«بسامد» مرکب از دو کلمهٔ «بس» و «آمد»، و برابر نهادهٔ لفظ فرانسوی frequence (انگلیسی: frequency) است، که خود فرکانس (که تلفظ فرانسوی آن است) نیز در فارسی به کار رفته و می‌رود. رایج‌ترین مترادف این لغت، «فراوانی» است، که به‌ویژه در دانش آمار کاربرد دارد. «بسامدی» نیز منسوب یا مربوط به «بسامد» است. بنابراین، «فرهنگ بسامدی» یعنی فرهنگی که بر مبنای بسامد لغات تهیه شده است؛ به عبارت دیگر، یعنی فرهنگی که همهٔ لغات یک پیکرهٔ زبانی (corpus) مشخص و تعریف شده را دربرداشته و بسامد تک‌تک لغات در آن ذکر شده باشد.

تاکنون چندین فرهنگ بسامدی فارسی ـ هم در ایران و هم در کشورهای دیگر ـ تدوین و منتشر شده است، که با رجوع به کتاب‌شناسی‌ها و فهرست‌های منتشر شده می‌توان به سادگی به نام و نشان آنها دست یافت. گرچه فعلاً شمار این گونه فرهنگ‌ها در برابر تعداد فرهنگ‌های عمومی (مانند فرهنگ‌های تک‌زبانه و دوزبانه و چند زبانه و نیز واژه‌نامه‌های تخصصی و نیمه تخصصی) اندک است، اما به‌ویژه در یکی دو دههٔ اخیر، چند فرهنگ واژه‌نما و بسامدی منتشر شده که هر چند در بعضی از آنها به اصول علمی تدوین چنین فرهنگ‌هایی عنایت کافی نشده، اما کوشش پدیدآورندگان آنها درخور تقدیر و اجرشان مشکور است.

دربارهٔ تهیهٔ انواع فرهنگ‌های بسامدی برای زبان فارسی نیز در جای‌های مختلف، چه در مقاله‌های انتقادی و چه در کتاب‌های ادبی و زبان‌شناختی، کم‌وبیش سخن گفته شده است؛ برخی با وسعت نظر، این کار را امری بسیار بایسته و مهم تلقی کرده‌اند، و بعضی دیگر هم از سر ناآگاهی ـ یا شاید کم‌التفاتی ـ آن را کاری کاملاً عبث و ماشینی شمرده‌اند.

به طور مختصر و ساده باید گفت که ضرورت تهیهٔ چنین فرهنگ‌هایی، مانند ضرورت آمارگیری‌های عمومی نظیر سرشماری نفوس و مسکن است؛ یعنی مثلاً اگر دولت از جمعیت کشورمان، نرخ باسوادی، میزان تحصیلات باسوادان، میانگین سنی افراد، میانگین شمار فرزندان هر خانوار، عدهٔ پناهندگان و اتباع خارجی، فراوانی مهاجرت‌های داخلی، و چندین فقرهٔ دیگر آگاهی دقیق و صحیح نداشته باشد، قطعاً قادر به برنامه‌ریزی درست و اصولی فرهنگی، اجتماعی، اقتصادی، و سیاسی نخواهد بود.

به همین قیاس، اگر ما از محدودهٔ واژگان (vocabulary) زبان فارسی، پربسامدترین لغات آن (واژگان پایه)، دامنهٔ طبقات دستوری کلمات، فراوانی لغات بسیط و مشتق و مرکب بومی و غیربومی، چگونگی ترکیب‌سازی، و نظایر اینها اطلاع کافی نداشته باشیم، چگونه می‌توانیم آن‌گونه که شایسته است، فرهنگ لغت و کتاب درسی تألیف کنیم؟ تاکنون تقریباً همهٔ مواد فرهنگ‌های تالیف شده برای زبان فارسی، براساس شمّ زبانی مؤلف یا مؤلفان، و یا به تقلید از

حمید حسنی

فرهنگ‌های پیشین گردآوری و ثبت شده‌اند. حتی در معدودی فرهنگ‌ها، که با تکیه بر شواهد استخراج شده از متون (معاصر یا قدیم و یا هر دو) تألیف شده‌اند، رگه‌های تقلید و پیروی از قاموس‌های کهن را به وضوح می‌توان دید.

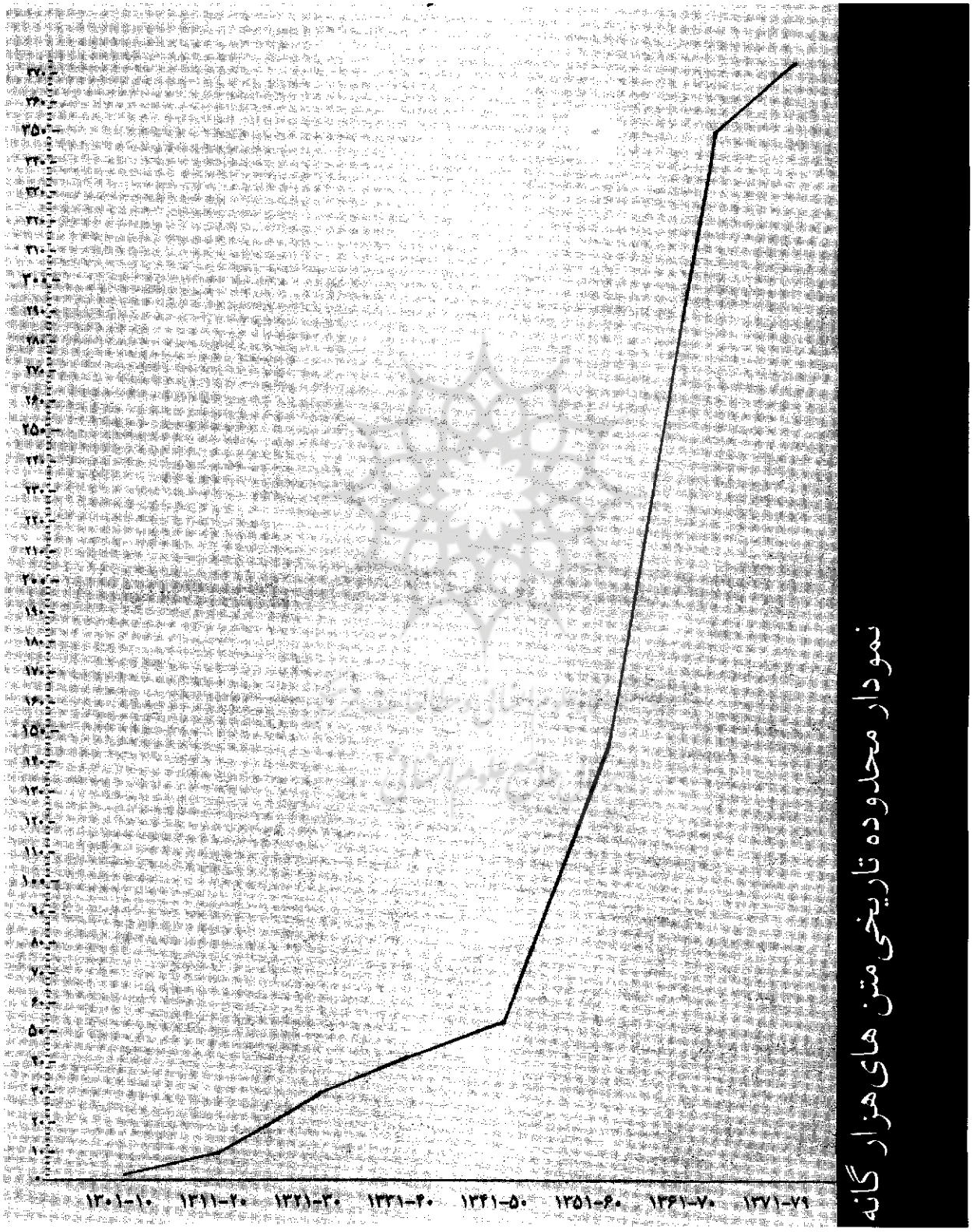
حال به طرح یک پرسش می‌پردازیم: اگر از دوستی خواهش کنیم که فهرستی از نام‌های پنج هزار شیء را که بارها آنها را دیده است و می‌بیند، برای ما تهیه کند، چگونه و از کجا باید آغاز کند؟ طبیعی‌ست که انتظار می‌رود در چنین فهرستی، لغات «آب» و «آینه» و «ابر» و «اشک» و «باران» و «برف» ثبت شده باشد، ولی دشواری کار گزینش، آن‌جا ظاهر می‌گردد که او به انتهای فهرست نزدیک شده باشد و مثلاً تنها پانصد لغت دیگر به اتمام آن مانده باشد. کافی است دقایقی به اطراف خود بنگرد، یا لحظه‌ای چند تأمل کند؛ از لوازم منزل، کدام‌ها را برگزیند و از ذکر کدام‌ها صرف‌نظر نماید؟ از آلات و ادوات موسیقی، کدام‌ها را انتخاب کند؟ نام چند گیاه و پرنده و حشره را بنگارد؟ آیا اگر فهرست پنج هزار لغتی به پایان برسد و لغاتی مثل «آسیاب»، «اکسیژن»، «بازو»، «پنبه»، «تسبیح»، «ثروت»، «جیب»، «چتر»، «حلقه»، «خرطوم»، «دست‌مال»، و چندین و چند لغت دیگر نانوشته مانده باشد، کدام لغت یا لغات ثبت شده را می‌تواند حذف کند تا لغات یاد شده را تک‌تک به جای آنها بگذارد؟ اصلاً معیار و ضابطهٔ حذف و افزودن لغات چیست؟

این مشکل برای شخصی چون فرهنگ‌نویس، دوچندان است؛ زیرا او علاوه بر نام‌های اشیا، لغاتی را که بر مفاهیم انتزاعی دلالت می‌کنند نیز باید ثبت کند. او چگونه می‌تواند از «آرامش»، «انتظار»، «بخت»، «پرسش»، «تصمیم»، و هزاران لغت دیگر، که همه روزه در روزنامه‌ها و سایر رسانه‌ها به کار برده می‌شوند و اهل زبان نیز در طول زندگی خود، بارها و بارها آنها را می‌شنوند و می‌گویند، چشم بپوشد؟ حتی اگر به جای پنج هزار لغت، پنجاه هزار یا صدهزار لغت را فهرست کند، باز هم نمی‌تواند با اطمینان ادعا کند که کار او دقیق و کم‌نقص است.

در **فرهنگ فارسی**، تألیف شادروان استاد دکتر محمد معین، بسیاری از عنوان‌های اصلی، صرفاً از متون قدیم استخراج شده‌اند و امروزه دیگر کاربردی ندارند، ولی همین انتخاب به شکلی بسیار ناقص و حتی غیرعلمی صورت گرفته است؛ چه لغات و ترکیب‌های فراوانی در آثار معروف گذشتگان (مانند **کلیله و دمنه**، **گلستان**، **تاریخ جهانگشای**، **دیوان صائب**، و بسیاری از کتاب‌های دیگر) می‌توان یافت که در فرهنگ یاد شده ثبت نشده‌اند. حتی در اثر عظیم اما ناتمام **لغت‌نامهٔ فارسی**، که در «سازمان لغت‌نامهٔ دهخدا» تدوین می‌شود نیز چنین کاستی‌هایی را ـ به خصوص در بخش ترکیب‌های آن ـ می‌توان مشاهده کرد[1].

نگارنده از سال‌ها پیش، این آرزو را در سر داشت که بتواند با استقصا و بسامدگیری دقیق از متون مختلف فارسی در موضوعات گوناگون، فهرستی از لغات پایهٔ این زبان فراهم آورد. سرانجام در اسفندماه ۱۳۷۸، به‌طور جدی تصمیم گرفت که به عنوان مقدمه‌ای بر طرح اصلی، یک پیکرهٔ صدهزار لغتی تهیه کند و نتیجهٔ آن را به صورت گزارشی مختصر ارائه کند. تهیهٔ این فهرست در دی ماه ۱۳۷۹ به اتمام رسید.

فهرست تهیه شده، براساس نمونه‌گیری از هزار متن فارسی معاصر (کتاب، مجله، و روزنامه، شامل حدود ۹۰۲۰۰ لغت، به علاوهٔ تکمله‌ای ۱۰۰۰۰ لغتی) با موضوعات گوناگون تهیه شده است. مراحل و روش کار به این ترتیب بوده است:



۱.انتخاب ده سطر اول نخستین بند صفحهٔ صد از کتاب‌ها، و ده سطر اول نخستین بند صفحهٔ ده از نشریات، و سپردن آنها به حافظهٔ کامپیوتر (محدودهٔ تاریخی متن‌ها از ۱۳۰۱ تا ۱۳۷۹ هـ. ش است.)

۲.استخراج لغات

۳.جدا کردن اعلام (اسم‌های خاص)

۴. جدا کردن لغاتی که در عبارت‌های عربی (مانند آیات و احادیث) به کار رفته‌اند.

۵. جدا کردن لغات و عباراتی که در متن‌ها با حروف لاتین یا یونانی ثبت شده‌اند.

۶. لماتیزه کردن (لمابندی) (lemmatize/lemmatiser) مواد، یعنی مرتب کردن آنها به صورتی که هر یک به عنوان یک «واحد واژگانی» (lexical unit) قابل ثبت در فرهنگ باشد (به استثنای انواع فعل‌های مرکب).

چنان که گفتیم، ده‌هزار سطر انتخاب شده جمعاً حدود ۹۰۲۰۰ لما (عنوان مستقل) را تشکیل می‌دهند. اما برای رساندن شمار لماها به ۱۰۰۰۰۰، از قسمتی از متن یکی از روزنامه‌های پرتیراژ کشور (۱۷ شماره از روزنامهٔ همشهری، ستون «گشتی در دنیای خبرها»، شامل نزدیک به ۱۰۰۰۰ لما) نیز استفاده شد، و به این ترتیب تعداد لماها از رقم صدهزار گذشت.

قابل ذکر است که بسیاری از مشکلات فنی کار، به لطف و یاری آقایان دکتر حسن انوری، دکتر محمد شادروی منش، و بهروز صفرزاده برطرف شد؛ آقای صفرزاده در بازخوانی نمونه‌ها نیز به من کمک کردند. مادرم نیز در مقابلهٔ بخشی از متن‌ها مرا یاری کردند. از همگی سپاسگزارم.

دوستان و همکاران عزیزی نیز با در اختیار گذاشتن منابع، با بنده همکاری کرده‌اند، که هر گاه متن کامل فهرست به چاپ برسد، نامشان به رسم امتنان ذکر خواهد شد.

چنان که در نمودار ملاحظه می‌شود، از میان ۱۰۰۰ متن مورد استفادهٔ ما، تنها ۲ متن از دههٔ نخست سدهٔ چهاردهم هـ. ش برگزیده شده است، یعنی ۰/۲٪ از کل ۱۰۰۰ متن. از دههٔ دوم نیز فقط ۸ متن انتخاب شده است (۰/۸٪ از ۱۰۰۰ متن)، از دههٔ بیست (دههٔ سوم) ۳۰ متن (۳/۰٪)، از دههٔ سی ۴۰ متن (۴/۰٪)، از دههٔ چهل ۵۳ متن (۵/۳٪)، از دههٔ پنجاه ۱۴۶ متن (۱۴/۶٪)، از دههٔ شصت ۳۴۸ متن (۳۴/۸٪)، و سرانجام از دههٔ هفتاد (کم‌تر از نه سال) ۳۷۳ متن (۳۷/۳٪) اختیار شده است.

پس بیش از ۷۲٪ از متون مورد استفاده متعلق به سال‌های ۱۳۶۱ و بعد از آن است، در حالی که تنها ۸٪ از منابع ما را متون چهار دههٔ نخستین سده (۱۳۰۱ـ۱۳۴۰) تشکیل می‌دهند.

ده لغتی که از لحاظ بسامد در صدر فهرست قرار دارند، عبارت‌اند از:

و (۵۲۱۰ بار) (۵/۲۱٪ از ۱۰۰۰۰۰)
بودن (۲۸۵۱) (۲/۸۵٪)
در (حرف اضافه) (۲۴۸۳) (۲/۴۸۳٪)
که (حرف ربط و موصول) (۲۴۸۲) (۲/۴۸۲٪)
به (۲۲۶۲) (۲/۲۶٪)
از (۲۱۳۷) (۲/۱۴٪)
را (۲۱۲۷) (۲/۱۳٪)
کردن (۱۸۴۰) (۱/۸۴٪)
شدن (۱۵۳۶) (۱/۵۴٪)
این/اینها (اسم اشاره و ضمیر اشاره) (۱۵۰۹) (۱/۵۱٪)[۲]

در میان نام‌های خاص، این ده نام در صدر هستند:
ایران/ایران زمین (۱۳۱ بار)
آمریکا/امریکا/ایالات متحدهٔ آمریکا (۵۶)
روسیه/روسیهٔ شوروی/شوروی/اتحاد شوروی/اتحاد جماهیر شوروی (۴۸)
انگلستان/انگلیس/بریتانیا (۴۰)
تهران/طهران (۳۹)
چین (۳۲)
محمد/پیامبر/پیامبر اسلام/پیغمبر/رسول/رسول اکرم/رسول خدا (۳۰)
ژاپن (۲۵)
استرالیا (۲۳)
اروپا (۲۲)

آمارهایی دیگر:

پرکاربردترین لغات عربی:
اما (۲۴۴ بار)
قرار (۱۶۵)
ولی (۱۵۸)
وجود (۱۳۱)
نظر (۱۱۲)
کتاب (۱۱۰)
انسان (۹۸)
استفاده (۹۶)
حرف (۸۶)
صدا (۸۳)

پرکاربردترین لغات معرب:
موسیقی (۵۵ بار)
فلسفه (۳۳)
قالب (۲۷)
کلیسا (۲۰)
آدم (۱۹)
الف (۱۸)
قند (۱۰)
اقیانوس (۷)
قهرمان (۷)

پرکاربردترین لغات اروپایی:
فیلم (۷۲ بار)
سینما (۶۵)
دکتر (۳۹)
پلیس (۳۰)
دلار (۲۹)
سیستم (۲۸)
اتومبیل (۲۵)
متر (۲۵)
تیم (و تیم ملی) (۲۴)

مدل (۲۳)

پرکاربردترین لغات ترکی:
اتاق/اطاق (۴۰ بار)
کمک (۳۸)
جلو (اسم، صفت، و قید) (۲۶)
توپ (۱۸)
تومان (۱۲)
قایق (۱۲)
اردو (۸)
پرچم (۸)

پرکاربردترین لغات مغولی:
آقا (۸۲ بار)
خانم (۳۹)
خان (۲۴)
میز (۱۰)

پرکاربردترین لغات اسپانیایی:
سیگار (۱۱ بار)
ریال (۷)

پرکاربردترین لغت آرامی ـ سریانی:
کشیش (۱۰ بار)

پرکاربردترین لغات یونانی، هلندی، هندی و چینی، به ترتیب عبارت‌اند از:
پول (۳۴ بار)، دلار (۲۹ بار)، چاپ (۲۴ بار)، و چای/چایی (۱۸بار)

اینک آمارهایی دیگر، که به لحاظ دستوری قابل اهمیت‌اند:

پرکاربردترین اسم‌ها (شامل اسم، اسم مصدر، و حاصل مصدر)[۲]:
سال (۳۵۵ بار)
کار (۲۶۸)
دست (۲۰۳)
کشور (۱۸۹)
قرار (۱۶۵)
زبان (۱۳۱)
وجود (۱۳۱)
شهر (۱۲۶)
زندگی (۱۱۷)
مرد (۱۱۷)

صفت‌ها:
چند (۱۶۰)
هر (۱۴۱)
بزرگ (۱۳۷)
چنین (۸۳)
تمام (۸۲)
همان (۸۰)
جدید (۷۴)
کوچک (۶۸)
همین (۶۶)
ممکن (۶۲)

قیدها:
وقتی (۸۰)
هنوز (۵۹)
چرا (۵۴)
فقط (۵۴)
آیا (۵۱)
البته (۴۷)
کاملاً (۴۰)
سپس (۳۳)
چه‌گونه (۳۲)
هنگامی که (۳۱)

ضمایر:
او (۵۳۹)
ما (۲۱۹)
تو (۱۳۵)
شما (۱۱۷)
کس (۹۵)
وی (۵۵)
یک‌دیگر (۳۶)
ایشان (۳۰)
هر چه (۲۱)
هر یک (۱۹)

حروف اضافه:
در (۲۴۹۸)
به (۲۲۶۲)
از (۲۱۳۷)
با (۱۰۰۴)
برای (۶۱۹)
بر (۲۵۳)
پس از (۱۴۴)
روی (۱۲۸)
دربارهٔ (۹۴)
مانند (۶۶)

حروف ربط:
یا (۳۷۶)
هم (۳۶۹)
اما (۲۴۴)
اگر (۱۹۴)
نیز (۱۸۴)
ولی (۱۵۸)
یعنی (۶۸)
حتی (۶۴)

بلکه (۵۸)
زیرا (۵۷)

حرف عطف:
و (۵۲۱۰)

حرف ندا:
ای (۲۴)

حرف نشانه:
را (۲۱۲۷)
ـو [= را] (۱۵)
رو [ro، = را] (۷)

مصدرهای بسیط، هم‌کردها، و فعل‌های ربطی:
بودن (۲۸۵۱)
کردن (۱۸۴۰)
شدن (۱۵۳۶)
داشتن (۸۵۸)
دادن (۵۳۴)
گفتن (۴۹۲)
توانستن (۳۴۷)
خواستن (۳۱۸)
گرفتن (۳۱۰)
رفتن (۲۷۹)

مصدرهای پیشوندی:
برداشتن (۴۶)
درآوردن (۳۵)
برگشتن (۳۱)
درآمدن (۳۰)
بازگشتن (۲۱)
برخاستن (۱۴)
دریافتن (۱۱)
فرو رفتن (۱۱)
برگرداندن/برگردانیدن (۸)
فرا گرفتن (۸)

صفت‌های شمارشی:
یک (۸۸۵ بار)
دو (۲۸۰)
سه (۸۷)
چهار (۶۲)
ده (۵۳)
پنج (۵۲)
شش (۴۰)
هفت (۳۵)
بیست (۳۰)
صد (۲۸)

صفت‌های ترتیبی:
اول (۵۸)
دوم (۴۲)
نخستین (۴۰)
اولین (۳۷)
سوم (۲۶)
چهارم (۱۸)
دومین (۱۴)
پنجم (۱۳)
هشتم (۸)
ششم (۸)

اینک فهرستی مقایسه‌ای میان لغات مربوط به مردان و زنان:
مرد (۱۱۷ بار)، زن (۱۱۴)[۴]
پدر (۹۶)، مادر(۶۲)
بابا (۱۴)، ماما/مامان (۴)
آقا (۸۲)، خانم (۳۹)
پسر (۴۴)، دختر (۵۱)
پسربچه (۵)، دختر بچه (۰)
پسرک (۳)، دخترک (۵)
برادر (۳۷)، خواهر (۱۳)
عمو (۴)، عمه (۵)
دایی (۴)، خاله (۳)
پدر بزرگ/بابابزرگ (۶)، مادر بزرگ (۴)
پدر شوهر (۰)، پدر زن (۱)
مادر شوهر (۲)، مادر زن(۱)
داماد (۰)، عروس (۲)

و اما از میان رنگ‌ها، اینها پرکاربردتر از همه‌اند:

سیاه/سیه (۲۸ بار)
سبز (۱۶)
سفید/سپید (۱۶)
قرمز/سرخ (۱۴)

پانوشتها:

۱ـ خوب است اشاره شود که آخرین ویرایش یکی از فرهنگ‌های یک جلدی بسیار معروف و معتبر دانشگاه آکسفورد، بر مبنای پیکره‌ای زبانی با رقم شگفت‌انگیز یک صد و چهل میلیون لغت (یک صدمیلیون لغت از متون بریتانیایی و چهل میلیون لغت از متون آمریکایی) تألیف شده است:

Oxford Advanced Learner's Dictionary of Current English, Great Britain, Oxford University Press, 5th Edition, 1995 (11th impression, with corrections, 1999)

۲ـ ده لغت اول، جمعاً ۲۴۴۳۷ بار به کار رفته‌اند، که بر روی هم حدود، ۲۴/۴۴٪ از کل متن‌ها را تشکیل می‌دهند.

۳ـ شامل لغاتی که طبقهٔ دستوری آنها لغزان است، نمی‌شود (یعنی لغاتی که ممکن است هم اسم باشند و هم صفت یا قید، و یا دارای سایر هویت‌های دستوری).

۴ـ شامل «زوجه» نیز هست.