**پژوهش‌نامه‌ی آموزش زبان فارسی به غیر فارسی‌زبانان**
**(علمی- پژوهشی)**
**سال سوم، شماره‌ی دوم (پیاپی ۷)، تابستان ۱۳۹۳**

# نقش واژگان بسامدی در ارزیابی مهارت واژگانی فارسی‌آموزان

**محمود بی‌جن‌خان**
استاد دانشگاه تهران
**عباس نصری**
دانشجوی کارشناسی ارشد دانشگاه تهران
**شهره جلایی**
دانشیار دانشگاه علوم پزشکی تهران

## چکیده

هدف این مقاله تبیین نقش واژگان بسامدی در مطالعه‌ی مهارت واژگانی فارسی‌آموزان است. این تحقیق ناظر به الگوی توزیع فراوانی واژه‌های فارسی، ارتباط آن با مهارت واژگانی و شیوه‌ی استخراج واژگان بسامدی از پیکره فارسی است. به این منظور، از روش تحقیق پیکره‌بنیاد استفاده شده است. بر این اساس، یک پیکره‌ی فارسی رسمی و محاوره‌ای انتخاب شد و توزیع آماری واژه‌ها در سطوح سه‌گانه‌ی قانون زیف‌مندلبرات به‌عنوان معیار مطمئنی برای ارزیابی مهارت واژگانی زبان‌آموزان به دست آمد. بر این اساس، معیار نمایه‌ی فراوانی واژگانی (لاوفر و نیشن، ۱۹۹۵) برای اندازه‌گیری غنای واژگانی معرفی شده که در آن به میزان به‌کارگیری واژه‌های پربسامد توجه ویژه‌ای شده است. استخراج واژگان پربسامد از پیکره‌ی رسمی و محاوره‌ای پس از اجرای عملیات بن‌واژه‌سازی انجام شده است. چهار آزمون خی‌دو بر روی ۸۶۱ کلمه‌ی مشترک، که برحسب چهار مقوله‌ی دستوری تقسیم‌بندی شدند، نشان داد که پیکره‌ی رسمی و محاوره‌ای از نظر واژگانی به دو جامعه‌ی آماری متفاوت تعلق دارند. یافته‌های تحقیق نشان می‌دهد که می‌توان از سطوح بسامدی واژگان رسمی و محاوره‌ایِ فارسی‌زبانان برای ارزیابی مهارت واژگانی فارسی‌آموزان استفاده کرد.

**کلیدواژه‌ها**: واژگان بسامدی، مهارت واژگانی، نمایه‌ی فراوانی واژگانی، پیکره‌ی زبانی

---

تاریخ دریافت مقاله: ۱۳۹۳/۱/۱۷ تاریخ پذیرش نهایی مقاله: ۱۳۹۳/۴/۱۵
رایانشانی نویسنده‌ی مسئول (محمود بی‌جن‌خان): mbjkhan@ut.ac.ir

## ۱. مقدمه

مهارت واژگانی یکی از مهم‌ترین مهارت‌های زبان‌آموزان است که جایگاه ویژه‌ای در ارزیابی مهارت‌های اصلی یعنی خواندن، نوشتن، صحبت کردن و شنیدن دارد. مهارت واژگانی، ذیل عناوینی چون غنای واژگانی یا تنوع واژگانی مطرح می‌شود. اهل زبان در تلفظ و ترکیب کلمات با یکدیگر مجبور هستند از قواعد و محدودیت‌های واجی، ساخت‌واژی و نحوی تبعیت کنند، اما در انتقال معنا به مخاطبان خود مختار هستند با توجه به بافت موقعیت و شرایط گفتمانی، کلمات مناسب و دلخواه خود را انتخاب کنند. بنابراین، مهارت واژگانی می‌تواند نشانگر میزان خلاقیت زبان‌آموزان نیز باشد. به همین دلیل، در آزمون‌سازی توجه ویژه‌ای به ارزیابی مهارت واژگانی می‌شود. مهارت واژگانی ناظر بر به‌کارگیری دانش واژگانی توسط زبان‌آموزان است که با استفاده از مدل‌ها یا الگوریتم‌های مشخص سنجیده می‌شود. دانش واژگانی مجموعه‌ای از شاخص‌های واژگانی است که می‌توان آن را از یک پیکره‌ی جامع اهل زبان به دست آورد.

کراسلی و همکاران (۲۰۱۰) در بررسی متون انگلیسی‌آموزان به این نتیجه رسیدند که فراوانی واژه و تنوع واژگانی در زمره‌ی شاخص‌های مهم واژگانی قرار دارند. بنابراین، استخراج واژگان هر زبان برحسب بسامد و تنوع دستوری نقش بنیادی در مطالعه و ارزیابی مهارت واژگانی زبان‌آموزان دارد. سؤالی که در نگاه اول به ذهن متبادر می‌شود این است که دایره‌ی لغات یک زبان‌آموز باید چه میزان باشد تا از حد معینی از دانش واژگانی برخوردار باشد. برای پاسخ به سؤال از روش تحقیق پیکره‌بنیاد استفاده کرده و به بررسی واژگان بسامدی فارسی‌زبانان پرداخته‌ایم تا در تحقیقات بعدی از یافته‌های آن برای ارزیابی مهارت واژگانی فارسی‌آموزان استفاده کنیم.

در این پژوهش، سه سؤال بنیادی در خصوص نقش واژگان بسامدی در ارزیابی مهارت واژگانی فارسی‌آموزان مطرح شده است؛ (۱) توزیع فراوانی کلمات فارسی از چه الگویی تبعیت می‌کند و این الگو چه ارتباطی با مهارت واژگانی دارد؟ (۲) در استخراج واژگان بسامدی زبان فارسی چه ملاحظاتی باید مد نظر داشت؟ (۳) آیا تفاوت معنی‌داری بین واژگان مستخرج از پیکره‌ی متنی رسمی و محاوره‌ای وجود دارد؟ برای پاسخ به این سؤالات، از یک پیکره‌ی رسمی و یک پیکره‌ی محاوره‌ای فارسی‌زبانان استفاده شده است.

## ۲. پیشینه‌ی پژوهش

لاوفر و نیشن (۱۹۹۵) یک معیار کمّی برای اندازه‌گیری مهارت واژگانی، تحت عنوان «نمایه‌ی فراوانی واژگانی»[1] ارائه کردند که در آن به میزان به‌کارگیری واژه‌های پربسامد و واژه‌های دانشگاهی در نوشته‌های

[1]. lexical frequency profile

زبان‌آموزان توجه شده است. آنان معتقد هستند که با استفاده از این نمایه می‌توان رابطه‌ی بین دانش واژگانی و به‌کارگیری واحدهای واژگانی را در متون نوشتاری زبان‌آموزان مشخص کرد. نمایه‌ی فراوانی واژگانی، نشانگر درصد کلماتی است که یک زبان‌آموز در سطوح مختلف واژگانی به‌کار می‌برد. هر واژگان دارای دو یا چند سطح فراوانی واژگانی است. آنان چهار سطح فراوانی واژگانی برای زبان‌آموزان در نظر گرفتند؛ ۱۰۰۰ کلمه‌ی پربسامد اولِ زبان، ۱۰۰۰ کلمه‌ی پربسامد دومِ زبان، ۳۱۰۰ کلمه از کتاب‌های درسی دانشگاهی که به ۵۷۰ موضوع تعلق دارند و کلمات کم‌بسامد زبان. فرضیه‌ی آنان این بود که هر چه مهارت واژگانی زبان‌آموز کمتر باشد، احتمال به‌کارگیری واژه‌های کم‌بسامد در نوشته‌های وی کمتر می‌شود؛ در حالی که مهارت واژگانی بیشتر زبان‌آموزان، احتمال به‌کارگیری واژه‌های کم‌بسامد را در نوشته‌ها بیشتر می‌کند. بنابراین، تغییر و رشد دانش واژگانیِ زبان‌آموزان همراه با تغییر نمایه‌ی فراوانی واژگانی است.

لاوفر و نیشن (۱۹۹۹) در یک آزمون رشدِ حجم واژگانیِ زبان‌آموزان، پنج سطح فراوانی برای واژه‌ها در نظر گرفتند و به بررسی اعتبار و روایی آنها پرداختند. اشمیت (۱۹۹۹) معتقد است اعتبار آزمون واحدهای واژگانی نقش برجسته‌ای در اعتبار سازه‌ها دارد. بر این اساس، وی واحدهای واژگانی شش آزمون تافل را به ۳۰ زبان‌آموز پیش‌دانشگاهی داد و طی یک مصاحبه، میزان دانش آنان را در خصوص ویژگی‌های دستوری و معنایی واژه‌ها و همایندها[1] بررسی کرد. نتیجه‌ی تحقیق نشان داد آزمون واحدهای واژگانی کفایت لازم را برای انعکاس میزان دانش زبان‌آموزان ندارد. ورمیر (۲۰۰۰) با استفاده از نسبت نوع به قطعه به مطالعه‌ی غنای واژگانی در گفتار فی‌البداهه‌ی اهل زبان و زبان‌آموزان هلندی پرداخت. این نسبت برابر با حاصل تقسیم تعداد کلمات متفاوت در متن به کل کلمات متن است. وی عدم کفایت این نسبت را برای اندازه‌گیری مهارت واژگانی به دست آورد و به این نتیجه رسید که شاخص مؤثر غنای واژگانی باید مبتنی بر فراوانی کلمات مشکل در متون زبان‌آموزان و سطح مهارت آنان باشد. رید و چپل (۲۰۰۱) معتقد هستند که تکالیف واژگانی زبان‌آموزان باید به گونه‌ای باشد که تحت شرایط بافتی، قوه‌ی استنتاج آنان در به کارگیری مفهوم واژگانی سنجیده شود.

یارویس (۲۰۰۲) به مقایسه‌ی دقت پنج فرمول در مدل‌سازیِ نسبت نوع به قطعه برای متون نوشتاری زبان‌آموزان و اهل زبان پرداخت. اسمیت (۲۰۰۴) معتقد است جدی‌ترین نقص نمایه‌ی فراوانی واژگانی این است که واژه‌های چندکلمه‌ای مانند افعال مرکب، حروف اضافه یا ربط مرکب، اصطلاحات و امثال آن، در سطوح چهارگانه‌ی فراوانی واژگانی جایی ندارند. دوران و همکاران (۲۰۰۴) یک معیار کمّی برای اندازه‌گیری تنوع یا غنای واژگانی تحت عنوان فرمول دی[2] معرفی کردند که صورت توسعه‌یافته‌ی نسبت نوع به قطعه است. هرچه مقدار دی برای یک متن بیشتر باشد، تنوع واژگانی آن بیشتر بوده و بنابراین، خالق آن متن

1. collocations
2. type to token ratio

دارای مهارت واژگانی بیشتری خواهد بود. میرا (۲۰۰۵) اعتبار نمایه‌ی فراوانی واژگانی را به عنوان ابزاری جهت ارزیابی غنای واژگانی زیر سؤال برد. وی معتقد است شیوه‌ی پردازش متن برای استخراج سطوح فراوانی واژگانی دقیق نیست، زیرا لاوفر و نیشن (۱۹۹۵) تعریف مشخصی برای کلمه ندارند. به‌عنوان مثال، معلوم نیست رشته‌های دستورالعملی[1]، مانند نام و نام خانوادگی افراد را باید یک کلمه یا بیش از یک کلمه به حساب آورد.

یو (۲۰۰۷) معتقد است در مقیاس‌های نمره‌دهی در آزمون‌های بین‌المللی همواره یک رابطه‌ی مثبت بین تنوع واژگانی، کیفیت کلی گفتمان‌های گفتاری و نوشتاری و مهارت زبانی در میان زبان‌آموزان گزارش شده است. در این تحقیق از شاخص *دی* برای اندازه‌گیری تنوع واژگانی استفاده شده است.

گاردنر (۲۰۰۷) معتقد است که تحقیقات پیکره‌بنیادِ واژگان و پردازش رایانه‌ایِ متون تأثیر عمیقی بر آموزش زبان انگلیسی داشته داشت. وی همانند میرا (۲۰۰۵) تعریف قاعده برای تعیین کلمه و شمارش آن را یک چالش جدی برای محققان می‌داند. سه حوزه‌ی مسأله‌ساز در این خصوص عبارتند از رابطه‌ی ساخت‌واژی بین کلمات، هم‌نامی[2]، چندمعنایی[3] و عبارت‌های چندکلمه‌ای[4]. کراسلی و همکاران (۲۰۱۰) برای اندازه‌گیری غنای واژگانی به شاخص‌های معنایی وزن بیشتری دادند. آنان به بررسی شیوه‌ی طبقه‌بندی متون زبان‌آموزانی پرداخته‌اند که دارای سطوح مهارتی متفاوت هستند. اُلافلین (۲۰۱۲) نشان داد زبان‌آموزانی که کتاب‌های درسی سه سطح آموزشی را در حوزه‌ی تجارت گذرانده‌اند، در معرض تعداد کمتری از اولین ۱۵۰۰ واژه‌ی نمایه‌ی فراوانی واژگانی قرار می‌گیرند. پرنت (۲۰۱۲) تمایز بین هم‌نامی، هم‌آوایی، هم‌نگارگی و چندمعنایی را توضیح داده و با استفاده از داده‌های پیکره‌بنیاد، پربسامدترین هم‌نام‌ها را معرفی کرده است. رایج‌ترین معنی بیشتر هم‌نام‌ها در ۹۰ درصد مواقع در داده‌ها یافت شده‌اند. کراسلی و همکاران (۲۰۱۴) مهارت واژگانی زبان‌آموزان انگلیسی را تجزیه و تحلیل کردند. در این تحقیق زبان‌آموزان مبتدی، متوسط و پیشرفته ۲۴۰ متن گفتاری و ۲۴۰ متن نوشتاری تولید کردند و داوران خبره، متون را برحسب تنوع واژگانی، فراوانی واژگانی و به‌کارگیری درست همایندها نمره دادند.

از مطالعات مهارت واژگانی این نتیجه به دست می‌آید که چهار شاخص اندازه‌گیری برای کمّی‌سازی غنای واژگانی زبان‌آموزان مورد توجه محققان قرار دارد: «نسبت نوع به قطعه»، «نمایه‌ی فراوانی واژگانی»، «فرمول دی» و «مؤلفه‌های معنایی». نکته‌ی مورد توجه در این مقاله این است که در محاسبه‌ی تمامی این

---

1. formulaic sequences
2. homonymy
3. polysemy
4. multiword expressions

شاخص‌ها استخراج واژگان بسامدیِ اهل زبان و زبان‌آموزان نقش کلیدی دارد. در این مقاله، با استفاده از روش تحقیق پیکره‌بنیاد به بحث درباره‌ی شیوه‌ی استخراج واژگان بسامدی فارسی‌زبانان پرداخته‌ایم.

## ۳. روش پژوهش

روش پژوهش در این مقاله از نوع پیکره‌بنیاد[۱] است. در آموزش و آزمون‌سازی دو نوع پیکره‌ی زبانی مطرح می‌شود: «پیکره‌ی مرجع» و «پیکره‌ی زبان‌آموز»[۲]. پیکره‌ی مرجع پیکره‌ای است که از متون نوشتاری و گفتاری افراد بالغ اهل زبان اول به دست می‌آید؛ مانند «پیکره‌ی ملی زبان انگلیسی بریتانیایی»[۳]، «پیکره‌ی ملی زبان انگلیسی آمریکایی»[۴]، «پایگاه داده‌های زبان فارسی»[۵] و پیکره‌ی متنی زبان فارسی معاصر (بی-جن‌خان و همکاران، ۲۰۱۱). «پیکره‌ی زبان‌آموز» پیکره‌ای است که از متون نوشتاری و گفتاری زبان‌آموزان به دست می‌آید؛ مانند «پیکره‌ی بین‌المللی انگلیسی‌آموزان»[۶].

فن‌روی و تربلانش (۲۰۰۹: ۲۴۱) با استفاده از تعدادی ویژگی زبانی به بررسی سطح دانش زبان انگلیسیِ دانشجویان اهل تسوانا در آفریقای جنوبی پرداختند. آنان برای این منظور به مقایسه‌ی پیکره‌ی زبانی انگلیسی‌آموزان اهل تسوانا با پیکره‌ی بین‌المللی انگلیسی‌آموزان، به‌عنوان پیکره‌ی مرجع، پرداختند. هر دو پیکره شامل مجموعه‌ای از مقالات با حجم واژگانی تقریباً یکسان است. برزینا و گابلاسوا (۲۰۱۳) داده‌های چهار پیکره‌ی زبان انگلیسی با حجم بیش از ۱۲ بیلیون کلمه را جمع‌آوری کردند. سپس هم‌پوشی واژگانی بین پیکره‌ها را در اولین ۳۰۰۰ کلمه‌ی پربسامد با استفاده از روش متوسط فراوانی کاهش‌یافته به دست آوردند. نتایج آماری نشان داد که ۲۱۲۲ کلمه‌ی ثابت، یعنی ۷۰/۷ درصد کلمات را می‌توان واژگان کانونی پیکره‌ها در نظر گرفت. گاردنر و دیویس (۲۰۱۳) به معرفی یک واژگان دانشگاهی پرداخته‌اند که از یک زیرپیکره به حجم ۱۲۰ میلیون کلمه استخراج شده است. زیرپیکره متعلق به پیکره‌ی زبان انگلیسی آمریکایی معاصر با حجم ۴۲۵ میلیون کلمه است.

### ۱،۳. پیکره‌ی پژوهش

در روش پیکره‌بنیاد ویژگی‌های آماریِ ساخت‌های زبانی، مانند واژگان در این مقاله، از طریق اجرای نرم‌افزارهای رایانه‌ایِ پردازش متن بر روی حجم بسیار زیادی از متون، به عنوان پیکره، انجام می‌شود. برای

---

[۱]. برای آشنایی با «روش پیکره‌بنیاد» رجوع شود به لیتوسلتی، ۲۰۱۰ (فصل پنجم).

[2]. learner corpus

[3]. **B**ritish **N**ational **C**orpus (=**BNC**)

[4]. **A**merican **N**ational **C**orpus (=**ANC**)

[5]. http://www.pldb.ihcs.ac.ir

[6]. http://www.uclouvain.be/en-cecl-icle.html

بررسی و استخراج واژگان بسامدی فارسی‌زبانان به منظور ارزیابی مهارت واژگانی فارسی‌آموزان، از پیکره‌ی زبان فارسی معاصر استفاده شده است (بی‌جن‌خان و همکاران، ۲۰۱۱). جامعه‌ی آماری پیکره شامل متون نوشتاری زبان فارسی است که از شروع انقلاب اسلامی تاکنون به چاپ رسیده‌اند. این پیکره در حال حاضر شامل حدود ۱۰۴ میلیون کلمه یا قطعه‌ی نوشتاری است.

برای این تحقیق، یک زیرپیکره از متون محاوره‌ای و یک زیرپیکره از متون رسمی فارسی انتخاب شد. زیرپیکره‌ی محاوره‌ای شامل ۵۱۶۴۵۸۳ کلمه در ۲۵۱ پرونده‌ی متنی است. موضوع هر متن، یا داستانی است یا خاطره‌ی شخصی و از اینترنت جمع‌آوری شده است. برای اطمینان از محاوره‌ای بودن متون، شیوه‌ی نگارشِ کلمات پربسامد فرهنگ بسامدی (بی‌جن‌خان و محسنی، ۱۳۹۱: ۳۸۳)، مانند کلمات دستوریِ «و»، «در»، «بودن»، «را»، «دیگر» و کلمات واژگانیِ «گفتن»، «توانستن»، «گرفتن» و «دست»، در داخل هر متن منتخب بررسی شده و در صورت مشاهده‌ی نگارش محاوره‌ایِ این کلمات، آن متن در زمره‌ی پیکره‌ی محاوره‌ای قرار گرفته است. متون پیکره را دو ویراستار خوانده و خطاهای املایی و مرزهای نادرست کلمات را اصلاح کرده‌اند. اما زیرپیکره‌ی رسمی شامل ۵۸۵۷۸۸۶ کلمه در ۲۱۲۰ پرونده‌ی متنی است. متون پیکره موضوعات متنوع علمی، سیاسی، اقتصادی، فرهنگی و مذهبی را پوشش می‌دهد و از روزنامه‌ها و کتاب‌های درسی و آموزشی استخراج شده‌اند. چند دانشجوی زبان‌شناسی متون پیکره را خوانده، خطاهای املایی و مرزهای نادرست کلمات را اصلاح کرده و بر اساس یک دستورالعمل از قبل تعریف شده به هر کلمه یک برچسبِ سلسله‌مراتبیِ صرفی-نحوی اختصاص داده‌اند.

### ۲،۳. مدخل‌های واژگانی

این زیربخش به بحث درباره‌ی مسائلی می‌پردازد که به استخراج مدخل‌های واژگانی از پیکره مربوط است. هر واژه یا مدخل واژگانی، کلمه‌ای است که از ریشه‌یابیِ کلمات و شمارش فراوانی آنها در متون پیکره به دست می‌آید. تعریف کلمه یا تعیین مرز کلمات با یکدیگر یکی از مهم‌ترین چالش‌ها در پردازش متون فارسی است. این چالش در پردازش متون سایر زبان‌ها نیز مطرح است (میرا، ۲۰۰۵ و گاردنر، ۲۰۰۷). در متون فارسی این چالش ناشی از فراوانیِ بالای «قطعه‌های چندواحدی»[1]، «واحدهای چندقطعه‌ای»[2] و عبارت‌های دستورالعملی است (شریفی آتشگاه و بی‌جن‌خان، ۱۳۸۸).

مسأله‌ی قطعه‌های چندواحدی، مانند صورت‌های تصریفیِ اسم‌ها و فعل‌ها و کلماتِ دارای واژه‌بست را می‌توان با روش بن‌واژه‌سازی حل کرد. اما برای واحدهای چندقطعه‌ای و عبارت‌های دستورالعملی باید

---

1. multi-unit tokens
2. multi-token units

راهبرد مناسبی انتخاب کرد. «واحدهای چندقطعه‌ایِ» پربسامد در متون فارسی عبارتند از فعل مرکب، مصدر مرکب، حرف اضافه و حرف ربط مرکب. «عبارت‌های دستورالعملی» پربسامد عبارتند از:

- عبارت (غیر) قابل X (مثل عبارت قابل قبول و غیر قابل قبول) و ترکیب‌های هم‌پایه‌ایِ آن (مثل غیر قابل استناد و بررسی)
- عبارت به‌طور X (مثل به‌طور موفقیت‌آمیزی) و ترکیب‌های هم‌پایه‌ای آن
- عبارت به‌طرز X (مثل به‌طرز جالبی) و ترکیب‌های هم‌پایه‌ای آن
- ضد X (مثل ضد ولایت فقیه) و ترکیب‌های هم‌پایه‌ای آن
- اعداد مرکب مثل دو هزار و هشتاد و یک
- عبارت عدد X (مثل دویست و پنجاه ساله و بیست و پنج سالگی)

سؤال اساسی این است که آیا واحدهای چندقطعه‌ای و عبارت‌های دستورالعملی را باید مدخل واژگانی به حساب آورد یا این که کلمات سازنده‌ی آنها را باید مدخل واژگانی در نظر گرفت و در سطح بالاتر از کلمه به‌عنوان واحدهای زبانی وارد محاسبه کرد. راهبردی که در این تحقیق دنبال شده این است که عبارت‌های دستورالعملی و آن دسته از واحدهای چندقطعه‌ای که ساختار زایا دارند، به کلمات سازنده‌شان تجزیه شده‌اند و کلمات سازنده‌ی مدخل واژگانی به حساب آمده‌اند. بنابراین، (۱) فراوانی کلمات سازنده به فراوانی همان کلمات که به‌طور منفرد در متن ظاهر شده‌اند، اضافه شده است. (۲) فراوانی یک واحد چندقطعه‌ای و عبارت دستورالعملی در سطح بالاتر از کلمه، که آن را می‌توان «سطح چندکلمه‌ای‌ها»[1] نامید، محاسبه شده است.

استدلالی که برای انتخاب این راهبرد می‌توان داشت این است که اگر واحدهای چندقطعه‌ای و عبارت‌های دستورالعملی در کنار سایر کلمات بسیط، اشتقاقی و مرکب به‌عنوان مدخل‌های واژگانی در نظر گرفته شوند، آن‌گاه از میزان فراوانی کلمات سازنده‌ی آنها به میزان بسیار زیادی کاسته می‌شود. به‌عنوان مثال، اگر افعال مرکب، مدخل واژگانی باشند، از میزان فراوانی افعال همکردی که جزئی از آنها هستند، مانند «کردن»، «زدن»، «خوردن» و امثال آن، در واژگانِ بسامدی به‌طور جدی کاسته می‌شود و این خلاف شمّ زبانی اهل زبان است. علاوه بر آن، واحدهای چندقطعه‌ای و عبارت‌های دستورالعملی هر کدام یک مجموعه‌ی باز را تشکیل می‌دهند که براساس قواعد زبانی ساخته می‌شوند. بنابراین، انتظار می‌رود فراوانی کلمات سازنده‌ی آنها بسیار بیشتر از فراوانی خود آنها باشد.

[1]. n-grams

## ۴. الگوی توزیع فراوانی کلمات: غنای واژگانی

اولین سؤال تحقیق این است که توزیع فراوانی کلمات فارسی از چه الگویی تبعیت می‌کند و این الگو چه ارتباطی با مهارت واژگانی دارد. اِلیس (۲۰۰۲ و ۲۰۰۶) نقش بسامد (فراوانی) کلمات را در پردازش‌های واجی، صرفی، نحوی و خواندن و نوشتن و تأثیر آن را در آموزش زبان دوم بررسی کرده و به این نتیجه رسیده است که تولید و درک، تابع فراوانی وقوع کلمات در زبان است. وی معتقد است کلمات و ساخت‌های کم‌بسامد در زبان‌آموزان پیشرفته، و کلمات و ساخت‌های پربسامد در زبان‌آموزان مبتدی مشاهده می‌شود. بر این اساس، آزمون‌های مهارت واژگانی مبتنی بر به‌کارگیری کلمات برحسب میزان فراوانی آنها است (آلدرسون و بنرجی، ۲۰۰۲: ۹۰). جورج کینگزلی زیف در زمره‌ی اولین کسانی بود که با مطالعه‌ی آماری متون نوشتاری متوجه یک رابطه‌ی ریاضی بین فراوانی هر کلمه و رتبه‌ی آن شد که به «قانون زیف» مشهور شد و در سایر علوم نیز به‌کار گرفته شد. قانون زیف در مطالعه‌ی آمارِ کلمات ناظر به این معنا است که اگر فراوانی هر کلمه را در یک پیکره‌ی زبانی با حجم بالا به دست آورده و آنها را به صورت نزولی مرتب کرده و کلمات را رتبه‌بندی کنیم، به‌طوری‌که پربسامدترین کلمه دارای رتبه‌ی ۱ باشد و با کاهش فراوانی، رتبه‌ی کلمات افزایش یابد، در این صورت اگر فراوانی کلمه را با $f$ و رتبه‌ی آن را با $r$ نشان دهیم، رابطه‌ی ریاضی زیر را خواهیم داشت ($k$ یک مقدار ثابت است که مقدار آن برای هر پیکره به صورت تجربی به دست می‌آید):

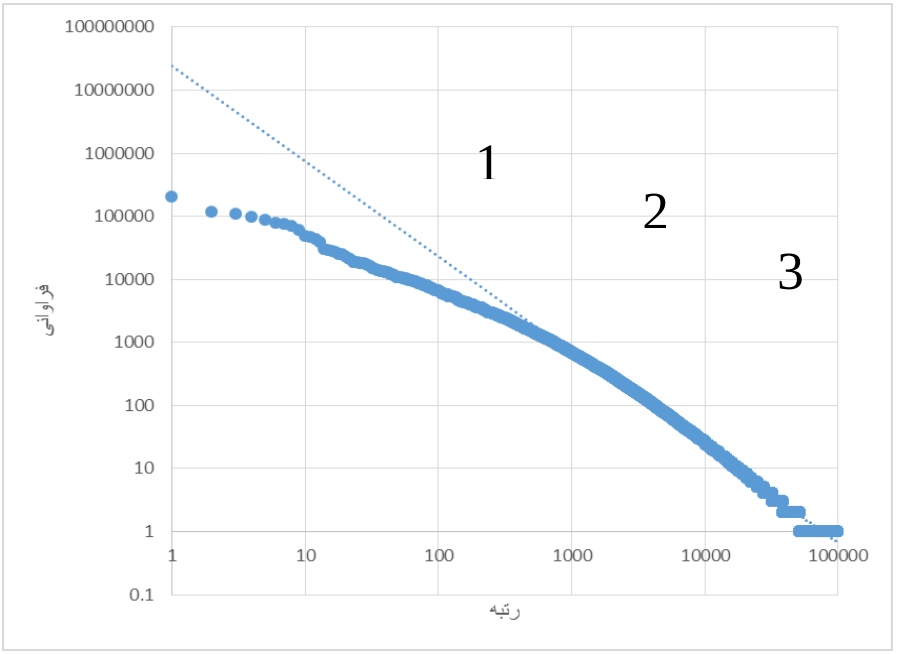
$$f * r = k$$

بعدها مندلبرات این رابطه را به صورت زیر اصلاح کرد (مانینگ و شوتز،۲۰۰۰: ۲۵):

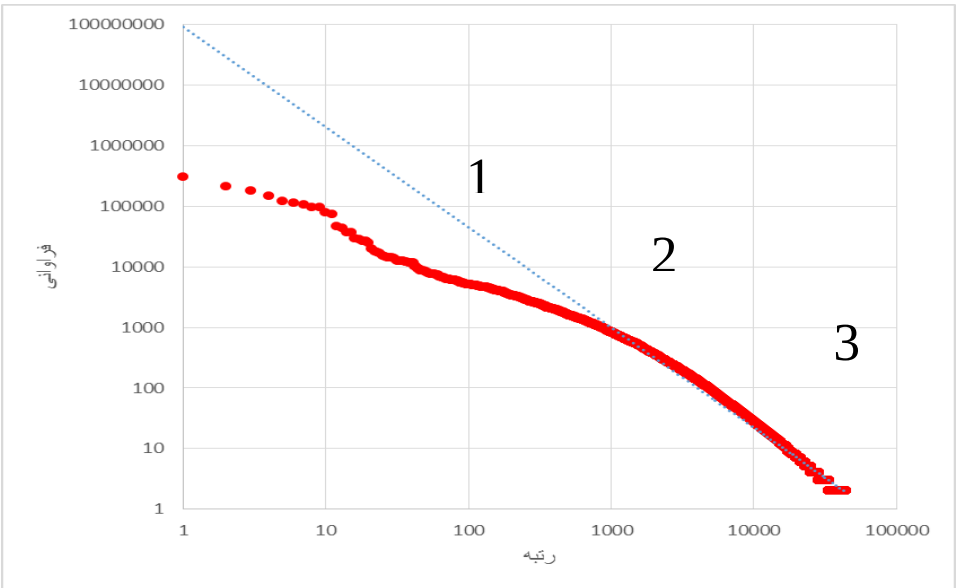
$$f = p * (r + c)^{-B}$$

پارامترهای $p$، $c$ و $B$ به صورت تجربی از پیکره‌ی زبانی به دست می‌آیند و ناظر به غنای واژگانی پیکره هستند.

شکل (۱) و (۲) منحنی مندلبرات را برای توزیع فراوانی کلمات فارسی محاوره‌ای و رسمی نسبت به رتبه‌ی آنها نشان می‌دهد.




شکل ۱. منحنی مندلبرات برای توزیع فراوانی کلمات فارسی محاوره‌ای نسبت به رتبه‌ی آنها




شکل ۲. منحنی مندلبرات برای توزیع فراوانی کلمات فارسی رسمی نسبت به رتبه‌ی آنها

همان‌طورکه مشاهده می‌شود منحنی از توزیع نرمال تبعیت نمی‌کند. اما یک الگوی واحد در هر دو شکل وجود دارد؛ این که می‌توان منحنی را به سه ناحیه (شماره‌های ۱، ۲ و ۳) تقسیم کرد؛ ناحیه‌ی کلمات پربسامد که تعداد بسیار کمی از کلمات را پوشش می‌دهد، ناحیه‌ی کلمات با بسامد متوسط که تعداد بیشتری از کلمات را در مقایسه با کلمات پربسامد شامل می‌شود و ناحیه‌ی کلمات کم‌بسامد که تعداد بسیار زیادی از کلمات را پوشش می‌دهد. زیف این پدیده را «اصل کم‌کوشی» نام نهاد؛ زیرا در یک ارتباط زبانی، گوینده تلاش دارد تا با استفاده از تعداد کمِ کلمات پربسامد، انرژی خود را ذخیره کند و شنونده تلاش می‌کند تا با استفاده از تعداد زیاد کلمات کم‌بسامد، از میزان ابهام معناییِ پیام دریافتی بکاهد (مانینگ و شوتز، ۲۰۰۰: ۲۵). نصری و همکاران (۱۳۹۳) با استفاده از منحنی نرمالِ معیار مرز بین ناحیه‌ی (۱) و (۲) را فراوانیِ ۷۵۰ و مرز بین ناحیه‌ی (۲) و (۳) را فراوانیِ ۱۶۰۰ به دست آورده است. الگوی توزیع فراوانی کلمات را می‌توان با دقت در شاخص‌های آماری منحنی مندلبرات تبیین کرد:

(۱) در هر دو پیکره، فراوانی ۵۰ درصد کلمات زیر عدد ۲ است.

(۲) فراوانی ۹۰ درصد کلمات در پیکره‌ی رسمی و محاوره‌ای به ترتیب زیر ۴۱ و ۲۵ است.

(۳) فراوانی ۹۹ درصد کلمات در پیکره‌ی رسمی و محاوره‌ای به ترتیب زیر ۱۰۰۴ و ۶۸۳ است.

(۴) کلمات دستوری شامل حروف اضافه و ربط، ضمایر و افعال کمکی و همکرد در پیکره‌ی رسمی و محاوره‌ای به ترتیب ۳۲ و ۴۹ درصد، اولین ۱۰۰ کلمه‌ی پربسامد را پوشش می‌دهند.

ویژگی‌های توزیع آماری کلمات در سطوح سه‌گانه‌ی قانون زیف-مندلبرات معیار مطمئنی برای ارزیابی مهارت واژگانی زبان‌آموزان به دست می‌دهد. این قانون فضای واژگانی را برحسب فراوانی کلمات به سه ناحیه افراز می‌کند و معیار لاوفر و نیشن نیز به چهار سطح واژگانی برای اندازه‌گیری مهارت یا غنای واژگانی بها می‌دهد. پس رابطه‌ی معنی‌داری بین واژگان بسامدی و معیار لاوفر و نیشن وجود دارد.

## ۵. استخراج واژگان پربسامد

انتخاب واژگان برای آموزش و آزمون زبان دوم از اهمیت فوق‌العاده‌ای برخوردار است؛ خصوصاً این که زبان‌آموزان همانند افراد بومی، زمان و فرصت کافی برای یادگیری واژگان زبان ندارند (لاوفر، ۲۰۱۴). بنابراین، در نگاه اول طبیعی به نظر می‌رسد که چنین واژگانی شامل کلماتی باشد که بسامد آنها در ارتباط اهل زبان بیشتر باشد. این واژگان را «واژگان پربسامد» می‌نامیم.

دومین سؤال تحقیق این است که در استخراج واژگان بسامدی زبان فارسی چه ملاحظاتی باید مد نظر داشت. گاردنر و دیویس (۲۰۱۳) معتقد هستند که تحقیقات پیکره-بنیاد باعث گسترش فرهنگِ لغت و واژگان شده و تأثیر عمیقی بر آموزش زبان انگلیسی داشته است. آنها مهم‌ترین چالش محققان را در این

خصوص، تعریف قواعدِ تعیین و تحدیدِ کلمه و شمارش آن می‌دانند. این چالش برای متون فارسی به سه دلیل بسیار جدی‌تر است.

اول این که در خط فارسی از حروف عربی استفاده می‌شود و اکثر حروف عربی به هم می‌چسبند و تغییر شکل می‌دهند. چون ساخت‌واژه‌ی فارسی از نوع پیوندی است و کلمات از طریق باهمایی تکواژها ساخته می‌شوند، بسته به این که در نوشتن کلمات، حروفِ مرزیِ تکواژها به هم چسبیده شوند یا نشوند، تنوع نوشتاری کلمات در متون فارسی فراوان یافت می‌شود.

دوم این که چون واکه‌های کوتاه در خط فارسی، به صورت فتحه، کسره و ضمه، به‌ندرت نوشته می‌شوند، هم‌نگاره‌های واژگانی، مانند مهر:[mehr]، مهر: [mahr] و مهر: [mohr] در متون فارسی کم نیستند.

سوم این که تکواژهای اشتقاقی و دستوری فارسی به صورت پسوند یا پی‌چسب به انتهای کلمه‌ی میزبان می‌چسبند و این فرایند باعث می‌شود تا تعداد هم‌نگاره‌های غیرواژگانی، مانند مهری[meh.ʾri]، به‌عنوان اسم خاص، و مهری[meh.riʿ]، به‌عنوان اسم عام نکره، در متون فارسی به‌طرز قابل توجهی افزایش یابند (علامتِ ʾ نشانگر تکیه‌بر بودن هجا است؛ مرز هجا با نقطه مشخص شده است) (بی‌جن‌خان و همکاران، ۲۰۱۱).

برای استخراج واژگان پربسامد از پیکره‌ی متون محاوره‌ای و رسمی باید عملیات بن‌واژه‌سازی[1] را انجام داد. با توجه به این که هر کلمه در متون الکترونیکی عبارت است از رشته‌ای از حروف که در دو طرف آن نویسه‌ی جای خالی و یا علائم نقطه‌گذاری وجود داشته باشد، بن‌واژه‌سازی دلالت بر حذف وندهای تصریفی و جداسازی واژه‌بست‌ها از صورت نوشتاری هر کلمه‌ی متن دارد تا از این رهگذر بن‌واژه‌ی کلمه به دست آید. بن‌واژه‌سازیِ پیکره‌ی رسمی با استفاده از برچسب صرفی-نحوی هر کلمه به صورت خودکار انجام شد و خطاهای بسیار اندک آن به‌طور دستی اصلاح گردید. از آن‌جا که یک نرم‌افزار بن‌واژه‌ساز[2] با دقت قابل قبول برای متون فارسی، که کلمات آن فاقد برچسب صرفی-نحوی هستند، وجود ندارد، عملیات بن‌واژه‌سازی پیکره‌ی محاوره‌ای برای استخراج واژگان بسامدی کاملاً به روش دستی انجام شد. این فعالیت طی شش مرحله صورت پذیرفت:

(۱) ابتدا توزیع فراوانی کلمات (یا تک‌کلمه‌ای‌ها)[3]، دوکلمه‌ای‌ها[4] و سه‌کلمه‌ای‌ها[5] برای متون هر کدام از پیکره‌ها به دست آمد و سپس با استفاده از نرم‌افزار *اکسل*، توزیع کلمات برحسب فراوانی به صورت نزولی مرتب شدند.

---

1. lemmatization
2. lemmatizer
3. monogram
4. bigrams
5. trigrams

(۲) کلماتی که دلالت بر نام اشخاص داشتند، حذف شدند.

(۳) تعداد ۵۳۷۰ کلمه که فراوانی آنها حداقل ۷۰ بود، انتخاب شد.

(۴) برای ابهام‌زدایی معنایی از کلماتی که عضو مجموعه‌ی هم‌نگاره‌های متون فارسی هستند، از فهرست سه‌کلمه‌ای‌ها (تا فراوانی ۱۰) و دوکلمه‌ای‌ها (تا فراوانی ۲) که دارای بافت هستند، استفاده شد و مقدار فراوانی کلمه‌ی مورد نظر اصلاح گردید. به‌عنوان مثال، فراوانی کلمه‌ی «نگین» در پیکره‌ی محاوره‌ای ۲۷۱۲ است. با بررسی سیاهه‌ی فراوانی دوکلمه‌ای‌ها و سه‌کلمه‌ای‌ها مشخص شد که ۱۵۰ بار این کلمه به‌صورت اسم خاص مونث و در بقیه‌ی موارد به صورت منفیِ التزامیِ دوم شخصِ جمعِ فعلِ «گفتن» به‌کار رفته است. بنابراین، عدد ۱۵۰ از فراوانی مدخل واژگانی «گفتن» کسر شد.

(۵) با استفاده از نرم‌افزار اکسل، توزیع کلمات برحسب ترتیب قاموسی مرتب شد تا امکان دسترسی و بررسی کلماتی که صورت نوشتاری یک مدخل واژگانی هستند، فراهم شود. با این روش، فراوانی هر مدخل واژگانی از جمع فراوانی صورت‌های تصریفی‌اش به دست آمد.

(۶) توزیع فراوانی مدخل‌های واژگانی به صورت نزولی مرتب شدند و یک واژگان پربسامد ۲۰۰۰ مدخلی به دست آمد. جدول (۱) و (۲)، ده واژه‌ی ابتدایی و انتهایی واژگان پربسامد ۲۰۰۰ مدخلیِ پیکره‌ی متنی رسمی و محاوره‌ای فارسی امروز را نشان می‌دهد.

با یک بررسی دیداری مشخص شد که بین فراوانی بعضی از واژه‌ها در دو واژگان تفاوت زیادی وجود دارد. بنابراین، واژه‌های دو واژگان را برحسب پربسامد بودن می‌توان در دو گروه قرار داد. گروه اول، واژه‌هایی که فراوانی آنها تفاوت زیادی با یکدیگر دارد. گروه دوم، واژه‌هایی که در مقایسه با گروه اول، فراوانی آنها تفاوت زیادی با یکدیگر ندارد. در بخش بعد به بررسی آماری این دو گروه واژگانی خواهیم پرداخت.

جدول ۱. ده واژه‌ی ابتدایی و انتهایی واژگان پربسامد ۲۰۰۰ مدخلی پیکره‌ی متنی رسمی

| ردیف | کلمه | فراوانی | ردیف | کلمه | فراوانی |
|---|---|---|---|---|---|
| ۱ | و | ۳۰۷۴۸۴ | ۱۹۹۱ | متناسب | ۳۴۹ |
| ۲ | به | ۱۷۸۷۴۹ | ۱۹۹۲ | بازنشسته | ۳۴۹ |
| ۳ | از | ۱۴۸۰۱۱ | ۱۹۹۳ | تشکل | ۳۴۹ |
| ۴ | که | ۱۲۱۳۲۶ | ۱۹۹۴ | اقامت | ۳۴۸ |
| ۵ | بودن | ۱۱۳۶۴۷ | ۱۹۹۵ | ارزشمند | ۳۴۸ |
| ۶ | این | ۱۰۶۶۱۳ | ۱۹۹۶ | پنهان | ۳۴۸ |
| ۷ | کردن | ۹۶۳۷۱ | ۱۹۹۷ | قدیم | ۳۴۸ |
| ۸ | را | ۹۵۰۷۶ | ۱۹۹۸ | ترویج | ۳۴۷ |
| ۹ | شدن | ۸۰۲۳۷ | ۱۹۹۹ | علل | ۳۴۷ |
| ۱۰ | با | ۷۳۰۱۷ | ۲۰۰۰ | تعبیر | ۳۴۷ |

جدول ۲. ده واژه‌ی ابتدایی و انتهایی واژگان پربسامد ۲۰۰۰مدخلی پیکره‌ی متنی محاوره‌ای

| ردیف | کلمه | فراوانی | ردیف | کلمه | فراوانی |
|---|---|---|---|---|---|
| ۱ | و | ۲۰۱۶۶۶ | ۱۹۹۱ | بریدن | ۸۷ |
| ۲ | کردن | ۱۸۲۳۲۸ | ۱۹۹۲ | باسرعت | ۸۷ |
| ۳ | به | ۱۴۰۱۳۳ | ۱۹۹۳ | آرشیو | ۸۷ |
| ۴ | بودن | ۱۳۱۹۷۵ | ۱۹۹۴ | اینترنت | ۸۷ |
| ۵ | که | ۱۲۵۶۱۹ | ۱۹۹۵ | اندام | ۸۷ |
| ۶ | را | ۱۱۵۵۷۳ | ۱۹۹۶ | اشتها | ۸۷ |
| ۷ | از | ۱۰۳۸۷۷ | ۱۹۹۷ | لجباز | ۸۶ |
| ۸ | گفتن | ۱۰۱۴۲۷ | ۱۹۹۸ | صادق | ۸۶ |
| ۹ | شدن | ۱۰۰۳۳۹ | ۱۹۹۹ | سرباز | ۸۶ |
| ۱۰ | من | ۸۱۶۹۱ | ۲۰۰۰ | زهر | ۸۶ |

## ۶. بررسی تفاوت پیکره‌ی متنی رسمی و محاوره‌ای

سومین سؤال تحقیق این است که آیا تفاوت معنی‌داری بین واژگان مستخرج از پیکره‌ی متنی رسمی و محاوره‌ای وجود دارد؟ پاسخ این سؤال برای ارزیابی مهارت واژگانی فارسی‌آموزان اهمیت دارد، زیرا این احتمال وجود دارد که تفاوت ناشی از این دو پیکره صرفاً ناشی از شیوه‌ی تلفظ کلماتی باشد که دارای بافت واجی معیّنی هستند و در متون نوشتاری نیز ممکن است منعکس شوند، مانند کلمه‌ی «آسمان» در پیکره‌ی رسمی و کلمه‌ی «آسمون» در پیکره‌ی محاوره‌ای. فرضیه‌ی ما این است که اگر واقعاً بین پیکره‌ی رسمی و محاوره‌ای تفاوت قابل توجهی وجود داشته باشد، این تفاوت باید در بین کلمات پربسامد مشترک در دو پیکره یافت شود. زیرا اگر این تفاوت در کلماتی بررسی شود که در یکی پربسامد و در دیگری

کم‌بسامد باشد، اثبات وجود تفاوت معنی‌دار در آن دو کاملاً بدیهی است. با وجود این، ابتدا به بررسی کلماتی که در پیکره‌ی رسمی، پربسامد و در پیکره‌ی محاوره‌ای، کم‌بسامد هستند، و بالعکس در پیکره‌ی رسمی، کم‌بسامد و در پیکره‌ی محاوره‌ای، پربسامد هستند، می‌پردازیم و سپس به بررسی واژگان پربسامد مشترک در دو پیکره خواهیم پرداخت.

یافته‌های هر دو بخش می‌تواند در ارزیابی مهارت واژگانی زبان‌آموزان مفید باشد. بنابراین، ابتدا واژگان‌های مستخرج از دو پیکره برحسب فراوانی و به ترتیب نزولی مرتب شدند. سپس مدخل‌های مشترک دو پیکره، که تفاوت زیادی بین فراوانی آنها مشاهده شد، به دست آمدند. جدول (۳) تعدادی از این کلمات را به‌عنوان «کلمات منتخب» نشان می‌دهد. نتیجه‌ی کلی که از داده‌های این جدول به دست می‌آید این است که بعضی از کلمات بار محاوره‌ای بیشتری در مقایسه با کلمات متناظرشان دارند. تناظر کلمات می‌تواند ناشی از ترادف معنایی باشد:

(۱) «مدرسه» بار محاوره‌ای بیشتری در مقایسه با «دبستان» دارد.

(۲) «توی» بار محاوره‌ای بیشتری در مقایسه با «در» (به معنای حرف‌اضافه‌ای) دارد.

(۳) «واسه» بار محاوره‌ای بیشتری در مقایسه با «برای» دارد.

(۴) «حرف زدن» بار محاوره‌ای بیشتری در مقایسه با «صحبت کردن» و «صحبت کردن» بار معنایی بیشتری در مقایسه با «گفتگو کردن» دارد.

(۵) «چطور» بار محاوره‌ای بیشتری در مقایسه با «چگونه» دارد.

(۶) «گوشی» بار محاوره‌ای بیشتری در مقایسه با «تلفن» دارد.

تناظر کلمات می‌تواند ناشی از رابطه‌ی معنایی ذهنی و عینی باشد:

(۱) «اسلام» و «اسلامی» به عنوان دو کلمه‌ی ذهنی بار محاوره‌ای بسیار کمتری در مقایسه با کلمات ذهنی «حس»، «احساس» و «باور» دارند.

(۲) بار محاوره‌ای افعال حسی چون «خندیدن» و «زنگیدن» (فعل جعلی) به مراتب بیشتر از افعالی چون «گذاردن»، «برانگیختن» و «گذاشتن» است.

(۳) به‌کارگیری کلمات عینی «ماما(ن)»، «بابا»، «بچه»، «بچه‌ها» (به معنای رفقا و دوستان)، «سلام»، «زنگ» و «گوش» در پیکره‌ی محاوره‌ای به مراتب بیشتر از پیکره‌ی رسمی است.

جدول ۳. فراوانی کلمات منتخب در واژگان مستخرج از پیکره‌ی محاوره‌ای و رسمی

| واژه | پیکره‌ی محاوره‌ای | | پیکره‌ی رسمی | |
|---|---|---|---|---|
| | فراوانی | رتبه | فراوانی | رتبه |
| دبستان | ۳۰ | ۱۰۰۸۳ | ۱۰۰ | ۴۶۲۱ |
| مدرسه | ۸۳۰ | ۹۱۹ | ۱۰۶۵ | ۷۸۸ |
| اسلامی | ۱۴ | ۱۵۰۲۰ | ۱۴۳۳۲ | ۲۷ |
| اسلام | ۲۴ | ۱۱۶۳۸ | ۳۴۷۳ | ۱۸۶ |
| توی (داخل) | ۱۳۸۴۵ | ۳۶ | ۳۲ | ۸۹۹۰ |
| در (حرف اضافه) | ۳۰۲۴۳ | ۱۶ | ۵۴۰۰۰ | ۱۲ |
| برای | ۲۷۱۸۶ | ۳۲ | ۳۶۹۳۱ | ۱۶ |
| واسه | ۶۴۲۷ | ۱۰۷ | ۱۶ | ۱۳۵۰۸ |
| ماما(ن) | ۱۲۱۶۶ | ۶۶ | ۱۹۵ | ۳۱۹۷ |
| بابا | ۱۲۱۷۱ | ۷۰ | ۱۹۰ | ۳۵۱۱ |
| زنگ | ۵۴۴۴ | ۱۲۱ | ۲۱۴ | ۲۷۵۷ |
| سلام | ۵۵۰۸ | ۱۲۲ | ۲۵۶ | ۲۳۷۰ |
| صحبت (کردن) | ۲۶۵۲ | ۲۸۲ | ۱۰۲۱ | ۷۷۹ |
| گفتگو (کردن) | ۳۶ | ۸۱۲۵ | ۲۵۳۲ | ۲۸۴ |
| حرف (زدن) | ۱۰۳۱۵ | ۵۶ | ۱۵۵۶ | ۴۸۲ |
| حس | ۲۶۴۷ | ۲۸۴ | ۳۹۴ | ۱۷۱۳ |
| احساس | ۳۵۶۰ | ۲۰۲ | ۱۸۲۰ | ۴۱۷ |
| بچه | ۲۴۳۰ | ۳۱۳ | ۹۶۱ | ۸۲۹ |
| بچه‌ها | ۲۱۸۱ | ۳۱۳ | ۵۶۹ | ۸۲۹ |
| گوشی | ۲۱۵۵ | ۳۵۷ | ۸۱ | ۵۲۴۲ |
| گوش | ۲۱۶۸ | ۳۵۶ | ۷۶۱ | ۱۰۰۱ |
| باور | ۲۱۳۸ | ۳۶۰ | ۷۲۱ | ۱۰۵۷ |
| تلفن | ۲۱۲۵ | ۳۶۲ | ۱۳۰۳ | ۶۱۰ |
| زنگیدن | ۱۰ | ۲۸۰۶۳ | - | - |
| چطور | ۱۹۳۸ | ۳۹۱ | ۳۰۹ | ۲۰۳۵ |
| چگونه | ۹۴ | ۴۳۹۸ | ۱۰۸۴ | ۷۲۸ |

همچنین واژگان پربسامد مشترک در پیکره‌ی رسمی و محاوره‌ای طی دو مرحله استخراج شد. ابتدا یک برنامه‌ی رایانه‌ای نوشته شد که درونداد، آن واژگان پربسامد ۲۰۰۰ مدخلیِ پیکره‌ی رسمی و محاوره‌ای، و برونداد آن واژگان پربسامد مشترک در آنها باشد. سپس با اجرای برنامه، یک واژگان مشترک ۸۶۱ مدخلی به دست آمد. جدول (۴)، ده واژه‌ی ابتدایی و انتهایی واژگان مشترکِ ۸۶۱ مدخلی پیکره‌های رسمی و محاوره‌ای فارسی امروز را نشان می‌دهد.

جدول ۴. ده واژه‌ی ابتدایی و انتهایی واژگان مشترک ۸۶۱ مدخلی پیکره‌های متنی رسمی و محاوره‌ای

| ردیف | کلمه | فراوانی | |
|---|---|---|---|
| | | پیکره رسمی | پیکره نوشتاری |
| ۱ | و | ۳۰۷۴۸۴ | ۲۰۱۶۶۶ |
| ۲ | به | ۱۷۸۷۴۹ | ۱۴۰۱۳۳ |
| ۳ | از | ۱۴۸۰۱۱ | ۱۰۳۸۷۷ |
| ۴ | که | ۱۲۱۳۲۶ | ۱۲۵۶۱۹ |
| ۵ | بودن | ۱۱۳۶۴۷ | ۱۳۱۹۷۵ |
| ۶ | این | ۱۰۶۶۱۳ | ۶۵۵۸۷ |
| ۷ | کردن | ۹۶۳۷۱ | ۱۸۲۳۲۸ |
| ۸ | را | ۹۵۰۷۶ | ۱۱۵۵۷۳ |
| ۹ | شدن | ۸۰۲۳۷ | ۱۰۰۳۳۹ |
| ۱۰ | با | ۷۳۰۱۷ | ۸۱۵۹۳ |
| ۸۵۲ | پرده | ۳۲۵ | ۲۴۷ |
| ۸۵۳ | اوج | ۳۲۳ | ۲۰۰ |
| ۸۵۴ | تلفنی | ۳۲۲ | ۲۱۴ |
| ۸۵۵ | بهانه | ۳۲۰ | ۶۱۴ |
| ۸۵۶ | برخاستن | ۳۱۹ | ۱۸۱ |
| ۸۵۷ | روحیه | ۳۱۸ | ۱۰۳ |
| ۸۵۸ | روحی | ۳۱۸ | ۱۲۲ |
| ۸۵۹ | توقف | ۳۱۸ | ۱۵۵ |
| ۸۶۰ | ستاره | ۳۱۷ | ۱۱۰ |
| ۸۶۱ | خیال | ۲۹۰ | ۱۹۰۹ |

فرضیه‌ی آماری تحقیق ناظر به این معناست که اگر کلمات مشترک، به یک نسبت در هر دو پیکره به‌کار رفته باشند، در آن صورت دو پیکره شبیه به یکدیگر هستند و بنابراین، متعلق به یک جامعه‌ی آماری واحد هستند و نمی‌توان استقلال آنها را از هم اثبات کرد. بنابراین، اگر حداقل یک کلمه‌ی مشترک یافت شود که به یک نسبت در دو پیکره به‌کار نرفته باشد، فرضیه‌ی تحقیق رد می‌شود (مانینگ و شوتز، ۲۰۰۰: ۱۷۱). به عبارت دیگر، در این صورت می‌توان با استفاده از واژگان، متون فارسی را به دو سبک محاوره‌ای و رسمی افراز کرد.

اگر حجم پیکره‌ها را با $N_1$ و $N_2$ و فراوانی مشاهده‌ی هر کلمه مانند w را در دو پیکره با $O_{w,1}$ و $O_{w,2}$ نشان دهیم، آن‌گاه فراوانی قابل انتظار کلمه‌ی w در پیکره‌ی iام (به ازای i برای ۱ و ۲)، که با $E_{w,i}$ نشان می‌دهیم، از رابطه‌ی زیر به دست می‌آید (کیلگاریف و رُز، ۱۹۹۸):

$$E_{w,i}= N_i* (O_{w,1} + O_{w,2}) / N_1 + N_2$$

واضح است که $N_1$ مجموع فراوانی کلمات پربسامد مشترک در پیکره‌ی رسمی و $N_2$ مجموع فراوانی کلمات پربسامد مشترک در پیکره‌ی محاوره‌ای است. با توجه به این که براساس قانون زیف، رتبه‌ی هر کلمه در محور افقی با خودِ کلمه هم‌ارز است و تابع توزیع فراوانی کلماتِ زبان از توزیع نرمال تبعیت نمی‌کند و این که فراوانی کلماتِ مشترک همگی بیشتر از عدد ۵ است (این فراوانی در پیکره‌ی رسمی و محاوره‌ای به ترتیب حداقل برابر با ۸۶ و ۳۵۰ است)، مقدار خی‌دو که از متوسط تفاضل فراوانی قابل انتظار از فراوانی مشاهده شده برای همه‌ی کلمات مشترک به دست می‌آید، با مقدار احتمال سطح آلفا برابر با ۰،۵ آزمون مناسبی برای اثبات استقلال دو پیکره از یکدیگر یا همگنی دو پیکره است. نتیجه‌ی چهار آزمون خی دو با استفاده از نرم‌افزار اس‌پی‌اس‌اس نسخه‌ی ۲۲ برای کلمات مشترک، که در چهار مقوله‌ی دستوری اسم، فعل، صفت/قید، حرفِ اضافه/ربط تفکیک شدند، با درجات آزادی ۵۳۸، ۴۹، ۱۲۹ و ۸۳ نشان داد که کلمات متعلق به هر کدام از مقولات دستوری به یک جامعه‌ی آماری واحد تعلق ندارند و بنابراین، تفاوت معنی‌داری بین پیکره‌ی رسمی و محاوره‌ای وجود دارد. جدول ۵ فراوانی مشاهده شده و قابل انتظار را برای کلمات منتخب نشان می‌دهد.

جدول۵. فراوانی مشاهده شده و قابل انتظار برای کلمات منتخب

| کلمه | فراوانی مشاهده شده | | فراوانی قابل انتظار | | کلمه | فراوانی مشاهده شده | | فراوانی قابل انتظار | |
|---|---|---|---|---|---|---|---|---|---|
| | رسمی | محاوره‌ای | رسمی | محاوره‌ای | | رسمی | محاوره‌ای | رسمی | محاوره‌ای |
| آخر | ۷۹۷ | ۳۴۵۰ | ۲۰۱۲ | ۲۲۳۴ | پیام | ۱۳۸۲ | ۶۲۶ | ۹۵۱ | ۱۰۵۶ |
| آزاد | ۱۸۹۴ | ۴۸۵ | ۱۱۲۷ | ۱۲۵۱ | حمایت | ۲۲۸۲ | ۱۱۲ | ۱۱۳۴ | ۱۲۵۹ |
| آزادی | ۲۱۱۱ | ۲۲۸ | ۱۱۰۸ | ۱۲۳۰ | خشک | ۳۵۸ | ۹۱۷ | ۶۰۴ | ۶۷۰ |
| اجازه | ۱۰۸۲ | ۲۳۰۷ | ۱۶۰۵ | ۱۷۸۳ | خودرو | ۱۳۳۸ | ۱۵۰ | ۷۰۵ | ۷۸۳ |
| اداره | ۲۴۹۱ | ۱۶۰ | ۱۲۵۶ | ۱۳۹۵ | دختر | ۱۳۵۳ | ۹۴۱۳ | ۵۱۰۰ | ۵۶۶۵ |
| ترس | ۳۷۹ | ۱۳۴۸ | ۸۱۸ | ۹۰۸ | رساندن | ۲۲۱۶ | ۴۹۹ | ۱۲۸۶ | ۱۴۲۸ |
| جمعیت | ۱۵۱۳ | ۳۳۴ | ۸۲۸ | ۹۱۹ | ساختن | ۴۸۸۵ | ۲۰۰ | ۲۴۰۹ | ۲۶۷۵ |
| جوان | ۳۳۵۵ | ۹۰۰ | ۲۰۱۵ | ۲۲۳۹ | و | ۳۰۷۴۸۴ | ۲۰۱۶۶۶ | ۲۴۱۲۲۳ | ۲۶۷۹۲۶ |
| آسمان | ۴۹۲ | ۵۷۸ | ۵۰۶ | ۵۶۳ | زیاد | ۳۳۲۹ | ۳۸۸۱ | ۳۴۱۵ | ۳۷۹۴ |
| بدون | ۳۰۷۷ | ۳۹۶۳ | ۳۳۳۵ | ۳۷۰۴ | کوچک | ۱۱۶۱ | ۱۲۸۹ | ۱۱۶۰ | ۱۲۸۹ |
| بردن | ۴۱۵۷ | ۴۲۰۲ | ۳۹۶۰ | ۴۳۹۸ | ماجرا | ۴۶۷ | ۵۵۰ | ۴۸۱ | ۵۳۵ |
| جدی | ۱۳۴۱ | ۱۴۱۲ | ۱۳۰۴ | ۱۴۴۸ | هرچند | ۵۶۳ | ۶۶۱ | ۵۸۰ | ۶۴۴ |

## ۷. بحث و نتیجه‌گیری

«واژگان بسامدی» یک ابزار اصلی برای اندازه‌گیریِ غنای واژگانیِ زبان‌آموزان است. بر این اساس، موضوع اصلی مقاله حول محور استفاده از واژگان بسامدی قرار گرفت تا در تحقیقات آتی از آن برای ارزیابی مهارت واژگانی فارسی‌آموزان استفاده شود. «واژگان بسامدی» حاوی ویژگی‌های آماریِ متونِ پیکره‌ی زبانی است، و پیکره‌ی فارسی رسمی و محاوره‌ای در شیوه‌ی به‌کارگیری واژه‌های فارسی اختلاف معنی‌داری نسبت به یکدیگر دارند. علاوه بر این، مهارت‌های واژگانیِ زبان‌آموزان از طریق داده‌های واقعی، موقعیت‌بنیاد و مستندِ زبانی سنجیده می‌شود. بنابراین، انتظار می‌رود که مواد آزمون‌ها و شیوه‌ی ارزیابی آزمون‌ها تا جایی که ممکن است دربردارنده‌ی ویژگی‌های آماری واژگان بسامدی باشد تا بتوان قضاوت درستی درباره‌ی میزان دانش واژگانی زبان‌آموزان داشت.

در این تحقیق، پیکره‌های زبانی برحسب گونه‌ی رسمی و محاوره‌ای متمایز شده‌اند و متون هر کدام از پیکره‌ها نه تنها از منابع مختلف از قبیل روزنامه، مجله، کتاب و اینترنت جمع‌آوری شده‌اند، بلکه شامل سیاق‌های مختلف، از قبیل سیاسی، اجتماعی، اقتصادی و امثال آن هستند. بنابراین، واژگان بسامدی و ویژگی‌های آماری به صورت مستقل از منبع و سیاق استخراج شده‌اند. این شیوه‌ی استخراج را می‌توان با این استدلال موجه دانست که اگر بنا باشد بر طبق یافته‌های کراسلی و همکاران (۲۰۱۰) دسترسی به واحدهای واژگانی کانونی یکی از شاخص‌های ارزیابی مهارت واژگانی باشد، انتظار می‌رود که واژگان بسامدی به علت عدم وابستگی به منبع و سیاق، بسیاری از واژه‌های کانونی را پوشش دهد (همان‌طورکه در جدول (۳) برحسب رسمی و محاوره‌ای بودن نمونه‌ی کوچکی از این واژه‌ها را آورده‌ایم). البته واژه‌های جدول (۳) را می‌توان برحسب سه ویژگی کانونی بودن، یعنی عینی‌تر بودن، قابل تصورتر بودن و آشناتر بودن ( کراسلی و همکاران، ۲۰۱۰: ۵۶۵) تفسیر کرد.

اکنون این سؤال مطرح می‌شود که چون هر داده‌ی مستند زبانی به یک سیاق و یک موقعیت خاص ارتباطی، مانند سخنرانی، سخن رودررو، مکالمه‌ی تلفنی، خاطره و امثال آن، وابسته است، چگونه می‌توان از واژگان بسامدی و ویژگی‌های آماری مستقل از سیاق به‌عنوان یک ابزار در ساخت سؤالات و ارزیابی آزمون استفاده کرد. این اشکال به نمایه‌ی فراوانی واژگانیِ لاوفر و نیشن (۱۹۹۵) نیز وارد است که در آن برای ارزیابی غنای واژگانیِ یک متن یا زبان‌آموز از دو واژگان بسامدی هزارکلمه‌ای و یک واژگان بسامدی از متون دبیرستانی و دانشگاهی استفاده می‌شود. پاسخی که براساس یافته‌های این مقاله قابل طرح است، ناظر به این معنا است که بر طبق «قانون زیف»، هر واژگانِ بسامدی دارای تعداد کمی واژه‌ی پربسامد است که انتظار می‌رود برحسب احتمال وقوع‌شان در هر ماده‌ی آزمون به‌کار روند و باید در ارزیابی مهارت واژگانی نیز مد نظر باشند. اما با توجه به این که فراوانی ۹۹ درصد کلمات در پیکره‌ی رسمی و محاوره‌ای به ترتیب زیر

۱۰۰۴ و ۶۸۳ است، که در زمره‌ی کلمات با بسامد متوسط و کم قرار می‌گیرند، بر طبق رویکرد «نمایه‌ی فراوانیِ واژگانی»، آگاهی بیشتر زبان‌آموز نسبت به واژه‌های کم‌بسامدِ موقعیت‌بنیاد در سؤالات آزمون نشانگر مهارت واژگانی بیشتر است. ملاحظه‌ی این نکته در ارزیابی مهارت واژگانی بسیار اهمیت دارد.

سؤال دیگر این است که واژگان بسامدی نه تنها به حجم پیکره‌ی زبانی وابسته است و با افزایش حجم پیکره محتوای آن تغییر می‌کند، بلکه واژگان بسامدیِ مستخرج از یک پیکره با واژگان بسامدیِ مستخرج از پیکره‌ی دیگر تفاوت دارد. بنابراین، چگونه می‌توان از واژگان بسامدی به‌عنوان یک معیار برای آموزش و آزمون زبان‌آموزان استفاده کرد؟ در پاسخ به این سؤال باید گفت که با افزایش حجم پیکره نه تنها به پوشش واژه‌های پربسامد با میزان فراوانی بیشتر افزوده می‌شود، بلکه بر تعداد واژه‌های با بسامد متوسط و کم نیز افزوده می‌شود و در نتیجه، امکان ارزیابی مهارت واژگانی بیشتر می‌شود؛ زیرا همان‌طورکه گفته شد آگاهی بیشتر زبان‌آموز نسبت به واژه‌های کم‌بسامدِ موقعیت‌بنیاد، نشانگر مهارت واژگانی بیشتر است. علاوه بر آن، همگرایی دو واژگان بسامدی که از دو پیکره‌ی متفاوت استخراج شده باشند، با افزایش حجم پیکره‌ها بیشتر می‌شود.

لاوفر (۲۰۱۴) این رویکرد را که واژگان بسامدی تنها معیار آموزش و آزمون مهارت واژگانی باشد، زیر سؤال برده است. وی معیار دیگری را تحت عنوان «قابلیت یادگیری» مطرح کرده که ناظر به سادگی یا دشواری یادگیری یک واژه‌ی مشخص است. به‌عنوان مثال، یک واژه‌ی کم‌بسامد در زبان دوم ممکن است هم‌خانواده‌ی یک کلمه در زبان اول یا کلمه‌ی قرضی از زبان اول باشد. یادگیری این گونه کلمات برای زبان‌آموز بسیار راحت است. این وضعیت برای عرب‌زبانان که فارسی یاد می‌گیرند کاملاً صادق است؛ زیرا کلمات عربیِ کم‌بسامد در متون فارسی بسیار زیاد یافت می‌شود. علاوه بر آن، مشتقات یک کلمه‌ی پربسامد ممکن است کم‌بسامد باشد، در حالی که یادگیری آنها دشوار نباشد. یا این که ممکن است یادگیری یک کلمه‌ی پربسامد که برابرنهاد معنایی در زبان اول ندارد، مشکل باشد.

## تشکر و قدردانی

از دفتر سیاست‌گذاری و برنامه‌ریزی امور پژوهشی وزارت علوم، تحقیقات و فناوری و معاونت پژوهشی دانشگاه تهران، که طی قرارداد شماره ۳/۱۱۲۰۴۹ تاریخ ۹۲/۸/۲۰ از طرح آزفان (آزمون‌سازی زبان فارسی برای غیرفارسی‌زبانان) حمایت کرده‌اند، کمال تشکر و قدردانی را داریم.

## منابع:


**بی‌جن‌خان، م. و محسنی، م.** (۱۳۹۱). *فرهنگ بسامدی: براساس پیکره‌ی متنی فارسی امروز*. تهران: انتشارات دانشگاه تهران.

**شریفی آتشگاه، م. و بی‌جن‌خان، م.** (۱۳۸۸). تجزیه و تحلیل پیکره‌بنیاد واحدهای چندقطعه‌ای در متون فارسی. *مجله‌ی بین‌المللی ارتباطات و فناوری اطلاعات*، دوره‌ی ۱، شماره ۳: ۲۶-۱۵.

**نصری، ع. و همکاران.** (۱۳۹۳). *قانون زیف و انتخاب واژگان مرجع برای تعیین سطح مهارت واژگانی فارسی‌آموزان*. مجموعه مقالات سومین همایش ملی زبان‌شناسی رایانشی (دی‌وی‌دی). دانشگاه صنعتی شریف، ۲۸ و ۲۹ آبان.

**Alderson, J. C.** & **Banerjee, J.** (2002). State-Of-The-Art Review: Language Testing and Assessment (Part 2). *Language Teaching*. 35: 79–113.

**Bijankhan, M. et al.** (2011). Lessons from Building a Persian Written Corpus: Peykare. *Language Resources and Evaluation*. 45: 143-164.

**Brezina, V.** & **Gablasova, D.** (2013). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*. doi:10.1093/applin/amt018.

**Crossley, S. A. et al.** (2010). Predicting the Proficiency Level of Language Learners Using Lexical Indices. *Language Testing*. 28(3): 561-580.

**Crossley, S. A. et al.** (2014). Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy. *Applied Linguistics*. (First published online February 10).

**Duran, P. et al.** (2004). Developmental Trends in Lexical Diversity. *Applied Linguistics*. 25(2): 220-242.

**Ellis, N. C.** (2002). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *SSLA*, 24: 143–188.

**Ellis, N. C.** (2006). Language Acquisition as Rational Contingency. Learning. *Applied Linguistics*. 27(1): 1-24.

**Gardner, D.** (2007). Validating the Construct of Word in Applied Corpus-Based Vocabulary Research: A Critical Survey. *Applied Linguistics*. 28(2): 241-265.

**Gardner, D.** & **Davies, M**. (2013). A New Academic Vocabulary List. *Applied Linguistics* (doi: 10.1093/applin/amt015).

**Jarvis, J.** (2002). Short Texts, Best-FItting Curves and New Measures of Lexical Diversity. *Language Testing*, 19(1): 57-84.

**Kilgarriff, A.** & **Rose, T.** (1998). *Metrics for Corpus Similarity and Homogeneity*. Manuscript, ITRI, University of Brighton.

**Laufer, B.** & **Nation, P.** (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*. 16(3): 307-322.

**Laufer, B.** & **Nation, P.** (1999). A Vocabulary-Size Test of Controlled Productive Ability. *Language Testing*, 16 (1): 33-51.

**Laufer, B.** (2014). Vocabulary in a Second Language: Selection, Acquisition, and Testing: A Commentary on Four Studies for JALT Vocabulary SIG. *Vocabulary Learning and Instruction*. 3(2): 38-46.

**Litosseliti, L.** (2010). *Research Methods in Linguistics*. New York: Continuum International Publishing Group.

**Manning, C. D.** & **Schutze, H**. (2000). *Foundations of Statistical Natural Language Processing* (2nd printing). Massachussuts: MIT Press.

**Meara, P.** (2005). Lexical Frequency Profiles: A Monte Carlo Analysis. *Applied Linguistics*, 26(1): 32–47.

**O'Loughlin, R.** (2012). Tuning Into Vocabulary Frequency in Coursebooks. *RELC Journal, 43: 255-269.*

**Parent, K.** (2012). The Most Frequent English Homonyms. *RELC Journal. 43: 69-81.*

**Read, J.** & **Chapelle, C. A.** (2001). A Framework for Second Language Vocabulary Assessment. *Language Testing*, 18(1): 1-32.

**Schmitt, N.** (1999). The Relationship between TOEFL Vocabulary Items and Meaning, Association, Collocation and Word-Class Knowledge. *Language Testing.*16(2): 189-216.

**Smith, R.** (2004). *The Lexical Frequency Profile: Problems and Use.* Conference Proceedings of JALT: 439-451. NARA, Tokyo.

**Van Rooy, B.** & **Terblanche, L**. (2009). A multi-dimensional analysis of a learner corpus. In A. Renouf & A. Kehoe (Eds.), *Corpus Linguistics: Refinements and Reassessments*. Amsterdam-New York: Editions Rodopi B.V.

**Vermeer, A.** (2000). Coming to Grips with Lexical Richness in Spontaneous Speech Data. *Language Testing*, 17(1): 65-83.

**Yu, G.** (2007). Lexical Diversity in MELAB Writing and Speaking Task Performances. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, Volume 5. English Language Institute, University of Michigan.