

فونیکس‌ستان بایب ادب فارسی

# مقاله

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



# 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 شاخه «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 متناهی 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.<sup>۱</sup> از کاربردهای مهم و اساسی پیکره، می‌توان به توصیف زبان - دست کم توصیف ویژگی‌های صرفی، لغوی، املائی، و آوایی - واجی -، به دست دادن آمارهای دقیق برای گردآوری واژه‌های پر کاربرد و نیز واژگان پایه (*basic vocabulary*) به‌منظور آموزش زبان، و تهیه «واژگان تعریف‌نگاری» (*defining vocabulary*) در امر فرهنگ‌نگاری اشاره کرد.

در اغلب فرهنگ‌های آموزشی ای که در کشورهای توسعه یافته تألیف شده‌اند - به‌ویژه فرهنگ‌های آموزشی یک‌زبان انگلیسی - و متوسط تعداد مدخل‌هایشان در حدود پنجاه‌هزار لغت است، «واژگان تعریف‌نگاری»، بر مبنای پیکره‌های بسیار عظیم اختصاصی، با حجم صد میلیون لغت و بیشتر<sup>۲</sup> استخراج شده‌اند. واژگان تعریف‌نگاری، نه تنها جزو ضمایم یا مقدمه 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.

۱- بنگرید به حسنی (۱۳۸۴، ص ۹)

۲- از بزرگترین پیکره‌های جهان، می‌توان «پیکره ملی بریتانیا» (*British National Corpus*)، یا به اختصار، BNC، با بیش از پانصد میلیون لغت (تهیه‌شده در دانشگاه بیرمنگام بریتانیا)، و «پیکره بین‌المللی کیمبرج» (در بریتانیا) *(Cambridge International Corpus)*، به اختصار، *CIC*، با حدود هفتصد میلیون لغت، را نام برد.

و اما «فرهنگ بسامدی»، عبارت از کتاب یا هر گونه مجموعه‌ای است که در آن، پس از تفکیک واحدهای واژگانی<sup>۳</sup> یک یا چند متن مشخص، اجزاء، پس از لماتیزه کردن (به تعبیر آقای بهروز صفرزاده، لمابندی<sup>۴</sup>)، هم به ترتیب الفبا و هم به ترتیب بسامد، در دست کم دو فهرست جداگانه، ثبت شده‌اند. در ضمن، هر یک از صورت‌های صرفی (از هر نوع کلمه، به ویژه فعل و اسم)، ذیل صورت قاموسی<sup>۵</sup> ثبت می‌شود. «پیکره‌نگار» موظف است دست کم فهرست الفبایی واحدهای واژگانی را، دقیقاً یا تقریباً به همان شکلی که در متن یا متنهای مادر به کار رفته‌اند، تهیه کند. البته پیکره‌نگاران یا مؤسسات پیکره‌نگاری معتبر، فهرست بسامدی را نیز در حاصل کار خود به دست می‌دهند. به‌طور بسیار ساده و خلاصه، باید گفت که در «پیکره»، برای مثال، مصدر «رفتن» و همه صورت‌های صرفی آن، دقیقاً به شکل به کار رفته و به‌طور مستقل ثبت می‌شوند؛ یعنی «رفتم»، «رفتگی»، «رفت»، ... «می‌روم»، «می‌روی»، «می‌رود»، ... «رفته‌ام»، ... «رفته بودم»، ... «رفته باشم»، ... «رفته بوده باشم»، ... «می‌رفتم»، ... «رفتند»، «رفته» (فعل و صفتی و صفت)، و امثال اینها، به‌علاوه صورت‌های منفی هر کدام، به‌صورت جداگانه، در «فهرست بسامدی» با ذکر بسامد مستقل، و در «فهرست الفبایی» در ردیف الفبایی خود ثبت می‌شوند. مصدر «رفتن»، نسبت به صورت‌های صرفی یادشده، در اصطلاح زبان‌شناسی، و به تبع آن فرهنگ‌نگاری و پیکره‌نگاری، «لِما» (lemma) نامیده می‌شود. «لماتیزه کردن» یا «لمابندی کردن» (lemmatization, to lemmatize) عبارت است از تبدیل صورت‌های صرفی به «لِما». مثالی دیگر: می‌دانیم که بسیاری از اسمها، دست کم به دو صورت جمع بسته می‌شوند، یعنی با دو پسوند «ها» و «ان»؛ گاهی نیز، به‌ویژه کلمات عربی، با دو پسوند دیگر «ات» و «ین». مثلاً دو کلمه «پسر» و «دختر»، به ترتیب، هم به‌صورت «پسرها» و «دخترها» جمع بسته می‌شوند و هم به‌شکل «پسران» و «دختران». درست است که مطمئنیم جمع این دو کلمه هیچ‌گاه به‌صورت «پسران» و «پسرین» / pesar-īn، و «دختران» و «دخترین» / doxtar-īn به کار نرفته‌است، اما به یقین هم نمی‌توانیم حکم کنیم که کاربرد «دختران» و «پسران» بیشتر است یا «دخترها» و «پسرها».

۳- Lexical units، یا به اختصار، LU (جمع LU).

۴- بنگرید به کوچرا، ص ۵۲ و ۵۳. درباره «لِما» و «لماتیزه کردن/لمابندی کردن» توضیح خواهیم داد.

۵- صورت قاموسی هر کلمه، شکلی از آن است که در یک فرهنگ لغت معیار (استاندارد)، مثلاً در فرهنگ بزرگ سخن یا فرهنگ فشرده سخن (دو جلدی) مدخل (entry) قرار گیرد.

تنها با گرفتن آمار و با استفاده از پیکره یا فرهنگ بسامدی و مقایسه پیکره‌ها و فرهنگهای بسامدی 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 صرفی «می» (مانند: می‌رود، می‌رفت، می‌رفته‌است،...)، «بِ / بُ» (مثل: برود، برفت، بُرو،...)، «نَ / نِ» (نظیر: نرفت، نمی‌رود، نمی‌رفته‌است،...)، و امثال اینها به کار رفته‌اند در زبان چه جایگاهی دارند، تنها از طریق استفاده از «پیکره» می‌تواند به مقصود خود دست یابد، یا دست کم، حصول به مقصود از طریق بهره بردن از «پیکره» راحت‌تر و دقیق‌تر است تا «فرهنگ بسامدی». 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.

۹۱

با این حال، «فرهنگهای بسامدی» نسبت به «پیکره» یا «پیکره‌ها» مزایایی دارند که مهم‌ترین آنها دقیق بودن اطلاعات جمع‌آوری‌شده و منظم بودن آنها از لحاظ دستوری (به‌ویژه صرفی یا اشتقاقی) و ریشه‌شناختی است. مثلاً، در فرهنگ بسامدی، گردآورنده موظف است که «بار» را، نه به تعداد معانی آن، بلکه به تعداد صورتهای متفاوتش از نظر ریشه‌شناختی، به‌طور جداگانه ثبت کند. حتی تکیه (مراد تکیه اصلی است) و جای آن نیز در فرهنگ بسامدی عامل تفکیک مدخل است؛ مثلاً در «مردی»، اگر هنگام تلفظ، تکیه بر هجای نخست آن قرار گیرد، به معنی «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) است و به معنی «مردانگی» خواهد بود، و در این صورت در مدخلی جداگانه ثبت خواهد شد، همان‌گونه که در فرهنگ لغت هم مدخلی جداگانه دارد. برعکس، در پیکره، «بار» و «مردی»، با هر نوع اشتقاق 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، کلمه saw تنها یک بار و یک جا با بسامدی معین شده‌است، در حالی که چندین معنا و اشتقاق متفاوت دارد:

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



گذشته فعل see (به معنی دیدن)، ارّه (اسم)، ارّه شدن، و ارّه کردن (به ترتیب، مصدر متعدی و لازم).

البته در نسخه‌های (versions) خاصی از «پیکره‌ها»، موسوم به «پیکره‌های کُدگذاری شده»، ریشه (اشتقاق) و تکیه از عوامل اصلی تفکیک مدخلهاست.

این جانب، در کتاب خود، واژه‌های پرکاربرد فارسی امروز، مثالهایی از پیکره یک میلیون لغتی خویش - که آن را به سرمایه و با حمایت اولیای محترم و فهیم «کانون زبان ایران» تهیه کرده‌ام - و نیز از پیکره دانشگاه براون، که آن هم دارای حدود یک میلیون لغت است و در اوایل دهه شصت میلادی تهیه شده و معروف‌ترین پیکره زبانی جهان است، ذکر کرده‌ام.

در پایان، باید این نکته را بیفزایم که از طریق تهیه انواع «پیکره‌ها» و «فرهنگهای بسامدی»، و نیز استفاده از آنها، مطمئن‌ترین، دقیق‌ترین، و مفیدترین اطلاعات زبانی را در جهت توصیف زبان می‌توان به دست آورد.

#### فهرست منابع:

حسینی، حمید: «ضرورت تهیه فرهنگهای بسامدی برای زبان فارسی»، مجله کتاب ماه (ادبیات و فلسفه)، سال پنجم، شماره ۱ (پیاپی: ۴۹)، آبان ۱۳۸۰، صص ۱۴۰-۱۴۵؛

حسینی، حمید: واژه‌های پرکاربرد فارسی امروز (بر مبنای پیکره یک میلیون لغتی؛ شامل بیش از ۸۰۰۰ لغت قاموسی و غیر قاموسی)، تهران، کانون زبان ایران، ۱۳۸۴؛

کوچرا، هنری: «ریاضیات زبان»، ترجمه بهروز صفرزاده، مجله کتاب ماه (ادبیات و فلسفه)، سال چهارم، شماره ۷ (پیاپی: ۴۳)، اردیبهشت ۱۳۸۰، صص ۵۰-۵۳.

*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 Digital 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

