فرهنگستان زبان و ادب فارسی

تشابهات و تفاوتهای «پیکره» و «فرهنگ بسامدی» از دیدگاه فرهنگ‌نگاری

فرهنگستان زبان و ادب فارسی

تشابهات و تفاوتهای «پیکره» و «فرهنگ بسامدی» از دیدگاه فرهنگ‌نگاری

حمید حسنی

حمید حسنی، پژوهشگر گروه فرهنگ‌نویسی فرهنگستان، متولد سال ۱۳۴۷، تحقیقات خود را به‌ویژه در چند سال اخیر، بر روی فرهنگ‌نگاری و نیز شاخهٔ «پیکره‌نگاری» متمرکز کرده و تازه‌ترین کتاب وی، به‌نام *واژه‌های پرکاربرد فارسی امروز*، در شهریورماه ۱۳۸۴ توسط کانون زبان ایران چاپ و منتشر شده‌است.

«پیکره»، در اصطلاح علم زبان‌شناسی و دانش یا فنّ فرهنگ‌نگاری، عبارت است از مجموعه‌ای از کلمات یا مجموعه‌ای از اطلاعات خام زبانی، که متن یا متنهایی مشخص را پدید آورده، و از طریق گردآوری متون یا تحقیق میدانی به دست آمده‌است. «پیکره»، معادل لفظ corpus در زبان انگلیسی یا فرانسه است[1]. از کاربردهای مهم و اساسی پیکره، می‌توان به توصیف زبان ــ دست‌کم توصیف ویژگیهای صرفی، لغوی، املایی، و آوایی ـ واجی ــ، به دست دادن آمارهای دقیق برای گردآوری واژه‌های پرکاربرد و نیز واژگان پایه (basic vocabulary) به‌منظور آموزش زبان، و تهیّهٔ «واژگان تعریف‌نگاری» (defining vocabulary) در امر فرهنگ‌نگاری اشاره کرد.

در اغلب فرهنگهای آموزشی‌ای که در کشورهای توسعه‌یافته تألیف شده‌اند ــ به‌ویژه فرهنگهای آموزشی یک‌زبانهٔ انگلیسی ــ و متوسط تعداد مدخلهایشان در حدود پنجاه‌هزار لغت است، «واژگان تعریف‌نگاری»، برمبنای پیکره‌های بسیار عظیم اختصاصی، با حجم صدمیلیون لغت و بیشتر[2] استخراج شده‌اند. واژگان تعریف‌نگاری، نه‌تنها جزو ضمایم یا مقدمهٔ این فرهنگها و به ترتیب الفبا فهرست شده‌اند، بلکه مؤلفان، در متن مدخلها، با افزودن نشانه‌ها، اعداد، یا با متمایز کردن مدخلها از طریق استفاده از رنگهای متفاوت با سایر مدخلها، شکلی جذاب و دیدنی به فرهنگ داده‌اند.

۱ـ بنگرید به حسنی (۱۳۸۴، ص ۹)

۲ـ از بزرگترین پیکره‌های جهان، می‌توان «*پیکرهٔ ملّی بریتانیا*» (*British National Corpus*)، یا به‌اختصار، BNC، با بیش از پانصدمیلیون لغت (تهیّه‌شده در دانشگاه بیرمنگام بریتانیا)، و «*پیکرهٔ بین‌المللی کیمبریج*» (در بریتانیا) (*Cambridge International Corpus*)، به اختصار ، *CIC*، با حدود هفتصدمیلیون لغت، را نام برد.

و امـا «فرهنگ بسـامدی»، عبارت از کتاب یا هرگونه مجموعه‌ای اسـت که در آن، پس‌از تفکیک واحدهای واژگانی[۳] یک یا چند متن مشخص، اجزا، پس از لِماتیزه کردن (به‌تعبیـر آقای بهروز صفرزاده، لِمابندی[۴])، هم به ترتیب الفبا و هم به ترتیب بسـامد، در دسـت‌کم دو فهرسـت جداگانه، ثبت شـده‌اند. در ضمن، هر یک از صورتهای صرفی (از هر نوع کلمه، به‌ویژه فعل و اسـم)، ذیل صورت قاموسـی[۵] ثبت می‌شود. «پیکره‌نگار» موظف اسـت دسـت‌کم فهرسـت الفبایی واحدهای واژگانی را، دقیقاً یا تقریباً به همان شـکلی که در متن یا متنهای مادر به‌کار رفته‌اند، تهیه کند. البته پیکره‌نگاران یا مؤسسات پیکره‌نـگاری معتبر، فهرسـت بسـامدی را نیز در حاصـل کار خود به دسـت می‌دهند. به‌طور بسـیار سـاده و خلاصه، باید گفت کـه در «پیکره»، برای مثـال، مصدر «رفتن» و همهٔ صورتهای صرفی آن، دقیقاً به شـکل به‌کاررفته و به‌طور مستقل ثبت می‌شوند؛ یعنی «رفتم»، «رفتی»، «رفت»، ... ، «می‌روم»، «می‌روی»، «می‌رود»، ... ، «رفته‌ام»، ...، «رفته‌بودم»، ... ، «رفته‌باشـم»، ...، «رفته‌بوده‌باشـم»، ... ، «می‌رفتم»، ...، «رفتنها»، «رفته» (فعلِ وصفی و صفت)، و امثال اینها، به‌علاوهٔ صورتهای منفیِ هر کدام، به‌صورت جداگانه، در «فهرست بسـامدی» با ذکر بسـامد مسـتقل، و در «فهرسـت الفبایی» در ردیف الفبایی خود ثبت می‌شـوند. مصدر «رفتن»، نسبت به صورتهای صرفی یادشده، در اصطلاح زبان‌شناسی، و به‌تبعِ آن فرهنگ‌نگاری و پیکره‌نگاری، «لِما» (lemma) نامیده می‌شود. «لِماتیزه کردن» یـا «لِمابندی کـردن» (lemmatization, to lemmatize) عبارت اسـت از تبدیل صورتهای صرفی به «لِما». مثالی دیگر: می‌دانیم که بسـیاری از اسـمها، دسـت‌کم به دو صورت جمع بسـته می‌شوند، یعنی با دو پسـوندِ «ـها» و «ـان»؛ گاهی نیز، به‌ویژه کلمات عربی، با دو پسـوند دیگـرِ «ـات» و «ـین». مثلاً دو کلمهٔ «پسـر» و «دختر»، به ترتیب، هم به‌صورت «پسـرها» و «دخترها» جمع بسـته می‌شوند و هم به‌شکل «پسـران» و «دختران». درسـت اسـت که مطمئنیم جمع این دو کلمه هیچ‌گاه به‌صورت «پسـرات» و «پسرین» / pesar-īn/ ، و «دخترات» و «دخترین» /doxtar-īn/ به کار نرفته‌است، اما به‌یقین هم نمی‌توانیم حکم کنیم که کاربرد «دختران» و «پسران» بیشتر است یا «دخترها» و «پسرها».

۳ـ Lexical units، یا به اختصار، LUs (جمعِ LU).

۴ـ بنگرید به کوچرا، ص ۵۲ و ۵۳. دربارهٔ «لِما» و «لِماتیزه کردن/ لِمابندی کردن» توضیح خواهیم داد.

۵ـ صورت قاموسی هر کلمه، شکلی از آن است که در یک فرهنگ لغت معیار (استاندارد)، مثلاً در *فرهنگ بزرگ سخن* یا *فرهنگ فشردهٔ سخن* (دوجلدی) مدخل (entry) قرار گیرد.

تنها با گرفتن آمار و با استفاده از پیکره یا فرهنگ بسامدی و مقایسهٔ پیکره‌ها و فرهنگهای بسامدیِ متعدد می‌توان نتیجه‌ای مطلوب حاصل کرد. پس، پیکره و فرهنگ بسامدی، در این زمینه اشتراک دارند.

اما هرگاه پژوهشگری بخواهد به بررسی این موضوع بپردازد که فعلهایی که با پیشوند صرفیِ «مـی ـ» (مانند: می‌رود، می‌رفت، می‌رفته‌اسـت،...)، «بِـ / بُـ» (مثـل: برود، برفت، بُـرو،...)، «نَـ/ نِـ» (نظیر: نرفت، نمی‌رود، نمی‌رفته‌اسـت، ...)، و امثـال اینها به‌کار رفته‌اند در زبـان چه جایگاهی دارند، تنها ازطریق اسـتفاده از «پیکـره» می‌تواند به مقصود خود دسـت یابد، یا دسـت‌کم، حصول به مقصـود ازطریق بهره بردن از «پیکـره» راحت‌تر و دقیق‌تر است تا «فرهنگ بسـامدی». برای اثبات فواید «پیکره» مثالهای بی‌شماری در این زمینه می‌توان زد، اما در این مختصر مجال بازگو کردن آنها نیست و تنها به ذکر مثالها و اشاراتی بسنده شد.

با این‌حال، «فرهنگهای بسـامدی» نسـبت به «پیکره» یا «پیکره‌هـا» مزایایی دارند که مهم‌ترین آنها دقیق بودن اطلاعات جمع‌آوری‌شـده و منظم بودن آنها از لحاظ دستوری (به‌ویژه صرفی یا اشـتقاقی) و ریشه‌شناختی است. مثلاً، در فرهنگ بسامدی، گردآورنده موظف اسـت که «بار» را، نه به‌تعداد معانی آن، بلکه به‌تعداد صورتهای متفاوتش از نظر ریشه‌شـناختی، به‌طور جداگانه ثبت کند[۶]. حتی تکیه (مراد تکیهٔ اصلی است) و جای آن نیز در فرهنگ بسـامدی عامل تفکیک مدخل اسـت؛ مثلاً در «مردی»، اگر هنگام تلفظ، تکیه بر هجای نخسـتِ آن قرار گیرد، به‌معنی «یک مرد» (با «ی» وحدت یا نکره) است، و در ذیل «مرد» آورده می‌شود؛ اما چنانچه تکیه بر هجای دومش قرار بگیرد، «ی» در آن مصدری (یا حاصل‌مصدری) اسـت و به‌معنی «مردانگی» خواهد بود، و دراین‌صورت در مدخلی جداگانه ثبت خواهد شـد، همان‌گونـه که در فرهنگ لغت هم مدخلی جداگانه دارد. برعکس، در پیکره، «بار» و «مردی»، با هر نوع اشـتقاق و تکیه‌ای، هر کدام تنها در یک مدخل ثبت می‌شوند؛ زیرا اشتقاق لغوی و جای تکیه در پیکره عامل تفکیک نیست. بر همین اساس است که در پیکره‌هایی نظیر BNC و Brown، کلمهٔ saw تنها یک بار و یک جا با بسـامدی معیّن شده‌اسـت، درحالی‌که چندین معنا و اشـتقاق متفاوت دارد:

۶ـ در *فرهنگ بزرگ سخن*، «بار» در هفت مدخل آمده‌است؛ اگر مدخلِ پس از مدخل هفتم («بار/ بارّ» عربی) را هم به حساب بیاوریم، هشت مدخل خواهد بود.

گذشــتهٔ فعل see (به‌معنی دیدن)، ارّه (اسم)، ارّه شــدن، و ارّه کردن (به‌ترتیب، مصدر متعدّی و لازم).

البتــه در نســخه‌های (versions) خاصــی از «پیکره»هــا، موســوم بــه «پیکره‌های کُدگذاری‌شده»، ریشه (اشتقاق) و تکیه از عوامل اصلی تفکیک مدخلهاست.

این‌جانــب، در کتاب خــود، *واژه‌های پرکاربرد فارســی امروز*، مثالهایــی از پیکرهٔ یک‌میلیون‌لغتی خویش ـ که آن را به ســرمایه و بــا حمایت اولیای محترم و فهیم «کانون زبان ایران» تهیه کرده‌ام ـ و نیز از پیکرهٔ دانشگاه براون، که آن هم دارای حدود یک‌میلیون لغت اســت و در اوایل دههٔ شصت میلادی تهیه شــده و معروف‌ترین پیکرهٔ زبانیِ جهان است، ذکر کرده‌ام.

در پایــان، بایــد این نکته را بیفزایم کــه ازطریق تهیهٔ انواع «پیکره»هــا و «فرهنگهای بســامدی»، و نیز اســتفاده از آنها، مطمئن‌ترین، دقیق‌ترین، و مفیدترین اطلاعات زبانی را درجهت توصیف زبان می‌توان به دست آورد.

فهرست منابع:

حســنی، حمید: «ضرورت تهیهٔ فرهنگهای بســامدی برای زبان فارســی»، *مجلّهٔ کتاب ماه (ادبیات و فلسفه)*، سال پنجم، شمارهٔ ۱ (پیاپی: ۴۹)، آبان ۱۳۸۰، صص ۱۴۰ـ ۱۴۵؛

حسنی، حمید: *واژه‌های پرکاربرد فارسی امروز (بر مبنای پیکرهٔ یک‌میلیون‌لغتی؛ شامل بیش از ۸۰۰۰ لغت قاموسی و غیرقاموسی)*، تهران، کانون زبان ایران، ۱۳۸۴؛

کوچِــرا، هنری: «ریاضیّات زبان»، ترجمهٔ بهروز صفرزاده، *مجلّهٔ کتاب ماه (ادبیات و فلســفه)*، ســال چهارم، شمارهٔ ۷ (پیاپی: ۴۳)، اردیبهشت ۱۳۸۰، صص ۵۰ـ۵۳.

Collins Cobuild English Language Dictionary, Editor-in-Chief: John Sinclair, Great Britain, 2001

Francis, W. Nelson & Henry Kucera: *Brown Corpus Manual (Manual of Information to accompany A Standard Corpus of Present-Day, Edited American English, for use with Dīgttal Computers)*, USA, 1964 (Revised 1971, Revised and Amplified 1979)

Hornby, A.: *Oxford Advanced Learner's Dictionary*, fifth edition, Editor: S. Jonathan Crowther, Great Britain, 2000

Longman Dictionary of Contemporary English, Director: Della Summers, Great Britain, 2003

MacMillan English Dictionary (for Advanced Learners of American English), Editor-in-Chief: Michael Rundell, Great Britain, 2002