ضرورت تألیف «فرهنگ جامع زبان فارسی»

دکتر علی اشرف صادقی
استاد دانشگاه تهران و عضو پیوستهٔ فرهنگستان زبان و ادب فارسی و
مدیرگروه فرهنگ‌نویسی و مجری اصلی طرح فرهنگ جامع زبان فارسی

پیش‌درآمد

توسعهٔ هر کشوری به یک زبان علمی استوار و زایا نیازمند است که بتواند برای همهٔ نیازهای فناوری، اقتصادی، سیاسی و فرهنگی، واژگان و تعابیر روشن و صحیح فراهم آورد. ارزش زبان فارسی آنگاه جلوهٔ حقیقی خود را می‌نمایاند که به زبان علم تبدیل گردد و طبیعی است که زبان علمی قوی با توسعهٔ کشور کاملاً مرتبط، بلکه از ارکان آن است.

نخستین گام برای تبدیل زبان فارسی به زبان محکم و زایا، تألیف فرهنگی جامع است؛ به‌شیوهٔ علمی که به زبان فارسی تمرکز و جهتی یگانه بخشد و شکل معیار آن را معین سازد و یک‌بار، برای همیشه، داده‌های واژگانی زبان فارسی را در یک مجموعهٔ منسجم و قابل گسترش در مراحل بعدی گرد آورد.

فرهنگ‌نویسی در ایران دارای پیشینه‌ای طولانی است. کهن‌ترین فرهنگی که از ایرانیان باقی مانده، فرهنگی است اوستایی- پهلوی، به‌نام *اویم- ایوک (Oīm-ēwak)* که برخی از محقّقان با استناد به شواهد و قراینی، اصل آن را به دورهٔ پیش از عصر ساسانی رسانده‌اند. پس از رواج زبان

فارسی دری به‌عنوان زبان نوشتار، از قرن سوم هجری به بعد و ظهور آثار برجستهٔ منثور و منظوم فارسی، تدوین فرهنگهای فارسی به فارسی نیز آغاز شد. ظاهراً *فرهنگ قطران تبریزی*، شاعر فارسی‌گوی قرن پنجم هجری، نخستین اثری است از این دست که اطلاعاتی از آن به‌دست ما رسیده است. اما کهن‌ترین فرهنگی که به‌دست ما رسیده، *لغت فرس* اسدی طوسی، شاعر و لغوی نامدار قرن پنجم هجری است. این فرهنگ مأخذ اصلی غالب فرهنگهایی بوده است که از آن پس در پهنهٔ جغرافیایی زبان فارسی، از آسیای صغیر گرفته تا هندوستان، تألیف شده‌اند.

فرهنگ‌نویسی در کشورهای پیشرفته، جریانی مستمر بوده که از دهها و بلکه صدها سال پیش آغاز گشته است. مثلاً در کشور فرانسه در اواسط قرن نوزدهم میلادی، یعنی حدود یک و نیم قرن پیش، فرهنگ‌نویسی به‌نام پیر لاروس (۱۸۷۵– ۱۸۱۷م) پدید آمد و فرهنگی را بنیاد نهاد که تا امروز به‌طور مستمر به حیات خود ادامه داده و همواره در مسیر ترقّی و پیشرفت بوده و همگام با تحوّلات، متحوّل گشته است، به‌طوری‌که طیف وسیع فرهنگهایی که امروز همچنان با نام *لاروس* منتشر می‌شوند، به هیچ‌وجه با فرهنگی که پیر لاروس در اواسط قرن نوزدهم چاپ کرد، قابل مقایسه نیستند. این امر مدیون وجود مؤسسه‌ای است به‌نام لاروس که دهها سال است به کار فرهنگ‌نویسی اشتغال دارد و پیوسته در حال بازبینی و روزآمد کردن انواع فرهنگهای ریز و درشت خود است.

متأسفانه، سنت فرهنگ‌نویسی در ایران به‌هیچ روی، چنین استمراری نداشته است، به‌طوری‌که مثلاً *لغت‌نامه*، اثر عظیم شادروان علی‌اکبر دهخدا و یارانش، پس از گذشت دهها سال همچنان به طریق افست و بدون هرگونه بازبینی و ویرایشی تجدید چاپ می‌شود. همچنین است وضع *فرهنگ فارسی* شادروان دکتر محمّد معین.

این وضع معلول نبودِ یک مؤسسهٔ فرهنگ‌نگاری پویا در کشورمان است. پس از تألیف *لغت‌نامه*، از یک‌سو متون فارسی بسیاری تصحیح و به چاپ رسیده و از سوی دیگر دانش فرهنگ‌نگاری، پیشرفتهای شگرفی داشته است. ازاین‌رو، امروزه دشوار می‌توان *لغت‌نامه* را پایهٔ فرهنگی جامع در نظر گرفت و آن را با منابع جدید و شیوه‌های نوین فرهنگ‌نویسی مبتنی بر دانش زبان‌شناسی تکمیل کرد و به چاپ رساند. بنابراین جا دارد هرچه زودتر یک مؤسسهٔ

فرهنگ‌نگاری پویا در کشورمان پایه‌گذاری شود تا کشور ما نیز همچون کشورهای پیشرفته، صاحب مؤسسه‌ای گردد که کار اصلی آن تألیف فرهنگی جامع برای زبان فارسی و در مرحلهٔ بعدی تدوین انواع فرهنگها با توجه به نیازهای گوناگون جامعه باشد. بی‌گمان، مناسب‌ترین گزینه برای اجرای چنین طرحهای عظیمی، فرهنگستان زبان و ادب فارسی است.

تشکیل گروه فرهنگ‌نویسی در فرهنگستان زبان و ادب فارسی و اقدامات آن

فکر تدوین *فرهنگ جامع زبان فارسی* در سال ۱۳۷۵ در فرهنگستان زبان شکل گرفت و از آن زمان به بعد با تشکیل جلسات متعدد، شیوه‌های مختلفی برای اجرای این طرح عظیم به بحث گذاشته شد و سرانجام قرار شد به‌جای استفاده از روش معمول سنتی، یعنی برگه‌نویسی و تعریف واحدهای واژگانی (LU=lexical unit) براساس برگه‌ها، نخست پیکرهٔ (Corpus) متونِ فارسی در رایانه تشکیل شود و براساس آن واحدهای واژگانی تعریف و فرهنگ تألیف گردد.

در اواخر سال ۱۳۷۷ طرح تألیف *فرهنگ جامع زبان فارسی*، به‌عنوان طرح ملی، به تصویب شورای پژوهشهای علمی کشور رسید که به شرح زیر تعریف و اهداف آن معیّن شد:

تعریف طرح: فرهنگ جامع بر مبنای یک بانک اطلاعاتی و واژگانی. این فرهنگ، حاوی کلیهٔ واژه‌هایی خواهد بود که در آثار مکتوب زبان فارسی- از قدیمی‌ترین نوشته‌ها تا زمان حاضر- به‌کار رفته‌اند و اطلاعات مختلفی را دربارهٔ هریک از آنها (از نظر تلفّظ، نوع دستوری، معانی، ریشه تاریخی و جز آنها) به‌دست می‌دهد.

اهداف و دستاوردهای طرح: هدف از تدوین این فرهنگ، ارائهٔ تصویری همه‌جانبه از واژگان زبان فارسی در طول تاریخ تکوین این زبان و ایجاد امکان دست‌یابی به انواع اطلاعات واژگانی آن است. این فرهنگ مأخذی خواهد بود برای هرگونه مطالعه‌ای درزمینهٔ واژه‌های زبان فارسی و می‌تواند منشأ تألیف یک رشته فرهنگهای خاص، قرار گیرد.

شورای علمی گروه فرهنگ‌نویسی، با پشتوانهٔ اعتباری که شورای پژوهشهای علمی کشور برای اجرای این طرح درنظر گرفته بود چگونگی تشکیل پیکرهٔ متون زبان فارسی را بررسی کرد و قرار شد به کمک دستگاه پویشگر (Scanner) از ۵۰۰ متن تعیین شده، تصویر گرفته شود و با

استفاده از نرم‌افزار OCR متن تحلیل شود و واحدهای واژگانی در دسترس قرار گیرند و فرهنگ جامع براساس شواهد متون تألیف گردد.

پس از مشورتهای مجری اصلی و همکاران طرح با متخصصان رایانه و بازدید از چند شرکت رایانه‌ای، که سابقهٔ کار در چنین زمینه‌هایی را دارا بودند، مشخصات نرم‌افزار فرهنگ جامع مدوّن شد که مهم‌ترین آنها عبارتند از:

۱. امکان ذخیره‌سازی متون تایپ‌شده، امکان عملیات ویرایشی و درنتیجه داشتن بانک اطلاعات لغات و متون زبان فارسی.

۲. امکان مقابله و پردازش اطلاعات و ثبت اطلاعات تکمیلی در پیکره.

۳. امکان جستجو در متن (متن اصلی و حواشی) و نیز در بانک واژگان به کمک موتور جستجو.

٤. امکان اجرای این نرم‌افزار به صورتهای مختلف از جمله بر روی شبکهٔ داخلی، لوح‌های فشردهٔ (CD) مستقل و اینترنت.

اما از نیمهٔ سال ۱۳۷۹ش شورای علمی فرهنگ‌نویسی به این نتیجه رسید که شیوهٔ تدوین فرهنگ جامع را تغییر دهد؛ زیرا تشکیل پیکرهٔ متون زبان فارسی در رایانه و تدوین فرهنگ جامع براساس آن، نیازمند اعتبار مالی گزافی بود که با اعتبار اختصاص‌یافته از سوی سازمان پژوهشهای علمی کشور و نیز بودجه‌ای که فرهنگستان زبان در اختیار گروه فرهنگ‌نویسی قرار داده بود، تناسب نداشت. به همین علّت و نیز به‌سبب برخی مشکلات اداری و فنی و مبهم بودن دورنمای کار رایانه‌ای، تصمیم گرفته شد که کار تدوین فرهنگ جامع با استفاده از روش مرسوم و سنتی سریعاً آغاز شود. با این‌همه، قرار شد تشکیل پیکرهٔ متون زبان فارسی در رایانه همچنان در دستور کار گروه فرهنگ‌نویسی قرار داشته باشد تا این‌کار مهم نیز در زمان مقتضی تحقّق پذیرد.

شایان ذکر است که در یکی دو سال گذشته قرار بود این طرح در «شورای گسترش زبان فارسی» به مرحلهٔ اجرا درآید و گامهای نخستین آن نیز برداشته‌شد، ولی به علل نامعلوم اجرای آن ظاهراً متوقّف شده است. بسیاری از کشورهای پیشرفته، به‌ویژه آمریکا و اروپا، پیکرهٔ کاملی از مجموعهٔ میراث ادبی خود را در رایانه دارند؛ حتی پیکره‌ای رایانه‌ای از غالب متون عربی نیز تشکیل شده و سالهاست که به‌صورت رایگان در اینترنت در دسترس پژوهشگران است

(وب‌گاه www.alwaraq.com)، ولی تا این زمان، مسئولان فرهنگی مملکت ما گویا هنوز به اهمّیت تشکیل پیکرهٔ متون زبان فارسی در رایانه پی نبرده‌اند.

تدوین فرهنگ جامع زبان فارسی به‌شیوهٔ سنتی معمول، دارای سه مرحلهٔ عمدهٔ زیر است:

۱. انتخاب متون کهن زبان فارسی و گزینش واحدهای واژگانی (کلمات، ترکیبات، اصطلاحات) آنها: گزینش همهٔ واحدهای واژگانی متون و شواهد و برگه‌نویسی آنها از یک‌سو بسیار وقت‌گیر و پرهزینه بود و از سوی دیگر تعریف‌نگار با انبوهی از شواهد غیرضروری مواجه می‌شد. ازاین‌رو، شورای علمی‌گروه فرهنگ‌نویسی در جلسات متعدّد، ضوابط ویژه‌ای را برای گزینش واحدهای واژگانی درنظر گرفت که مهم‌ترین آنها تشخّص داشتن و نامتعارف بودن واحد واژگانی بود؛ بدین معنی که پژوهشگران، آن دسته از واحدهای واژگانی‌ای را گزینش می‌کردند که از نظر تلفّظ، معنی، املا یا نقش دستوری نسبت به زبان فارسی معیار امروزی متمایز باشند. برای دسترسیِ تعریف‌نگاران به واحدهای واژگانی دیگر قرار شد، همهٔ واحدهای واژگانی- چه متعارف و چه نامتعارف- برخی از آثار مهم زبان فارسی در حوزه‌های گوناگون و قرنهای مختلف گزینش شوند. از سوی دیگر از مدّتها پیش، همهٔ مداخل اصلی و فرعی برخی از فرهنگهای معتبر فارسی مانند *آنندراج*، *فرهنگ نفیسی*، *لغت‌نامهٔ دهخدا*، *فرهنگ معین* و *فرهنگ عامیانه* ابوالحسن نجفی تایپ شده بود که مجموعهٔ این مداخل به‌علاوهٔ مداخلی که از متون استخراج خواهند شد، مداخل فرهنگ جامع را تشکیل خواهند داد.

۲. برگه‌نویسی واحدهای واژگانی: پژوهشگران اطلاعات زیر را به‌ترتیب بر روی برگه می‌نویسند: سرمدخل، زیرمدخل، شاهد، نام کتاب، سال تألیف کتاب، شمارهٔ جلد و صفحه.

۳. تعریف واحدهای واژگانی براساس برگه‌ها.

تغییر شیوهٔ تألیف فرهنگ جامع از روش سنتی برگه‌نویسی به برگه‌نویسی رایانه‌ای

گروه فرهنگ‌نویسی با پشتوانهٔ اعتبار سازمان پژوهشهای علمی کشور و بودجهٔ جاری فرهنگستان تا تابستان ۱۳۸۲ش توانست گزینش واحدهای واژگانی از حدود ۱۷٦ متن زبان فارسی به تعداد حدود ۸٤٦۰۰ صفحه را به پایان برد و از آن میان ۱۰۸ متن را برگه‌نویسی کند و در مجموع

حدود ۹۴۳۰۰۰ برگه فراهم آورد. امّا از اوایل همین سال، شورای فرهنگ‌نویسی به این نتیجه رسید که برای امنیت بیشتر برگه‌ها و سهولت و سرعت در کارهای مقدماتی فرهنگ جامع، به‌ویژه الفبایی کردن برگه‌ها، از رایانه استفاده شود. به همین منظور، بخش رایانه در گروه فرهنگ‌نویسی متشکّل از چهار حروف‌نگار و یک متخصص رایانه به‌عنوان سرپرست تشکیل شد و قرار شد تا همهٔ برگه‌هایی که تاکنون فراهم شده است، با استفاده از برنامهٔ Excel وارد رایانه شوند. افزون بر این، همهٔ پژوهشگرانی که تا آن زمان به‌صورت دستی برگه‌نویسی می‌کردند، زیر نظر بخش رایانه آموزش دیدند و از آن پس واحدهای واژگانی متون و شواهد و ارجاع آنها را بی‌واسطهٔ برگه، مستقیماً از متن وارد رایانه می‌کنند.

تاکنون واحدهای واژگانی ۲۹۷ عنوان (۴۱۳ جلد) گزینش شده و واحدهای واژگانی ۱۸۳ کتاب (در ۲۲۹ جلد) به تعداد ۱۷۰۰۰۰۰ واحد وارد رایانه شده است.

از سوی دیگر، با توجه به اینکه هر شاهد می‌تواند شاهدی برای همهٔ واژه‌های خود نیز باشد و به‌طور میانگین هر شاهد حاوی حدود ۲۰ واژه است، پیکره‌ای که تاکنون تشکیل شده ۳۴۰۰۰۰۰۰ واحد واژگانی و شواهد مربوط به آن را دربردارد.

قابلیتهای پیکرهٔ واحدهای واژگانی و شواهد آنها در رایانه برای پژوهشهای گوناگون

گروه فرهنگ‌نویسی برای نخستین بار پیکره‌ای از واحدهای واژگانی متون و شواهد و ارجاع آنها در رایانه فراهم آورده است. این پیکره گذشته از آنکه منبعی غنی برای تألیف *فرهنگ جامع زبان فارسی* است، راه را برای هرگونه جستجوی واحدهای واژگانی، الفاظ، اشخاص، مکانهای جغرافیایی، نام مذاهب و فرقه‌ها، اصطلاحات و تعبیرات شاخه‌ها و رشته‌های گوناگون (نجوم، پزشکی، گیاه‌شناسی، تصوف و جز آنها) در مداخل و شواهد هموار کرده است. برخی از مهم‌ترین قابلیتهای پیکره به شرح زیر است:

۱. قابلیت طبقه‌بندی و جستجوی ترکیبی در همهٔ سرمدخلها، زیرمدخلها، شواهد، نام کتابها و تاریخ تألیف آنها؛ مثلاً کاربرد واژه «تیغ» در متون قرن پنجم هجری.

۲. قابلیت مرتب‌سازی همهٔ سرمدخلها، زیرمدخلها و شواهد براساس تاریخ تألیف کتابها یا حروف الفبا؛ مثلاً کاربرد واژهٔ «استوار» از متون قرن چهارم تا دوازدهم هجری.

۳. قابلیت جستجو و جایگزینی هرگونه حرف، واژه، عبارت و جمله در سرمدخلها، زیرمدخلها و متن شواهد منثور و منظوم.

٤. بررسی تغییر و تحوّلات واژه‌ها به لحاظ معنایی در طول تاریخ زبان فارسی؛ مثلاً تحول معنایی واژهٔ «زخم».

۵. تعیین تاریخ تقریبی ورود برخی واژه‌ها به زبان فارسی و احیاناً خروج آنها.

٦. قابلیت جستجو، طبقه‌بندی و استخراج پیشوندها، پسوندها، میانوندها و هرگونه عنصر واژگانی و دستوری؛ مثلاً واژه‌هایی که پسوند «ناک» دارند.

۷. قابلیت جستجو و استخراج همهٔ مصادر بسیط و مرکب فارسی. براساس این پیکره می‌توان میزان کاربرد افعال بسیط و مرکب را با یکدیگر سنجید.

۸. قابلیت طبقه‌بندی تاریخی سرمدخلها، زیرمدخلها و شواهد منثور و منظوم در هر دورهٔ زمانی؛ مانند: طبقه‌بندی بین دو سال مختلف، طبقه‌بندی از سال موردنظر به بعد، طبقه‌بندی پیش از سال موردنظر، طبقه‌بندی از قرن موردنظر به بعد، طبقه‌بندی قبل از قرن مورد نظر.

۹. قابلیت هرگونه تحلیل آماری و بسامدی واژه‌ها و عبارتها با طبقه‌بندیهای مختلف زمانی و غیره.

۱۰. قابلیت پژوهش در نام اشخاص؛ مثلاً کاربرد نام «سلیمان» در متون مختلف.

۱۱. قابلیت پژوهش در نامهای جغرافیایی به‌کار رفته در متون؛ مانند کاربرد شهر «همدان».

۱۲. قابلیت پژوهش در نامهای اقوام مختلف به‌کار رفته در متون؛ مثلاً کاربرد قوم «تازی» و فرقهٔ «زندیق» در متون.

۱۳. بررسی هم‌نشینی واژه‌ها با یکدیگر در متون فارسی و از این لحاظ جستجو در پیکره می‌تواند به مصححان متون فارسی کمک شایانی کند؛ مثلاً به بیت زیر از *شاهنامه* توجه فرمایید:

پرستنده پنجاه و خادم چهـل بر او برگذشتند **شادان** به **دل**

در مصراع دوم، بیشتر نسخه‌های *شاهنامه* ضبط «شاداب‌دل» دارند و مصحح به‌راستی در گزینش **شادان به دل** و **شاداب‌دل** تردید می‌کند، ولی با مراجعه به پیکرهٔ متون می‌تواند با اطمینان

بیشتری ضبط درست اقلیت نسخه‌ها یعنی **شادان** به **دل** را انتخاب کند؛ زیرا در این پیکره همه‌جا در متون کهن **شادان** با **دل** آمده است نه **شاداب** با **دل**.

۱۴. محقّقان فرهنگ مردم (فولکلور) می‌توانند در پژوهشهای خود از این پیکره به‌خوبی بهره برند؛ مثلاً می‌توانند به سادگی ارتباط میان «چشم زخم» و «اسپند» را در متون منثور و منظوم فارسی دریابند.

۱۵. بررسی نقش دستوری برخی واژه‌ها در طول تاریخ زبان فارسی؛ مثلاً فعل مرکب «تغییر کردن» در متون کهن متعدّی است و در دوران متأخّر به فعل لازم تبدیل شده است.

۱۶. قابلیت جستجوی همهٔ واژه‌های بیگانه و تعیین تقریبی زمان ورود آنها به زبان فارسی و احیاناً خروج آنها. در این پیکره همهٔ واژه‌های بیگانه با * مشخص شده‌اند.

۱۷. امکان آگاهی از افزایش یا کاهش بسامد کاربرد واژه‌ها در طول تاریخ زبان فارسی، مثلاً این پیکره نشان می‌دهد که واژهٔ پرکاربرد «ایراد» به‌معنی خرده‌گیری در فارسی معاصر، در متون کهن کمتر کاربرد داشته است، و یا با استناد به پیکره می‌توان گفت که واژهٔ پرکاربرد «قدردان» در فارسی معاصر، در گذشته چندان کاربرد نداشته است.

۱۸. جستجوی یک جمله، عبارت، بیت، مصراع یا بخشی از مصراع برای یافتن عنوان کتاب یا نام نویسنده یا گویندهٔ آن.

۱۹. امکان کشف اقتباسها و سرقتهای ادبی. همچنین براساس این پیکره می‌توان به سادگی بیتها یا اشعاری را یافت که به دو یا چند شاعر مختلف منسوب‌اند؛ مثلاً یک بیت در یک منبع به **لبیبی** منسوب شده و در منبع دیگر به **زینبی**. هم‌چنین از مقایسهٔ این دو بیت و تأمّل در اختلافهای میان آنها می‌توان از ضبط درست «رباب‌زنی» در بیت منسوب به زینبی به‌جای «سماع زمانی» و در بیت منسوب به لبیبی آگاهی یافت.

۲۰. امکان پژوهش در آواشناسی زبان فارسی.

۲۱. قابلیت ایجاد و ورود هرگونه سیستم رمزگذاری (coding) به منظور طبقه‌بندی دستوری و صرفی واژه‌ها، ترکیبات و عبارات.

احتمال انتقال همهٔ اسناد فرهنگ فارسی ناتمام مرکز نشر دانشگاهی به فرهنگستان زبان

نخستین پایه‌های تدوین فرهنگ فارسی مرکز نشر دانشگاهی در سال ۱۳۶۵ گذاشته شد و تا اوایل دههٔ هفتاد به تهیهٔ چارچوب کار و مواد خام اوّلیه گذشت. پس از آن، کار تدوین فرهنگ و تعریف‌نگاری واژه‌های عمومی معاصر و نیز واژه‌های علمی و تخصصی به سرویراستاری مهندس علی کافی آغاز شد. پیکرهٔ واژه‌های عمومی براساس بیش از ۲۰۰ رمان و اثر ادبی از نویسندگانِ معاصر تشکیل شده است و برجسته‌ترین استادان و صاحب‌نظران هفتاد رشتهٔ علمی نیز زیر نظر سرویراستار فرهنگ، واژه‌های تخصصی مربوط به رشتهٔ خود را گزینش و تعریف کرده‌اند. گذشته از اینها، تعریف‌نگاری حرفه‌ها و پیشه‌ها و ریشه‌شناسی واژه‌های فارسی از دیگر کارهای انجام یافته در گروه فرهنگ فارسی مرکز نشر بوده است. تا اواسط سال ۱۳۸۳، که تدوین و تکمیل این فرهنگ بنا به دلایلی متوقّف شد، بیش از ۶۵ درصد از واژه‌های عمومی معاصر، واژه‌های علمی، تخصّصی، حرفه‌ها و پیشه‌ها تعریف‌نگاری شده و برخی از آنها ویرایش شده و شماری دیگر در مرحلهٔ ویراستاری است.

همان‌گونه که پیشتر گفته شد، گروه فرهنگ‌نویسی فرهنگستان، متون فارسی را از آغاز زبان فارسی تا سال ۱۳۰۰ش، در دستور کار خود دارد و قرار بود متون پس از این تاریخ نیز به‌تدریج وارد رایانه شوند؛ ولی با تلفیق فرهنگ ناتمام مرکز نشر با فرهنگ فرهنگستان، که در این مورد توافق اولیه بین این دو نهاد صورت گرفته است، تألیف فرهنگ جامع زبان فارسی با شتاب بیشتری به انجام خواهد رسید.

نتیجه

کسانی‌که در زبان فارسی تخصّص دارند، بر این باورند که همهٔ فرهنگهای فارسی که تاکنون تألیف یافته‌اند، مانند *لغت‌نامهٔ دهخدا*، *فرهنگ معین* و جدیدترین آنها *فرهنگ فارسی سخن* (به سرپرستی دکتر حسن انوری) از جنبه‌های مختلف ناقصند و کل واحدهای واژگانی زبان فارسی را از آغاز تا امروز دربرندارند؛ ولی گروه فرهنگ‌نویسی فرهنگستان برای تألیف فرهنگی جامع

برای زبان فارسی تاکنون گامهای استواری برداشته و شیوه‌های نوین و بی‌نظیری را برای اجرای آن به‌کار بسته است.

گروه فرهنگ‌نویسی از میان متون زبانی فارسی تاکنون واحدهای واژگانی ۲۹۷ متن (۴۱۳ جلد) را به پایان رسانده و واحدهای واژگانی ۱۸۳ متن (۲۲۹ جلد) به تعداد ۱۷۰۰۰۰۰ واحد را همراه با شواهد و ارجاع آنها وارد رایانه کرده است. از آنجا که هر شاهد می‌تواند شاهدی برای همهٔ واژه‌های خود نیز باشد و چنانچه به‌طور میانگین هر شاهد ۲۰ واژه را دربر داشته باشد، با حذف شواهد تکراری، پیکره‌ای که تاکنون تشکیل شده، مشتمل است بر حدود ۱۷۰۰۰۰۰۰ واحد واژگانی و شواهد مربوط به آن. هم‌اکنون این پیکره، تحول شگرفی را در تحقیقات مربوط به زبان و ادب فارسی پدید آورده و راه را برای انواع پژوهشها هموار کرده است.

گروه فرهنگ‌نویسی گزینش واحدهای واژگانی ۶۰۰ متن فارسی را، از آغاز رواج زبان فارسی تاکنون، و ورود آنها را به رایانه در دستور کار خود دارد و چنانچه اعتبار کافی برای این طرح در نظر گرفته شود، حداکثر تا پایان سال ۱۳۸۵ اجرای این طرح تکمیل می‌گردد. در آن صورت پیکره‌ای بالغ بر ۵۶۰۰۰۰۰ واحد واژگانی و شواهد آنها در رایانه فراهم می‌آید. از آنجا که هر شاهد می‌تواند شاهدی برای همه واژه‌های خود نیز باشد، و به‌طور میانگین هر شاهد ۲۰ واژه را دربردارد، با حذف شواهد تکراری این پیکره مشتمل خواهد بود بر حدود ۵۶۰۰۰۰۰۰ واحد واژگانی و شواهد و ارجاع آنها در رایانه. مجموعهٔ مذکور گذشته از آنکه مبنای تألیف فرهنگ جامع زبان فارسی قرار خواهد گرفت، می‌تواند به‌عنوان یک بانک اطلاعات بسیار غنی، آمادهٔ تهیه و تدوین هرگونه نرم‌افزاری برای پردازشهای مختلف باشد. این پیکره می‌تواند هم به‌صورت مستقل (لوح فشرده: CD) و هم در قالب یک پیکرهٔ زبانی در شبکه‌های جهانی در دسترس همگان قرار گیرد. شایان ذکر است که گروه فرهنگ‌نویسی از آغاز سال ۱۳۸۴ تعریف‌نگاری واحدهای واژگانی را همراه با ارائهٔ کار تکمیل پیکرهٔ واحدهای واژگانی آغاز خواهد کرد. پیش‌بینی می‌شود تعریف‌نگاری و تدوین و تألیف فرهنگ جامع زبان فارسی ۱۰ سال، از ۱۳۸۴ تا پایان ۱۳۹۳ به‌طول انجامد و در ۲۰ جلد منتشر شود.