (در «از تهران تا اصفهان») برچسب «اسم» خورده بود، اصلاح شده و به «حرف» تغییر یافت). سوم، ریشهٔ هر واژه، بازنگری و در صورت لزوم اصلاح شد (اشتباه در جداسازی وندهای اشتقاقی و ثبت «کارمند» در زیرمجموعهٔ «کار»، اشتباه در به دست آوردن مصدر فعل وموارد مشابه). چهارم، با توجه به اصلاحات انجام‌شده، جدول «فراوانی واژه‌ها» شامل واژه‌های پربسامد (واژه‌هایی با فراوانی بالاتر از ۵۰) به دست آمد. پنجم، اصلاح‌های نهایی از سوی مدیر گروه، بر روی این جدول اعمال شد (شامل اصلاح واژه‌ها، برچسب‌ها و توضیح لازم برای ابهام‌زدایی از برخی واژه‌ها) و جدول نهایی «واژه‌های پایه» (شامل ۲۱۵۰ واژه با فراوانی ۵۰ و بالاتر) در حوزه‌های سیاسی، اجتماعی، اقتصادی، ادبی، علمی، فرهنگی و ورزشی به دست آمد. این فهرست براساس فراوانی وقوع واژه‌ها در متن، به ترتیب از پربسامدترین واژه تا کم‌بسامدترین آن، مرتب شده و نوع هر واژه از لحاظ ساخت در مقابل آن آورده شده‌است. دویست ردیف نخست این فهرست، در جدول (۷) آمده‌است.

جدول ۱: دویست واژهٔ نخست فهرست واژه‌های پایهٔ فارسی

ردیف	واژه	فراوانی
۱	و	۶۱۰۸۱
۲	در	۴۸۷۹۵
۳	به	۳۸۳۱۲
۴	از	۳۳۴۲۳
۵	که	۲۹۴۳۰
۶	این	۲۹۳۸۰

ردیف	واژه	فراوانی
۷	بودن	۲۶۰۴۶
۸	را	۲۱۴۳۸
۹	با	۱۹۹۰۹
۱۰	آن	۱۰۴۵۴
۱۱	برای	۱۰۳۳۰
۱۲	یک	۷۹۱۳
۱۳	گفتن	۶۷۳۲
۱۴	سال	۶۴۲۴
۱۵	داشتن	۶۳۱۷
۱۶	خود	۵۹۱۹
۱۷	هم	۵۵۸۰
۱۸	کشور	۵۵۲۵
۱۹	تا	۵۴۶۰
۲۰	بر	۵۲۴۰
۲۱	شدن	۴۹۴۳
۲۲	توانستن	۴۸۰۴
۲۳	اما	۴۴۵۴
۲۴	باید	۳۹۲۷
۲۵	نیز	۳۸۳۴
۲۶	دیگری	۳۸۰۱
۲۷	ما	۳۶۳۹
۲۸	دو	۳۵۲۰
۲۹	تیم	۳۳۷۶
۳۰	هر	۳۲۵۳
۳۱	او	۳۰۴۰
۳۲	روز	۲۸۸۱
۳۳	کار	۲۸۰۱
۳۴	وی	۲۷۸۹
۳۵	یا	۲۷۴۱
۳۶	دولت	۲۷۱۲
۳۷	بیش	۲۶۸۵
۳۸	فیلم	۲۵۹۹
۳۹	فرهنگ	۲۱۶۷
۴۰	کتاب	۲۱۴۰
۴۱	بخش	۲۱۲۳
۴۲	اثر	۲۱۲۰
۴۳	درصد	۲۰۸۷
۴۴	مردم	۲۰۷۹
۴۵	اگر	۲۰۵۷

ردیف	واژه	فراوانی
۴۶	بسیار	۲۰۵۳
۴۷	حال	۲۰۱۳
۴۸	من	۱۹۸۰
۴۹	همه	۱۹۴۸
۵۰	گذشته	۱۹۳۶
۵۱	وجودداشتن	۱۹۰۰
۵۲	سینما	۱۸۶۹
۵۳	روی	۱۸۳۸
۵۴	گزارش	۱۸۱۵
۵۵	سازمان	۱۸۱۳
۵۶	درباره	۱۸۰۹
۵۷	ملی	۱۷۹۸
۵۸	رسیدن	۱۷۶۷
۵۹	زمان	۱۷۶۶
۶۰	عنوان	۱۷۶۲
۶۱	پس	۱۷۶۱
۶۲	فرد	۱۷۴۰
۶۳	مورد	۱۷۴۰
۶۴	برنامه	۱۷۲۷
۶۵	برخی	۱۶۷۷
۶۶	بازی	۱۶۶۱
۶۷	اعلام‌کردن	۱۶۵۷
۶۸	قیمت	۱۶۴۸
۶۹	ماه	۱۶۴۸
۷۰	انجام‌دادن/شدن	۱۶۱۵
۷۱	همین	۱۵۹۰
۷۲	حضور	۱۵۴۸
۷۳	چند	۱۵۳۷
۷۴	افزودن	۱۵۰۷
۷۵	بعد	۱۵۰۴
۷۶	خواستن	۱۴۹۶
۷۷	افزایش	۱۴۹۵
۷۸	موضوع	۱۴۷۱
۷۹	گروه	۱۴۶۶
۸۰	جدید	۱۴۶۴
۸۱	پیش	۱۴۵۳
۸۲	اول	۱۴۴۷
۸۳	دوره	۱۴۴۲
۸۴	استفاده‌کردن	۱۴۲۲

ردیف	واژه	فراوانی
۸۵	بزرگ	۱۴۰۲
۸۶	همچنین	۱۳۹۳
۸۷	جا	۱۳۷۸
۸۸	دلیل	۱۳۷۸
۸۹	توجه‌داشتن/ کردن	۱۳۷۶
۹۰	دانستن	۱۳۶۸
۹۱	مشکل	۱۳۶۲
۹۲	شهر	۱۳۳۳
۹۳	بین	۱۳۰۶
۹۴	چه	۱۳۰۵
۹۵	شرط	۱۳۰۵
۹۶	مهم	۱۲۹۶
۹۷	رئیس‌جمهوری	۱۲۹۴
۹۸	بانک	۱۲۷۵
۹۹	مرکز	۱۲۶۹
۱۰۰	برگزارکردن	۱۲۴۲
۱۰۱	تولید	۱۲۳۹
۱۰۲	بازار	۱۲۳۱
۱۰۳	البته	۱۲۱۲
۱۰۴	اجرا	۱۲۱۱
۱۰۵	سوی	۱۲۰۰
۱۰۶	شرکت	۱۱۹۷
۱۰۷	طرح	۱۱۹۵
۱۰۸	خوب	۱۱۹۳
۱۰۹	هیچ	۱۱۶۶
۱۱۰	کم	۱۱۵۴
۱۱۱	مختلف	۱۱۴۷
۱۱۲	تأکیدکردن/ داشتن	۱۱۴۴
۱۱۳	وقت	۱۱۲۰
۱۱۴	صورت	۱۱۱۶
۱۱۵	فوتبال	۱۱۱۲
۱۱۶	قانون	۱۱۱۲
۱۱۷	بار	۱۱۱۰
۱۱۸	میلیون	۱۱۰۸
۱۱۹	ایرانی	۱۰۹۶
۱۲۰	جهان	۱۰۹۳
۱۲۱	رئیس	۱۰۸۶
۱۲۲	جشنواره	۱۰۷۸
۱۲۳	زیاد	۱۰۷۴

ردیف	واژه	فراوانی
۱۲۴	دَه	۱۰۷۰
۱۲۵	نخست	۱۰۶۸
۱۲۶	تنها	۱۰۶۷
۱۲۷	تمام	۱۰۶۶
۱۲۸	نظر	۱۰۶۳
۱۲۹	حتی	۱۰۳۸
۱۳۰	جامعه	۱۰۲۸
۱۳۱	مسئله	۱۰۲۷
۱۳۲	هفته	۱۰۲۱
۱۳۳	زندگی	۱۰۰۸
۱۳۴	مدیر	۱۰۰۸
۱۳۵	شعر	۱۰۰۴
۱۳۶	خودرو	۹۸۸
۱۳۷	گرفتن	۹۸۴
۱۳۸	سه	۹۸۲
۱۳۹	دست	۹۸۱
۱۴۰	تومان	۹۷۹
۱۴۱	مسئول	۹۷۳
۱۴۲	حوزه	۹۶۶
۱۴۳	منطقه	۹۶۰
۱۴۴	نوع	۹۵۹
۱۴۵	اقتصادی	۹۵۷
۱۴۶	وزیر	۹۵۷
۱۴۷	بازیکن	۹۵۲
۱۴۸	نسبت	۹۵۲
۱۴۹	خبر	۹۴۸
۱۵۰	رفتن	۹۴۸
۱۵۱	نقش	۹۴۵
۱۵۲	مسابقه	۹۳۷
۱۵۳	نام	۹۳۷
۱۵۴	دیدار	۹۳۴
۱۵۵	قرارگرفتن	۹۳۴
۱۵۶	آینده	۹۳۳
۱۵۷	امروز	۹۲۸
۱۵۸	واردکردن	۹۲۶
۱۵۹	پایان	۹۲۰
۱۶۰	جهانی	۹۱۷
۱۶۱	زمینه	۹۱۴
۱۶۲	میان	۹۱۳

ردیف	واژه	فراوانی
۱۶۳	گفت‌وگو	۹۰۸
۱۶۴	انتخاب	۹۰۱
۱۶۵	نرخ	۸۹۹
۱۶۶	نشان‌دادن	۸۹۷
۱۶۷	مرحله	۸۹۵
۱۶۸	دیدن	۸۸۵
۱۶۹	راه	۸۸۴
۱۷۰	نوشتن	۸۸۲
۱۷۱	نُه	۸۷۶
۱۷۲	جوان	۸۷۵
۱۷۳	اشاره	۸۷۴
۱۷۴	آمدن	۸۷۴
۱۷۵	دنیا	۸۷۲
۱۷۶	داستان	۸۶۳
۱۷۷	ویژه	۸۵۸
۱۷۸	هرکس/هرکسی	۸۵۶
۱۷۹	هرطور/هرطوری	۸۵۵
۱۸۰	خانه	۸۵۴
۱۸۱	نفر	۸۵۲
۱۸۲	نمایشگاه	۸۴۸
۱۸۳	مجموع	۸۳۹
۱۸۴	کاهش	۸۳۸
۱۸۵	بیان	۸۳۱
۱۸۶	اساس	۸۳۰
۱۸۷	حدود	۸۱۸
۱۸۸	کودک	۸۱۷
۱۸۹	حق	۸۱۵
۱۹۰	عضو	۸۱۵
۱۹۱	چون (برای بیان علت)	۸۱۲
۱۹۲	مذاکره	۸۱۰
۱۹۳	نتیجه	۸۰۹
۱۹۴	وزارت	۸۰۴
۱۹۵	شما	۸۰۳
۱۹۶	درخصوص	۸۰۰
۱۹۷	معاون	۷۹۹
۱۹۸	اصلی	۷۹۶
۱۹۹	بالا	۷۸۸
۲۰۰	هدف	۷۸۵

همچنین ۵۰ اسم، ۵۰ حرف، ۵۰ صفت و ۲۰ قید پایهٔ زبان فارسی به ترتیب در جدول‌های ۲ تا ۶ ارائه شده‌است. روشن است که در این جا منظور، مقولهٔ واژه است و نه نقش آن در جمله. در مواردی که واژه‌ای در نقش و مقولهٔ یکسان ظاهر شده (مانند «این» و «آن» که گاهی صفت اشاره و گاهی ضمیر است) در بیان نوع مقوله، هر دو مقوله آمده‌است. نکتهٔ دیگر اینکه ضمایر در این جدول‌ها جزء اسم‌ها به شمار آمده‌اند و گونه‌های مختلف هر یک از مقوله‌ها، نیز در یک مقولهٔ کلان گروه‌بندی شده‌اند. برای نمونه، انواع صفت در مقولهٔ صفت آورده شده‌اند.

جدول ۲: پنجاه اسم پایهٔ فارسی

ردیف	واژه	نوع واژه	فراوانی
۱	این	صفت/اسم	۲۹۳۸۰
۲	آن	صفت/اسم	۱۰۴۵۴
۳	یک	اسم	۷۹۱۳
۴	سال	اسم	۶۴۲۴
۵	خود	اسم	۵۹۱۹
۶	کشور	اسم	۵۵۲۵
۷	ما	اسم	۳۶۳۹
۸	تیم	اسم	۳۳۷۶
۹	او	اسم	۳۰۴۰
۱۰	روز	اسم	۲۸۸۱
۱۱	کار	اسم	۲۸۰۱
۱۲	وی	اسم	۲۷۸۹
۱۳	دولت	اسم	۲۷۱۲
۱۴	بیش	اسم	۲۶۸۵
۱۵	فیلم	اسم	۲۵۹۹
۱۶	فرهنگ	اسم	۲۱۶۷
۱۷	کتاب	اسم	۲۱۴۰
۱۸	بخش	اسم	۲۱۲۳
۱۹	اثر	اسم	۲۱۲۰
۲۰	درصد	اسم	۲۰۸۷
۲۱	مردم	اسم	۲۰۷۹
۲۲	حال	اسم	۲۰۱۳
۲۳	من	اسم	۱۹۸۰
۲۴	همه	اسم	۱۹۴۸
۲۵	سینما	اسم	۱۸۶۹
۲۶	گزارش	اسم	۱۸۱۵
۲۷	سازمان	اسم	۱۸۱۳
۲۸	زمان	اسم	۱۷۶۶

ردیف	واژه	نوع واژه	فراوانی
۲۹	عنوان	اسم	۱۷۶۲
۳۰	فرد	اسم	۱۷۴۰
۳۱	مورد	اسم	۱۷۴۰
۳۲	برنامه	اسم	۱۷۲۷
۳۳	بازی	اسم	۱۶۶۱
۳۴	قیمت	اسم	۱۶۴۸
۳۵	ماه	اسم	۱۶۴۸
۳۶	همین	صفت/اسم	۱۵۹۰
۳۷	حضور	اسم	۱۵۴۸
۳۸	افزایش	اسم	۱۴۹۵
۳۹	موضوع	اسم	۱۴۷۱
۴۰	گروه	اسم	۱۴۶۶
۴۱	دوره	اسم	۱۴۴۲
۴۲	جا	اسم	۱۳۷۸
۴۳	دلیل	اسم	۱۳۷۸
۴۴	مشکل	اسم	۱۳۶۲
۴۵	شهر	اسم	۱۳۳۳
۴۶	شرط	اسم	۱۳۰۵
۴۷	رئیس‌جمهوری	اسم	۱۲۹۴
۴۸	بانک	اسم	۱۲۷۵
۴۹	مرکز	اسم	۱۲۶۹
۵۰	تولید	اسم	۱۲۳۹

جدول ۳: پنجاه حرف پایهٔ فارسی

ردیف	واژه	نوع واژه	فراوانی
۱	و	حرف	۶۱۰۸۱
۲	در	حرف	۴۸۷۹۵
۳	به	حرف	۳۸۳۱۲
۴	از	حرف	۳۳۴۲۳
۵	که	حرف	۲۹۴۳۰
۶	را	حرف	۲۱۴۳۸
۷	با	حرف	۱۹۹۰۹
۸	برای	حرف	۱۰۳۳۰
۹	هم	حرف	۵۵۸۰
۱۰	تا	حرف	۵۴۶۰
۱۱	بر	حرف	۵۲۴۰
۱۲	اما	حرف	۴۴۵۴
۱۳	نیز	حرف	۳۸۳۴

ردیف	واژه	نوع واژه	فراوانی
۱۴	هر	حرف	۳۲۵۳
۱۵	یا	حرف	۲۷۴۱
۱۶	اگر	حرف	۲۰۵۷
۱۷	رویِ	حرف	۱۸۳۸
۱۸	درباره	حرف	۱۸۰۹
۱۹	پس	حرف	۱۷۶۱
۲۰	بعد	حرف	۱۵۰۴
۲۱	همچنین	حرف	۱۳۹۳
۲۲	بین	حرف	۱۳۰۶
۲۳	البته	حرف	۱۲۱۲
۲۴	سویِ	حرف	۱۲۰۰
۲۵	حتیِ	حرف	۱۰۳۸
۲۶	چون (برای بیان علت)	حرف	۸۱۲
۲۷	چرا	حرف/ادات پرسش	۷۷۱
۲۸	فقط	حرف	۷۶۸
۲۹	مانند	حرف	۷۴۷
۳۰	درواقع	حرف/قید	۷۳۲
۳۱	ولی	حرف	۷۲۳
۳۲	در مقابل	حرف	۶۸۶
۳۳	بدون	حرف	۶۷۶
۳۴	توسط	حرف	۶۷۰
۳۵	بنابراین	حرف	۵۷۹
۳۶	طی	حرف	۵۳۲
۳۷	اکنون	حرف/قید	۴۹۰
۳۸	بلکه	حرف	۴۹۰
۳۹	از طریق	حرف	۴۸۱
۴۰	گاهی	حرف/قید	۴۵۳
۴۱	تاکنون	حرف/قید	۴۴۹
۴۲	آیا	حرف	۴۰۸
۴۳	خارج‌از	حرف	۳۸۲
۴۴	زیرا	حرف	۳۷۹
۴۵	درضمنِ	حرف/قید	۳۷۳
۴۶	مثل	حرف	۳۶۹
۴۷	علاوه‌براین	حرف	۳۶۱
۴۸	حالا	حرف/قید	۳۵۴
۴۹	همچنان	حرف	۳۴۵
۵۰	آخر	صفت/حرف	۳۲۸

جدول ۴: پنجاه صفت پایهٔ فارسی

ردیف	واژه	نوع واژه	فراوانی
۱	این	صفت/اسم	۲۹۳۸۰
۲	آن	صفت/اسم	۱۰۴۵۴
۳	دیگری	صفت	۳۸۰۱
۴	دو	صفت	۳۵۲۰
۵	بسیار	صفت	۲۰۵۳
۶	گذشته	صفت	۱۹۳۶
۷	ملی	صفت	۱۷۹۸
۸	برخی	صفت	۱۶۷۷
۹	همین	صفت/اسم	۱۵۹۰
۱۰	چند	صفت	۱۵۳۷
۱۱	جدید	صفت	۱۴۶۴
۱۲	پیش	صفت	۱۴۵۳
۱۳	اول	صفت	۱۴۴۷
۱۴	بزرگ	صفت	۱۴۰۲
۱۵	چه	صفت	۱۳۰۵
۱۶	مهم	صفت	۱۲۹۶
۱۷	خوب	صفت	۱۱۹۳
۱۸	هیچ	صفت	۱۱۶۶
۱۹	کم	صفت	۱۱۵۴
۲۰	مختلف	صفت	۱۱۴۷
۲۱	ایرانی	صفت	۱۰۹۶
۲۲	زیاد	صفت	۱۰۷۴
۲۳	دَه	صفت	۱۰۷۰
۲۴	نخست	صفت	۱۰۶۸
۲۵	تنها	صفت	۱۰۶۷
۲۶	تمام	صفت	۱۰۶۶
۲۷	سه	صفت	۹۸۲
۲۸	اقتصادی	صفت	۹۵۷
۲۹	جهانی	صفت	۹۱۷
۳۰	نُه	صفت	۸۷۶
۳۱	جوان	صفت	۸۷۵
۳۲	ویژه	صفت	۸۵۸
۳۳	اصلی	صفت	۷۹۶
۳۴	بالا	صفت	۷۸۸
۳۵	قبل	صفت	۷۸۱
۳۶	چنین	اسم/صفت	۷۷۰
۳۷	بین‌المللی	صفت	۷۶۲

ردیف	واژه	نوع واژه	فراوانی
۳۸	اسلامی	صفت	۷۵۴
۳۹	اجتماعی	صفت	۷۲۸
۴۰	اخیر	صفت	۷۲۱
۴۱	همان	صفت/اسم	۷۱۹
۴۲	هنوز	صفت/قید	۷۱۷
۴۳	برابر	صفت	۶۹۵
۴۴	خیلی	صفت	۶۶۹
۴۵	کارشناس	صفت	۶۱۴
۴۶	مناسب	صفت	۵۹۳
۴۷	خارجی	صفت	۵۵۹
۴۸	سیاسی	صفت	۵۵۴
۴۹	شاعر	اسم/صفت	۵۴۳
۵۰	خاص	صفت	۵۳۴

جدول ۵: پنجاه فعل پایهٔ فارسی

ردیف	واژه	نوع واژه	فراوانی
۱	بودن	فعل	۲۶۰۴۶
۲	گفتن	فعل	۶۷۳۲
۳	داشتن	فعل	۶۳۱۷
۴	شدن	فعل	۴۹۴۳
۵	توانستن	فعل	۴۸۰۴
۶	وجودداشتن	فعل	۱۹۰۰
۷	رسیدن	فعل	۱۷۶۷
۸	اعلام کردن	فعل	۱۶۵۷
۹	انجام‌دادن/شدن	فعل	۱۶۱۵
۱۰	افزودن	فعل	۱۵۰۷
۱۱	خواستن	فعل	۱۴۹۶
۱۲	استفاده کردن	فعل	۱۴۲۲
۱۳	توجه‌داشتن/ کردن	فعل	۱۳۷۶
۱۴	دانستن	فعل	۱۳۶۸
۱۵	برگزار کردن	فعل	۱۲۴۲
۱۶	تأکید کردن/ داشتن	فعل	۱۱۴۴
۱۷	گرفتن	فعل	۹۸۴
۱۸	رفتن	فعل	۹۴۸
۱۹	قرار گرفتن	فعل	۹۳۴
۲۰	وارد کردن	فعل	۹۲۶
۲۱	نشان‌دادن	فعل	۸۹۷
۲۲	دیدن	فعل	۸۸۵

ردیف	واژه	نوع واژه	فراوانی
۲۳	نوشتن	فعل	۸۸۲
۲۴	آمدن	فعل	۸۷۴
۲۵	اضافه‌کردن	فعل	۷۷۶
۲۶	بررسی‌کردن	فعل	۷۷۵
۲۷	تغییردادن	فعل	۷۳۲
۲۸	فروختن	فعل	۶۶۸
۲۹	آغازکردن	فعل	۶۱۸
۳۰	دادن	فعل	۶۰۸
۳۱	ادامه‌دادن	فعل	۶۰۷
۳۲	اتفاق‌افتادن	فعل	۵۶۷
۳۳	خواندن	فعل	۵۵۲
۳۴	قراردادن	فعل	۵۵۱
۳۵	خبردادن	فعل	۵۳۶
۳۶	منتشرکردن	فعل	۵۱۷
۳۷	گذشتن	فعل	۵۰۴
۳۸	دنبال‌کردن	فعل	۴۹۸
۳۹	حرکت‌کردن	فعل	۴۷۹
۴۰	گذاشتن	فعل	۴۵۳
۴۱	صورت‌گرفتن	فعل	۴۵۲
۴۲	پرداختن	فعل	۴۵۰
۴۳	زدن	فعل	۴۴۳
۴۴	صحبت‌کردن	فعل	۴۴۰
۴۵	معتقدبودن	فعل	۴۳۴
۴۶	ایجادکردن	فعل	۴۲۶
۴۷	مطرح‌کردن	فعل	۴۲۳
۴۸	حضورداشتن	فعل	۴۱۷
۴۹	قراربودن	فعل	۴۱۰
۵۰	تشکیل‌دادن/شدن	فعل	۴۰۶

جدول ۶: پنجاه قید پایهٔ فارسی

ردیف	واژه	نوع واژه	فراوانی
۱	باید	قید	۳۹۲۷
۲	درواقع	حرف/قید	۷۳۲
۳	هنوز	صفت/قید	۷۱۷
۴	شاید	قید	۵۷۴
۵	اکنون	حرف/قید	۴۹۰
۶	گاهی	حرف/قید	۴۵۳
۷	تاکنون	حرف/قید	۴۴۹

ردیف	واژه	نوع واژه	فراوانی
۸	همیشه	صفت/قید	۴۲۰
۹	درضمنِ	حرف/قید	۳۷۳
۱۰	حالا	حرف/قید	۳۵۴
۱۱	همزمان	قید/صفت	۳۱۰
۱۲	دوباره	صفت/قید	۲۶۹
۱۳	معمولاً	صفت/قید	۲۶۹
۱۴	درنهایت/نهایتاً	صفت/قید	۲۴۱
۱۵	همواره	صفت/قید	۲۱۶
۱۶	اغلب	صفت/قید	۱۸۸
۱۷	مثلاً	صفت/قید	۱۸۷
۱۸	سالانه	صفت/قید	۱۸۶
۱۹	تقریباً	قید/صفت	۱۷۸
۲۰	هم‌اکنون	صفت/قید	۱۵۷

سپس واژه‌ها بر اساس نوع متن، جداسازی شده و جدول‌های «فراوانی واژه‌ها بر اساس نوع متن» به دست آمد. در پایان واژه‌های مشترک بین همهٔ گونه‌های متن به دست آمده و از جدول‌های «فراوانی واژه‌ها بر اساس نوع متن» حذف شد تا جدول‌های «واژه‌های پایهٔ اختصاصی هر موضوع» (شامل ۵۰۰ واژهٔ اختصاصی هر موضوع با فراوانی ۱۲ یا بالاتر) به دست آید.

۴. بحث و نتیجه‌گیری

در این پژوهش، ۲۱۵۰ واژهٔ پایهٔ زبان فارسی از متن‌های مطبوعاتی (از روزنامه‌های پرشمارگان ایران) به دست آمد. همچنین علاوه بر به‌دست‌آمدن فهرست واژه‌های پایهٔ زبان فارسی به شکل عمومی و تخصصی، پیکره‌ای با یک میلیون و ۲۰۳ هزار و ۵۹۸ واژه (۲۴۰۱ متن) نیز در دسترس قرار گرفت. از این دو یافته، می‌توان در پژوهش‌های گوناگونی بهره برد. به منظور بهره‌وری هرچه بیشتر در زمینهٔ فهرست واژه‌های پایهٔ زبان فارسی، پیشنهادهای زیر در این زمینه ارائه می‌شود:

الف) ساخت سامانه‌ای که در آن بتوان با جستجوی یک واژه از درجهٔ آن در فهرست‌های مختلف بسامدی و مقدار فراوانی آن واژه آگاه شد، می‌تواند برای انواع پژوهش‌های زبانی، به ویژه در آموزش زبان فارسی به غیرفارسی‌زبانان بسیار مفید باشد. جستجوی واژه‌ها و یافتن ریشهٔ واژه‌ها (مانند جستجوی یک فعل و یافتن مصدر آن) نیز می‌تواند این سامانه را قوی‌تر کند.

ب) ساخت سامانه‌ای که در آن بتوان با وارد کردن یک متن فارسی از میزان پایه بودن واژه‌های آن آگاه شد، می‌تواند برای اهداف آموزشی مفید باشد.

پ) متن‌های سطح‌بندی‌شدهٔ موجود را که با نام گروه‌های سنی الف، ب، ج و د در بازار کتاب موجود است، می‌توان بر اساس یافته‌های پژوهش حاضر، مورد نقد و بررسی قرار داد. ارزیابی‌های

اولیه نشان می‌دهد، بسیاری از این متون، پشتوانهٔ پیکره‌ای ندارند. روشن است که سطح‌بندی بدون چنین پشتوانه‌ای از اعتبار قابل دفاعی برخوردار نیست.

فهرست منابع

ایمن، لیلی (۱۳۵۷). *فهرست پربسامدترین واژه‌های خردسالان ۶ تا ۱۲ سال*. تهران: کمیتهٔ ملی پیکار جهانی با بی‌سوادی.

بدره‌ای، فریدن (۱۳۵۰). *واژگان نوشتاری کودکان دبستانی ایران*. تهران: فرهنگستان زبان ایران.

براهنی، محمد تقی (۱۳۵۴). *بررسی میزان فراوانی واژه‌ها و تأثیر آن بر خواندن*. تهران: پژوهشکدهٔ تعلیم و تربیت.

بی‌جن‌خان، محمود (۱۳۹۰). *فرهنگ بسامدی بر اساس پیکرهٔ متنی زبان فارسی امروز*. تهران: مؤسسه انتشارات دانشگاه تهران.

تحریریان، محمد حسن (۱۳۷۳). «نگاهی به فراوانی واژه‌ها و واژگان پایه در فارسی». *مجموعه مقالات دانشگاه علامه طباطبائی*. به کوشش سید علی میرعمادی. تهران: دانشگاه علامه طباطبایی. شمارهٔ ۸۰. صص ۴۷-۷۰.

حسنی، حمید (۱۳۸۴). *واژه‌های پرکاربردِ فارسیِ امروز بر مبنای پیکرهٔ یک‌میلیون لغتی شاملِ بیش از ۸۰۰۰ لغتِ قاموسی و غیرِقاموسی*. تهران: کانون زبان ایران.

دبیرمقدم، محمد (۱۳۷۶). «فعل مرکب در زبان فارسی». *زبان‌شناسی*. سال . شمارهٔ ۱. صص ۲-۴۶.

درودی احسان، هما برادران‌هاشمی، ابوالفضل آل‌احمد، امیر حسین حبیبیان، محمد علی زارع، فرزاد مهدی‌خانی، آزاده شاکری و مسعود رهگذر (۱۳۸۷). *محک استاندارد برای تحقیقات بازیابی اطلاعات وب فارسی*، تهران: گزارش فنی گروه تحقیقاتی پایگاه داده ها دانشگاه تهران، شماره: DBRG-TR-138702.

صحرایی، ر ضا مراد و شهناز احمدی‌قادر (۱۳۹۴). «آموزش مستقیم واژه در متن: مقایسهٔ تأثیر دو رویکرد یادگیری مستقیم و تصادفی در یادگیری واژه». *زبان‌پژوهی*. دورهٔ ۷. شمارهٔ ۱۷. صص ۹۷-۱۰۰.

صحرایی، رضا مراد (۱۳۹۱). «انگارهٔ زایشی فراگیری زبان». *زبان پژوهی*. دورهٔ ۴. شمارهٔ۷. صص ۱۶۶-۱۶۹.

صفارپور، عبدالرحمن (۱۳۷۰). *واژگان پایهٔ کلاس‌های اول تا پنجم دبستان*. تهران: انتشارات کمک آموزشی وزارت آموزش و پرورش.

غروی قوچانی، مهدی (۱۳۸۵). *تعیین واژگان پایهٔ فارسی معیار گفتاری در بزرگسالان و تجزیه و تحلیل تواتر آن‌ها بر اساس قانون زیف* (پایان‌نامهٔ کارشناسی ارشد چاپ‌نشده). دانشگاه پیام نور تهران.

نعمت‌زاده، شهین، محمد دادرس، مهدی دستجردی‌کاظمی و محرم منصوری‌زاده (۱۳۹۰). *واژه‌های پایه فارسی از زبان کودکان ایرانی*. تهران: موسسه فرهنگی مدرسه برهان (انتشارات مدرسه).

References

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

Badreee, F. (1971). *Writing vocabulary of Iranian elementary school children.* Tehran: Academy of Iranian Language [In Persian].

Barnett, B., Lehmann, H., & Zoeppritz, M. (1986). A word database for natural language processing. *Proceedings of the 11th International Conference on Computational Linguistics COLING86* (pp. 435-440). Bonn, West Germany, Agust 1986.

Bijan Khan, M. (1994). A look at the abundance of basic words and vocabulary in Farsi. In S. A. Miremadi (Ed.), *Proceedings of Allameh Tabataba'i University* (vol. 80. pp. 47-70). Tehran: Allameh Tabataba'i University [In Persian].

Bijnkhan, M. (2011). *Frequency dictionary based on the textual corpus of Contemporary Persian language.* Tehran: Tehran University Press [In Persian].

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book.* Boston: Houghton Mifflin.

Carter, R. (2002). Vocabulary. In R. Carter & D. Nunan (Eds.), *The Cambridge guide to teaching English to speakers of other languages* (pp. 42-47). Cambridge: Cambridge University Press.

Cook, V. (1991). *Second language learning and language teaching*. London: Edward Arnold.

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

Dabirmoghaddam, M. (1997). Compound verb in Persian language. *Journal of Linguistics, 12*(1), pp. 2-46 [In Persian].

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). A method of semantic description. In D. D. Steinberg & L. A. Jakobovits. *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology* (pp. 436-471). Cambridge: Cambridge University Press.

Dolch, E. W. (1936). A Basic sight vocabulary. *Elementary School Journal*, 36, 456-460.

Doroodi, A., BaradaranHashemi, H., AleAhmad, A., ZareBidaki, A., Habibian, A., Mehdikhani, F., Shakeri A. & Rahgozar M. (2009). *dorIR collection for Persian web retrieval.* Tehran: Technical Report of the Databases Research Group, University of Tehran, No. DBRG-TR-138702[In Persian].

Fry, E. B., Kress, J. E., & Fountoukidis, D. L. (2000). *The reading teachers book of lists* (4nd ed). London: Pearson PTR.

Gourori Gouchani, M. (2006). *Determination of basic Persian words speech standard in adults and analysis of their continuity based on Ziff law* (Unpublished Master's thesis), Payam-e Nour University, Tehran, Iran [In Persian].

Hasani, H. (2005). *The most widely used words in contemporary Persian based on a one-million-word corpus of more than 8,000 lexical and non-lexical words*. Tehran: Iran Language Institute [In Persian].

Imen, L. (1978). *The most frequent words list for 6 to 12 children years old*. Tehran: National Committee of the World Campaign on Illiteracy [In Persian].

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 cannot 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.

Meara, P. (1980). Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching. Language Teaching, 13(*3-4), 221-246.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59–82.

Nematzadeh, S., Dadras, M., DastjerdiKazemi, M., & Mansourizadeh, M. (2011). *Basic Persian words from Iranian childre.* Tehran: Borhan School of Cultural Education (Madrese Publishing) [In Persian].

Osgood, C. E. (1976). *Focus on meaning: explorations in semantic space*. New York: Mouton.

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

Safarpour, A. (1991). *Basic words of first to fifth grade.* Tehran: Educational assistance Publications, Ministry of Education [In Persian].

Sahraee, R. (2012). The generative model for language acquisition rethinking the nature of core grammar and its representation in child language. *Zabanpazhuhi*, *4*(7), 145-176 [In Persian].

Sahraee, R. M., & Ahmadiye Ghader, Sh. (2016). Direct teaching of vocabulary in context: the comparison of effect of direct and incidental teaching in learning vocabulary. *Zabanpazhuhi*, *7*(17), 97-100 [In Persian].

Shin, D. & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339-348 .

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) ,113-139.

West, M. (1953). *A general service list of English words*. London: Longman.

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

Wilkins, D. A. (1976). *Notional Syllabuses*. Oxford: Oxford University Press.

Persian Basic Words based on Newspaper Texts

Reza Morad Sahraei[1]
Amirhossein Mojiri Foroushani[2]
Morvarid Talebi[3]

Received: 11/06/2018
Accepted: 01/01/2019

Abstract

No information can be transmitted without familiarity with language words. Teaching vocabulary is one of the most important components of foreign language teaching that can affect all the four main language skills (listening, speaking, reading and writing). The first step in teaching vocabulary is to access the list of basic words. According to the studies conducted in the related area, high-frequency words as well as the basic vocabularies are significant in language teaching since they are easy to learn and frequently used in everyday conversations. Frequently-used words list or frequency dictionary is a set of words that have more repetition in a collection of texts (corpus). Basic words are generally extracted from the architecture of the corpus-based researches, and the output of each linguistic corpus can be a basic words list (depending on the language and type of the corpus texts).

From 1897 there are many lists of basic words in different languages of the world. English language researches is more than any other researches in this field. Since the year 1971, researches have also been conducted in Persian to extract frequent words.

Thorndike (1921) presented 30,000 basic English words. In 1923, Ogden and Richards listed 850 basic words of English. Dolch (1936) listed 220 and West (1953) presented 2,000 basic English words. Coxhead (2000) has derived basic words in four areas of art, commerce, law and science. In 2001, Verlinde and Selva provided a list of frequent words in French. Also, 100 and 1000 English basic words were extracted by Fry et al. (2000). Jones and Tschirner (2006) extracted 4,307 frequent German words. Davies and Gardner (2010) extracted 1,000 to 5,000 basic English words. The list of 100 frequent words of Oxford English dictionary and the 3,000 basic words of Langman's dictionary are other instances.

In Iran, Barahani (1975), Imen (1978), Safarpour (1991), Tahriryan (1994), Hasani (2005), Gharavi Qouchani (2006), Doroody et al. (2008), Alahmad et al.

[1] PhD in Linguistics, Associate Professor at Department of Linguistics, Allameh Tabataba'i University, (Corresponding Author); sahraei@atu.ac.ir
[2] MA in Teaching Persian to Foreign Language Learners, Researcher at Sa'di Foundation; amojiry@saadifoundation.ir
[3] MA in Teaching Persian to Foreign Language Learners, Allameh Tabataba'i University; morvarid_talebi@atu.ac.ir

(2009), Bijankhan (2011), Nematzadeh et al. (2011) were among the scholars of studying basic words.

This research has two main stages: a) extracting texts and registering in database: 8 persons within 100 working days, each day extracted 3 passages with an average of 500 words from three newspapers in one of the seven different areas (including culture, society, politics, sports, fiction, economics and science), resulting in a corpus with 1,203,589 words (2401 texts).

The software used for this project was written specifically for this research using the PHP programming language and is a web-based software. Types of words in the software are "name", "verb" (and in particular "compound verb"), "preposition", "proper noun", "adjective" and "adverb". Also, in this corpus, each text has metadata of "type of text" (cultural, social, political, etc.), the name of the newspaper, the date of printing, the date of the text typing, the date of frequency extraction and the name of the researcher. A list of the collections was also made to identify broken plurals. For example, "اثر: آثار". Also, a list of inflectional affix (prefixes and suffixes) and rules governing them was provided. This list specifies what suffixes or prefixes any type of word can take. For example, verb can start with the "می" prefix, the "ات"suffix can be added to the word "آیه" but before that, the letter "ه" should be deleted from the end of the word.

Each word in the software also has attributes like prefix, suffix, word root, word category, text numbers, word frequency, and main word (the word used in the text).

b) Labeling Words: Labeling involves specifying the word category (noun, adjective, verb or preposition), and the lemma. Words with derivational affixes and without inflectional affixes and clitics were recorded. The verbs were recorded as infinitive. Compound verbs were recorded in an infinitive form, and their nominal and verbal parts were not separated.

After the end of the previous stage, the words of the texts (including 1,203,598 words) were obtained along lemma of each word. After corrections such as label correction and lemma correction, the "Basic words" table (including 2,150 words with a frequency of 50 and above) was obtained. In addition, the list of 50 most frequent names, prepositions, adjectives and 20 most frequent adverbs in Persian were also obtained. Then, the specific base words of each topic (including 500 specific words of each topic with a frequency of 12 and above) were obtained.

Keywords: corpus-based research, newspaper texts corpus, teaching vocabulary, basic words, high-frequency Persian words