اوّلین نمونه از دست‌آوردهای زبان‌شناسیِ رایانه‌ای در زبان فارسی

احمد طاهریان

زبانِ علم زبانی در درون زبان است. زبان علم آن گونهٔ کاربردی زبان فارسی است که در علم به کار می‌رود. علم شاخه‌های گوناگونی دارد و هر شاخه مجهز به دستگاه اصطلاحاتی و سنت‌های زبانی مختص به خود است. بنابر این، در هر زبان، نه با یک گونهٔ کاربردی برای علم به مفهوم مطلق، بلکه با گونه‌های کاربردی متعدّد برای علوم گوناگون سر و کار داریم، یعنی در هر زبان برای هر علمی زبانی داریم.

داشتن زبان علم از مهم‌ترین شرایط کسب توانایی در علوم است، زیرا زبان علم مهم‌ترین و نخستین ابزار یادگیری و کاربرد و گسترش آن است. بدون در دست داشتن زبان علم، تفاهم علمی و تبادل اطلاعات به آسانی میسر نیست. اما آنچه زبان‌های علوم متعدّد را از یک‌دیگر متمایز می‌سازد واژگان آنهاست. غنا، شفافیت و یک‌دستیِ مجموعهٔ واژگان و اصطلاحات و ترکیباتِ تخصصی در زبان علم، تفاهم علمی و انتقال سریع اطلاعات را سهل‌تر می‌کند. بدین لحاظ، تدوین واژه نامه‌ها و فرهنگ‌های تخصصی برای معیارسازیِ صورت و معنا و کاربرد واژگان گامی در جهت تقویت زبان علم است.

در زمینهٔ زبان‌شناسی تاکنون چند فرهنگ توصیفی و واژه‌نامه تدوین شده است. آخرین آنها واژگان گزیدهٔ زبان‌شناسی است که به دو صورتِ نرم‌افزار رایانه‌ای و چاپی ارائه

شده است. واژگان گزیده نخستین فرهنگ در زبان فارسی است که با استفاده از دست‌آوردهای زبان‌شناسی رایانه‌ای فراهم آمده است. مبنای نظری کار نظریهٔ متن‌شناسی مقابله‌ای است که بر اساس آن مهم‌ترین عامل در تهیهٔ فرهنگ بافت‌های موقعیتی و آن شرایطی است که فرهنگ عملاً در آنها به کار می‌رود. به سخن دیگر، شکل و محتوای فرهنگ را کاربران و شرایط کاربردی آنها مشخص می‌کند. گروه‌های کاربران نیز متنوع‌اند. از آن جمله‌اند: نویسندگانِ متون تخصصی، اصطلاح شناسان، روزنامه‌نگاران، واژه‌شناسان، کارشناسان، دانشجویان، مترجمان، معلمان و نیز زبان‌آموزانِ زبان دوم یا تخصصی. واژگان گزیده برای دانشجویان و مترجمان متون زبان‌شناسی تهیه شده است، اما از صورت نرم‌افزاری آن دیگر گروه‌های کاربردی نیز می‌توانند استفاده کنند.

گردآورندگان واژگان گزیده متخصصان زبان‌شناسی، فرهنگ‌نگاری و رایانه هستند. مقصود آنها از تدوین این واژه‌نامه کمک به کاربران در درک (در مقابل تولید) متونِ زبان‌شناسی است. بدین لحاظ، کارکرد این واژه‌نامه توصیفی و هنجارگذار است. زبان به کار رفته در واژگان گزیده زبان‌نوشتاری و رسمی است و همهٔ واژه‌ها، ترکیبات و اصطلاحاتی از علم زبان شناسی را در بر می‌گیرد که از منابع معتبر و شناخته شده استخراج شده‌اند. در تهیهٔ واژگان گزیده واژه‌سازی انجام نشده (به جز یک مورد در خارج از بدنهٔ اصلی واژه‌نامه) بلکه کلیهٔ برابرهای موجود در پیکرهٔ زبانی ارائه شده است. تنها در شکل چاپی، یکی از برابرها «گزیده» شده است.

پیکرهٔ زبانیِ واژگان گزیده شامل حدود ۱۵۰ مقاله، کتاب، ترجمه و مجموعهٔ مقالات است که مباحث گوناگون زبان‌شناسی شامل آواشناسی، واج‌شناسی، ساخت واژه، دستور، گویش‌شناسی، معناشناسی، روان‌شناسی زبان، جامعه‌شناسی زبان، گفتمان، تاریخ زبان‌شناسی، زبان‌شناسی رایانه‌ای، روش‌شناسی، زبان‌شناسی مقابله‌ای، زبان‌شناسی کاربردی، ترجمه، کارکرد شناسی، آموزش زبان، زبان‌های باستان و مسائل خط را شامل می‌شود. بدین ترتیب، به نظر می‌رسد که نمونه‌ای از هر شاخهٔ زبان‌شناسی در این پیکرهٔ زبانی درج شده است.

از این پیکره، با استفاده از رایانه و به روش بسامدی، بیش از ۴۸۰۰۰ واژهٔ انگلیسی و فارسیِ زبان‌شناختی استخراج شده است که حدود ۲۲۰۰۰ واژه از آن نامکررند و از این

تعداد ۹۳۰۰ واژهٔ انگلیسی و ۱۲۸۰۰ واژه برابرهای فارسی آنها هستند.

معیارسازی واژه‌های انگلیسی بیشتر به منظور یکسان کردن املای آنها و، در ضمن، برای حصول اطمینان از تخصصی بودن واژه‌ها در علم زبان‌شناسی با استفاده از چهار فرهنگ تخصصیِ زبان‌شناسی و عمومی انجام گرفته است.

تهیهٔ سیاهه‌ها، یعنی فهرست آماری و فهرست بسامدی واژه‌ها، از زبان انگلیسی به فارسی انجام گرفته زیرا فرض شده است که اصطلاحات پیکره در اصل به زبان انگلیسی بوده و در اینجا معادل‌های فارسی آنها آمده است. واژگان گزیده انتخاب معادل را به کاربر وا می‌گذارد. برای واژه‌ها، همهٔ برابرهای به کار رفته در پیکره همراه با منبع آنها آمده است. تنها در شکل چاپی، برای کاربرانِ خاص (دانشجویان و مترجمان)، یکی از برابرها بر اساس ملاک‌های ده‌گانه انتخاب شده است. به گفتهٔ گرد آورندگان، کوشش شده است تا، با توجه به این ملاک‌ها، مناسب‌ترین برابرها انتخاب شود. اما خود گردآورندگان واقف‌اند که این ملاک‌ها «نیاز به توضیح و احتمالاً بازنگری دارند» و هم انتخاب برابرها باید «به شکل گروهی» انجام گیرد.

کارهای پژوهشی اوّلیهٔ واژگان گزیده در سال‌های ۶۰-۱۳۵۷ انجام گرفته، اما مراحل مختلف تهیهٔ این واژه‌نامه در سال ۱۳۶۹ شروع شده است. طی سال‌های مذکور، کارهای برنامه‌ریزی، اجراء، آزمایش و تدوین نرم‌افزاری مناسب برای تحلیل و استخراج اطلاعات از داده‌های زبانی به انجام رسید. این نرم‌افزار با استفاده از بسته برنامهٔ Oxford Concordance Program[1] و نیز Snobol[2] تهیه شده است، اما نرم‌افزار ارائه شده به کاربران با زبانِ C[3] نوشته شده است.

واژگان گزیده اولین نمونه از الگویی است که یکی از گردآورندگان آن در سخن‌رانی خود در دومین کنفرانس زبان‌شناسی[4] ترسیم کرد. بدین لحاظ، می‌توان گفت که اهمیت واژگان گزیده در روش گردآوری و تدوین آن است نه در شکل و محتوای آن، با این تبصره

---

۱) برنامه‌ای برای تهیه فهرست واژه‌ها و نیز استخراج جملات آنها از متن‌های مختلف.

۲) String Oriented Symbolic Language؛ زبان برنامه‌نویسی سطح بالایی که در آن از روش‌های تحلیل خطی استفاده می‌شود.

۳) زبان برنامه‌نویسی سطح بالایی که بیشتر به منظور نوشتن برنامه‌های ساختار یافته به کار می‌رود.

۴) عاصی: ۱۳۷۲

که این واژه‌نامه اولین نمونه از کاری است که با استفاده از امکانات رایانه تهیه شده است.

از امکانات و روش‌های رایانه‌ای می‌توان برای مقاصد متعددی استفاده کرد. از برجسته‌ترین ویژگی‌های این روش گستردگی و تنوع پیکره است. گستردگی و تنوع پیکره در شکل‌های سنتی آن محدودیت‌های فراوانی به همراه دارد؛ زیرا اگر حجم پیکره از حد مشخصی بگذرد دیگر سازمان‌دهی و استفاده از آن مشکل و چه بسا ممتنع گردد، اما این مشکل با استفاده از امکانات رایانه حل شدنی است. همچنین، در صورت نیاز، روز آمد کردن پیکره به سرعت و سهولت انجام می‌گیرد. بسیاری از فعالیت‌های علمی در حوزهٔ زبان، ادبیات و زبان‌شناسی به داده‌های مشابهی نیاز دارند. پس می‌توان از پیکره‌ای معیّن یا بخش‌هایی از آن برای پژوهش‌های متنوع در زمینه‌های گوناگون استفاده کرد. بدین ترتیب، در جمع‌آوری پیکره دوباره کاری انجام نمی‌گیرد. با استفاده از امکانات رایانه می‌توان پیکره را در سطوح متعددی توزیع کرد و هر بخش را در پژوهشی جداگانه به کار گرفت. این تقسیم‌بندی می‌تواند بر اساس ادوار تاریخی، نوع اثر، موضوع اثر، نویسنده، سبک و یا حتی واژگان صورت گیرد.

رایانه این امکان را در اختیار ما قرار می‌دهد که پیکره‌ای عظیم و با گستردگی و گوناگونی‌های بسیار اما دارای ساختاری به سامان و منطقی در اختیار داشته باشیم تا هر گونه جست‌وجو و دست‌رسی سریع به اطلاعات مورد نیاز در هر زمان به سهولت فراهم گردد.

بر پایهٔ هر یک از اقلام اطلاعاتی یا ویژگی‌های مربوط به آنها می‌توان انواع پژوهش‌ها را در پیکره انجام داد. برای مثال جست‌وجوی واژگانی بر پایهٔ یک یا چند واژه، جست‌وجوی مفهومی بر پایهٔ مفهوم یا معنای مورد نظر، جست‌وجوی هم بافت بر پایهٔ واژه‌های هم‌آیند و یا بافت‌های وابسته و دیگر جست‌وجوها. نیز می‌توان انواع فهرست‌ها، شامل فهرست‌های واژگانی، آماری، بسامدی، نمایه‌ها و واژه‌نماها را، به شکل اطلاعات موردی و یا به شکل فرهنگ واژه‌نما همراه با فهرستِ صورت‌های کاربردی واژه‌ها و اطلاعاتی مربوط به بافتِ زبانی آنها، شمارهٔ سطر و صفحهٔ متن، نام نویسنده، مشخصات متن، تاریخ، بسامد و مانند آن، از پیکره استخراج کرد. از این گونه اطلاعات، در گسترهٔ وسیعی از پژوهش‌های زبانی، به عنوان مواد خام استفاده می‌شود. اما برای استفادهٔ گسترده از رایانه مسائلی نیز باید حل شود. مهم‌ترین آنها، که به نظر

می‌رسد گردآورندگان واژگان گزیده نیز با آن رو به رو بوده‌اند، مسئلهٔ خط فارسی است. نگاهی به برابرهای فارسیِ واژگان گزیده ابعاد این مشکل را به خوبی نمایان می‌سازد. در مورد نیاز به شیوهٔ املای معیار و مناسب برای کار با رایانه در اینجا و آنجا حروف‌هایی زده‌اند. اما برای به کارگیریِ گسترده‌ترِ رایانه در پژوهش‌های زبانی باید سریعاً به نتیجه‌ای رسید. این کار استفاده از رایانه را در پژوهش‌های زبان فارسی چند گام بلند به پیش می‌برد.

منابع

حق‌شناس، ع. م، «در جست و جوی زبان علم»، مجموعه مقالات سمینار زبان فارسی و زبان علم، تهران ۱۳۷۲.

عاصی، م، «طرحی برای تهیه فرهنگ‌های تخصصی با کمک کامپیوتر». مجموعه مقالات دومین کنفرانس زبان‌شناسی، تهران ۱۳۷۲.

عاصی، م، «پایگاه داده‌های زبان فارسی»، مجموعه مقالات سومین کنفرانس زبان‌شناسی، ۱۳۷۴ (در دست چاپ)

عاصی، م و م. عبدعلی، واژگان گزیده زبان‌شناسی، تهران ۱۳۷۵.